Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(9): 6072–6100.
DOI: 10.3934/era.2026268
Received: 31 March 2026
Revised: 12 June 2026
Accepted: 26 June 2026
Published: 14 July 2026

Research article

Compute-conscious multi-view neural framework with efficient state space models and reduced-order cross-attention for rolling bearing fault diagnosis

Guangpeng Sun[1], **Peiju Chang**[1,2,*], **Shaojuan Ma**[1,2,*] **and Yucong Li**[1]

[1] School of Mathematics and Information Science, North Minzu University, Yinchuan 750021, China
[2] Ningxia Key Laboratory of Intelligent Information and Big Data Processing, Yinchuan 750021, China

* **Correspondence:** Email: chang_peiju@nmu.edu.cn, sjma@nmu.edu.cn.

Abstract: Rolling bearing fault diagnosis under practical compute constraints and long time series remains challenging in industrial settings. This study presented a compact same-source multi-view framework that combines a multi-scale gated Mamba sequence branch, a residual network with a convolutional block attention module (ResNet-CBAM) image branch, and lightweight low-rank cross-view fusion. Continuous wavelet transform (CWT) images were used as the default image view in the main experiments because they provide the most suitably maintained same-condition representation among the evaluated image views. Under a unified three-seed protocol, the proposed model reached 99.20% on Case Western Reserve University (CWRU) and 99.61% on Southeast University (SEU) in same-condition diagnosis, and attained the strongest cross-condition results within the internal comparison, with 27.39% on CWRU and 30.43% on SEU. The model used about 5.46×10^5 parameters and 5.64×10^8 multiply–accumulate operations (MACs). These results indicate a balanced trade-off between relative cross-condition generalization and computational cost, while also showing that broader operating-condition diversity, industrial validation, and deployment-oriented central processing unit (CPU) or edge profiling remain necessary.

Keywords: bearing fault diagnosis; multi-view fusion; state space model; Mamba gating; low-rank cross-attention; deep learning

1. Introduction

Rolling element bearings are critical components in rotating machinery, and their health strongly affects the safety and reliability of aircraft engines, wind turbines, rail systems, and other industrial equipment [1, 2]. In practice, bearing fault diagnosis must cope with long vibration sequences, background noise, changing operating conditions, and limited computational budgets [3, 4]. These con-

straints make it insufficient to pursue accuracy alone; practical methods also need compact models and efficient feature fusion.

Recent bearing diagnosis studies can also be understood from the perspective of input representation. Time-domain methods operate directly on raw vibration sequences, whereas frequency-domain and time–frequency methods rely on transformed views such as spectra, wavelet maps, or other image-like encodings derived from the same underlying signal [5–7]. This distinction matters because different representations highlight different fault cues. Raw sequences retain local temporal dynamics, while two-dimensional encodings can reveal transition structure, correlation patterns, and scale-dependent energy variations that are often easier to exploit with image backbones.

Beyond single-view learning, recent studies have explored richer relations among views, sensors, and temporal segments. For example, a spatial–temporal hypergraph model has been used to capture higher-order relations in aero-engine bearing diagnosis [8], while adaptive multi-view hypergraph learning has also been studied for cross-condition bearing diagnosis [9]. These studies suggest that the diagnostic pipeline should be described not only by the learning architecture, but also by how the signal is represented before learning.

At the model level, deep learning has substantially improved automatic feature extraction for bearing diagnosis. Convolutional neural network (CNN)-based methods remain effective for image-like representations, and multi-scale convolution designs improve the capture of local and global patterns [10]. Transformer-based models strengthen long-range dependency modeling, but their quadratic attention cost can become restrictive for long sequences or resource-constrained deployment [7, 11]. Structured state space models (SSMs), including S4 and Mamba, provide a more efficient alternative for long-sequence processing [12–14]. Meanwhile, lightweight modules such as the convolutional block attention module (CBAM) and low-rank parameterization remain attractive for controlling the cost of image feature enhancement and cross-branch fusion [15, 16].

Despite this progress, three issues remain insufficiently clarified for compute-constrained bearing diagnosis. First, studies that combine raw signals with image-like encodings derived from the same vibration source are often discussed together with heterogeneous multimodal sensing, even though the learning setting and fusion objective are different. Second, long-sequence modeling and cross-view fusion are frequently studied separately, leaving the balance between long-range dependency capture and lightweight interaction cost insufficiently examined. Third, low-rank attention ideas provide a useful efficiency prior, but in same-source cross-view diagnosis their contribution is more appropriately framed as a task-oriented parameterization than as a fundamentally new attention family.

Accordingly, this study investigates a compact fusion framework that combines two synchronized views derived from the same vibration source: the raw time-series signal and a derived image representation. This setting is therefore described as *same-source multi-view* rather than heterogeneous multi-sensor multimodality, because both branches originate from one sensor stream. In the main experiments, the image branch uses CWT-based time–scale images, while alternative Gramian angular field (GAF) and Markov transition field (MTF) encodings are evaluated separately in the image-representation ablation. The objective is to clarify this problem setting and to improve the balance between relative cross-condition generalization and computational cost under unified same-condition and cross-condition protocols, while keeping the empirical claims aligned with the available evidence and deployment scope.

The main contributions of this study are as follows:

1) To address the difficulty of jointly capturing local fault transients and long-range dependencies in long vibration sequences, a multi-scale gated Mamba branch is introduced to combine parallel local convolutions with linear-time state-space modeling.

2) To reduce redundancy in cross-view interaction between same-source signal and image representations, existing low-rank attention ideas are adapted into a task-oriented fusion module, dynamic low-rank multi-head cross-attention (DyLoMHCA), rather than being framed as a fundamentally new attention family.

3) To provide a clearer and more compact formulation of same-source dual-representation learning for bearing diagnosis, an integrated framework is established to couple raw vibration sequences with CWT images and to examine the resulting accuracy–efficiency trade-off, especially under cross-condition evaluation.

The remaining sections are organized as follows: Section 2 introduces the proposed multi-view fault diagnosis model and its components. Section 3 describes the datasets, split protocols, and evaluation settings. Section 4 presents the comparative results under the unified same-condition and cross-condition settings. Section 5 reports the component-wise ablation and parameter sensitivity analyses. Finally, Section 6 summarizes the main findings and limitations.

2. Methodology

2.1. Problem definition

In this study, each sample is represented by two synchronized views derived from the same vibration source: a time–scale image and its corresponding time series signal. The learning problem is denoted by $\mathcal{D}_{all} = \{x^j_{img}, x^j_{tseq}, z^j_{target}\}$, $j = 1, 2, \ldots, n_t$, where x^j_{img} denotes the image representation, x^j_{tseq} denotes the time series sample, and z^j_{target} denotes the fault label. The term *multi-view* or *same-source dual-representation* learning is therefore used rather than heterogeneous independent-sensor multimodality. In the main experiments, a continuous wavelet transform (CWT) is adopted as the default image representation because the same-condition image-representation ablation identifies the CWT as the most suitable default among the CWT/GAF/MTF choices across the CWRU and SEU datasets. Specifically, each 1024-point signal segment is transformed with the Morlet wavelet using geometrically spaced scales from 1 to 128, the coefficient magnitude is mapped through $\log(1 + |\cdot|)$, and the resulting time–scale map is resized to 224×224. This representation preserves localized energy patterns across time and scale, while alternative GAF and MTF encodings are revisited in Section 5.

2.2. Framework

The proposed multi-view fusion model consists of three key components: 1) deep feature extraction and attention enhancement for image features using ResNet-CBAM, 2) long-sequence data representation with multi-scale gated Mamba, and 3) fusion of global information from both branches through a dynamic low-rank multi-head cross-attention mechanism. This framework balances scalability with learning efficiency and representational capacity. The framework is shown in Figure 1.

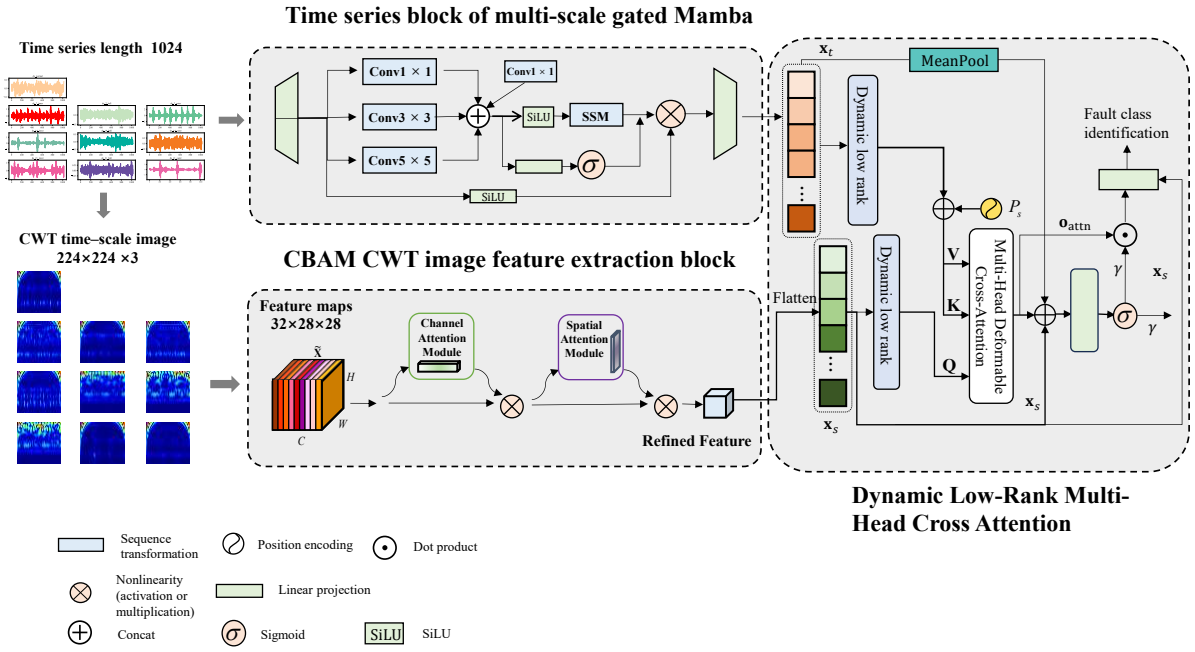

Figure 1. Overview of the proposed same-source multi-view fusion framework.

2.3. ResNet + CBAM for image feature extraction

In computer vision, convolutional neural networks (CNNs) have been widely adopted for image feature extraction. As network depth increases, traditional CNN models often encounter issues such as vanishing gradients and increased training difficulty. These problems lead to situations where the model's performance on the training and validation sets does not improve or even decreases as the network depth increases. To mitigate these challenges, the residual network (ResNet) architecture has been proposed and widely used in image recognition tasks. ResNet alleviates the vanishing gradient problem by introducing identity skip connections, allowing the model to construct deeper network structures. Additionally, to improve the model's feature representation capability, the convolutional block attention module (CBAM) [15] is introduced into the ResNet architecture. The CBAM dynamically adjusts the weight distribution of the feature map along the channel and spatial dimensions through the sequential application of channel attention (CA) and spatial attention (SA).

From a theoretical perspective, a 2D convolution with kernel $\mathbf{W} \in \mathbb{R}^{C_{out}\times C_{in}\times k_h\times k_w}$ and bias $\mathbf{b} \in \mathbb{R}^{C_{out}}$ acts on an input feature map $\mathbf{X} \in \mathbb{R}^{B\times C_{in}\times H\times W}$ as

$$Y_{b,c,i,j} = \sum_{c'=1}^{C_{in}} \sum_{u=1}^{k_h} \sum_{v=1}^{k_w} W_{c,c',u,v}\, X_{b,c',\,i+u-1,\,j+v-1} + b_c. \tag{2.1}$$

In ResNet, a residual block replaces a direct mapping with

$$\mathbf{h}^{(\ell+1)} = \mathbf{h}^{(\ell)} + \mathcal{F}(\mathbf{h}^{(\ell)}; \Theta^{(\ell)}), \tag{2.2}$$

where $\mathcal{F}(\cdot)$ is a shallow transformation (e.g., two 3×3 convolutions with normalization and activation). This formulation stabilizes gradients and preserves information flow during deep stacking.

In this attention mechanism, the CA module generates channel-level attention weights through global average pooling and global max pooling operations, thereby allowing the network to identify and assign higher weights to feature channels. Meanwhile, the SA module obtains spatial-level attention weights by applying average and max pooling across the channel dimension, guiding the model to focus on discriminative regions in the feature map. This dual attention mechanism refines the feature map's representation and can improve performance in image recognition tasks.

To mitigate the vanishing gradient problem while maintaining model depth, the image branch uses ResNet as the backbone. For the input CWT image $\mathbf{x}_{img} \in \mathbb{R}^{B\times 3\times H\times W}$, after passing through convolutional layers and batch normalization layers, the output feature map $\mathbf{f}_0 \in \mathbb{R}^{B\times C\times H'\times W'}$ is obtained. Then, several residual blocks (Basic Block) are stacked:

$$\mathbf{f}_{k+1} = \text{BasicBlock}(\mathbf{f}_k), \quad k = 0, 1, \ldots. \tag{2.3}$$

Each residual block has an identity skip connection, which alleviates the training difficulty caused by deep networks. After extracting multi-scale features with the residual network, this model further incorporates the CBAM. In the CA module, each channel is weighted, while the SA module weights the spatial locations of the feature map. The core formulas are as follows:

1) CA: For the input feature map $\mathbf{X} \in \mathbb{R}^{B\times C\times H'\times W'}$, global average pooling and global max pooling are first applied to obtain $\mathbf{X}_{avg}, \mathbf{X}_{max} \in \mathbb{R}^{B\times C\times 1\times 1}$, respectively. Fully connected layers then map each of them back to the original feature-map dimension, and sum them elementwise. Finally, the CA matrix $\mathbf{M}_c(\mathbf{X})$ is derived via the sigmoid function:

$$\mathbf{M}_c(\mathbf{X}) = \sigma\big(f(\mathbf{X}_{avg}) + f(\mathbf{X}_{max})\big), \tag{2.4}$$

where σ represents the element-wise sigmoid function, and $f(\cdot)$ denotes the CA mapping.

2) SA: For the feature map $\mathbf{X}'$ weighted by the CA, average pooling and max pooling are applied across the channel dimension, and the results are then concatenated to form $\mathbf{X}'_{avg\_max} \in \mathbb{R}^{B\times 2\times H'\times W'}$. The 7×7 convolution is applied, followed by a sigmoid activation to obtain the final SA:

$$\mathbf{M}_s(\mathbf{X}') = \sigma\big(\text{Conv2D}(\text{Concat}(\text{Avg}(\mathbf{X}'), \text{Max}(\mathbf{X}')))\big). \tag{2.5}$$

Taken together, residual shortcuts maintain stable gradient propagation in deeper stacks, CBAM reweights features along channels and spatial locations to enhance saliency without heavy architectural overhead, and global pooling followed by lightweight heads yields compact yet discriminative image representations.

2.4. Multi-scale gated Mamba for time-series feature modeling

The evolution of sequence modeling has been driven by the need to efficiently capture both local patterns and long-range dependencies in temporal data. The discussion therefore begins with the continuous-time state-space model (SSM), which provides the theoretical framework for efficient sequence modeling. A continuous-time SSM is mathematically defined by the system of differential equations:

$$\frac{d\mathbf{s}(t)}{dt} = \mathbf{A}\,\mathbf{s}(t) + \mathbf{B}\,\mathbf{u}(t), \quad \mathbf{y}(t) = \mathbf{C}\,\mathbf{s}(t) + \mathbf{D}\,\mathbf{u}(t), \tag{2.6}$$

where $\mathbf{s}(t)$ represents the latent state, $\mathbf{u}(t)$ denotes the input signal, and $\mathbf{y}(t)$ is the corresponding output. For practical implementation in discrete-time systems, the continuous model is discretized using a sampling step Δ, yielding:

$$\mathbf{s}_{k+1} = \overline{\mathbf{A}}\,\mathbf{s}_k + \overline{\mathbf{B}}\,\mathbf{u}_k, \quad \mathbf{y}_k = \mathbf{C}\,\mathbf{s}_k + \mathbf{D}\,\mathbf{u}_k, \tag{2.7}$$

where the discrete matrices are given by $\overline{\mathbf{A}} = e^{\mathbf{A}\Delta}$ and $\overline{\mathbf{B}} = \int_0^{\Delta} e^{\mathbf{A}\tau} d\tau\,\mathbf{B}$. Modern SSM implementations, such as S4 and Mamba, enhance stability by parameterizing the matrix $\mathbf{A}$ in log-space and employ efficient convolutional or scanning algorithms to achieve linear-time complexity in sequence modeling.

Existing sequence modeling approaches encounter significant limitations in industrial fault diagnosis applications. Recurrent neural networks (RNNs, LSTMs, and GRUs) suffer from vanishing gradient problems when processing long sequences, while Transformer architectures exhibit quadratic computational complexity $O(L^2)$ that becomes computationally prohibitive for extended time-series data. In contrast, structured state space models (SSMs) address some of these limitations by providing linear-complexity sequence modeling, making them suitable for long-range dependency modeling in industrial scenarios.

To address the specific challenges of same-source multi-view bearing fault diagnosis, the multi-scale gated Mamba architecture systematically integrates local pattern recognition with global temporal modeling. The design centers on three complementary mechanisms that capture fault signatures across multiple temporal scales while maintaining computational efficiency.

The architecture operates through a structured information flow: for input sequence $\mathbf{x}_{seq} \in \mathbb{R}^{B\times L\times d_{model}}$, the representation is decomposed into parallel processing pathways that capture local patterns at different temporal resolutions. These multi-scale local features are then integrated with selective state-space updates to model long-range dependencies. An adaptive gating mechanism balances the contributions of local and global features to produce stable fault representations.

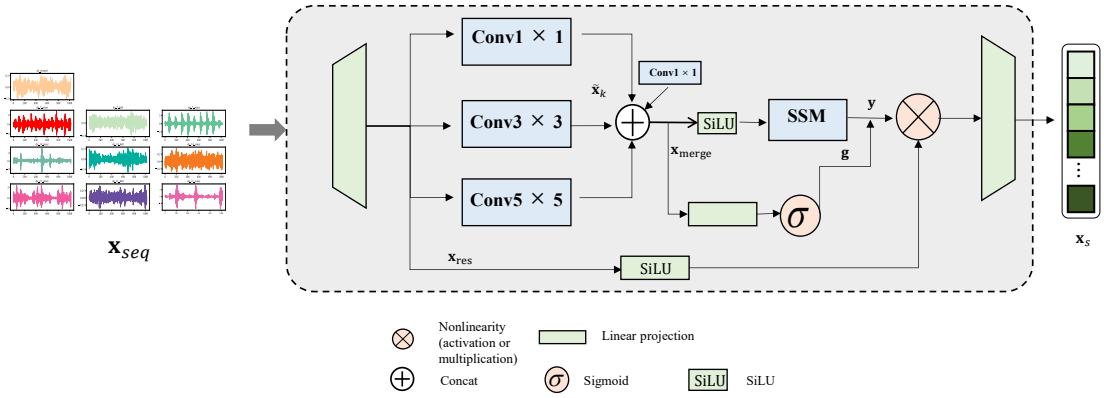

Figure 2. Overview of the multi-scale gated Mamba sequence branch.

The sequence branch contains three key design elements: 1) multi-scale depthwise convolution that efficiently captures temporal patterns across different scales using parallel branches with kernel sizes $k \in \{1, 3, 5\}$; 2) selective state-space modeling that employs input-dependent parameters for adaptive long-range dependency modeling; and 3) gated residual fusion that combines local convolution features with global state-space representations through learnable gating mechanisms.

This integrated design enables the proposed model to simultaneously capture high-frequency transient fault events and low-frequency operational trends, while maintaining linear computational com-

plexity $O(L)$ that scales favorably with sequence length. The following subsection presents the specific implementation of these mechanisms.

Figure 2 gives a compact overview of the multi-scale gated sequence branch used in the proposed model.

2.4.1. Multi-scale depthwise convolution and gating mechanism

Building upon the architectural overview, the specific implementation of the multi-scale gated Mamba mechanism is detailed below. The processing pipeline consists of three interconnected stages: multi-scale local feature extraction, selective state-space modeling, and gated fusion.

Input decomposition and multi-scale convolution: For the input sequence $\mathbf{x}_{seq} \in \mathbb{R}^{B\times L\times d_{model}}$, a linear transformation decomposes it into two complementary pathways: a convolution branch $\mathbf{x}_{conv}$ for local pattern extraction and a residual branch $\mathbf{x}_{res}$ for information preservation. The convolution branch is reshaped to $\mathbb{R}^{B\times d_{inner}\times L}$ to enable efficient convolution operations:

$$[\mathbf{x}_{conv}, \mathbf{x}_{res}] = \text{Linear}(\mathbf{x}_{seq}) \in \mathbb{R}^{B\times L\times(2d_{inner})}, \tag{2.8}$$

$$\tilde{\mathbf{x}}_k = \text{DepthwiseConv}_k(\mathbf{x}_{conv}), \quad k \in \{1, 3, 5\}. \tag{2.9}$$

The multi-scale convolution design captures temporal patterns at different resolutions: kernel size 1 focuses on point-wise transformations, kernel size 3 extracts short-term local patterns characteristic of bearing defects, and kernel size 5 identifies medium-range temporal correlations. These three parallel outputs are concatenated and integrated through a 1×1 convolution to produce the unified multi-scale feature $\mathbf{x}_{merge} \in \mathbb{R}^{B\times d_{inner}\times L}$:

$$\mathbf{x}_{merge} = \text{Conv}_{1\times1}(\text{Concat}(\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_3, \tilde{\mathbf{x}}_5)). \tag{2.10}$$

Selective state-space modeling: The unified multi-scale features are processed through an activation function (SiLU) and input to the selective state-space model for global temporal dependency modeling. The state-space parameters are dynamically generated from the input features to enable selective information processing:

$$[\delta, \mathbf{B}, \mathbf{C}] = \text{Linear}(\mathbf{x}_{merge}) \in \mathbb{R}^{B\times L\times(d_{t\_rank}+2d_{ff})}, \tag{2.11}$$

where δ is mapped to $\mathbb{R}^{B\times L\times d_{inner}}$ through a softplus function to control the state decay rate. The learnable parameters $\mathbf{A}$ (obtained by exponentiating $\mathbf{A}_{log}$) and $\mathbf{D}$ ensure numerical stability in logarithmic space during selective scanning.

Gated residual fusion: To balance the contributions of local multi-scale features and global state-space representations, an adaptive gating mechanism is employed. A gating vector $\mathbf{g}$ is computed from the multi-scale features to regulate the influence of the state-space model output:

$$\mathbf{g} = \sigma\left(\text{Linear}(\mathbf{x}_{merge})\right). \tag{2.12}$$

The final output integrates the gated state-space features with the residual pathway through element-wise operations:

$$\mathbf{y}_{final} = (\mathbf{g} \odot \mathbf{y}_{ssm}) \odot \text{SiLU}(\mathbf{x}_{res}), \tag{2.13}$$

where $\mathbf{y}_{ssm}$ represents the output from the selective state-space model, $\sigma(\cdot)$ denotes the sigmoid function, and the $\odot$ operator represents element-wise multiplication. The result is then projected back to the original dimension $\mathbb{R}^{B\times L\times d_{model}}$.

This integrated architecture achieves three objectives: 1) multi-scale depthwise convolutions capture diverse local temporal patterns with reduced computational overhead; 2) the log-parameterized selective state-space model enables stable, linear-time modeling of long-range dependencies; and 3) the adaptive gating mechanism balances local and global feature contributions to improve robustness to noise and strengthen fault discrimination.

2.5. Dynamic low-rank multi-head cross-attention for feature fusion

In multi-view scenarios, traditional cross-attention mechanisms require large matrix multiplications in high-dimensional spaces, leading to high computational costs and noise susceptibility. These limitations restrict model efficiency and robustness. To address these challenges, low-rank attention techniques are adapted to a cross-attention fusion module, termed dynamic low-rank multi-head cross-attention (DyLoMHCA). Rather than introducing a fundamentally new attention mechanism, DyLoMHCA builds on existing low-rank attention methods (e.g., low-rank adaptation, LoRA) and tailors them to cross-view fusion with dynamic rank selection and gated integration, aiming to reduce computational complexity while maintaining modeling flexibility.

The DyLoMHCA module employs multi-head attention to partition the hidden space into multiple subspaces, where each subspace independently learns query, key, and value representations to capture information from different feature domains. A dynamic low-rank projection mechanism [16] is incorporated into each attention head, introducing learnable dynamic rank selectors that enable adaptive rank selection through differentiable weighting of the base matrix columns.

To enhance cross-view feature integration, a global temporal summary is obtained by averaging the time-series features and is then injected into the image-space features. Finally, a gated fusion mechanism [17] performs element-wise fusion between the original image features and the cross-attention outputs. Through the learned gating vector γ, time-series information is incorporated into the image features, improving fusion efficiency and attention to critical information.

Theoretical formulation: Let the base low-rank factors per head be $\mathbf{U} \in \mathbb{R}^{d_{in} \times r_{\max}}$, $\mathbf{V} \in \mathbb{R}^{r_{\max} \times d_{out}}$. A differentiable weight vector $\mathbf{w} = \mathrm{softmax}(\boldsymbol{\alpha})$ induces a diagonal matrix $\mathbf{W} = \mathrm{diag}(\mathbf{w})$, yielding dynamic factors $\mathbf{U}_{dyn} = \mathbf{UW}$, $\mathbf{V}_{dyn} = \mathbf{WV}$. The projected queries, keys, and values are then

$$\mathbf{Q} = \mathbf{X}_s \, \mathbf{U}_{Qdyn} \, \mathbf{V}_{Qdyn}, \quad \mathbf{K} = \mathbf{X}_t \, \mathbf{U}_{Kdyn} \, \mathbf{V}_{Kdyn}, \quad \mathbf{V} = \mathbf{X}_t \, \mathbf{U}_{Vdyn} \, \mathbf{V}_{Vdyn}. \tag{2.14}$$

Standard scaled dot-product attention is applied per head, followed by concatenation and an output projection. Finally, a gating network combines the space feature, attention output, and a global temporal summary to obtain the fused representation.

2.5.1. Traditional multi-head low-rank projection

Let there be M attention heads (here num_heads = 4), and for each head, the input includes: image space features $\mathbf{x}_s \in \mathbb{R}^{d_s}$ and time-series features $\mathbf{x}_t \in \mathbb{R}^{L \times d_t}$, where L is the number of time steps and d_t is the feature dimension per time step. For the h-th head, three pairs of learnable base matrices $\mathbf{U}_Q^h, \mathbf{V}_Q^h, \mathbf{U}_K^h, \mathbf{V}_K^h, \mathbf{U}_V^h, \mathbf{V}_V^h$ are defined, whose dimensions are given by:

$$\mathbf{U}_Q^{(h)} \in \mathbb{R}^{d_s \times r_{\max}}, \quad \mathbf{V}_Q^{(h)} \in \mathbb{R}^{r_{\max} \times d_h}, \tag{2.15}$$

$$\mathbf{U}_K^{(h)} \in \mathbb{R}^{d_t \times r_{\max}}, \quad \mathbf{V}_K^{(h)} \in \mathbb{R}^{r_{\max} \times d_h}, \tag{2.16}$$

$$\mathbf{U}_V^{(h)} \in \mathbb{R}^{d_t \times r_{\max}}, \quad \mathbf{V}_V^{(h)} \in \mathbb{R}^{r_{\max} \times d_h}, \tag{2.17}$$

where $r_{\max}$ is the maximum rank and $d_h = \frac{d_{\text{hidden}}}{M}$ is the dimension of each head's hidden space. Taking $\mathbf{Q}$ as an example, the traditional low-rank projection can be written as:

$$\mathbf{Q}_h = \mathbf{x}_s \mathbf{U}_Q^{(h)} \mathbf{V}_Q^{(h)} \in \mathbb{R}^{d_h}. \tag{2.18}$$

In practice, the suitable rank can vary with the data distribution. If $r_{\max}$ is set too large, it may lead to overfitting or redundancy, while setting it too small may restrict the captured information. Therefore, dynamic rank selection is introduced in each head.

2.5.2. Dynamic rank selection and the multi-head attention mechanism

The implementation of the dynamic rank method follows these steps: for each head, the columns $\mathbf{U}$ and $\mathbf{V}$ are weighted in a differentiable manner to obtain the dynamic low-rank matrices $\mathbf{U}_{\text{dyn}}$ and $\mathbf{V}_{\text{dyn}}$. Let $\alpha \in \mathbb{R}^{r_{\max}}$ be a learnable parameter. The softmax function is applied to obtain $\mathbf{w}$, and then a diagonal matrix $\mathbf{W}$ is constructed, followed by weighted summation over the columns $\mathbf{U}$ and $\mathbf{V}$, as follows:

$$\mathbf{w} = \text{softmax}(\alpha) \in \mathbb{R}^{r_{\max}}, \quad w \geq 0, \quad \|\mathbf{w}\|_1 = 1, \tag{2.19}$$

$$\mathbf{W} = \text{diag}(\mathbf{w}) \in \mathbb{R}^{r_{\max} \times r_{\max}}, \tag{2.20}$$

$$\mathbf{U}_{\text{dyn}} = \mathbf{U}\mathbf{W} \in \mathbb{R}^{d_{\text{in}} \times r_{\max}}, \quad \mathbf{V}_{\text{dyn}} = \mathbf{W}\mathbf{V} \in \mathbb{R}^{r_{\max} \times d_{\text{out}}}. \tag{2.21}$$

This allows for adaptive rank allocation during training, enabling soft rank selection, where the contribution of each column can be increased or decreased. For the h-th head, the calculation of $\mathbf{Q}_h$, $\mathbf{K}_h$, and $\mathbf{V}_h$ is as follows:

$$\mathbf{U}_{Q,\text{dyn}}^{(h)} = \mathbf{U}_Q^{(h)} \mathbf{W}_Q^{(h)}, \quad \mathbf{V}_{Q,\text{dyn}}^{(h)} = \mathbf{W}_Q^{(h)} \mathbf{V}_Q^{(h)}, \tag{2.22}$$

$$\mathbf{Q}_h = \mathbf{x}_s \, \mathbf{U}_{Q,\text{dyn}}^{(h)} \, \mathbf{V}_{Q,\text{dyn}}^{(h)}, \tag{2.23}$$

$$\mathbf{K}_h = (\mathbf{x}_t)_{\text{flat}} \mathbf{U}_{K,\text{dyn}}^{(h)} \mathbf{V}_{K,\text{dyn}}^{(h)}, \tag{2.24}$$

$$\mathbf{V}_h = (\mathbf{x}_t)_{\text{flat}} \mathbf{U}_{V,\text{dyn}}^{(h)} \mathbf{V}_{V,\text{dyn}}^{(h)}, \tag{2.25}$$

where $(\mathbf{x}_t)_{\text{flat}} \in \mathbb{R}^{(B \times L) \times d_t}$ represents the flattened batch and time dimensions, after which the result is reshaped to $[B, L, d_h]$ for $\mathbf{K}_h$ and $\mathbf{V}_h$. Once $\mathbf{Q}_h$, $\mathbf{K}_h$, and $\mathbf{V}_h$ are obtained, the attention output for the h-th head is computed using the standard dot-product attention formula:

$$\text{Attn}_h(\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h) = \text{softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^T}{\sqrt{d_h}}\right) \mathbf{V}_h, \tag{2.26}$$

and the output of all heads is concatenated as:

$$\mathbf{O}_{mh} = \text{Concat}(\text{Attn}_1, \ldots, \text{Attn}_M) \in \mathbb{R}^{B \times d_{\text{hidden}}}. \tag{2.27}$$

Finally, a linear transformation is applied to map the output back to the image feature dimension d_s, as follows:

$$\mathbf{W}_{\text{out}} \in \mathbb{R}^{d_{\text{hidden}} \times d_s}, \tag{2.28}$$

$$\mathbf{o}_{\text{attn}} = \mathbf{O}_{mh}\, \mathbf{W}_{\text{out}} \in \mathbb{R}^{B \times d_s}, \tag{2.29}$$

In the cross-view scenario, to further inject the global trend of the time-series, the mean $\mathbf{m}_{\text{time}}$ of the time-series features is computed. Then, the original image features $\mathbf{x}_s$, attention outputs $\mathbf{o}_{\text{attn}}$, and $\mathbf{m}_{\text{time}}$ are concatenated into a single vector $\mathbf{z}$. Through a gating network (Gate FC), a gating vector $\gamma \in \mathbb{R}^{d_s}$ is learned, as shown in the following equations:

$$\mathbf{m}_{\text{time}} = \text{MeanPool}(\mathbf{x}_t) \in \mathbb{R}^{d_t}, \tag{2.30}$$

$$\mathbf{z} = \text{Concat}(\mathbf{x}_s, \mathbf{m}_{\text{time}}, \mathbf{o}_{\text{attn}}) \in \mathbb{R}^{d_s + d_t + d_s}, \tag{2.31}$$

$$\gamma = \sigma\left(\text{MLP}(\mathbf{z})\right), \tag{2.32}$$

$$\mathbf{x}_{\text{fused}} = (1 - \gamma) \odot \mathbf{x}_s + \gamma \odot \mathbf{o}_{\text{attn}}. \tag{2.33}$$

where $\sigma(\cdot)$ represents the element-wise sigmoid function, and $\odot$ denotes element-wise multiplication. This balances the original features and the injected cross-attention information. In summary, dynamic low-rank projections adapt the effective rank per head to reduce redundancy, the multi-head design partitions representation subspaces to capture complementary cross-view cues, and gated fusion balances original spatial features with cross-view injection to improve robustness. DyLoMHCA is therefore intended as an application-oriented adaptation of low-rank attention for cross-view bearing fault diagnosis rather than a fundamentally new attention paradigm.

3. Case study

3.1. Dataset description

In this study, two benchmark bearing fault datasets, Case Western Reserve University (CWRU) [18] and the Southeast University (SEU) drivetrain dataset used in prior transfer-learning diagnosis work [19], are employed for model evaluation. The CWRU dataset consists of vibration signal data with a sampling frequency of 12 kHz under 0 HP load from the drive end. The experimental setup of the CWRU bearing test rig is shown in Figure 3. Unnecessary outer race fault and other loading conditions are removed, leaving only the outer race fault at the 6 o'clock direction. The final dataset includes fault data for three types of faults (inner race, rolling element, and outer race) at defect sizes of 0.007, 0.014, and 0.021, as well as normal data, totaling 10 categories, as shown in Figure 4. The SEU

dataset is from Southeast University, where the speed system load is set to 20 Hz-0 V. The experimental setup of the SEU bearing test rig is shown in Figure 5. The bearing fault conditions are cracks on the rolling elements, cracks on the inner race, cracks on the outer race, and cracks on both the inner and outer races, totaling four types of faults, and normal motor vibration data, resulting in 5 categories, as shown in Figure 6.

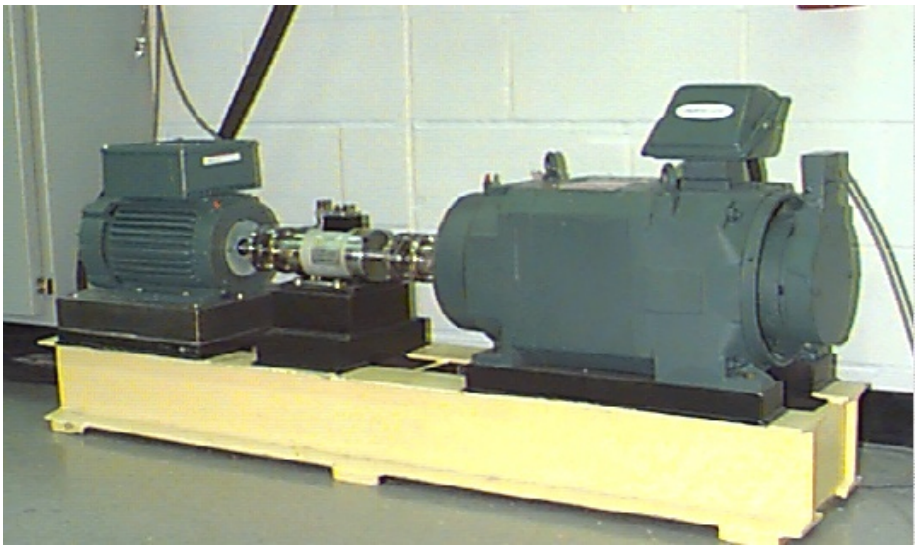

Figure 3. Experimental setup of the CWRU bearing test rig.

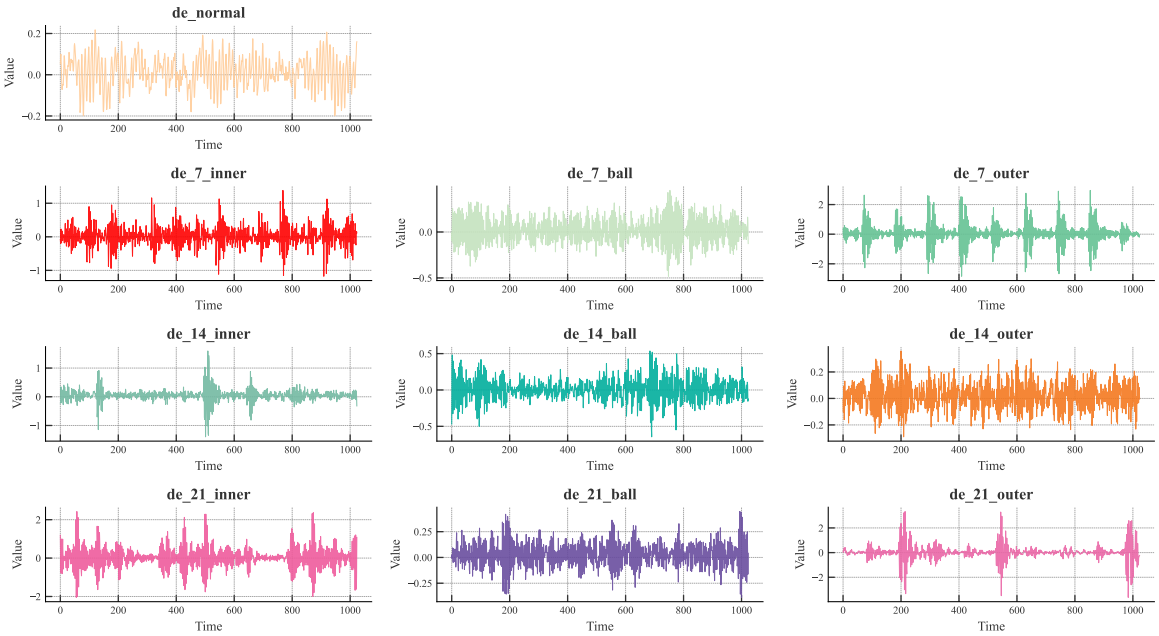

Figure 4. Representative CWRU vibration sequences for the considered fault categories (first 1024 points).

All experiments use non-overlapping datasets rebuilt from the raw signals. Each sample is a 1024-point vibration window. In the same-condition suite, each case is built from one operating condition only: the raw signal is split in chronological order at the raw-signal level into 70% training, 10% validation, and 20% test partitions, and only then segmented into windows. In the cross-condition suite, train, validation, and test data come from disjoint operating conditions or source–target condition pairs. For the grouped CWRU transfer cases and the SEU transfer cases, the source conditions are split chronologically into 90% training and 10% validation partitions, while the unseen target condition is reserved as the full test set. For the two CWRU single-transfer cases, one full operating condition is used for training, one for validation, and one unseen condition for testing. All cases use per-source train min–max normalization.

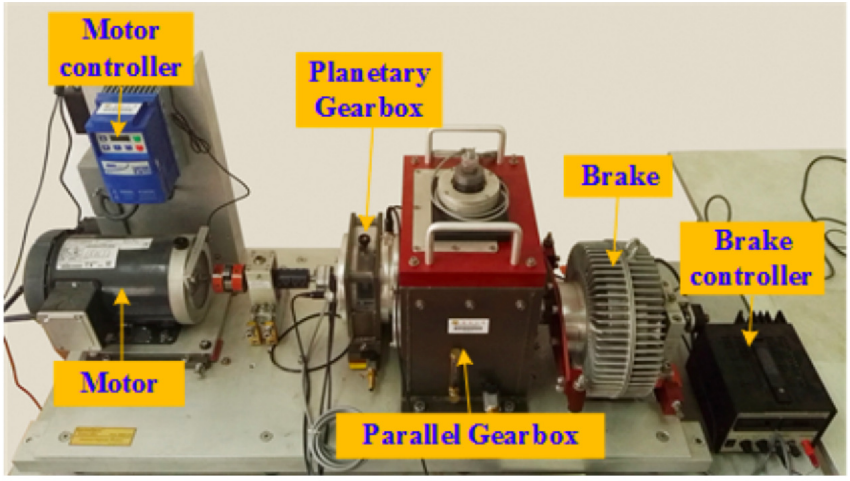

Figure 5. Experimental setup of the SEU bearing test rig.

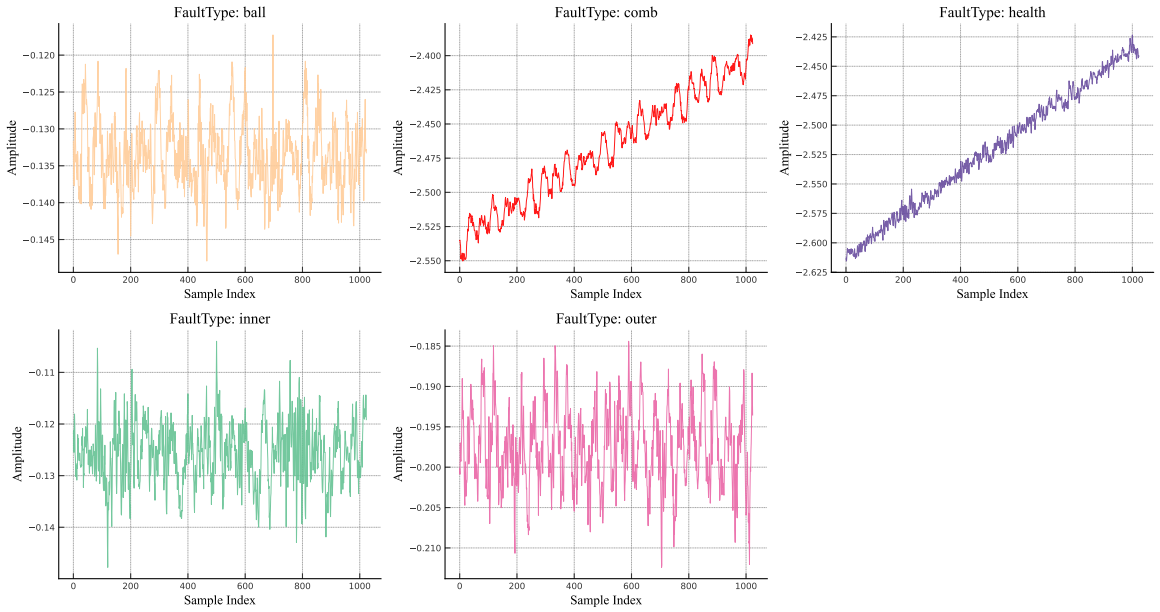

Figure 6. Representative SEU vibration sequences for the considered fault categories (first 1024 points).

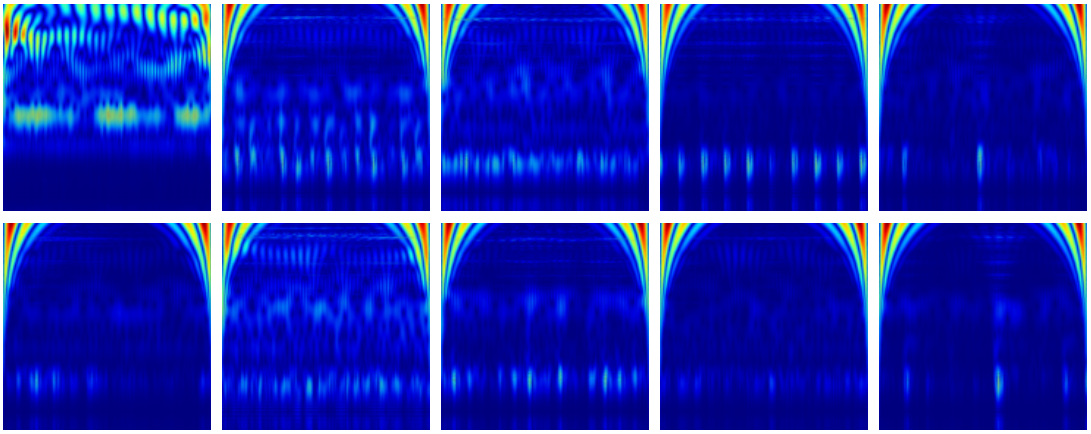

Figure 7. Representative CWT-based time–scale images from the CWRU bearing dataset.

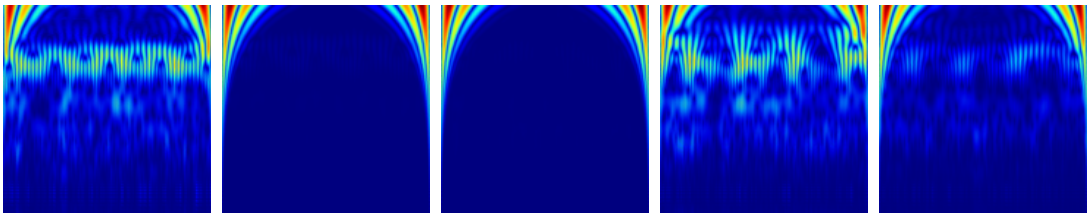

Figure 8. Representative CWT-based time–scale images from the SEU bearing dataset.

The same-condition and cross-condition cases are summarized in Tables 1 and 2. The CWRU cases contain 10 fault categories, including normal condition, inner-race faults at three defect sizes, rolling-element faults at three defect sizes, and outer-race faults at three defect sizes. The SEU cases contain 5 categories, namely, normal, inner-race fault, outer-race fault, rolling-element fault, and compound fault. All cases are class-balanced at the case level because each label is windowed and split with the same deterministic protocol inside each source condition.

Table 1. Same-condition evaluation cases and split sizes.

Case	Dataset	Operating condition	Split protocol	Train	Val	Test
S1_cwru_condition_0	CWRU	Condition 0	ordered raw 70/10/20	810	110	230
S2_cwru_condition_1	CWRU	Condition 1	ordered raw 70/10/20	810	110	230
S3_cwru_condition_2	CWRU	Condition 2	ordered raw 70/10/20	810	110	230
S4_cwru_condition_3	CWRU	Condition 3	ordered raw 70/10/20	810	110	230
S5_seu_20_0	SEU	20 Hz–0 V	ordered raw 70/10/20	3580	510	1020
S6_seu_30_2	SEU	30 Hz–2 V	ordered raw 70/10/20	3580	510	1020

For the main experiments, the image branch uses CWT-based time–scale images. Each standardized 1024-point vibration segment is transformed by a continuous wavelet transform with the Morlet wavelet, using geometrically spaced scales from 1 to 128. The coefficient magnitude is converted by $\log(1 + |\cdot|)$, mapped to a color image, and resized to 224×224. All train, validation, and test signals are processed identically, and representative CWT examples from the CWRU and SEU benchmarks

are shown in Figures 7 and 8. Alternative GAF and MTF representations are evaluated separately in the image-representation ablation.

Table 2. Cross-condition evaluation cases and split sizes.

Case	Dataset	Transfer setting	Split protocol	Train	Val	Test
g1_cwru_grouped_01_to_3	CWRU	conditions 0,1 → 3	source ordered raw 90/10, target full test	2100	220	1170
g2_cwru_grouped_23_to_0	CWRU	conditions 2,3 → 0	source ordered raw 90/10, target full test	2100	220	1170
g3_cwru_single_0_1_3	CWRU	train 0, val 1, test 3	disjoint full conditions	1170	1170	1170
g4_cwru_single_3_2_0	CWRU	train 3, val 2, test 0	disjoint full conditions	1170	1170	1170
g5_seu_20_0_to_30_2	SEU	20 Hz–0 V → 30 Hz–2 V	source ordered raw 90/10, target full test	4605	510	5115
g6_seu_30_2_to_20_0	SEU	30 Hz–2 V → 20 Hz–0 V	source ordered raw 90/10, target full test	4605	510	5115

3.2. Experimental settings

The experiments follow a unified multi-seed evaluation protocol. Unless otherwise stated, all main comparison models are trained with AdamW, learning rate 5×10^{-4}, weight decay 1×10^{-4}, no scheduler, 100 training epochs, and early stopping patience 25. Three random seeds $\{7, 42, 2024\}$ are used for each case. To keep the optimization budget comparable while avoiding out-of-memory failures across model families, the implementation uses model-aware safe per-device batch sizes together with gradient accumulation when necessary. The same-condition suite evaluates the 6 cases listed in Table 1, and the cross-condition suite evaluates the 6 transfer cases listed in Table 2. Table 3 summarizes the unified preprocessing, optimization, and hardware settings used in the experiments. The image branch of the main experiments uniformly uses 224×224 CWT images. The implementation is based on PyTorch, torchvision, pyts, scikit-learn, pandas, NumPy, THOP, and PyWavelets.

Table 3. Unified data processing, optimization, and hardware settings used in the experiments.

Item	Setting
Signal window length	1024 points
Window overlap	0%
Same-condition split	ordered raw-signal split within one operating condition: 70% train / 10% val / 20% test
Cross-condition grouped transfer	source condition(s) split in order: 90% train / 10% val; unseen target condition used as full test set
Cross-condition single transfer	one full operating condition for train, one for val, and one unseen condition for test
Normalization	per-source train min–max normalization
Default image representation	CWT with Morlet wavelet, scales 1–128, and log-magnitude mapping
Image size	224×224 RGB
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	1×10^{-4}
Scheduler	none
Maximum epochs	100
Early stopping	patience 25
Random seeds	$\{7, 42, 2024\}$
Batch policy	model-aware safe batch size with gradient accumulation when required
Precision	fp16
Software stack	PyTorch, torchvision, pyts, scikit-learn, pandas, NumPy, THOP, PyWavelets
Hardware	Linux workstation with 6 NVIDIA Tesla V100 GPUs (1×32 GB and 5×16 GB)

Architecture settings. The proposed model uses $d_{model} = 32$, expand = 4, $d_{ff} = 4$, branch kernels $\{1, 3, 5\}$, fusion hidden dimension = 64, 4 attention heads, and $r_{\max} = 8$. Unless otherwise stated, the Mamba branch uses 4 stacked residual blocks and dropout 0.3. The same core configuration is used across the same-condition, cross-condition, image-representation, and sensitivity experiments.

3.3. Evaluation metrics

To assess both predictive performance and computational efficiency, accuracy (Acc), precision (Prec), recall (Rec), macro F1-score (F1), Matthews correlation coefficient (MCC), parameter count (Params), multiply–accumulate operations (MACs), mean batch latency, and throughput/FPS are reported. Params and MACs serve as the primary architecture-level efficiency indicators because they are independent of runtime batching. Latency and throughput are reported as supplementary end-to-end profiling results produced by a unified evaluation pipeline under the same hardware environment. However, to avoid out-of-memory failures across heterogeneous model families, that pipeline uses model-aware safe per-run evaluation batch sizes rather than one fixed batch size for every baseline. Accordingly, latency/FPS should be interpreted as practical system-level measurements rather than strict kernel-level microbenchmarks. In the comparison tables, parameter counts are written as CWRU / SEU when the classifier head differs between the 10-class CWRU label space and the 5-class SEU label space. Peak GPU memory is not used as a primary cross-model comparison metric in the main text because it is not consistently retained in all summary files. Let TP, FP, FN, and TN denote true positives, false positives, false negatives, and true negatives, respectively. For multi-class classification, macro-averaging over classes is employed. The evaluation metrics are defined as follows:

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}, \tag{3.1}$$

$$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{3.2}$$

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{3.3}$$

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Prec} \cdot \mathrm{Rec}}{\mathrm{Prec} + \mathrm{Rec}}. \tag{3.4}$$

Here, TP represents true positives (correctly predicted positive samples), TN represents true negatives (correctly predicted negative samples), FP represents false positives (incorrectly predicted as positive), and FN represents false negatives (incorrectly predicted as negative). For multi-class classification problems, macro-averaging is used to compute these metrics across all fault classes. In addition, MCC is used because it remains informative under class imbalance and low-accuracy cross-condition settings. Params counts the total learnable parameters, MACs counts multiply–accumulate operations, mean batch latency measures the average inference time per evaluation batch, and throughput/FPS reports processed samples per second.

4. Comparative methods

This section compares the proposed model against five baseline models under the unified protocol. The same-condition suite covers the 6 cases in Table 1, while the cross-condition suite covers the 6

transfer cases in Table 2. Tables 4 and 5 report the aggregated same-condition and cross-condition results, respectively. Each case is evaluated with three random seeds, and the reported values are aggregated means and standard deviations. For bounded metrics such as accuracy, the standard-deviation term summarizes variability across runs and should not be interpreted as a feasible accuracy interval. Figures 9–12 summarize the same-condition, cross-condition, convergence, and efficiency comparisons.

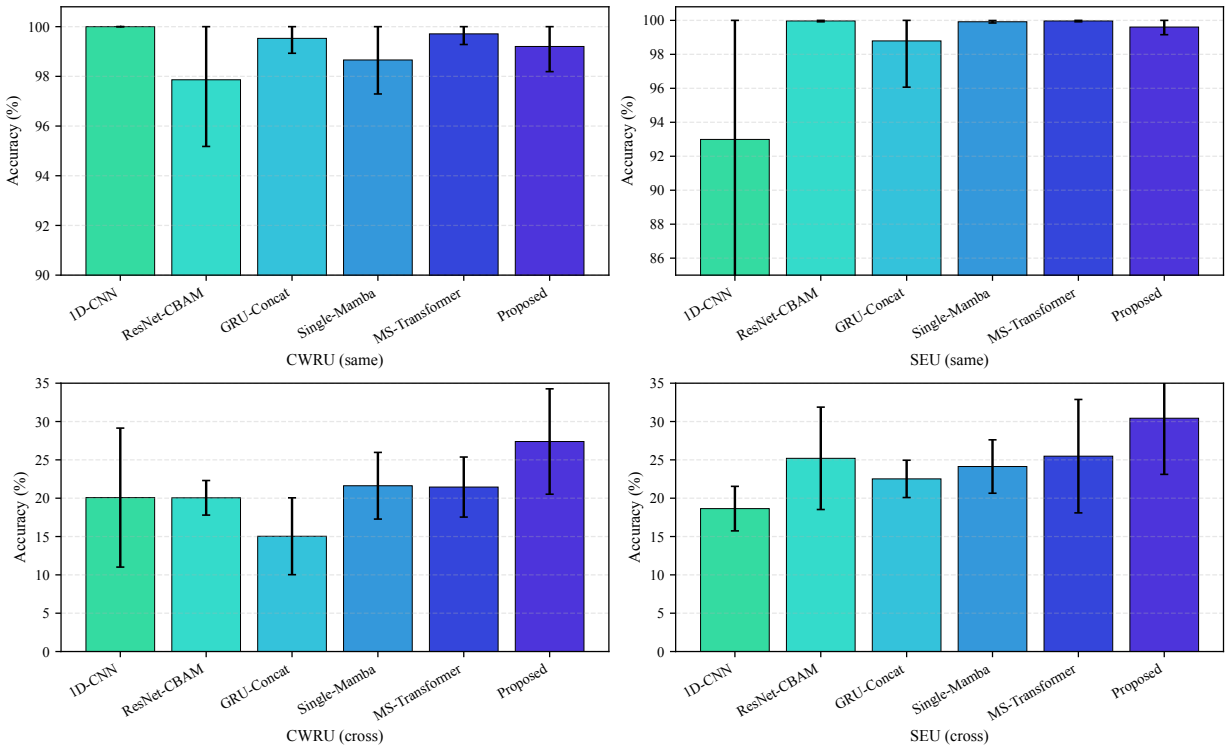

Figure 9. Accuracy comparison of the baseline models under same-condition and cross-condition evaluation.

Table 4. Same-condition main comparison aggregated over all cases and seeds.

Diagnosis method	CWRU Acc (%)	SEU Acc (%)	CWRU MCC	SEU MCC	Params	MACs	CWRU Lat. (ms)	SEU Lat. (ms)	CWRU FPS	SEU FPS
1D-CNN	100.00 ± 0.00	92.99 ± 8.37	1.0000	0.9190	61,610 / 60,965	20,825,472	0.97	1.12	122,331.3	118,689.9
ResNet-CBAM	97.86 ± 2.68	99.97 ± 0.05	0.9768	0.9996	48,082 / 47,917	234,113,720	39.55	41.32	2,926.8	3,097.9
GRU-Concat	99.53 ± 0.60	98.79 ± 2.72	0.9948	0.9857	440,402 / 438,957	246,688,952	22.60	23.99	2,570.1	2,660.0
Single-Mamba	98.66 ± 1.37	99.92 ± 0.07	0.9853	0.9990	547,664 / 547,284	487,943,964	41.02	40.59	702.7	785.4
MS-Transformer	99.71 ± 0.43	99.97 ± 0.05	0.9968	0.9996	556,970 / 556,805	426,467,000	35.03	35.01	822.8	911.4
Proposed	99.20 ± 1.01	99.61 ± 0.45	0.9912	0.9951	545,962 / 545,797	563,961,528	65.17	30.87	919.7	1,032.7

Table 5. Cross-condition main comparison aggregated over all transfer cases and seeds.

Diagnosis method	CWRU Acc (%)	SEU Acc (%)	CWRU MCC	SEU MCC	Params	MACs	CWRU Lat. (ms)	SEU Lat. (ms)	CWRU FPS	SEU FPS
1D-CNN	20.08 ± 9.06	18.64 ± 2.90	0.1489	-0.0337	61,610 / 60,965	20,825,472	0.97	1.03	122,924.5	127,877.6
ResNet-CBAM	20.05 ± 2.25	25.20 ± 6.68	0.1387	0.1009	48,082 / 47,917	234,113,720	37.98	41.50	3,086.2	3,081.1
GRU-Concat	15.04 ± 5.01	22.52 ± 2.43	0.0898	0.0452	440,402 / 438,957	246,688,952	23.25	23.95	2,659.1	2,673.1
Single-Mamba	21.62 ± 4.35	24.13 ± 3.48	0.1571	0.0918	547,664 / 547,284	487,943,964	44.77	42.47	709.8	753.5
MS-Transformer	21.45 ± 3.92	25.48 ± 7.40	0.1517	0.0985	556,970 / 556,805	426,467,000	37.98	36.93	837.8	866.6
Proposed	27.39 ± 6.87	30.43 ± 7.32	0.2189	0.1730	545,962 / 545,797	563,961,528	61.07	33.28	1,008.9	961.5

Table 6. Case-wise same-condition accuracy on the CWRU operating conditions.

Diagnosis method	S1 (cond. 0)	S2 (cond. 1)	S3 (cond. 2)	S4 (cond. 3)
1D-CNN	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
ResNet-CBAM	97.25 ± 1.81	99.13 ± 0.87	96.38 ± 5.19	98.70 ± 1.15
GRU-Concat	99.57 ± 0.75	99.71 ± 0.25	100.00 ± 0.00	98.84 ± 0.50
Single-Mamba	96.81 ± 1.65	99.42 ± 0.66	99.28 ± 0.25	99.13 ± 0.43
MS-Transformer	99.71 ± 0.25	100.00 ± 0.00	99.71 ± 0.50	99.42 ± 0.66
Proposed	99.71 ± 0.25	99.86 ± 0.25	99.13 ± 0.87	98.12 ± 1.40

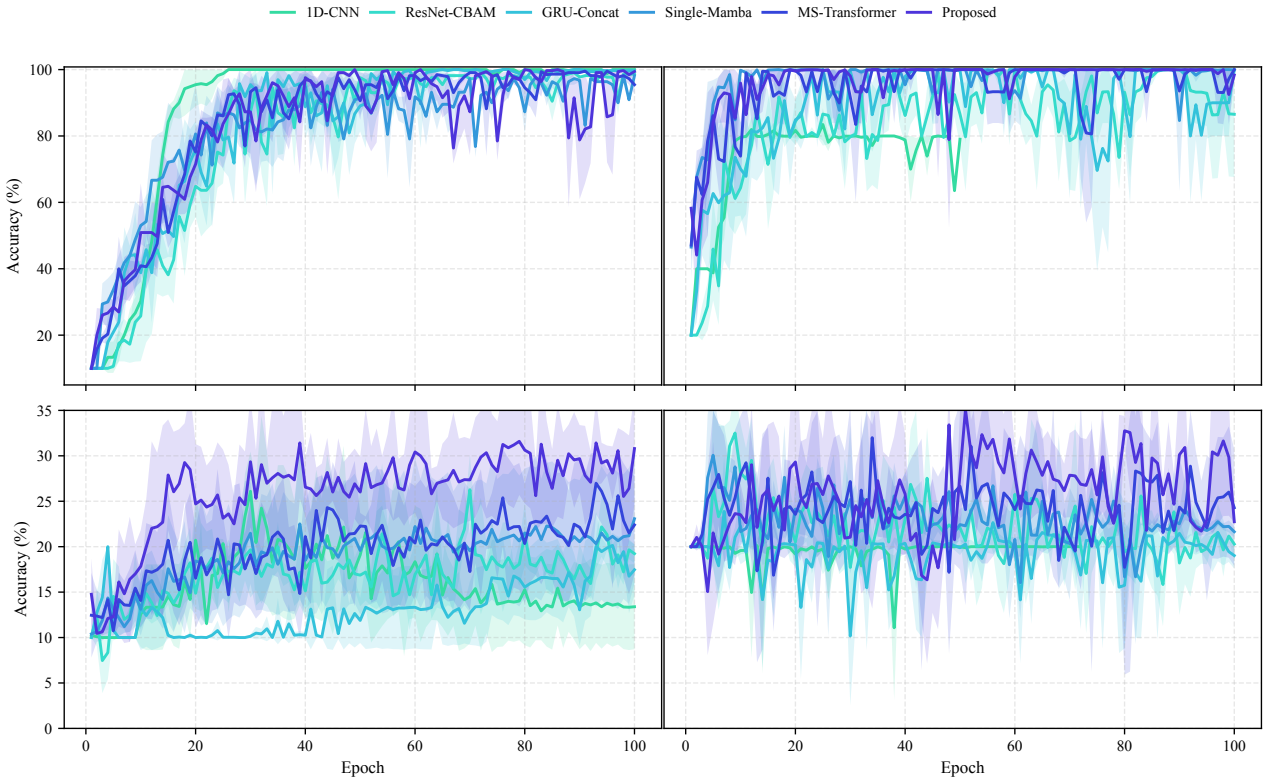

Figure 10. Three-seed mean accuracy trajectories of the baseline models on representative same-condition and cross-condition cases. Solid lines denote the mean over seeds {7,42,2024}, and shaded bands denote ±1 standard deviation.

The comparison leads to a nuanced conclusion. Under same-condition evaluation, the proposed model is high-performing but not uniformly best. In Table 4, it reaches 99.20% mean accuracy on CWRU with a standard deviation of 1.01%, below the higher same-condition averages of 1D-CNN and MS-Transformer; on SEU, it reaches 99.61% mean accuracy with a standard deviation of 0.45%, remaining competitive but slightly below the leading image-centric and transformer baselines. These results indicate that the model should be interpreted as competitive rather than as a uniformly leading same-condition classifier.

The cross-condition results provide more direct but still limited evidence for the proposed design. As reported in Table 5, the proposed model ranks first on all 6 transfer cases and also achieves the

highest dataset-level averages on both CWRU and SEU. Its cross-condition accuracy reaches 27.39% ± 6.87% on CWRU and 30.43% ± 7.32% on SEU, exceeding the second-highest baselines by 0.86–10.26 percentage points depending on the transfer case. Although these absolute cross-condition accuracies remain modest for all compared baselines, the proposed model improves the relative transfer results under condition shift more consistently than the comparison baselines. In other words, although the model is not the most accurate same-condition classifier in every setting, it shows more stable cross-condition behavior within the matched comparison when the operating condition shifts. Robust cross-condition bearing diagnosis nevertheless remains unresolved.

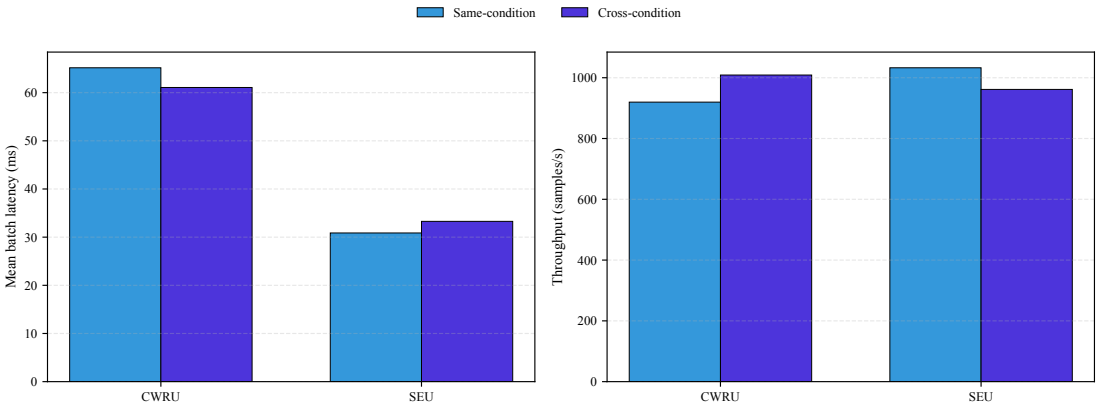

Figure 11. Latency and throughput of the proposed model under same-condition and cross-condition evaluation.

Table 7. Case-wise same-condition accuracy on the SEU operating conditions.

Diagnosis Method	S5 (20 Hz–0 V)	S6 (30 Hz–2 V)
1D-CNN	86.08 ± 5.63	99.90 ± 0.10
ResNet-CBAM	99.97 ± 0.06	99.97 ± 0.06
GRU-Concat	97.68 ± 3.85	99.90 ± 0.17
Single-Mamba	99.90 ± 0.10	99.93 ± 0.06
MS-Transformer	100.00 ± 0.00	99.93 ± 0.06
Proposed	99.61 ± 0.39	99.61 ± 0.60

Table 8. Case-wise cross-condition accuracy on the CWRU transfer cases.

Diagnosis Method	g1 (01→3)	g2 (23→0)	g3 (0/1/3)	g4 (3/2/0)
1D-CNN	13.39 ± 5.72	19.97 ± 9.92	20.00 ± 0.00	26.95 ± 13.52
ResNet-CBAM	17.75 ± 2.31	22.79 ± 1.59	19.60 ± 0.73	20.06 ± 0.20
GRU-Concat	19.77 ± 2.49	10.34 ± 3.45	13.45 ± 5.45	16.58 ± 4.04
Single-Mamba	20.85 ± 7.57	22.17 ± 4.58	23.13 ± 1.81	20.34 ± 3.90
MS-Transformer	22.02 ± 7.22	19.34 ± 0.44	21.54 ± 2.92	22.91 ± 3.64
Proposed	32.28 ± 6.00	23.65 ± 4.03	25.30 ± 9.78	28.35 ± 6.86

To avoid ambiguity about where these aggregated values come from, Tables 6–9 report the case-wise accuracies for all same-condition and cross-condition settings. This finer-grained view shows that

the proposed model remains near the ceiling on S1 and S2, but its CWRU same-condition average is pulled down by a modest drop on S4 and the remaining variance across CWRU conditions, although all four CWRU same-condition cases stay above 98%. On SEU, the two same-condition cases are much more stable for the proposed model, while the large variance of simpler baselines is concentrated in S5 rather than S6. Under cross-condition evaluation, the proposed model is the only compared baseline that stays in first place on all four CWRU transfer cases and both SEU transfer cases, which is the main empirical reason for emphasizing its transfer behavior rather than claiming same-condition dominance.

Figure 12. Normalized radar-chart summary of the baseline models over same-condition accuracy, cross-condition accuracy, cross-condition MCC, FPS, low-latency preference, and low-MAC preference. Higher values indicate better normalized performance on each axis.

The training-curve overview in Figure 10 provides complementary evidence about optimization behavior under the three-seed protocol. In same-condition diagnosis, most baselines converge rapidly to high validation accuracy, especially on SEU, which is consistent with the weak separation among the leading models in Table 4. In contrast, the cross-condition panels show a much larger gap between source-domain fitting and target-domain test accuracy, confirming that transfer robustness rather than

raw optimization difficulty is the main challenge. Within the displayed representative transfer cases, the proposed model tends to preserve relatively high mean target-domain trajectories and stable late-epoch accuracy bands among the baseline models, which is consistent with its higher aggregated cross-condition accuracy.

Table 9. Case-wise cross-condition accuracy on the SEU transfer cases.

Diagnosis method	g5 (20 Hz–0 V→30 Hz–2 V)	g6 (30 Hz–2 V→20 Hz–0 V)
1D-CNN	17.32 ± 3.98	19.96 ± 0.07
ResNet-CBAM	19.84 ± 1.31	30.56 ± 4.86
GRU-Concat	21.18 ± 2.24	23.86 ± 2.10
Single-Mamba	21.64 ± 0.77	26.62 ± 3.34
MS-Transformer	24.20 ± 7.88	26.76 ± 8.35
Proposed	28.01 ± 10.28	32.85 ± 3.29

Table 10. Aggregated noise-robustness summary across the same-condition and cross-condition suites.

Scenario	Method	Clean (%)	20 dB (%)	16 dB (%)	12 dB (%)	10 dB (%)	Noisy mean (%)
CWRU same	Proposed	99.20	27.21	19.20	14.93	13.22	18.64
CWRU same	MS-Transformer	99.71	28.51	21.01	16.74	14.96	20.31
CWRU same	GRU-Concat	99.53	46.78	33.15	27.72	25.69	33.33
CWRU same	1D-CNN	100.00	63.26	42.90	29.93	23.91	40.00
SEU same	Proposed	99.61	48.07	41.44	31.26	26.21	36.74
SEU same	MS-Transformer	99.97	54.08	42.06	32.12	28.07	39.08
SEU same	GRU-Concat	98.79	62.39	60.85	52.66	42.99	54.72
SEU same	Image-Only	99.97	53.07	47.08	38.92	31.70	42.69
CWRU cross	Proposed	27.39	25.75	24.05	22.18	20.95	23.23
CWRU cross	MS-Transformer	21.45	19.91	18.45	17.83	17.68	18.47
CWRU cross	1D-CNN	20.08	19.92	19.58	18.28	18.04	18.96
CWRU cross	Image-Only	20.05	19.72	18.48	17.04	16.37	17.90
SEU cross	Proposed	30.43	20.63	19.99	21.81	21.11	20.88
SEU cross	Image-Only	25.20	21.25	21.28	20.41	20.19	20.78
SEU cross	GRU-Concat	22.52	21.20	20.40	20.22	19.79	20.40
SEU cross	MS-Transformer	25.48	20.86	18.68	19.62	19.89	19.76

From the efficiency perspective, the proposed model is neither the smallest nor the fastest baseline under the evaluation settings used here. Its batch-size-independent footprint is about 5.46×10^5 parameters and 5.64×10^8 MACs, which are larger than those of the lightweight 1D-CNN and image-only baselines. The latency and FPS numbers in Tables 4 and 5, together with Figure 11, should be read as supplementary end-to-end profiling results produced by the same evaluator on the same hardware but with model-aware safe evaluation batch sizes, rather than as strict fixed-batch microbenchmarks. Under this practical profiling view, the extra cost of the proposed model is accompanied by improved cross-condition accuracy. Figure 12 summarizes this point visually: once the baseline models are normalized across accuracy, generalization, and efficiency dimensions, the proposed method does not dominate every axis, but it occupies a balanced overall profile. Therefore, the main comparison is best

interpreted as an accuracy–efficiency–generalization trade-off rather than a broad superiority claim.

4.1. Noise robustness under matched evaluation

To further examine performance degradation under signal corruption, the trained checkpoints were additionally evaluated on clean data and four additive Gaussian-noise levels (20, 16, 12, and 10 dB). This noise benchmark does not retrain the models. Instead, it reuses the same trained checkpoints from the same-condition and cross-condition suites, applies a shared noise seed, and evaluates all variants with the same evaluation pipeline. Figure 13 and Table 10 summarize the aggregated accuracy trends.

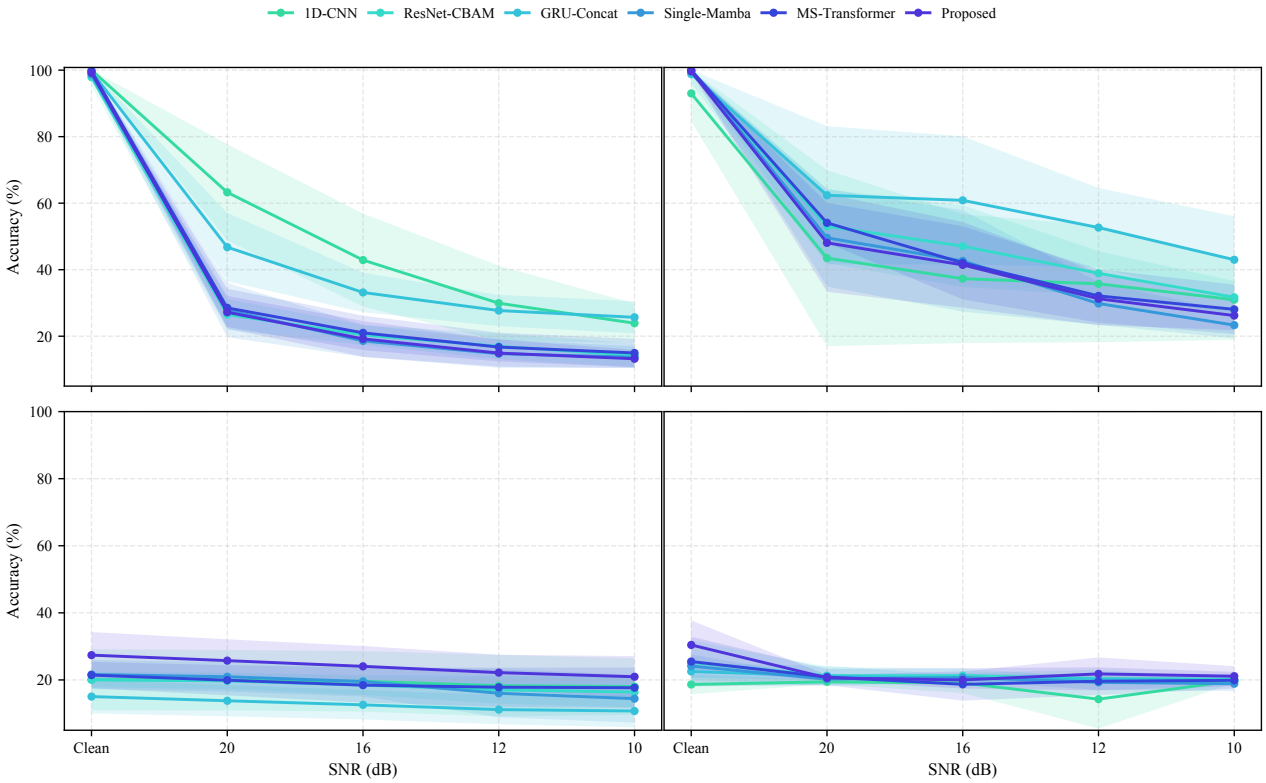

Figure 13. Three-seed mean noise-robustness trends of the baseline models under clean, 20, 16, 12, and 10 dB evaluation.

The noise results show that robustness depends strongly on the evaluation regime. Under same-condition diagnosis, the proposed model preserves the near-ceiling clean accuracy already observed in the main tables, but it is not the most noise-tolerant baseline after substantial corruption. On CWRU same-condition cases, the lightweight 1D-CNN yields the highest noisy mean accuracy, while on SEU same-condition cases, the GRU-Concat baseline yields the highest noisy mean. This indicates that the proposed same-source multi-view model should not be treated as the default anti-noise reference for easier within-condition recognition.

The cross-condition noise results are more consistent with the main interpretation of this study. On CWRU cross-condition transfers, the proposed model remains first in both clean and noisy averages and preserves a visible margin over the comparison baselines across the evaluated signal-to-noise ratio (SNR) range. On SEU cross-condition transfers, the margin is narrower, but the proposed model still

reaches the highest clean average and a comparable noisy mean. When operating-condition shift and additive noise are considered together, the proposed model continues to exhibit a balanced relation between cross-domain generalization and noise resilience among the baseline models. This is consistent with the broader view that the main value of the proposed framework lies in stable cross-condition behavior rather than in saturating easier same-condition settings.

4.2. Comparison with published methods

Cross-paper comparison should be interpreted cautiously. Different studies often adopt different split strategies, preprocessing pipelines, image representations, and reporting conventions, and several published methods do not provide directly comparable complexity or runtime details. For this reason, no ranked claim is made against heterogeneous literature baselines here. Instead, the proposed method is positioned as a compact same-source multi-view model whose main evidence lies in the internally matched cross-condition comparison.

5. Ablation and parameter sensitivity study

To systematically evaluate the contribution of each component and the robustness of the adopted structural settings, a series of ablation experiments and supplementary parameter sensitivity studies are conducted. The ablations assess the individual roles of the multi-scale gating Mamba module, the DyLoMHCA module, and the same-source multi-view framework in bearing fault diagnosis. The sensitivity study further examines how the performance changes when key structural parameters in the sequence branch and fusion module are perturbed around a fixed anchor configuration.

5.1. Ablation results on different datasets

The image-representation ablation is reflected not only in the final accuracy averages but also in the three-seed mean training trajectories. Table 11 and Figures 14 and 15 show that the CWT retains the strongest overall validation behavior among the CWT/GAF/MTF choices. The separation is especially clear on CWRU, where the CWT maintains a visible margin over GAF and MTF, while on SEU, all three encodings remain near the ceiling and the CWT still gives the highest dataset-level average. Taken together, these observations motivate the use of the CWT as the default image view in the main experiments.

Table 11. Image-representation ablation of the proposed model under same-condition evaluation.

Image type	CWRU Acc (%)	SEU Acc (%)	CWRU MCC	SEU MCC	CWRU Thr.	SEU Thr.
CWT	98.51 ± 1.68	99.92 ± 0.16	0.9837	0.9990	858.0	838.8
GAF	88.41 ± 7.09	99.56 ± 0.30	0.8726	0.9945	812.0	696.4
MTF	87.39 ± 4.94	99.74 ± 0.27	0.8615	0.9967	893.5	776.2

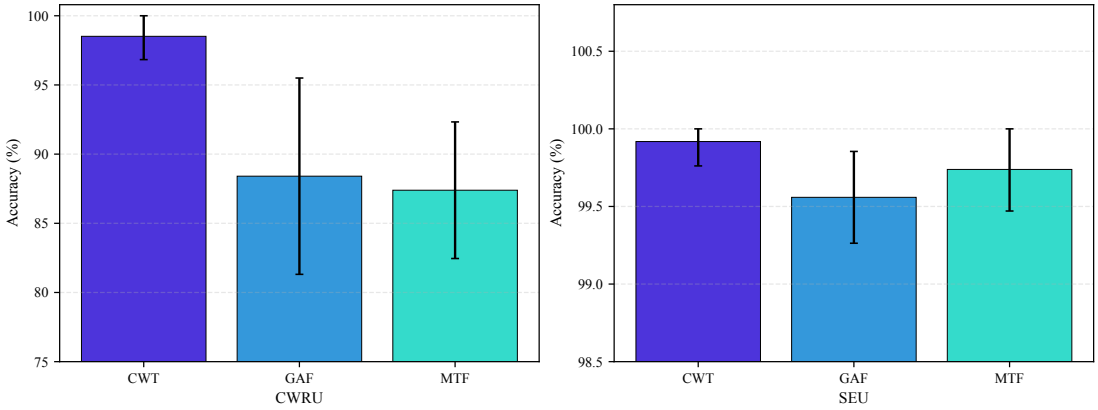

Figure 14. Same-condition accuracy summary of the CWT, GAF, and MTF image representations.

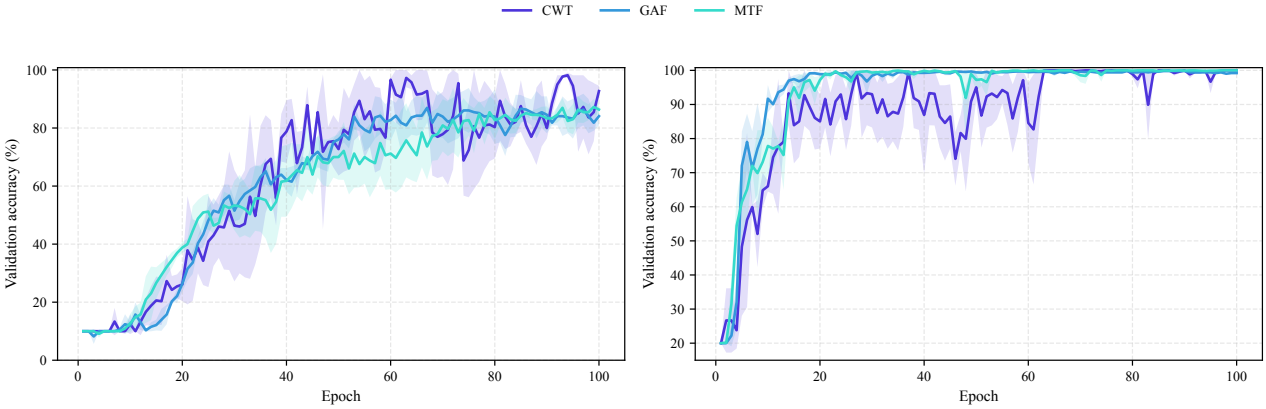

Figure 15. Three-seed mean validation-accuracy trajectories of the proposed model under the CWT, GAF, and MTF image-representation ablation settings. Solid lines denote the mean over seeds {7,42,2024}, and shaded bands denote ±1 standard deviation.

5.1.1. Effect and role of multi-scale gating Mamba

Table 12. Representative structural ablation summary aggregated over all same-condition cases and seeds.

Group	Variant	CWRU Acc (%)	SEU Acc (%)	Params	MACs	CWRU Lat.	SEU Lat.	Main Observation
Framework	Signal-only Mamba	82.75 ± 6.56	97.17 ± 1.92	218 k	56 M	32.63	27.64	lower sequence-only baseline
Framework	Signal-only MSGM	90.33 ± 4.21	98.27 ± 1.24	488 k	330 M	43.75	36.02	improved sequence branch summary
Framework	Image-only ResNet-CBAM	94.31 ± 9.89	99.98 ± 0.04	48 k	234 M	37.42	44.21	high SEU image-only accuracy
Framework	Fusion + Concat	98.91 ± 1.09	99.89 ± 0.28	536 k	564 M	69.61	83.40	fusion remains above single branches
Fusion	Concat	98.99 ± 0.97	99.84 ± 0.26	536 k	564 M	51.37	89.94	simple lower-cost fusion
Fusion	DyLoMHCA	99.35 ± 0.73	99.97 ± 0.05	546 k	564 M	83.89	122.52	highest CWRU fusion average
Fusion	FullRank MHCA	99.28 ± 0.99	99.98 ± 0.04	547 k	568 M	98.16	99.25	similar accuracy with slightly higher cost
MSGM	Single-scale	97.25 ± 2.23	99.97 ± 0.05	279 k	292 M	51.95	56.40	lower CWRU average than multi-scale
MSGM	Multi-scale	99.46 ± 0.38	99.97 ± 0.05	546 k	564 M	66.18	92.36	highest MSGM average in this summary
MSGM	Kernels {1,3}	98.80 ± 1.16	99.92 ± 0.11	477 k	494 M	80.68	96.29	reduced scale diversity lowers CWRU average
MSGM	Kernels {3,5}	99.38 ± 0.60	99.80 ± 0.28	479 k	496 M	74.77	103.76	competitive subset variant

The structural ablation summary in Table 12 shows that the multi-scale sequence branch contributes

differently across datasets. When the basic signal-only Mamba is replaced by the multi-scale gated version, the average CWRU accuracy rises from 82.75% to 90.33%, while the SEU signal-only accuracy rises from 97.17% to 98.27%. This pattern suggests that multi-scale local modeling is especially relevant on the more difficult CWRU same-condition cases, where local transients and longer temporal dependencies both matter.

5.1.2. Effect and role of the DyLoMHCA module

The fusion ablation shows that DyLoMHCA does not produce the same gain over every alternative, but it provides a practical balance between accuracy and complexity. On the same-condition suite, DyLoMHCA reaches 99.35% ± 0.73% on CWRU and 99.97% ± 0.05% on SEU. It improves over simple concatenation in this summary, yields the highest average CWRU fusion accuracy, and remains essentially tied with the slightly heavier full-rank MHCA on SEU. At the same time, DyLoMHCA avoids the MAC increase of the full-rank variant, while its latency behavior remains case-dependent rather than universal. Overall, this behavior is consistent with describing DyLoMHCA as a task-oriented low-rank fusion choice rather than a fundamentally new attention mechanism.

5.1.3. Effect and role of the multi-view framework

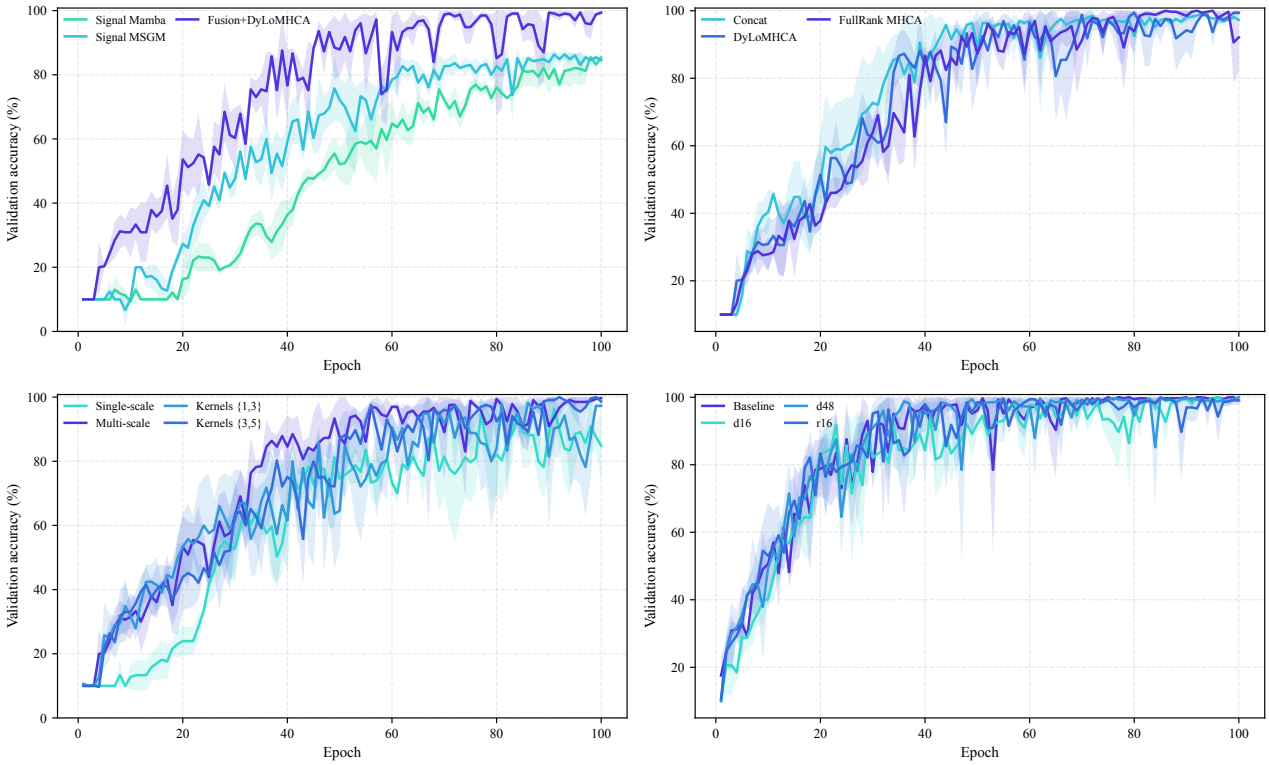

Figure 16. Three-seed mean validation-accuracy overview of the main structural ablation groups on representative CWRU cases. Solid lines denote the mean over seeds {7,42,2024}, and shaded bands denote ±1 standard deviation.

The proposed framework fuses two synchronized representations derived from the same sensor data: the time-domain signal and its CWT image. The ablation shows that the fused framework improves

over the individual signal-only branches and remains competitive with the image-only branch. On CWRU, the fusion models rise well above both signal-only Mamba and image-only ResNet-CBAM. On SEU, the image branch is already very strong, but the fused framework remains competitive while also linking more naturally to the cross-condition evidence in Section 4. These observations are consistent with the usefulness of same-source dual-representation fusion, especially when the objective extends beyond easier same-condition classification.

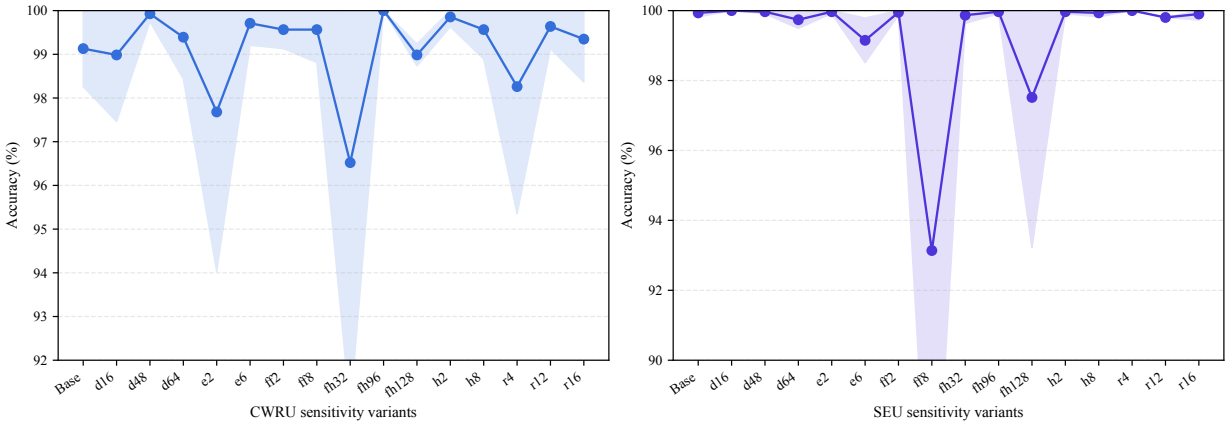

Figure 17. Sensitivity of same-condition accuracy to representative hyperparameter changes in the proposed model.

The structural-ablation overview in Figure 16 adds an optimization-level view to the summary tables while avoiding a proliferation of separate figures. The framework panel shows that the fused model maintains higher validation trajectories than the signal-only branches on the more difficult CWRU cases. The fusion panel indicates that DyLoMHCA and full-rank MHCA reach similar final operating regions, but DyLoMHCA preserves comparable convergence without the additional full-rank cost. The MSGM panel further shows that removing scale diversity mainly affects CWRU convergence stability rather than only lowering the terminal average, which is consistent with the adopted multi-scale design. Figure 17 likewise shows that the adopted setting lies in a stable neighborhood rather than at a fragile isolated optimum.

5.2. Parameter sensitivity analysis

To further examine the robustness of the adopted structural settings, one-at-a-time sensitivity experiments were conducted on the same-condition suite using the same default CWT image input and the same training budget as the structural-ablation experiments. The reference configuration is $d_{model} = 32$, $expand = 4$, $d_{ff} = 4$, fusion hidden dimension = 64, 4 attention heads, $r_{\max} = 8$, branch kernels $\{1, 3, 5\}$, AdamW, no scheduler, and the default safe batch setting used in the experiments. Unlike a single-seed sensitivity summary, the summary below is aggregated over all completed same-condition cases and seeds.

Table 13 suggests that the main configuration lies in a broad high-performing region rather than at an isolated optimum. On CWRU, increasing d_{model} from 32 to 48 gives the highest pooled average, whereas reducing *expand* to 2 or shrinking the fusion hidden dimension to 32 leads to lower performance. On SEU, most variants remain close to ceiling performance, but more aggressive changes such

as $d_{ff} = 8$ or fusion hidden dimension 128 are accompanied by weaker stability. The rank and head settings also remain fairly stable near the default choice: $r_{\max} = 8$, $r_{\max} = 12$, and 2–8 heads all remain competitive, whereas smaller hidden width and some width changes are more harmful. Taken together, these results are consistent with the adopted $d_{model} = 32$, $expand = 4$, $d_{ff} = 4$, fusion hidden dimension 64, 4 heads, and $r_{\max} = 8$ settings as a reasonable practical compromise.

Table 13. Parameter sensitivity of the proposed model under the same-condition sensitivity suite.

Category	Variant setting	CWRU accuracy (%)	SEU accuracy (%)
Reference setting	$d_{model} = 32$, $expand = 4$, $d_{ff} = 4$, heads = 4, $r_{\max} = 8$, kernels = $\{1, 3, 5\}$	99.13 ± 0.87	99.93 ± 0.11
Width	$d_{model} = 16$	98.99 ± 1.52	100.00 ± 0.00
Width	$d_{model} = 48$	99.93 ± 0.18	99.97 ± 0.06
Width	$d_{model} = 64$	99.39 ± 0.95	99.74 ± 0.25
Width	$expand = 2$	97.68 ± 3.65	99.97 ± 0.05
Width	$expand = 6$	99.71 ± 0.50	99.15 ± 0.64
Width	$d_{ff} = 2$	99.57 ± 0.43	99.95 ± 0.08
Width	$d_{ff} = 8$	99.57 ± 0.75	93.14 ± 9.63
Fusion	hidden = 32	96.52 ± 5.65	99.87 ± 0.24
Fusion	hidden = 96	100.00 ± 0.00	99.97 ± 0.05
Fusion	hidden = 128	98.99 ± 0.25	97.52 ± 4.30
Fusion	heads = 2	99.86 ± 0.22	99.97 ± 0.06
Fusion	heads = 8	99.57 ± 0.67	99.93 ± 0.11
Fusion	$r_{\max} = 4$	98.26 ± 2.91	100.00 ± 0.00
Fusion	$r_{\max} = 12$	99.64 ± 0.51	99.80 ± 0.00
Fusion	$r_{\max} = 16$	99.35 ± 0.98	99.90 ± 0.17

The sensitivity behavior is shown directly in Figure 17. The higher-performing sensitivity variants tend to converge within similar epoch ranges, while the weaker settings mainly differ by producing noisier validation trajectories or a lower late-epoch accuracy band. This behavior is consistent with interpreting the adopted parameter setting not as a fragile point estimate, but as a stable operating choice inside a broader acceptable region.

6. Conclusions

This study presents a compute-conscious same-source multi-view bearing fault diagnosis model based on a multi-scale gated Mamba branch and DyLoMHCA fusion. Under the unified CWT-based protocol, the model remains highly competitive in same-condition diagnosis and reaches the highest cross-condition averages in the internal comparison on both CWRU and SEU. The comparison is therefore best read as a balanced trade-off between relative cross-condition generalization and computational cost among the internally matched baselines, rather than as evidence of uniform same-condition superiority or absolute low complexity. The ablation and sensitivity analyses are likewise consistent with positive contributions from the multi-scale sequence branch, the low-rank fusion design, and the dual-representation framework.

At the same time, the claims of this study should be interpreted with appropriate caution. First, although cross-condition evaluation strengthens the empirical basis, the present validation is still limited to benchmark datasets rather than real industrial deployment data. Second, the two branches are derived

from the same vibration source, so the framework is more accurately described as same-source multi-view fusion than as heterogeneous multi-sensor multimodality. Third, absolute cross-condition accuracies remain modest for all compared models, so robust cross-condition bearing diagnosis remains unresolved. Fourth, the model is not the smallest or the fastest baseline, so the compute-conscious framing of this study is better understood as a practical trade-off than as an absolute minimum-cost solution.

Future work should therefore extend the evaluation to real industrial deployment data, broader operating-condition diversity, stricter split strategies, and cross-dataset transfer, together with fuller deployment-oriented efficiency measurements such as CPU/edge-device latency and memory footprint. It will also be valuable to analyze the CWT configuration choices more explicitly, to further test low-rank fusion under harsher domain shifts, and to expand the current benchmark evidence toward harsher and noisier industrial conditions.

Code availability

The code used in this study is available from the corresponding author upon reasonable request.

Use of AI tools declaration

During the preparation of this work, the authors used GPT-4o for language refinement and code-related assistance. The authors reviewed and edited all generated material and take full responsibility for the final content.

Acknowledgments

This work was supported by grants from the Key Project of the Natural Science Foundation of Ningxia (No. 2024AAC02033), National Natural Science Foundation (No. 12362005), Ningxia Higher Education First-Class Discipline Construction Funding Project (NXYLXK2017B09), Major Special Project of North Minzu University (No. ZDZX201902), Science and Technology Innovation Team of Ningxia (2025CXTD023), and 2025 Graduate Innovation Project of North Minzu University under grant (CYX25143).

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. P. Gupta, M. Pradhan, Fault detection analysis in rolling element bearing: A review, *Mater. Today Proc.*, **4** (2017), 2085–2094. https://doi.org/10.1016/j.matpr.2017.02.054
2. M. Hakim, A. A. B. Omran, A. N. Ahmed, M. Al-Waily, A. Abdellatif, A systematic review of rolling bearing fault diagnoses based on deep learning and transfer learning: Taxonomy, overview, application, open challenges, weaknesses and recommendations, *Ain Shams Eng. J.*, **14** (2023), 101945. https://doi.org/10.1016/j.asej.2022.101945

3. B. Peng, Y. Bi, B. Xue, M. Zhang, S. Wan, A survey on fault diagnosis of rolling bearings, *Algorithms*, **15** (2022), 347. https://doi.org/10.3390/a15100347
4. A. U. Rehman, W. Jiao, Y. Jiang, J. Wei, M. Sohaib, J. Sun, et al., Deep learning in industrial machinery: A critical review of bearing fault classification methods, *Appl. Soft Comput.*, **171** (2025), 112785. https://doi.org/10.1016/j.asoc.2025.112785
5. W. Zhang, C. Li, G. Peng, Y. Chen, Z. Zhang, A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load, *Mech. Syst. Signal Process.*, **100** (2018), 439–453. https://doi.org/10.1016/j.ymssp.2017.06.022
6. L. Wen, X. Li, L. Gao, Y. Zhang, A new convolutional neural network-based data-driven fault diagnosis method, *IEEE Trans. Ind. Electron.*, **65** (2018), 5990–5998. https://doi.org/10.1109/TIE.2017.2774777
7. X. Dai, K. Yi, F. Wang, C. Cai, W. Tang, Bearing fault diagnosis based on POA-VMD with GADF-Swin transformer transfer learning network, *Measurement*, **238** (2024), 115328. https://doi.org/10.1016/j.measurement.2024.115328
8. P. Bao, W. Yi, Y. Zhu, Y. Shen, B. X. Chai, STHFD: Spatial–temporal hypergraph-based model for aero-engine bearing fault diagnosis, *Aerospace*, **12** (2025), 612. https://doi.org/10.3390/aerospace12070612
9. Y. Li, K. H. Bwar, R. Chai, K. M. Tse, B. X. Chai, Adaptive multi-view hypergraph learning for cross-condition bearing fault diagnosis, *Mach. Learn. Knowl. Extr.*, **7** (2025), 147. https://doi.org/10.3390/make7040147
10. P. Kumar, I. Raouf, J. Song, H. S. Kim, Multi-size wide kernel convolutional neural network for bearing fault diagnosis, *Adv. Eng. Software*, **198** (2024), 103799. https://doi.org/10.1016/j.advengsoft.2024.103799
11. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Advances in Neural Information Processing Systems*, **30** (2017).
12. A. Gu, K. Goel, C. Re, Efficiently modeling long sequences with structured state spaces, in *International Conference on Learning Representations*, (2022), 1–27.
13. A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, in *First Conference on Language Modeling*, (2024), 1–32.
14. P. Wang, Y. Song, X. Wang, Q. Xiang, MD-BiMamba: An aero-engine inter-shaft bearing fault diagnosis method based on Mamba with modal decomposition and bidirectional features fusion strategy, *Measurement*, **242** (2025), 115870. https://doi.org/10.1016/j.measurement.2024.115870
15. S. Woo, J. Park, J. Y. Lee, I. S. Kweon, CBAM: Convolutional block attention module, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
16. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, et al., Lora: Low-rank adaptation of large language models, in *International Conference on Learning Representations*, **1** (2022), 3.
17. J. Arevalo, T. Solorio, M. Montes-y-Gómez, F. A. González, Gated multimodal units for information fusion, preprint, arXiv:1702.01992.

18. *Case Western Reserve University Bearing Data Center*, *Welcome to the Case Western Reserve University Bearing Data Center Website*, 2026. Available from: https://engineering.case.edu/bearingdatacenter/welcome.

19. S. Shao, S. M. McAleer, R. Yan, P. Baldi, Highly accurate machine fault diagnosis using deep transfer learning, *IEEE Trans. Ind. Inf.*, **15** (2019), 2446–2455. https://doi.org/10.1109/TII.2018.2864759

AIMS Press