fractal and fractional

MDPI

Article

Fractional-Order-Enhanced Dual-View Representation and VibrMamba–VMamba Collaborative Modeling for Gearbox Fault Diagnosis

Fengyun Xie [1,2,*], Kang Niu [2], Zeyan Song [2], Shulei Wang [2], Huihang Chen [2] and Ying Cao [2]

[1] Key Laboratory of Intelligent Operation and Maintenance Technology and Equipment for Transportation, Ministry of Education, East China Jiaotong University, Nanchang 330013, China
[2] School of Mechanical, Electrical, and Vehicle Engineering, East China Jiaotong University, Nanchang 330013, China; 19856112979@163.com (K.N.); 18086732725@163.com (Z.S.); wangtxue1014@163.com (S.W.); 18370203569@163.com (H.C.); 19567141626@163.com (Y.C.)
* Correspondence: 1493@ecjtu.edu.cn

Abstract

Gearbox fault diagnosis under controlled bench-test conditions with known speed variations and noise interference remains challenging because nonstationarity, background noise, and operating-condition fluctuations can easily submerge weak localized fault features. To address this issue, this study proposes a fault diagnosis method based on a fractional-order-enhanced dual-view representation and VibrMamba–VMamba collaborative modeling. First, this study introduces a Grünwald–Letnikov fractional-order differential enhancement module with a fractional order of $\alpha = 0.6$ to strengthen fault-sensitive impulsive components and improve the representation of nonstationary vibration signals. The framework then uses the enhanced signal to construct dual-view inputs: a fractional-order-enhanced one-dimensional vibration sequence and a fractional-order-enhanced synchrosqueezing transform (SST) time–frequency image. Subsequently, the framework constructs a VibrMamba temporal branch and a VMamba visual branch to extract dynamic temporal features and global structural features, respectively. Instead of using simple feature concatenation, this study designs a sample-adaptive collaborative fusion mechanism with gated weighting and cross-branch residual enhancement to integrate complementary temporal–visual representations. Bench-level experiments show that the proposed method achieves 98.90% diagnostic accuracy under clean test conditions and maintains 91.52% accuracy at -5 dB signal-to-noise ratio (SNR). These results should be interpreted as bench-level validation under controlled laboratory conditions rather than as direct evidence of field-level generalization. This framework provides a methodological solution that integrates fractional-order signal enhancement, dual-view representation, and Mamba-style collaborative state-space modeling for gearbox fault classification under controlled laboratory conditions with known speed variations and noise disturbances.

Keywords: fractional calculus; fractional-order differential enhancement; synchrosqueezing transform; gearbox fault diagnosis; dual-view learning; state-space model

Check for updates

Academic Editors: José de Jesús Rangel Magdaleno and Oscar Martínez-Fuentes

Received: 22 April 2026
Revised: 15 May 2026
Accepted: 16 May 2026
Published: 19 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Gearboxes serve as key power transmission components in modern rotating machinery. Their operational health directly affects the efficiency, safety, and maintenance overhead of the entire mechanical system. In actual industrial environments, gearboxes frequently

operate under nonstationary conditions characterized by shifting rotational speeds, severe load fluctuations, and heavy background noise. Highlighting these challenges, Li et al. [1] reviewed how integrating prior knowledge can benefit data-driven methods under such complex scenarios. Neupane et al. [2] systematically summarized the broader bottlenecks in machinery fault diagnosis, while Ahmad et al. [3] specifically emphasized the difficulties of feature extraction in planetary gearboxes. When early localized gearbox faults—such as pitting or slight wear—occur, the resulting impulsive responses can be weak, especially under strong disturbances and nonstationary operating conditions. Environmental disturbances and the dominant meshing harmonics of healthy gears can easily submerge these weak fault signatures. Consequently, developing a diagnostic framework capable of effectively capturing weak localized fault features under noisy conditions has become an important research problem, as recently echoed by Liu et al. [4].

In recent years, data-driven deep learning methods have become widely used in machinery fault diagnosis. For instance, Yuan et al. [5] developed an improved multi-scale Convolutional Neural Network (CNN) integrated with lightweight convolutional attention to capture diverse fault patterns. To further optimize feature aggregation, Nguyen et al. [6] designed a vibration-based diagnostic network utilizing depth-wise feature concatenation. Addressing the need for model efficiency, Cheng et al. [7] combined channel attention with transfer learning, while Dou et al. [8] employed a CBAM-DSC-ResNeXt50 structure to enhance feature extraction under complex conditions. To overcome the information bottleneck of single-modality inputs, researchers have increasingly shifted toward multi-sensor fusion. Xie et al. [9] proposed a deep spatiotemporal feature representation method, whereas Zhang et al. [10] constructed a cross-domain fusion network specifically tailored for variable operating conditions. Exploring sequence alignment, Zhang et al. [11] introduced a time-segment-wise fusion Transformer. Xiao et al. [12] also investigated multi-level architectures, and Li et al. [13] utilized adaptive weighting mechanisms to balance multisensor contributions. From the perspective of reliability, Mao et al. [14] developed an explainable fusion approach designed for noisy environments and limited sample conditions.

Beyond one-dimensional signal analysis, converting time-domain vibrations into two-dimensional (2D) time–frequency images has proven highly effective for revealing nonstationary dynamics. The synchrosqueezing transform (SST), originally mathematically formalized by Daubechies et al. [15], improves energy concentration through instantaneous frequency reassignment. Building on this foundation, Hazra and Narasimhan [16] pioneered the application of SST for gearbox fault detection, demonstrating clearer fault trajectories in the time–frequency plane. Zhu et al. [17] further refined this approach by developing a geometry-based SST with a shifted instantaneous frequency estimator. Hoang et al. [18] subsequently incorporated time–frequency representations into a data-driven framework for nonstationary conditions, and Bekaddour Benattia et al. [19] successfully combined time–frequency analysis with transfer learning for planetary gearbox diagnostics. Nevertheless, SST also introduces an additional preprocessing cost compared with direct one-dimensional sequence inputs or simpler time–frequency transforms such as the short-time Fourier transform (STFT), because it requires wavelet-domain analysis and energy reassignment. In the current implementation, SST-based representation construction is performed prior to network inference rather than being optimized as a dedicated real-time streaming module. Therefore, its preprocessing latency should be considered when evaluating practical deployment feasibility, and the computational costs of both the learning model and the full preprocessing–inference pipeline are further discussed in Section 5.9.

Although CNN- and Transformer-based models have achieved promising results in machinery fault diagnosis, they still have limitations when processing high-sampling-rate vibration signals. CNNs are effective at extracting local patterns but may require

deep stacking or large kernels to capture long-range temporal dependencies. Transformers can model global dependencies through self-attention, but their computational and memory costs increase quadratically with the sequence length, which becomes less efficient for long-window vibration sequences generated by high-frequency sampling. In contrast, Mamba-style state-space models provide linear-complexity sequence modeling while maintaining the ability to capture long-range dependencies through selective state propagation. This property makes them particularly suitable for vibration fault diagnosis, where fault-related impulses are often sparse, temporally distributed, and embedded in long nonstationary sequences.

Recently, the Mamba architecture proposed by Gu and Dao [20] and its visual counterpart, VMamba, proposed by Liu et al. [21], have provided promising alternatives for efficient long-range representation learning. Demonstrating their potential in machinery diagnosis, Yi et al. [22] developed a lightweight VibrMamba model for rotating machinery. Xia et al. [23] designed a 2D–1D fusion network based on a bidirectional multi-granularity Transformer–Mamba architecture. Xu et al. [24] applied a Mamba model combined with contrastive learning to complex motor bearing scenarios, and Zhang et al. [25] proposed a multi-scale group Mamba network with structural attention for efficient multisensor fault diagnosis.

However, relying solely on advanced deep architectures is insufficient if the raw input is heavily corrupted. If noise heavily masks fault-related impulsive components at the physical signal level, even sophisticated deep networks may struggle to learn robust decision boundaries. To alleviate this limitation, recent studies have increasingly explored fractional-calculus-inspired signal processing and intelligent diagnosis methods for rotating machinery. Unlike standard integer-order operators, fractional-order formulations provide additional flexibility for representing nonlocal and memory-related characteristics. Related studies have shown that introducing fractional or fractional-domain representations can improve the extraction of weak fault-sensitive components in vibration signals. For example, Li et al. [26] proposed a fractional-order Siamese deep residual shrinkage network for bearing fault diagnosis under small-sample conditions. Xie et al. [27] developed a rolling bearing diagnosis method based on the fractional constant-Q nonstationary Gabor transform and VMamba-Conv. Xie et al. [28] employed fractional Fourier transform information fusion combined with a lightweight neural network for early bearing fault diagnosis, and Wang et al. [29] showed that fractional envelope analysis can enhance weak bearing fault feature extraction under noisy conditions. More recently, Yang et al. [30] introduced hierarchical fractional-order attention entropy into an improved Transformer-based diagnosis framework for weak fault impact signals under complex operating conditions, while Jiao et al. [31] incorporated fractional-order gradient descent into a neural-network-based bearing fault diagnosis method. These studies suggest that fractional-order or fractional-domain processing can enhance fault-related transient components and improve the representation of weak fault signatures.

In addition to machinery fault diagnosis, related studies on memory-related dynamical systems, fractional-order neural modeling, and fractional integral inequalities also provide useful theoretical and methodological references. Bohner et al. [32] investigated the qualitative properties of integro-delay differential equations, including asymptotic stability, uniform stability, integrability, and boundedness, which provides a broader mathematical background for understanding memory-related dynamic systems. Caputo and Fabrizio [33] proposed a fractional derivative with a non-singular kernel, and Losada and Nieto [34] further discussed the corresponding fractional integral and related properties. More recently, Yasin et al. [35] established Hermite–Hadamard type inequalities for nonconvex functions using the Caputo–Fabrizio fractional integral, further demonstrating the

mathematical development of fractional integral operators. Gomolka et al. [36] introduced a fractional-order neural network for detecting process deviations in optical fiber cable manufacturing, where fractional derivatives were incorporated into an LSTM-based architecture using the Grünwald–Letnikov method. These studies further support the potential of fractional-order modeling for characterizing complex temporal dependencies and subtle deviations in nonlinear industrial signals.

Building on these studies, this work primarily makes a methodological and algorithmic framework contribution by integrating these components, rather than proposing a new mathematical definition of fractional calculus, a new SST theory, or a new fundamental Mamba/state-space block. The proposed method is not intended to simply stack fractional-order preprocessing, SST representation, and Mamba-style backbones. Instead, its novelty lies in organizing these components into a unified fractional-order-enhanced dual-view representation and collaborative state-space modeling framework.

Specifically, different from existing fractional-order diagnosis studies that mainly focus on fractional-order Siamese networks [26], fractional-domain time–frequency transforms [27,28], fractional envelope analysis [29], fractional-order entropy features [30], or fractional-order optimization strategies [31], this study introduces the Grünwald–Letnikov fractional-order differential operator as a controllable signal-level enhancement module. The same enhanced signal is then used to construct two strictly paired representations: a fractional-order-enhanced one-dimensional vibration sequence and a fractional-order-enhanced SST time–frequency image. This paired construction ensures that the temporal branch and the visual branch are driven by the same fault-sensitive enhanced signal source, rather than by loosely coupled heterogeneous inputs.

In addition, the proposed collaborative fusion module is algorithmically different from simple feature concatenation or score-level averaging. It uses sample-adaptive gated weighting to dynamically balance the contributions of the VibrMamba temporal branch and the VMamba visual branch, and further introduces cross-branch residual enhancement to preserve complementary information from the jointly constructed dual-view features. Therefore, the main novelty of this study lies in the fractional-order-enhanced paired 1D–2D representation strategy and the adaptive collaborative fusion mechanism for Mamba-style temporal–visual fault modeling under controlled laboratory conditions with known speed variations and noise disturbances.

Based on the above motivation and novelty clarification, this paper proposes a framework called fractional-order-enhanced dual-view representation and VibrMamba–VMamba collaborative modeling for gearbox fault classification under controlled bench-test conditions with known speed variations. The main contributions are clarified as follows:

(1) A fractional-order-enhanced dual-view representation strategy is proposed. Instead of using fractional-order processing only as an isolated preprocessing step, the Grünwald–Letnikov fractional-order differential operator is employed to generate a unified enhanced signal source, from which both the one-dimensional temporal sequence and two-dimensional SST time–frequency image are constructed. This design enables the model to jointly exploit enhanced impulsive temporal information and concentrated time–frequency structural information.

(2) A collaborative VibrMamba–VMamba modeling framework is developed for paired 1D–2D fault representations. The VibrMamba branch models long-range temporal dependencies in the fractional-order-enhanced vibration sequence, whereas the VMamba branch extracts global structural features from the corresponding enhanced SST image. Unlike simple feature concatenation, the proposed framework integrates the two branches through sample-adaptive gated weighting and cross-branch residual

enhancement, allowing complementary temporal and visual representations to be selectively fused for each input sample.

(3) The relationship between the proposed method and existing fractional-order diagnosis studies is explicitly discussed and distinguished. Unlike previous studies focusing on fractional-order Siamese networks, fractional-domain transforms, fractional envelope analysis, fractional-order entropy, or fractional-order optimization, fractional-order processing is not used in this work as an isolated module. Instead, Grünwald–Letnikov fractional-order enhancement is embedded into the construction of paired temporal and time–frequency representations and further couples these representations with adaptive Mamba-style temporal–visual collaborative modeling.

(4) The proposed method is experimentally validated under controlled laboratory conditions using a bench-level gearbox dataset with known speed variations. The current dataset does not include healthy baseline data at 600 and 900 r/min; therefore, the results should be interpreted as multi-state fault classification under known speed-varying bench-test conditions rather than strict cross-speed generalization. Future work should focus on further validation with complete healthy and faulty data across multiple speed conditions.

The remainder of this paper is organized as follows. Section 2 introduces the basic principles of fractional-order differential enhancement, SST, VibrMamba, and VMamba. Section 3 presents the proposed fractional-order-enhanced dual-view fault diagnosis framework, including the engineering scenario, overall diagnostic procedure, and collaborative fusion module. Section 4 describes the experimental platform, dataset construction, baseline models, training settings, and evaluation metrics. Section 5 reports the experimental results, including convergence behavior, visualization results, comparative experiments, noise robustness analysis, ablation studies, fractional-order sensitivity analysis, and computational-cost evaluation. Section 6 discusses the effectiveness, limitations, and practical implications of the proposed method. Finally, Section 7 concludes the paper and outlines future research directions.

2. Basic Principles

This section briefly introduces the basic principles required for the proposed framework, including Grünwald–Letnikov fractional-order differential enhancement, SST-based time–frequency representation, and Mamba-style sequence and visual modeling. To avoid repeating well-established details, the discussion focuses on how these principles support the fractional-order-enhanced dual-view diagnosis framework developed in Section 3.

2.1. Principle of Fractional-Order Differential Enhancement

Fractional calculus is a mathematical generalization of traditional integer-order integration and differentiation. Unlike integer-order derivatives that depend only on the local properties of a function, fractional-order operators inherently possess nonlocality and historical memory, which makes them useful for characterizing nonstationary mechanical dynamics. In the broader mathematical literature, different fractional operators have been developed from different perspectives, including classical Grünwald–Letnikov, Riemann–Liouville, and Caputo formulations, as well as non-singular-kernel operators such as the Caputo–Fabrizio derivative and its associated fractional integral. Related studies on fractional integral inequalities, such as Hermite–Hadamard type inequalities involving the Caputo–Fabrizio fractional integral, further enrich the theoretical foundation of fractional calculus.

In this study, the Grünwald–Letnikov (G-L) definition is adopted for fractional-order differential enhancement because it provides a direct and convenient discrete implemen-

tation for sampled vibration signals. Therefore, the Caputo–Fabrizio fractional integral and related inequalities are introduced here as theoretical background references, whereas the actual signal enhancement module is implemented using the G-L fractional-order differential operator.

For a continuous signal $x(t)$, the Grünwald–Letnikov fractional derivative of order α with lower terminal a can be defined as

$$ {}_aD_t^{\alpha}x(t) = \lim_{Ts\to 0} T_s^{-\alpha} \sum_{k=0}^{\lfloor (t-a)/T_s \rfloor} (-1)^k \binom{\alpha}{k} x(t-kT_s), \tag{1} $$

where $0 < \alpha < 1$ denotes the fractional order used for signal enhancement, T_s denotes the time step in the Grünwald–Letnikov approximation, a is the lower terminal of the fractional derivative, $\lfloor \cdot \rfloor$ denotes the floor operator, and $\binom{\alpha}{k}$ is the generalized binomial coefficient.

In practical vibration-signal processing, the measured signal is a discrete sequence $x[n]$, $n = 0, 1, \ldots, N-1$ with the sampling period $T_s = 1/f_s$. Therefore, the discrete Grünwald–Letnikov fractional-order differential enhancement used in this study is implemented as

$$ y[n] = D^{\alpha}x[n] \approx T_s^{-\alpha} \sum_{k=0}^{Ln} c_k^{(\alpha)} x[n-k], \tag{2} $$

where $y[n]$ is the fractional-order-enhanced signal, $L_n = \min(n, L)$ is the effective memory length used in the truncated implementation, and $c_k^{(\alpha)}$ denotes the fractional-order difference coefficient. The coefficient is computed recursively as

$$ c_0^{(\alpha)} = 1,\ c_k^{(\alpha)} = \left(1 - \frac{\alpha+1}{k}\right) c_{k-1}^{(\alpha)},\ k \geq 1. \tag{3} $$

The use of Equation (2) as the numerical approximation of the Grünwald–Letnikov fractional derivative is justified by the direct correspondence between the continuous G-L definition and its finite-difference form for uniformly sampled vibration signals. In the G-L formulation, the fractional derivative is naturally expressed as a weighted summation of historical signal samples, where the weights are determined by generalized binomial coefficients. Therefore, Equation (2) preserves the nonlocal memory property of the fractional derivative while providing a direct discrete implementation for measured vibration sequences. Compared with conventional integer-order finite-difference operators, which mainly depend on adjacent samples, the G-L finite-difference approximation retains historical memory information through the weighted accumulation of past samples. Compared with Caputo- or Riemann–Liouville-type integral quadrature methods, the G-L approximation is more straightforward for uniformly sampled engineering signals because it does not require additional interpolation or explicit discretization of singular integral kernels. Compared with frequency-domain spectral implementations or rational-filter approximations, the time-domain G-L approximation can be directly applied to finite-length vibration segments and is convenient for constructing sample-wise enhanced one-dimensional sequences. In practical implementation, the infinite memory of the theoretical G-L derivative is truncated to an effective memory length L, which provides a balance between approximation fidelity and computational efficiency. Therefore, Equation (2) is adopted as a valid and computationally convenient G-L finite-difference approximation for the fractional-order enhancement of gearbox vibration signals in this study.

In this formulation, T_s is used consistently in both the continuous Grünwald–Letnikov definition and the discrete implementation. This avoids ambiguity between the limiting step size in the theoretical definition and the sampling period in practical signal processing.

From an analytical signal-processing perspective, the Grünwald–Letnikov fractional-order differential operator can be interpreted as a fractional high-pass filtering operator. For a continuous signal $x(t)$, the Fourier transform of its fractional-order derivative can be expressed as

$$F\{D^{\alpha}x(t)\} = (j\Omega)^{\alpha}X(\Omega), \tag{4}$$

where Ω denotes the angular frequency and $X(\Omega)$ is the Fourier transform of $x(t)$. Therefore, the magnitude response of the fractional-order differential operator is proportional to $|\Omega|^{\alpha}$, while its phase response is approximately $\alpha\pi/2$ for positive frequencies. This indicates that the operator attenuates low-frequency trends and relatively enhances high-frequency transient components. Compared with the standard first-order derivative whose magnitude response is proportional to $|\Omega|$, the fractional-order derivative with $0 < \alpha < 1$ increases more smoothly with frequency. Thus, it can enhance impulsive fault-related components while reducing the risk of excessive high-frequency noise amplification.

For the discrete Grünwald–Letnikov implementation used in this study, the corresponding z-domain transfer function can be written as

$$H_{\alpha}(z) = T_s^{-\alpha}(1 - z^{-1})^{\alpha}. \tag{5}$$

By setting $z = e^{j\omega}$, the frequency response becomes

$$H_{\alpha}(e^{j\omega}) = T_s^{-\alpha}(1 - e^{-j\omega})^{\alpha}, \tag{6}$$

and its magnitude response is

$$\left|H_{\alpha}(e^{j\omega})\right| = T_s^{-\alpha}[2|sin(\omega/2)|]^{\alpha}. \tag{7}$$

This discrete frequency response further shows that the gain is close to zero near $\omega = 0$, which suppresses slow-varying baseline components, and gradually increases as ω approaches the high-frequency region, which helps highlight transient impulses caused by localized gearbox faults. Since α controls the slope of the magnitude response, the fractional-order operator provides a tunable compromise between low-frequency trend suppression, transient impulse enhancement, and high-frequency noise sensitivity.

The above frequency-domain interpretation explains the physical significance of applying fractional-order differentiation to gearbox vibration signals. For weak localized gearbox faults, the localized impacts typically appear as transient, relatively high-frequency impulses embedded within low-frequency baseline fluctuations, meshing-related components, and background noise. When the fractional order is properly selected, the operator tends to enhance fault-induced transient components while suppressing slow-varying trends, thereby improving the visibility of weak impulsive signatures before dual-view feature construction.

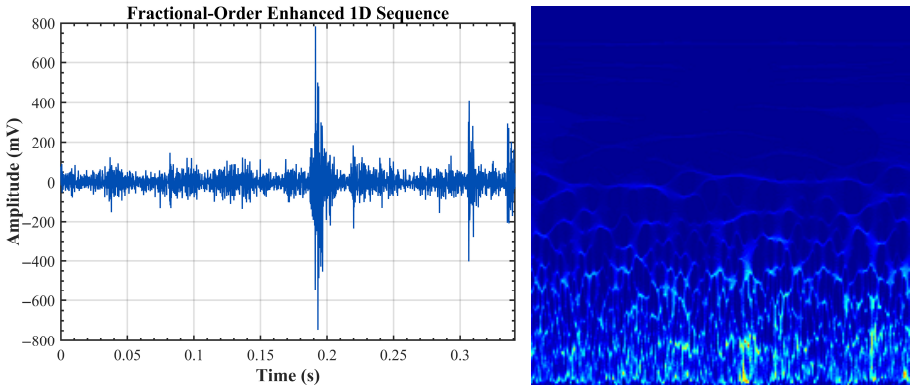
Compared with traditional first-order differentiation ($\alpha = 1$), the fractional-order derivative exhibits a smoother frequency-dependent amplification behavior. This property helps retain useful nonlinear dynamic information while avoiding the excessive amplification of high-frequency noise. In this sense, the fractional-order formulation provides a more flexible and controllable trade-off between transient feature enhancement and noise sensitivity than fixed integer-order differential operators.

Therefore, in this study, the fractional-order-enhanced signal $x_{\alpha}(t)$ (and its discrete form $x_{\alpha}[n]$) is used as the fundamental input for the subsequent dual-view feature construction.

2.2. Principle of the Synchrosqueezing Transform

The synchrosqueezing transform (SST) is adopted in this study to convert the fractional-order-enhanced vibration signal into a high-concentration time–frequency image. Compared with conventional time–frequency transforms, SST improves the concentration of nonstationary components by reassigning wavelet-domain energy to more accurate instantaneous frequency locations. This property is useful for gearbox vibration signals because fault-related impulses and modulation sidebands often vary with operating conditions and may be blurred in conventional spectra.

To provide a direct example of the representation process used in this study, Figure 1 shows an actual gearbox vibration signal after fractional-order differential enhancement and its corresponding fractional-order-enhanced SST time–frequency representation. This figure directly shows how the enhanced one-dimensional signal is mapped into the two-dimensional visual input used by the VMamba branch.



(**a**) (**b**)

Figure 1. Example of fractional-order differential enhancement and SST-based time–frequency transformation for a gearbox vibration signal: (**a**) fractional-order-enhanced vibration signal; (**b**) corresponding fractional-order-enhanced SST time–frequency representation. In panel (**b**), the color intensity represents the normalized time–frequency energy magnitude.

In this study, the fractional-order-enhanced vibration signal, denoted as $x_\alpha(t)$, is used as the input, and its continuous wavelet transform can be expressed as follows

$$W_{x_\alpha}(a,b) = \int x_\alpha(t) \frac{1}{\sqrt{a}} \psi^* \left(\frac{t-b}{a} \right) dt, \tag{8}$$

where a denotes the scale factor, b denotes the time-shift parameter, ψ denotes the mother wavelet, and $*$ denotes the complex conjugate operation. The continuous wavelet transform can reveal the local characteristics of the signal at different times and scales. However, since its representation is initially distributed in the scale domain and is affected by the window function and scale dilation, the time–frequency energy often suffers from dispersion, making localized fault features difficult to distinguish under complex background conditions.

The key idea of SST is first to estimate the local instantaneous frequency using the wavelet coefficients and then remap the corresponding energy from the scale axis onto the frequency axis. For points satisfying $W_{x_\alpha}(a,b) \neq 0$, the corresponding local instantaneous frequency estimate is generally expressed as

$$\hat{\omega}_{x_\alpha}(a,b) = -i(W_{x_\alpha}(a,b))^{-1} \partial_b W_{x_\alpha}(a,b), \tag{9}$$

where $\hat{\omega}_{x_\alpha}(a,b)$ denotes the estimated local instantaneous frequency of the fractional-order-enhanced signal at scale a and time shift b, ∂_b denotes the partial derivative with respect to b,

and i is the imaginary unit. It should be noted that $\hat{\omega}_{x_\alpha}(a,b)$ is not an independent frequency variable, but the estimated frequency location to which the wavelet coefficient $W_{x_\alpha}(a,b)$ is reassigned in the SST process. In the final synchrosqueezed representation $T_{x_\alpha}(\omega,b)$, ω denotes the frequency coordinate of the time–frequency plane. Therefore, Equation (9) describes how the reassignment frequency is estimated from the phase variation of the wavelet coefficient.

After the instantaneous frequency has been estimated, SST no longer preserves the original scale-domain distribution; instead, the wavelet coefficients are reassigned to their corresponding frequency locations, thereby yielding the synchrosqueezed time–frequency representation

$$T_{x_\alpha}(\omega,b) = \int W_{x_\alpha}(a,b)\delta(\omega - \omega_{x_\alpha}(a,b))a^{-3/2}da, \tag{10}$$

where $\delta(\cdot)$ denotes the Dirac delta function and the integral is performed over the scale parameters. Equations (8)–(10) summarize the SST procedure used in this study: the fractional-order-enhanced signal is first transformed into the wavelet domain, the local instantaneous frequency is then estimated from the phase variation of the wavelet coefficients, and the dispersed scale-domain energy is finally reassigned to the frequency axis. Through this reassignment process, fault-related impulses, modulation sidebands, and nonstationary frequency components can be represented as more concentrated time–frequency structures. Therefore, the SST image of the fractional-order-enhanced signal is used as the visual-view input of the proposed framework, providing complementary structural information for the VMamba branch.

2.3. Principle of VibrMamba

VibrMamba is adopted as the temporal branch in this study because it is designed for one-dimensional vibration-signal modeling and provides an efficient state-space alternative to attention-based long-sequence modeling. For high-sampling-rate gearbox vibration signals, fault-related impulses may be sparse and distributed over long temporal windows. The bidirectional state-space structure of VibrMamba helps capture contextual dependencies before and after local impacts, while its lightweight design reduces the computational burden compared with attention-based sequence models. The principle is illustrated in Figure 2.

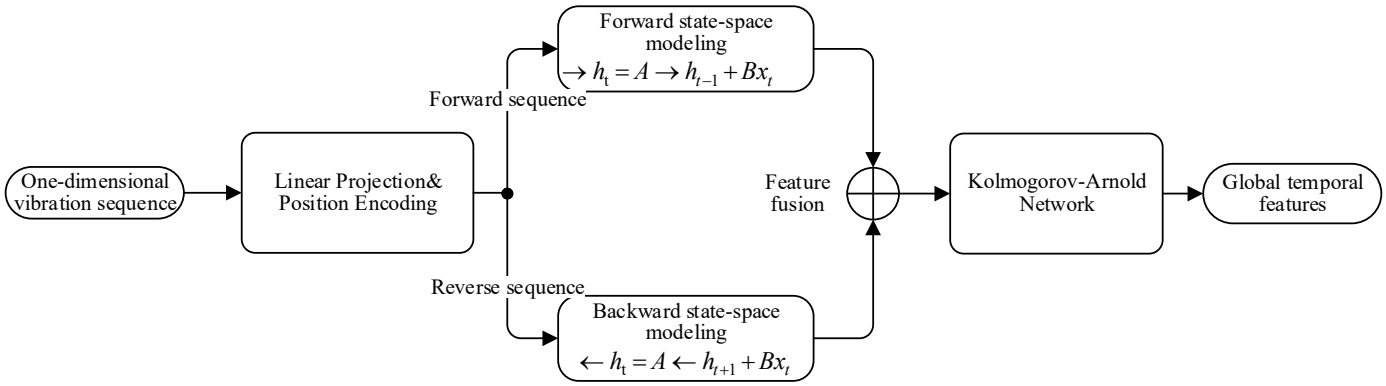


Figure 2. Schematic diagram of the VibrMamba network architecture and feature extraction process.

In the proposed framework, the input to the VibrMamba branch is not the raw vibration signal but the fractional-order-enhanced one-dimensional sequence. This design allows the temporal branch to model long-range dependencies based on a signal representation in which weak fault-related transients have already been strengthened. The basic discrete state-space formulation can be written as

$$h_t = Ah_{t-1} + Bx_t, \tag{11}$$

$$y_t = Ch_t + Dx_t, \tag{12}$$

where x_t, h_t, and y_t denote the input, hidden state, and output at time step t, respectively, and A, B, C, and D are the state transition, input projection, output projection, and skip-connection terms. In Mamba-style models, part of the state-space parameters can be conditioned on the input, enabling selective propagation of informative components and suppression of redundant fluctuations.

For the gearbox diagnosis task considered in this study, this mechanism is useful because fault-related impulses and modulation patterns may be temporally sparse and distributed over long signal segments. The bidirectional structure of VibrMamba further allows the model to exploit contextual information before and after local impacts. Therefore, the fractional-order-enhanced one-dimensional sequence is fed into the VibrMamba branch to extract long-range temporal dependencies and impulsive dynamic patterns.

In addition, VibrMamba replaces conventional linear projection layers with Kolmogorov–Arnold Network (KAN) layers. Unlike a standard multilayer perceptron (MLP), which mainly relies on fixed activation functions applied to linear transformations, KAN introduces learnable univariate functions on network edges. A simplified KAN mapping can be expressed as

$$z_j = \sum_{i=1}^{d} \phi_{ij}(u_i), \tag{13}$$

where u_i denotes the i-th input feature, $\phi_{ij}(\cdot)$ is a learnable univariate function, and z_j denotes the j-th output feature. In practical implementations, $\phi_{ij}(\cdot)$ is commonly parameterized using spline-based basis functions, allowing KAN to learn more flexible nonlinear mappings than fixed activation functions in standard MLPs. In this study, this property is used to improve the nonlinear feature mapping capability of the VibrMamba temporal branch for enhanced gearbox vibration sequences.

Overall, VibrMamba is used in this study to model the temporal-view representation generated from the fractional-order-enhanced signal. Its linear-complexity sequence modeling, bidirectional context aggregation, and KAN-based nonlinear mapping make it suitable for extracting temporal dependencies and impulsive patterns from long vibration sequences. The extracted temporal features are then combined with the visual time–frequency features learned by the VMamba branch.

2.4. Principle of VMamba

VMamba is adopted as the visual branch to extract global structural features from fractional-order-enhanced SST time–frequency images. Compared with the vision transformer (ViT)-style self-attention, VMamba extends selective state-space modeling to visual feature maps through two-dimensional selective scanning, thereby maintaining a global receptive field with linear-complexity visual modeling. This property is suitable for SST images because fault-related ridges, modulation bands, and transient energy concentrations may appear at different time–frequency locations. The basic visual state-space block and the two-dimensional selective scanning process used in the VMamba branch are illustrated in Figures 3 and 4, respectively.

In the proposed framework, the fractional-order-enhanced SST image is first divided into image patches and then processed by stacked visual state-space blocks, as illustrated in Figure 3. The two-dimensional selective scan mechanism unfolds the feature map along multiple directions, applies state-space scanning to the resulting sequences, and merges the scanned representations back into the image feature space, as shown in Figure 4.

In this way, VMamba captures global time–frequency structural dependencies without relying on quadratic self-attention. Therefore, it serves as the visual-view feature extractor complementary to the VibrMamba temporal branch.

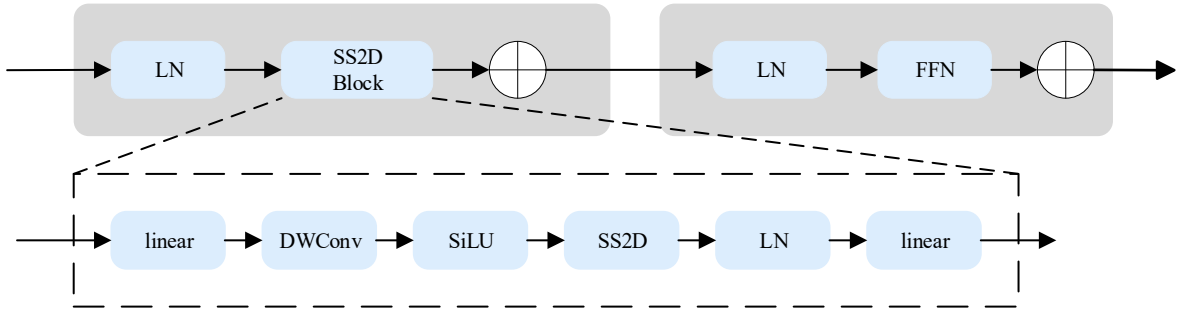


Figure 3. Structure of the visual state-space (VSS) block in VMamba. The arrows indicate the direction of feature flow. The gray regions denote the two residual sublayers, consisting of layer normalization (LN) followed by the SS2D block and layer normalization followed by the feed-forward network (FFN), respectively. The dashed box shows the internal structure of the SS2D block, where the input features are processed sequentially through linear projection, depth-wise convolution (DWConv), SiLU activation, 2D selective scan (SS2D), layer normalization, and linear projection.

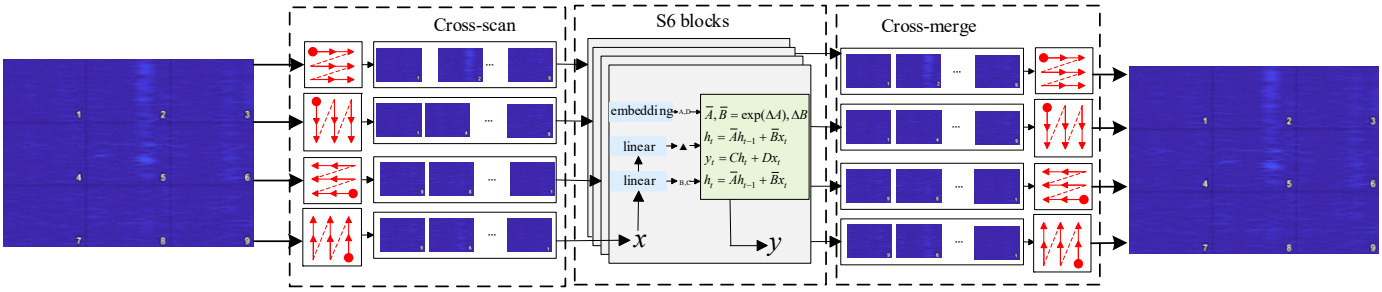


Figure 4. Schematic diagram of the 2D selective scanning process in VMamba. The dashed boxes indicate the three main stages: cross-scan, S6 blocks, and cross-merge. The red arrows denote the four directional scanning paths used to unfold the two-dimensional feature map into one-dimensional sequences. The numbers on the feature maps indicate spatial positions in the original and reconstructed feature maps. After directional sequence modeling by the S6 blocks, the cross-merge operation restores and integrates the scanned features back into the two-dimensional feature space. The colors are used only to distinguish feature maps and scanning paths and do not represent quantitative values.

3. Proposed Fractional-Order-Enhanced Dual-View Fault Diagnosis Framework

3.1. Engineering Scenario and Problem Definition

Variable-speed gearbox operating scenarios are often accompanied by rotational speed fluctuations, load variations, structural coupling effects, and external disturbances, which make the acquired vibration signals strongly nonstationary. Under such conditions, fault-related impulsive signatures may become weak and can be easily submerged by background interference and operating-condition variations. Therefore, developing a diagnosis framework that can capture weak localized fault features under known speed variations and noisy conditions remains an important research problem.

This study is limited to controlled laboratory conditions. The investigated object is not a field-measured gearbox from a specific industrial machine, but a laboratory bench-level approximation of variable-speed gearbox operation scenarios. In this bench-test scenario, four gearbox health conditions are considered: normal, broken tooth, pitting, and wear. By setting these health states and different rotational speed conditions on a

controllable test platform, representative localized damage characteristics are reproduced in a laboratory environment, and the corresponding vibration responses are collected for diagnostic analysis. This design allows the proposed method to be evaluated under controlled bench-test conditions involving known speed variations and noisy disturbances while maintaining clear labels and controllable experimental factors. However, because the healthy baseline is currently available only at 1200 r/min, this setting should not be interpreted as a complete cross-speed generalization benchmark.

Accordingly, the present results should be interpreted as bench-level validation under controlled laboratory conditions, rather than direct evidence of field-level generalization to a specific industrial machine or operating system. Although the platform can capture representative vibration evolution patterns under different fault and speed conditions, it does not fully cover coupled load variations, installation differences, long-term degradation effects, environmental disturbances, and multi-source interferences that may arise in practical field environments.

Based on this scenario definition, the problem addressed in this study can be stated as follows: under controlled laboratory conditions with known speed fluctuations, weak localized fault features of gearbox vibration signals are difficult to extract effectively because they are easily contaminated by noise, operating-condition fluctuations, and nonstationary coupling effects. Therefore, a fractional-order-enhanced pattern recognition and feature fusion method capable of jointly characterizing temporal dynamics and time–frequency structural information is required to improve fault identification under the current bench-test setting. The present task can be considered as multi-state fault classification under known speed fluctuations rather than strict generalization to unseen speed conditions.

3.2. Overall Intelligent Diagnosis Framework

Figure 5 shows the overall intelligent diagnosis framework of the proposed method under the current bench-test setting. The framework is based on fractional-order-enhanced dual-view representation and a VibrMamba–VMamba dual-branch collaborative fusion network.

The diagnostic procedure consists of four main stages:

(1) Signal acquisition and segmentation. Raw gearbox vibration signals are collected under different health conditions and rotational speed settings in the current bench-test platform. The continuous recordings are segmented into fixed-length samples, and strict raw file-level isolation is used in the subsequent dataset partition to avoid homologous data leakage.

(2) Fractional-order-enhanced dual-view construction. Each raw vibration segment is first processed by the Grünwald–Letnikov fractional-order differential operator to enhance weak impulsive components. The same enhanced segment is then used to construct two paired inputs: a one-dimensional enhanced vibration sequence for temporal modeling and a two-dimensional enhanced SST image for visual time–frequency modeling.

(3) Dual-branch feature extraction and collaborative fusion. The enhanced one-dimensional sequence is fed into the VibrMamba branch to model long-range temporal dependencies, whereas the enhanced SST image is fed into the VMamba branch to extract global time–frequency structural features. The aligned branch features are then integrated through the adaptive collaborative fusion module described in Section 3.3.

(4) Classification and evaluation. The fused representation is sent to the classifier for fault identification. The trained model is evaluated using accuracy, F1 score, Kappa coefficient, visualization analysis, ablation experiments, noise robustness tests, and computational-cost analysis.

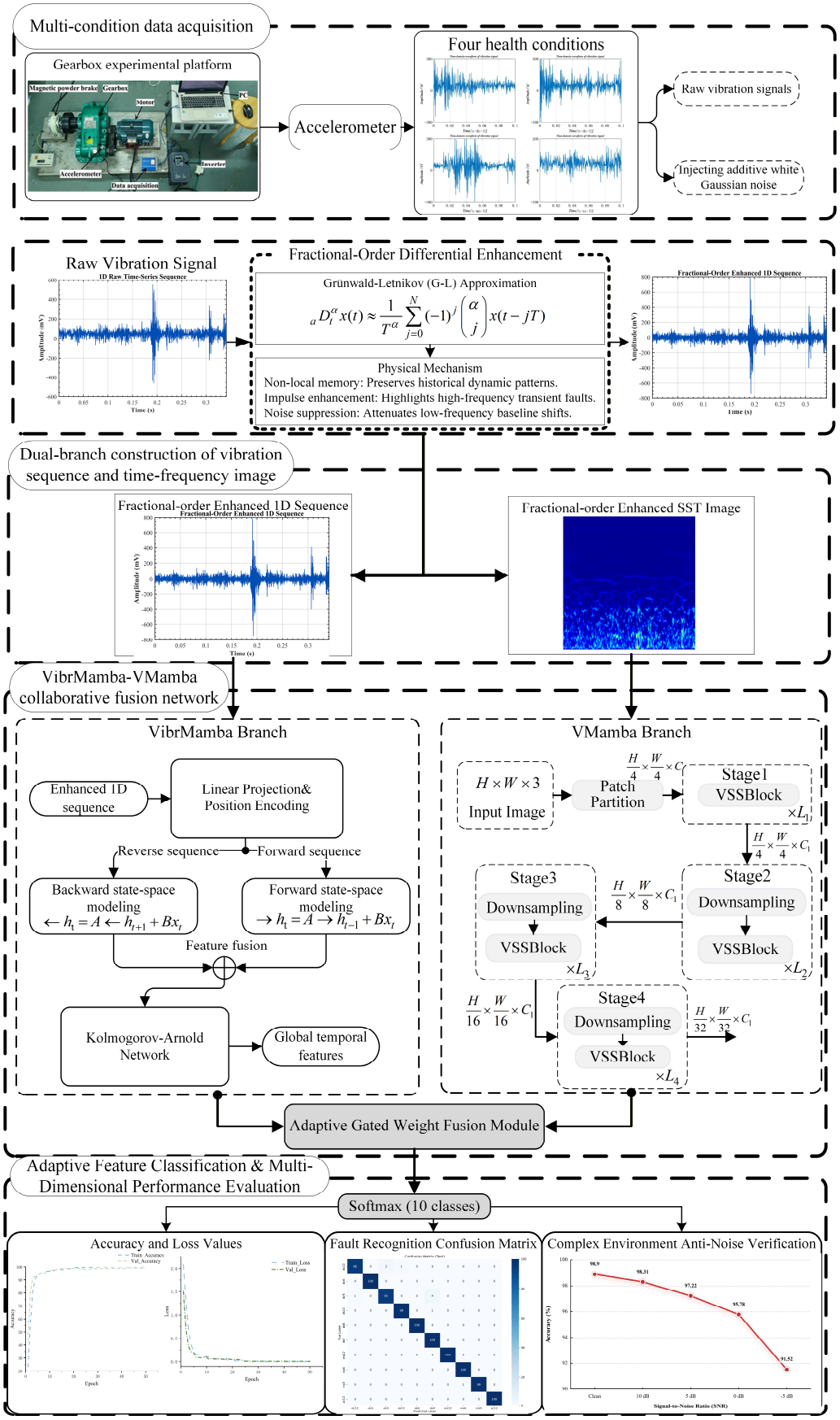


Figure 5. Overall framework of the proposed fractional-order-enhanced sequence–SST and VibrMamba–VMamba collaborative fault diagnosis model.

3.3. Collaborative Fusion Module

To further clarify the internal logic of the collaborative fusion module shown in Figure 5, this subsection provides the mathematical formulation of the adaptive gated weighting and cross-branch residual enhancement mechanisms. Let $\boldsymbol{f}_{1D} \in \mathbb{R}^d$ and $\boldsymbol{f}_{2D} \in \mathbb{R}^d$ denote the aligned feature vectors extracted from the VibrMamba temporal branch and the VMamba visual branch, respectively. The two branch features are first concatenated to form a joint dual-view representation:

$$\boldsymbol{f}_c = [\boldsymbol{f}_{1D}; \boldsymbol{f}_{2D}], \tag{14}$$

where $[\cdot;\cdot]$ denotes feature concatenation.

Based on this joint representation, an adaptive gated weighting vector is generated as

$$\boldsymbol{g} = \sigma(\boldsymbol{W}_g \boldsymbol{f}_c + \boldsymbol{b}_g), \tag{15}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, and $\boldsymbol{W}_g$ and $\boldsymbol{b}_g$ are learnable parameters. The gated complementary feature is then obtained by adaptively balancing the temporal and visual representations:

$$\boldsymbol{f}_g = \boldsymbol{g} \odot \boldsymbol{f}_{1D} + (1 - \boldsymbol{g}) \odot \boldsymbol{f}_{2D}, \tag{16}$$

where $\odot$ denotes element-wise multiplication. In this way, the model can dynamically adjust the contributions of the one-dimensional temporal branch and the two-dimensional visual branch for each sample.

To avoid the loss of branch-specific information during gated fusion, a residual enhancement mechanism is further introduced:

$$\boldsymbol{f}_r = \boldsymbol{f}_g + \boldsymbol{W}_r \boldsymbol{f}_c, \tag{17}$$

where $\boldsymbol{W}_r$ is a learnable projection matrix that maps the concatenated dual-view feature to the same dimension as $\boldsymbol{f}_g$. It should be noted that this residual enhancement mechanism is not a standard single-branch skip connection. Instead, it is a cross-branch residual projection that reintroduces the jointly concatenated temporal and visual features into the gated fused representation. This design helps preserve complementary information from both branches and improves the stability of the collaborative fusion process.

Finally, the enhanced fused representation is fed into the classifier to obtain the predicted fault category:

$$\hat{\boldsymbol{y}} = \mathrm{Softmax}(\boldsymbol{W}_o \boldsymbol{f}_r + \boldsymbol{b}_o), \tag{18}$$

where $\boldsymbol{W}_o$ and $\boldsymbol{b}_o$ are the classifier parameters, and $\hat{\boldsymbol{y}}$ denotes the predicted class probability distribution.

The above formulation shows that the proposed collaborative fusion module is not equivalent to direct concatenation of the two branch features. Direct concatenation treats temporal and visual features with fixed importance, whereas the gated weighting vector in Equation (15) enables sample-specific adjustment of the two branches according to the input condition. Moreover, the cross-branch residual term in Equation (17) reintroduces the jointly constructed dual-view representation into the fused feature space, which helps reduce the loss of branch-specific information during adaptive fusion. Therefore, the algorithmic distinction of the proposed fusion module lies in combining adaptive branch weighting with cross-view residual enhancement, rather than merely increasing the feature dimension through concatenation.

4. Experimental Setup

4.1. Experimental Platform and Data Acquisition

To verify the effectiveness of the proposed method for gearbox fault diagnosis under controlled laboratory conditions, this study established a gearbox fault diagnosis experimental platform. Vibration data were collected from a laboratory bench-test platform rather than from field-deployed industrial gearboxes. The platform primarily comprises a three-phase asynchronous motor (Model: YE3-100L2-4; Zhejiang Jinsu Motor Co., Ltd., Taizhou, China), gearbox reducer (Model: JZQ 250; Shandong Zibo Xinyuan Machinery Co., Ltd., Zibo, China), data acquisition card (Model: YE6231; Sinocera Piezotronics Inc., Yangzhou, China), frequency converter (Model: G7R5/P011-T4; Zhejiang Xintuo New Energy Co., Ltd., Taizhou, China), piezoelectric acceleration sensor (Model: CAYD051V; Xiamen Niell Electronics Co., Ltd., Xiamen, China), and magnetic powder brake (Model: FZ-A-12; Zhejiang Xinduode Transmission Components Co., Ltd., Zhejiang, China), as shown in Figure 6. This experimental platform can simulate the vibration response characteristics of a gearbox under different health conditions in a mechanical transmission system, thereby providing an experimental basis for subsequent fault diagnosis research.

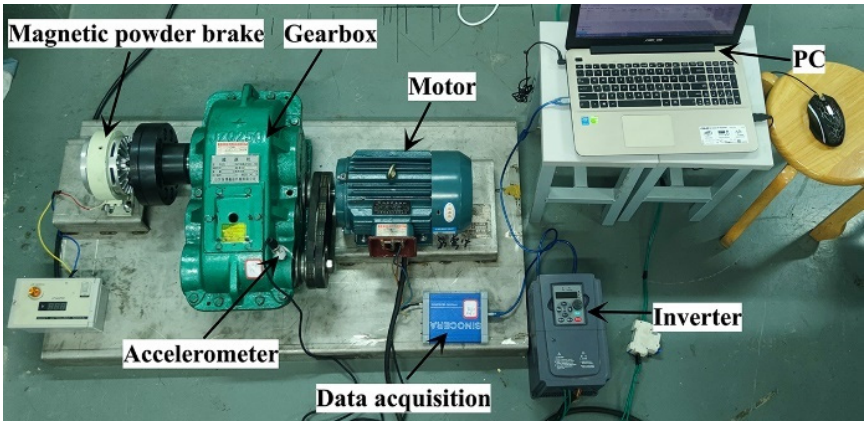


Figure 6. Experimental platform for gearbox fault diagnosis under known speed variations.

In the experiment, a gearbox was utilized as the research object, and acceleration sensors were installed at the corresponding measuring points to acquire vibration signals under different health conditions. According to the common fault types of transmission systems, four gearbox health conditions were considered in this study: normal, broken tooth, pitting, and wear. Typical vibration signals of these four health conditions under the operating condition of 1200 r/min are shown in Figure 7.

4.2. Dataset Construction and Partition

A total of 10 groups of gearbox vibration signals were collected in this study under the current bench-test setting. The sampling frequency was set to 6 kHz. Each sample contained 2048 sampling points, and 500 samples were collected for each data group. Specifically, the dataset consisted of one normal-state group collected at 1200 r/min and nine fault-state groups formed by the combinations of three representative fault types (broken tooth, pitting, and wear), with three rotational speed conditions: 600, 900, and 1200 r/min. After signal segmentation and dual-view feature construction, all samples were divided into training, validation, and test sets at a ratio of 6:2:2. Table 1 details the sample distributions under different health conditions.

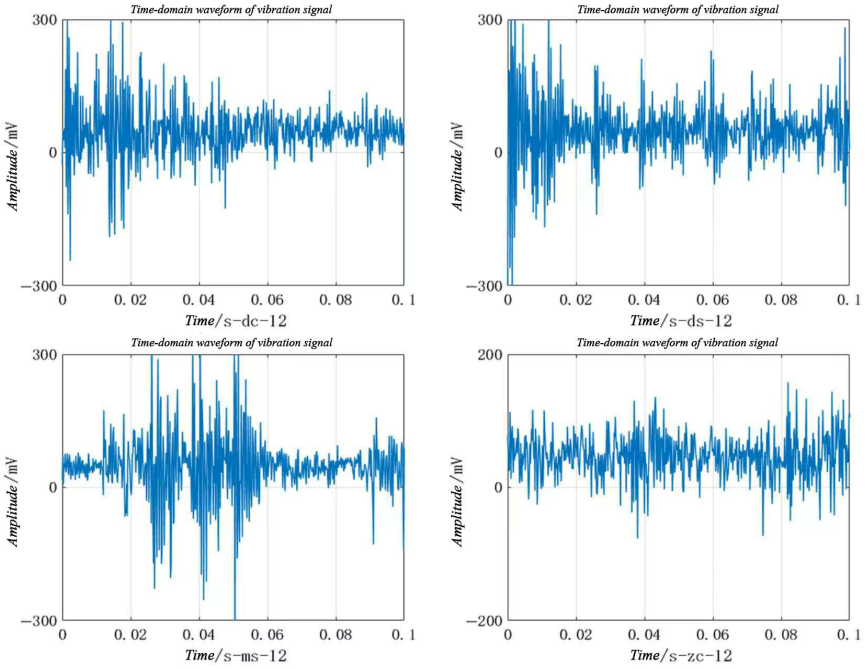


Figure 7. Typical vibration signals of the gearbox under different health conditions.

Table 1. Sample distribution under different health conditions.

Fault Type	Speed (r/min)	Label	Training Samples	Validation Samples	Test Samples
Normal	1200	zc_12	300	100	100
Broken-Tooth Fault	600	dc_6	300	100	100
	900	dc_9	300	100	100
	1200	dc_12	300	100	100
Pitting Fault	600	ds_6	300	100	100
	900	ds_9	300	100	100
	1200	ds_12	300	100	100
Wear Fault	600	ms_6	300	100	100
	900	ms_9	300	100	100
	1200	ms_12	300	100	100

Note that the 10 data groups in this study were constructed by jointly considering fault discriminability and operating-condition sensitivity. The normal group serves as a baseline healthy condition at the rated speed, whereas the nine fault groups are used to investigate the feature differences of different fault mechanisms under speed variations. Because baseline healthy data at 600 and 900 r/min are unavailable, this study primarily focused on multi-state fault classification under known speed fluctuations rather than strict cross-speed or cross-condition generalization. Therefore, a rigorous cross-speed testing protocol, in which the model is trained on one set of speed conditions and tested on entirely unseen speed conditions with complete healthy and faulty categories, is not fully supported by the present dataset design. The current evaluation instead focuses on whether the proposed framework can distinguish different fault states and speed-related classes under the available controlled bench-test conditions. Furthermore, to rigorously prevent homologous data leakage and artificially inflated accuracy, the 6:2:2 dataset partition is executed using strict physical isolation at the raw file level. Specifically, the continuous vibration recordings for each condition were collected through multiple independent runs. The training, validation, and test sets are derived from completely independent raw data

files rather than randomly splitting samples from a single continuous recording. This strict isolation strategy ensures that no overlapping temporal windows exist between the subsets, thereby guaranteeing the authenticity and reliability of the model evaluation.

4.3. Fractional-Order Parameter Settings and Training Parameters

For the fractional-order differential enhancement module, the fractional order α is a crucial parameter that controls the enhancement intensity of transient impulsive components and the degree of high-frequency amplification. In this study, we did not treat $\alpha = 0.6$ as a fixed empirical constant or optimize it separately for each individual fault type or rotational speed condition. Instead, we evaluated candidate values of α under the same data partition, network architecture, and training configuration and selected the final value according to the average diagnostic performance across all health-condition and speed-related classes. Based on the sensitivity analysis reported in Section 5.8, $\alpha = 0.6$ achieved the best overall performance and provided a favorable balance between fault-related impulse enhancement and high-frequency noise amplification. Therefore, $\alpha = 0.6$ should be understood as a global compromise value for the current bench-test dataset and experimental configuration, rather than as a universally optimal fractional order for all gearbox diagnosis tasks.

To ensure the fairness and comparability of the experimental results, all methods were trained and tested under a unified experimental environment. The experiments were implemented on Ubuntu 22.04 using the PyTorch deep learning framework version 2.8.0+cu128, with CUDA 12.8 acceleration. During training, the batch size was set to 32, the number of training epochs was set to 50, the AdamW optimizer was adopted, the initial learning rate was set to 0.0001, and a cosine annealing strategy was used to adjust the learning rate dynamically.

Meanwhile, to ensure the objectivity of the comparative experiments, all comparison methods adopted the same data partition strategy and training configuration. During sample partitioning, strict physical isolation was carried out at the level of raw files to avoid homologous samples appearing simultaneously in the training and test sets. For fair comparison, all baseline models were trained with the same data partition, number of epochs, optimizer, and learning rate adjustment strategy. For one-dimensional models, fractional-order-enhanced vibration sequences with a unified input length of 2048 were used; for two-dimensional models, fractional-order-enhanced SST time–frequency images of the same size were used; and for dual-branch models, the same sample pairing method and training/validation/test partition scheme as the proposed method were adopted.

4.4. Baseline Models and Comparative Settings

To verify the effectiveness of the proposed fractional-order-enhanced dual-view framework, several representative methods were selected for comparison, including classical deep learning models (WDCNN, ResNet-50, and ViT), state-space-based sequence and visual models (VibrMamba and VMamba), and a multimodal fusion baseline (1D-CNN + ViT).

Furthermore, to specifically validate the contribution of the fractional-order differential enhancement module, a variant without fractional-order preprocessing (i.e., Original SST + VibrMamba + VMamba) is included as a core baseline. Additionally, a comparison model using STFT instead of SST, FO + VibrMamba + VMamba (STFT), was included to evaluate the necessity of the high-resolution SST representation.

4.5. Evaluation Metrics

To quantitatively evaluate the diagnostic performance of the proposed model and the baseline methods from multiple perspectives, the accuracy (Acc), F1 score, and the Kappa coefficient were adopted as the core evaluation metrics. In addition, to assess

the computational complexity and practical relevance of the algorithm, the number of model parameters (Params), floating-point operations (FLOPs), and inference time are also reported. For the comparative, ablation, and sensitivity analyses, the corresponding results are reported as mean ± standard deviation where repeated experiments are involved. For the noise robustness tests, diagnostic accuracy is reported under different signal-to-noise ratio (SNR) conditions.

4.6. Noise Robustness Test Settings

To further verify the robustness of the proposed method under controlled noisy conditions, this study conducted additive white Gaussian noise (AWGN)-based noise robustness experiments. Additive white Gaussian noise (AWGN) was selected as the initial noise model for three reasons. First, AWGN provides a standard and reproducible benchmark for evaluating the influence of different noise intensities under clearly defined SNR levels. Second, compared with more complex field noise, AWGN allows the noise level to be controlled independently while keeping the fault labels and operating conditions unchanged, which is suitable for isolating the effect of noise intensity on diagnostic performance. Third, AWGN has been widely used as a baseline disturbance model in machinery fault diagnosis studies, making the results easier to compare and interpret.

Note that AWGN is only a simplified representation of background noise and cannot fully cover practical industrial disturbances. In real gearbox monitoring scenarios, vibration signals may also be affected by non-Gaussian impulsive noise, colored noise caused by structural transmission paths, and nonstationary disturbances induced by load variations or environmental changes. Since this study was limited to controlled laboratory conditions, these more complex noise distributions were not included in the current robustness test. Subsequently, the same fractional-order enhancement and dual-view feature construction procedure as that used for the original samples was performed on the AWGN-corrupted signals, and the trained model was then used for testing.

The following signal-to-noise ratio (SNR) levels were considered in this study: clean, denoting the noise-free condition; 10 dB and 5 dB, representing mild-to-moderate noise interference; and 0 dB and -5 dB, representing severe noise interference. To provide a fair comparative evaluation, the same AWGN-corrupted test sets were used for the proposed method and the selected baseline models. All models were trained using the original training configuration and then tested under different SNR levels without retraining for each noise condition. For each method, its corresponding preprocessing and input construction procedures were applied to the noisy test signals before inference.

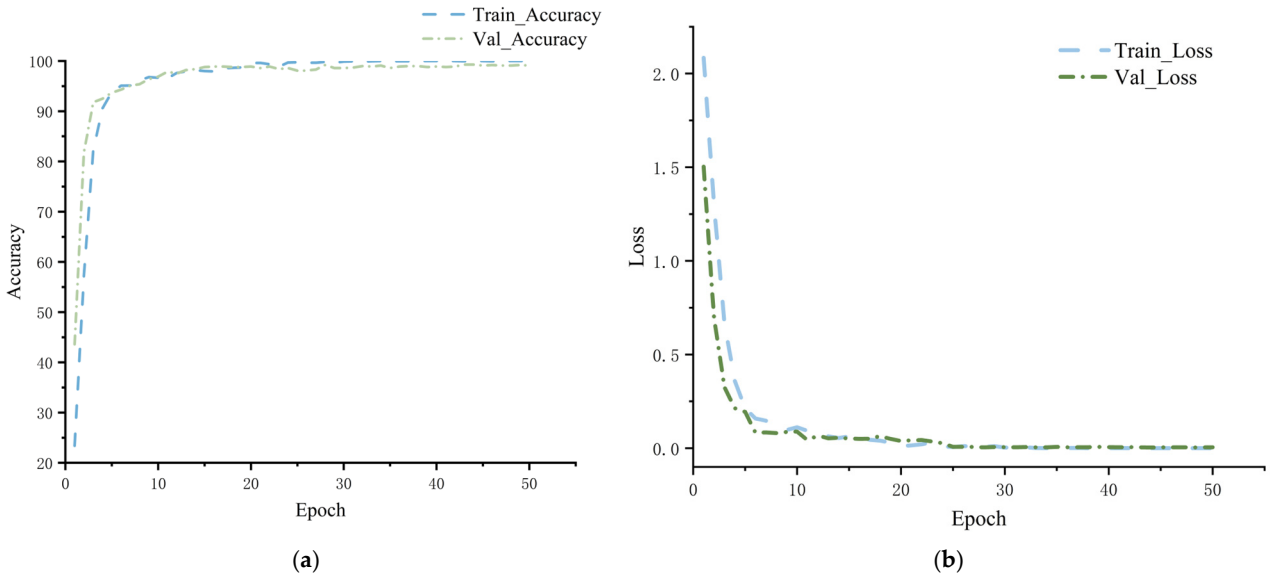
5. Results

5.1. Convergence Behavior

To analyze the stability and convergence characteristics of the proposed fractional-order-enhanced dual-view model during the training process, the accuracy (Acc) curves and loss (Loss) curves on the training and validation sets are plotted separately, as shown in Figure 8. Specifically, Figure 8a presents the curves of training and validation accuracy as a function of the number of training epochs, while Figure 8b shows the curves of training and validation loss as a function of the number of training epochs.

As shown in Figure 8a, the model accuracy increases rapidly during the early stage of training and gradually stabilizes as the number of epochs increases. After approximately 15 epochs, both the training and validation accuracy curves reach a relatively steady state. At the end of 50 training epochs, both curves maintain high recognition accuracy. As shown in Figure 8b, the loss values of the training and validation sets decrease rapidly at the beginning of training and then gradually converge to stable values. The overall

variation trends of the two curves are generally consistent, and no obvious divergence is observed. Although the validation loss exhibits slight fluctuations in the early stage, the amplitude of these fluctuations is relatively small and does not affect the overall convergence trend. Overall, the results in Figure 8 indicate that the proposed model achieves stable optimization behavior and satisfactory convergence during training.



(**a**) (**b**)

Figure 8. Iterative curves of the proposed model: (**a**) Accuracy curves for the training and validation sets; (**b**) loss curves for the training and validation sets.

5.2. Visualization of Fractional-Order-Enhanced Time–Frequency Representations

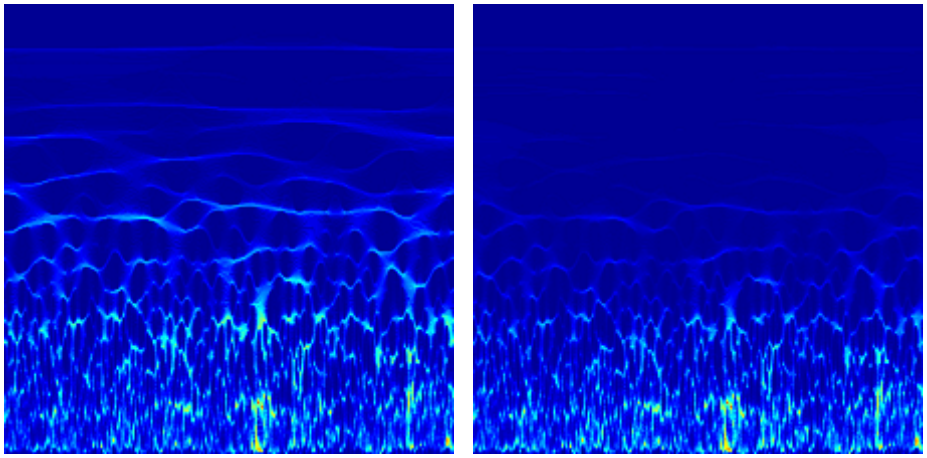
To intuitively demonstrate the signal enhancement capability of the fractional-order differential operator, the time–frequency representations before and after fractional-order preprocessing are visualized. As shown in Figure 9, the original SST image of the raw signal suffers from significant background noise and complex amplitude modulations, making the weak localized fault ridges difficult to distinguish. In contrast, in the fractional-order-enhanced SST image, the low-frequency background interference appears to be alleviated, and the high-frequency transient impacts associated with weak localized gearbox faults are more clearly highlighted. This observation is consistent with the frequency-domain interpretation in Section 2.1, where the fractional-order differential operator shows low gain near the low-frequency region and gradually increasing gain toward higher frequencies. This visual comparison suggests that the fractional-order enhancement module helps highlight fault-sensitive components, thereby providing a more informative structural input for the subsequent VMamba visual branch.

5.3. Feature Visualization Results

To visually evaluate the discriminative capability of the proposed model under different fault conditions, t-SNE is used to project both the raw features and the learned deep features of the test samples into a low-dimensional space, as shown in Figure 10.

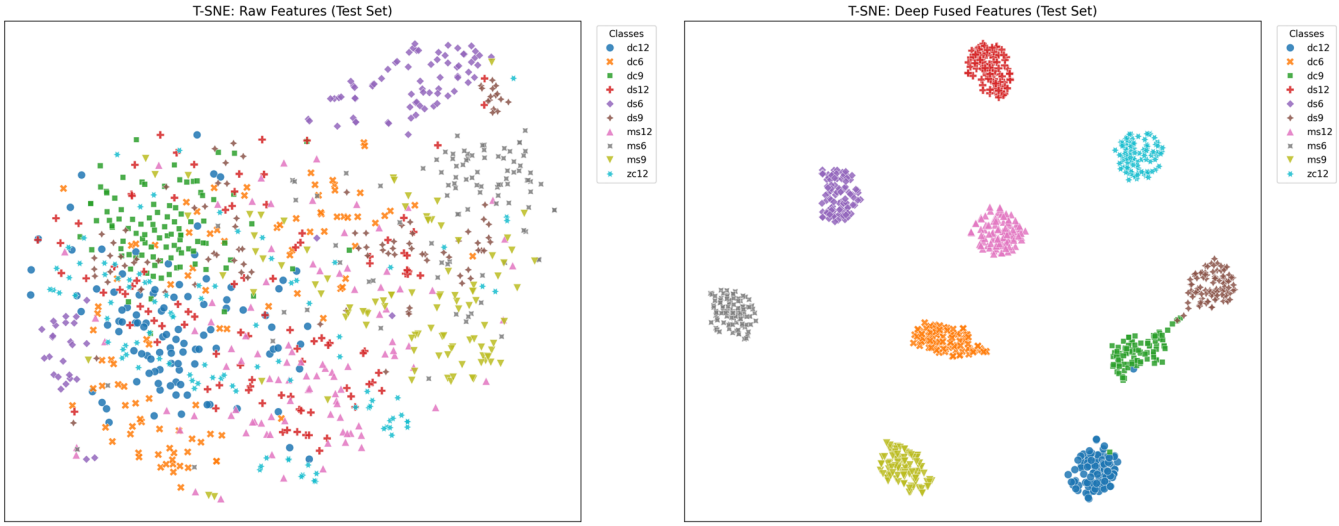
As shown in Figure 10a, the raw features are distributed in a highly scattered manner, and severe overlap exists among different fault categories, especially among samples belonging to the same fault type under different rotational speed conditions. The class boundaries are not clearly separable in the low-dimensional space. By contrast, Figure 10b shows that the deep features extracted and fused by the proposed fractional-order-enhanced VibrMamba–VMamba dual-branch model exhibit clear clustering behavior. Samples from the same category become more compact, whereas different categories show clearer visual

separation. This observation is also consistent with the classification results shown in the confusion matrix.



(**a**) (**b**)

Figure 9. Comparison of time–frequency representations: (**a**) original SST image; (**b**) fractional-order-enhanced SST image. The color intensity represents the relative time–frequency energy magnitude, with brighter regions indicating stronger energy concentrations.



(**a**) (**b**)

Figure 10. t-SNE visualization results: (**a**) 2D distribution of the raw features; (**b**) 2D distribution of the deep features. Different colors represent different health conditions and rotational speeds.

5.4. Confusion Matrix and Per-Class Recognition Results

To further evaluate the recognition performance of the proposed model for each specific fault condition, the confusion matrix of the test set is shown in Figure 11. The confusion matrix provides an intuitive description of the classification accuracy and misclassification distribution of different categories.

Based on the test results for 1000 samples, the model achieves high recognition accuracy across most categories. Six categories—zc_12, dc_6, ds_6, ds_9, ms_12, and ms_6—reached 100% classification accuracy. For dc_12, three samples were misclassified as dc_9, and one sample as ds_12. For dc_9, one sample was misclassified as dc_12, and four samples as ds_9. For ds_12, one sample was misclassified as ds_9. In addition, one sample of ms_9 was misclassified as ms_6. Overall, 11 samples were misclassified in the test set, corresponding to an overall accuracy of 98.90%.

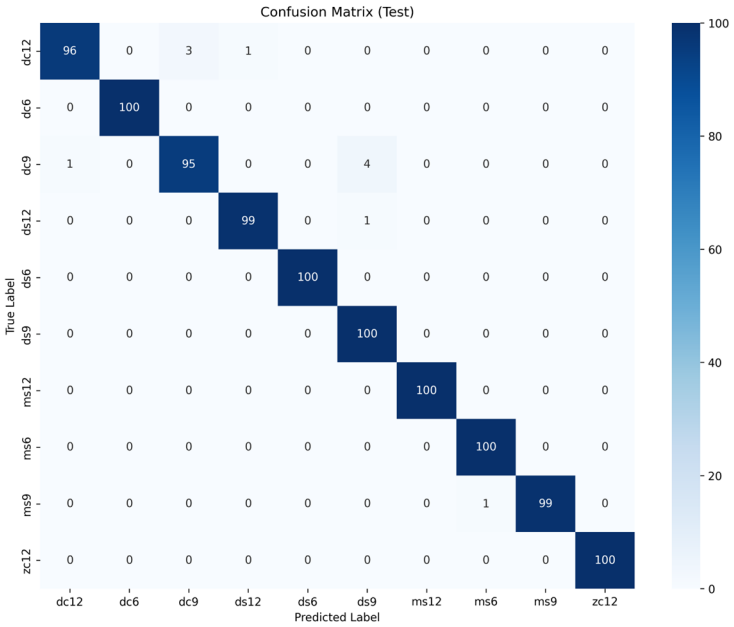


Figure 11. Confusion matrix.

The misclassification between dc_12 and dc_9 can be physically interpreted from the similarity of broken-tooth fault responses under adjacent high-speed conditions. Both dc_12 and dc_9 correspond to the same broken-tooth fault mechanism, and their vibration responses are mainly characterized by periodic impulsive components and meshing-related modulation. When the rotational speed increases, the time interval between consecutive fault-induced impacts becomes shorter, and the corresponding modulation sidebands and time–frequency ridges may become closer in the SST representation. Therefore, the high-speed broken-tooth samples at 900 r/min and 1200 r/min may share similar impulsive patterns, resonance-band excitations, and time–frequency distributions. This feature overlap makes a small number of dc_12 and dc_9 samples difficult to distinguish, especially when the fault-type-related features are more dominant than the speed-related differences.

These misclassifications indicate that the remaining errors are mainly associated with the overlap between fault mechanism similarity and rotational speed-related modulation, rather than random classification failures. At the same time, these errors also show that speed-related feature overlap remains a challenging factor in the current dataset. Therefore, the high overall accuracy should not be interpreted as evidence that the model has fully solved cross-speed generalization, especially under unseen speed conditions or incomplete healthy baselines. This also suggests that introducing speed-normalized features or order-tracking-based representations may further reduce speed-related confusion in future work.

5.5. Comparative Results

To evaluate the effectiveness of the proposed method for multi-state gearbox fault classification under the current bench-test setting with known speed variations, several representative methods were selected for comparison, including classical deep learning models (WDCNN, ResNet-50, and ViT), state-space-based models (VibrMamba and VMamba), and a multimodal fusion baseline (1D-CNN + ViT). In addition, two additional baselines were included: a model without fractional-order enhancement, Original SST + VibrMamba + VMamba, and a model in which SST is replaced with STFT, FO + VibrMamba + VMamba (STFT). The corresponding results are listed in Table 2, and the comparison is visualized in Figure 12.

Table 2. Comparison of performance evaluation metrics for different diagnostic methods.

Method	Accuracy (%)	F1 Score (%)	Kappa (%)
WDCNN	88.54 ± 0.42	88.42 ± 0.44	87.26 ± 0.47
ResNet-50	91.32 ± 0.35	91.25 ± 0.37	90.35 ± 0.40
ViT	93.45 ± 0.31	93.38 ± 0.32	92.72 ± 0.35
VibrMamba	95.80 ± 0.24	95.75 ± 0.26	95.33 ± 0.28
VMamba	96.25 ± 0.22	96.18 ± 0.23	95.83 ± 0.25
1D-CNN + ViT	97.46 ± 0.18	97.40 ± 0.19	97.17 ± 0.20
SST + VibrMamba + VMamba	97.86 ± 0.16	97.82 ± 0.17	97.65 ± 0.18
FO + VibrMamba + VMamba (STFT)	97.81 ± 0.17	97.75 ± 0.18	97.56 ± 0.19
FO + SST + VibrMamba + VMamba	98.90 ± 0.12	98.91 ± 0.11	98.78 ± 0.13

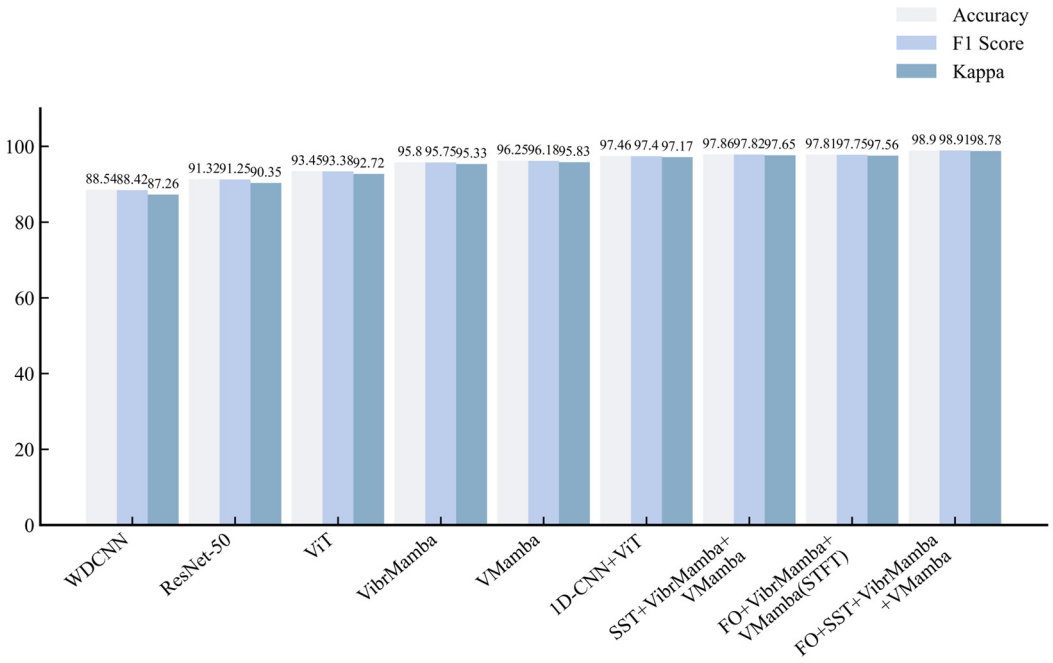


Figure 12. Bar chart comparing the diagnostic performance of different methods.

As shown in Table 2 and Figure 12, the proposed fractional-order-enhanced dual-branch method achieves the best overall performance among all methods compared. Specifically, the full model attains 98.90 ± 0.12% accuracy, 98.91 ± 0.11% F1 score, and 98.78 ± 0.13% Kappa. Compared with the dual-view baseline without fractional-order preprocessing (SST + VibrMamba + VMamba), the proposed method improves the accuracy by 1.04 percentage points. When SST is replaced with STFT, the accuracy decreases to 97.81 ± 0.17%. These results indicate that both fractional-order enhancement and the high-resolution SST representation contribute to the final performance gain under the current bench-test setting.

5.6. Comparative Noise Robustness Results

To evaluate whether the proposed fractional-order-enhanced dual-view framework provides a relative robustness advantage under noisy conditions, comparative AWGN-based noise tests were conducted for both the proposed method and selected baseline models. Additive white Gaussian noise with different intensities was added to the original vibration signals of the test set, and all methods were evaluated on the same noisy test samples under the SNR levels of clean, 10 dB, 5 dB, 0 dB, and −5 dB. The selected baselines include the multimodal fusion baseline 1D-CNN + ViT, dual-branch model without

fractional-order enhancement, and single-branch variants with fractional-order enhancement. The corresponding results are listed in Table 3.

Table 3. Comparative noise robustness results of different methods under different SNR levels.

Method	Clean	10 dB	5 dB	0 dB	−5 dB	Acc. Drop
1D-CNN + ViT	97.46	95.82	93.41	89.36	83.72	13.74
SST + VibrMamba + VMamba	97.86	96.41	94.56	91.02	86.34	11.52
FO + VibrMamba	96.85	95.87	94.03	90.72	85.91	10.94
FO + SST + VMamba	97.15	96.23	94.61	91.38	87.06	10.09
FO + SST + VibrMamba + VMamba	98.90	98.31	97.22	95.78	91.52	7.38

As shown in Table 3, all methods exhibit a performance decline as the SNR decreases, indicating that stronger noise interference makes the classification task more difficult. However, the proposed FO + SST + VibrMamba + VMamba model consistently achieves the highest accuracy across all tested SNR levels. Under clean conditions, the proposed method obtains an accuracy of 98.90%. At −5 dB SNR, it still maintains an accuracy of 91.52%, with an accuracy drop of 7.38 percentage points, which is smaller than those of the compared baseline models.

Compared with SST + VibrMamba + VMamba, which removes fractional-order enhancement, the proposed method improves the accuracy from 86.34% to 91.52% at −5 dB. This indicates that the fractional-order differential enhancement module helps strengthen fault-related transient components before dual-view feature construction, thereby improving robustness against noise interference. Compared with the single-branch variants FO + VibrMamba and FO + SST + VMamba, the full dual-view model also shows better performance under all tested SNR levels. This suggests that the collaborative use of enhanced temporal sequences and enhanced SST time–frequency images provides complementary information under noisy conditions.

In addition, the proposed method also outperforms the multimodal fusion baseline 1D-CNN + ViT, especially under severe noise conditions. At −5 dB, the accuracy of 1D-CNN + ViT decreases to 83.72%, whereas the proposed method maintains 91.52% accuracy. This comparison further supports the robustness advantage of the proposed fractional-order-enhanced dual-view Mamba-style framework.

These results demonstrate that the noise robustness of the proposed method benefits from the coordinated effect of fractional-order transient enhancement, high-concentration SST representation, and collaborative temporal–visual feature fusion. Nevertheless, since the current noise test is based on AWGN, the results should still be interpreted as comparative robustness verification under controlled noise settings rather than complete evidence of robustness under complex industrial noise.

5.7. Ablation Study on Fractional-Order Enhancement

Ablation experiments were conducted to further evaluate the contribution of different components, especially the proposed fractional-order enhancement (FO) module. The comparison covers the baseline model without FO, single-branch models with FO, and the full proposed model. The ablation results are presented in Table 4.

As shown in Table 4, the dual-view baseline without fractional-order preprocessing (SST + VibrMamba + VMamba) achieves 97.86 ± 0.16% accuracy. When only the 1D temporal branch is retained with fractional-order enhancement (FO + VibrMamba), the accuracy is 96.85 ± 0.21%. When only the 2D visual branch is retained (FO + SST + VMamba), the accuracy reaches 97.15 ± 0.20%. After integrating both branches with fractional-order enhancement,

the full model achieves the highest accuracy of 98.90 ± 0.12%. These results indicate that the performance gain comes from the coordinated effect of fractional-order preprocessing and dual-view collaborative modeling, rather than from either single branch alone.

Table 4. Results of ablation experiments.

Method	Accuracy (%)	F1 Score (%)	Kappa (%)
SST + VibrMamba + VMamba	97.86 ± 0.16	97.82 ± 0.17	97.65 ± 0.18
FO + VibrMamba	96.85 ± 0.21	96.80 ± 0.22	96.42 ± 0.24
FO + SST + VMamba	97.15 ± 0.20	97.10 ± 0.21	96.85 ± 0.22
FO + SST + VibrMamba + VMamba	98.90 ± 0.12	98.91 ± 0.11	98.78 ± 0.13

5.8. Sensitivity Analysis of the Fractional Order

The fractional order α determines the filtering and enhancement characteristics of the preprocessing module. A smaller α may preserve more low-frequency information but may be insufficient to enhance weak impulsive fault components, whereas a larger α tends to strengthen high-frequency components but may also amplify noise and unstable fluctuations. Therefore, α controls the trade-off between transient feature enhancement and noise sensitivity. To justify the selection of α, a sensitivity analysis was performed by varying α from 0.2 to 1.0 with a step size of 0.2 under the same data partition, network architecture, and training parameters. The experimental results are listed in Table 5.

Table 5. Impact of different fractional orders (α) on diagnostic performance.

Fractional Order (α)	Accuracy (%)	F1 Score (%)	Kappa (%)
0.2	97.35 ± 0.19	97.30 ± 0.21	97.12 ± 0.20
0.4	98.12 ± 0.16	98.05 ± 0.17	97.88 ± 0.18
0.6	98.90 ± 0.12	98.91 ± 0.11	98.78 ± 0.13
0.8	98.45 ± 0.14	98.40 ± 0.15	98.25 ± 0.16
1.0	97.65 ± 0.18	97.61 ± 0.19	97.42 ± 0.20

To further link the numerical sensitivity results with the frequency-domain behavior of the fractional-order operator, the magnitude responses of the discrete Grünwald–Letnikov fractional-order differential operator are plotted in Figure 13. According to the frequency-domain analysis in Section 2.1, the magnitude response can be expressed as $\left|H_\alpha(e^{j\omega})\right| = T_s^{-\alpha}[2|sin(\omega/2)|]^\alpha$, where ω denotes the discrete angular frequency. In Figure 13, the proportional factor $T_s^{-\alpha}$ is omitted for clearer comparison of the relative frequency-dependent amplification behavior under different fractional orders. This comparison is intended to illustrate how different fractional orders change the frequency-dependent amplification behavior of the Grünwald–Letnikov operator, thereby linking the numerical sensitivity results to the physical enhancement mechanism of gearbox fault impulses.

As shown in Figure 13, the response with $\alpha = 0.2$ exhibits a relatively shallow overall increase across the frequency range, indicating weak frequency selectivity and insufficient enhancement of fault-related transient components. In contrast, $\alpha = 1.0$, corresponding to the standard first-order derivative, exhibits the strongest high-frequency amplification and may therefore amplify high-frequency noise and irrelevant fluctuations. The response with $\alpha = 0.6$ lies between these two cases, providing a moderate frequency-selective enhancement behavior. This observation explains why $\alpha = 0.6$ achieves better balance between fault-related transient enhancement and noise sensitivity in the sensitivity analysis.

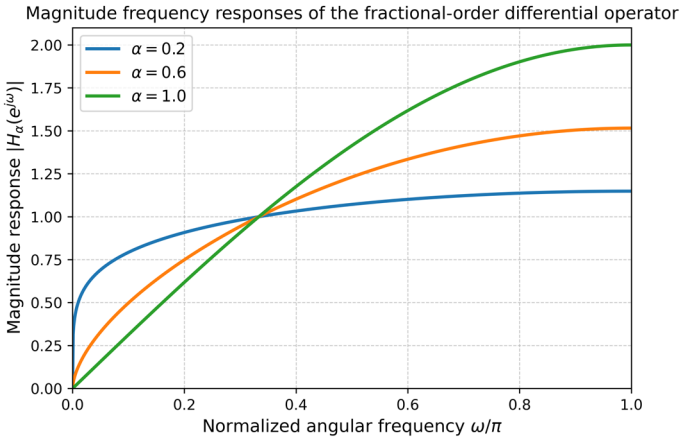


Figure 13. Magnitude frequency responses of the discrete Grünwald–Letnikov fractional-order differential operator with $\alpha = 0.2$, $\alpha = 0.6$, and $\alpha = 1.0$.

As shown in Table 5, the fractional order α has a clear influence on the final diagnostic performance. When $\alpha = 0.2$, the enhancement effect is relatively weak, and the model achieves an accuracy of $97.35 \pm 0.19\%$. This indicates that an excessively small fractional order may not sufficiently strengthen the weak impulsive components related to localized gearbox faults. When α increases from 0.2 to 0.6, the diagnostic performance gradually improves, and the highest accuracy, F1 score, and Kappa coefficient are obtained at $\alpha = 0.6$, reaching $98.90 \pm 0.12\%$, $98.91 \pm 0.11\%$, and $98.78 \pm 0.13\%$, respectively.

However, when α further increases to 0.8 and 1.0, the performance begins to decline. In particular, $\alpha = 1.0$ corresponds to standard first-order differentiation, which can excessively amplify high-frequency components, including noise and irrelevant fluctuations. This trend is consistent with the frequency response comparison in Figure 13. Specifically, compared with $\alpha = 0.2$ and $\alpha = 1.0$, $\alpha = 0.6$ provides a more balanced frequency-dependent amplification behavior, thereby enhancing fault-related transient impulses while avoiding excessive high-frequency noise amplification. Therefore, this study selects $\alpha = 0.6$ because it achieves the best quantitative performance and provides a suitable balance between fault-related transient impulse enhancement and noise suppression under the current dataset and experimental settings. This value serves as a global compromise identified for the present bench-test dataset and parameter configuration, rather than a universally optimal fractional order for all gearbox diagnosis tasks.

5.9. Computational Cost of the Model and Full Preprocessing–Inference Pipeline

For practical condition-monitoring tasks involving gearbox vibration signals, computational cost should be evaluated from two perspectives: the model-level inference cost and the full preprocessing–inference pipeline cost. The former reflects the efficiency of the trained neural network during the forward pass, whereas the latter includes all essential steps required by the proposed method, including fractional-order differential preprocessing, SST time–frequency image generation, and network inference. Therefore, this section first reports the model-level computational cost in Table 6 and then further quantifies the end-to-end processing time of the complete pipeline in Table 7.

As shown in Table 6, the self-attention-based fusion baseline (1D-CNN + ViT) has the highest computational overhead, with 12.58 M parameters, 3.52 G FLOPs, and 18.9 ± 0.5 ms of inference time. In comparison, the proposed model requires 9.86 M parameters and 2.35 G FLOPs, and its average network forward-pass inference time is 13.7 ± 0.4 ms per sample on the current GPU platform. These results suggest that the proposed framework achieves a better balance between diagnostic performance and model-level computational cost than the Transformer-based fusion baseline.

Table 6. Model-level computational complexity and network forward-pass inference time.

Method	Params (M)	FLOPs (G)	Inference Time (ms)
1D-CNN + ViT	12.58	3.52	18.9 ± 0.5
VibrMamba	3.15	0.85	4.6 ± 0.2
VMamba	6.62	1.48	8.8 ± 0.3
FO + SST + VibrMamba + VMamba	9.86	2.35	13.7 ± 0.4

Table 7. End-to-end processing time of the proposed full preprocessing–inference pipeline.

Component	Symbol	Time per Sample (ms)
Fractional-order differential preprocessing	T_{FO}	2.67 ± 0.33
SST time–frequency image generation	T_{SST}	60.57 ± 2.09
Network forward-pass inference	T_{net}	13.70 ± 0.40
Total end-to-end processing	T_{total}	76.94 ± 2.16

The inference time reported in Table 6 refers only to the average forward-pass time of the trained network. It does not include the fractional-order differential preprocessing or the SST time–frequency image generation steps. Therefore, readers should interpret Table 6 as a model-level efficiency comparison rather than as the computational cost of the full diagnostic pipeline. To provide a more complete evaluation, the end-to-end processing time of the proposed method is further analyzed. In this study, the end-to-end processing time refers to the average time required to process one 2048-point vibration sample from raw signal input to final classification output, including fractional-order differential preprocessing, SST time–frequency image generation, and network forward-pass inference. It is defined as

$$T_{\mathrm{total}} = T_{\mathrm{FO}} + T_{\mathrm{SST}} + T_{\mathrm{net}}, \tag{19}$$

where T_{FO}, T_{SST}, and T_{net} denote the fractional-order differential preprocessing time, SST time–frequency image generation time, and network forward-pass inference time, respectively. The fractional-order preprocessing and SST image generation times were measured using the same MATLAB R2021b implementation as that used to construct the time–frequency images. The file loading and image saving times were excluded from the measurement because they mainly depend on hardware and storage configurations rather than on the diagnostic algorithm itself. The measured results are listed in Table 7.

As shown in Table 7, fractional-order differential preprocessing requires only 2.67 ± 0.33 ms per sample, indicating that the fractional-order enhancement module introduces a relatively small computational burden. In contrast, SST time–frequency image generation requires 60.57 ± 2.09 ms per sample and accounts for the largest proportion of the full preprocessing–inference pipeline. After adding the network forward-pass inference time of 13.70 ± 0.40 ms, the total end-to-end processing time reaches 76.94 ± 2.16 ms per sample.

These results indicate that the trained dual-branch learning model itself remains computationally efficient, but the SST-based representation construction is the dominant time-consuming step in the complete diagnostic pipeline. Therefore, the computational advantage of the proposed Mamba-style model should be understood mainly at the model-inference level. For future real-time online deployment, the preprocessing latency introduced by SST should be carefully considered. Lightweight SST implementation, parallel preprocessing, GPU acceleration, or replacing SST with a faster time–frequency approximation may be necessary in practical edge-monitoring scenarios.

6. Discussion

6.1. Why Fractional-Order Enhancement Works

As demonstrated in the comparative and ablation experiments, the performance gain associated with the fractional-order differential preprocessing module can be interpreted from both the memory properties of fractional calculus and the frequency-domain behavior of the operator. As analyzed in Section 2.1, after omitting the constant factor $T_s^{-\alpha}$, the magnitude response of the discrete Grünwald–Letnikov fractional-order differential operator is proportional to $[2|sin(\omega/2)|]^{\alpha}$. Therefore, the operator has low gain near the low-frequency region and gradually increasing gain toward higher frequencies. This behavior helps suppress slow-varying baseline components while enhancing transient impulsive components related to localized gearbox faults.

Unlike standard first-order differentiation, the fractional-order derivative with $0 < \alpha < 1$ provides a smoother frequency-dependent amplification behavior. Therefore, it can enhance fault-related impulses without excessively amplifying high-frequency noise. In the present study, this preprocessing step appears to strengthen fault-related impulsive components while retaining part of the nonlinear dynamic details of the signal, thereby providing more informative inputs for both the 1D temporal branch and 2D visual branch. The sensitivity analysis further indicates that $\alpha = 0.6$ achieves the best diagnostic performance under the current experimental setting, which supports the interpretation that a moderate fractional order can provide a suitable balance between transient enhancement and noise sensitivity.

From the perspective of gearbox vibration dynamics, localized gear faults usually generate short-duration impulsive responses whose energy is distributed over relatively high-frequency components and modulated by rotational speed-related dynamics. A very small fractional order may not sufficiently enhance these weak impulsive signatures, whereas a very large order, particularly $\alpha = 1.0$, behaves similarly to first-order differentiation and may excessively amplify high-frequency noise. The value $\alpha = 0.6$ provides intermediate frequency-dependent amplification behavior, which is consistent with both the sensitivity results and the frequency-response comparison in Figure 13. Therefore, the results suggest that $\alpha = 0.6$ provides a trade-off between enhancing localized fault-induced transients and maintaining robustness against high-frequency noise under the current bench-test setting with known speed variations, rather than serving as a fixed order that is universally suitable for all gearbox fault diagnosis scenarios.

6.2. Effectiveness of Dual-View Representation and Collaborative Modeling

The results presented above indicate that the proposed method benefits from the collaborative use of fractional-order-enhanced vibration sequences and fractional-order-enhanced SST time–frequency images. The one-dimensional sequence preserves temporal dynamics, impulsive responses, and long-range dependencies, whereas the SST time–frequency image provides a more structured description of nonstationary fault evolution in the time–frequency domain. By combining these two complementary views, the model can simultaneously exploit enhanced temporal information and structural distribution information.

The comparative and ablation results further support this observation. Both Vibr-Mamba and VMamba achieve satisfactory performance when used separately; however, the dual-branch collaborative framework produces higher accuracy than either single-branch configuration. This suggests that the proposed architecture can exploit the complementary strengths of sequential modeling and visual feature extraction, thereby improving fault identification under the current bench-test conditions with known speed variations.

More importantly, the improvement should not be interpreted as simply increasing the number of feature channels. The paired dual-view construction ensures that both branches derive from the same fractional-order-enhanced signal source. At the same time,

the adaptive gated fusion mechanism assigns different branch contributions according to each input sample. Together with cross-branch residual enhancement, this design explains why the proposed method outperforms the single-branch variants and supports its methodological distinction from a loose engineering combination of existing components.

6.3. Why Fractional-Order SST Improves Fault Characterization

The advantage of the fractional-order SST over STFT in the current study may be attributed to its stronger time–frequency concentration capability. For weak localized gearbox fault features, characteristic components are often sparsely distributed and may be easily submerged by noise or masked by variations in rotational speed. The fractional-order-enhanced SST can improve the concentration of fault-related energy around more informative time–frequency trajectories, making the resulting image representation more suitable for subsequent feature extraction by the visual branch.

The comparison between the full model and STFT-based variant also supports this point. Although both methods adopt the same network architecture and fractional-order preprocessing, replacing SST with STFT leads to a measurable performance drop. This suggests that the quality and resolution of the time–frequency representation itself play a substantial role in the final diagnosis performance, rather than the network architecture alone.

6.4. Analysis of Misclassification Patterns Under Similar Operating Conditions

The confusion matrix shows that most misclassifications are concentrated among categories with similar rotational speeds or similar fault mechanisms, especially between fault classes under adjacent speed conditions. This phenomenon suggests that these classes may share similar impulsive modulation patterns and time–frequency distributions under comparable operating conditions. For example, broken-tooth samples at adjacent high-speed conditions may exhibit similar periodic impact patterns, meshing-related modulation sidebands, and resonance-band excitations, which can lead to partial overlap in the learned feature space.

From a diagnostic perspective, such errors are understandable because weak fault signatures under similar speed conditions may produce overlapping response characteristics in both the signal domain and the time–frequency domain. Therefore, the observed misclassification pattern reflects not only the remaining difficulty of the task but also the intrinsic similarity among some categories in the current dataset.

The present dataset does not constitute a strict cross-speed generalization benchmark. Since healthy baseline data are available only at 1200 r/min, whereas fault-condition data are available at 600, 900, and 1200 r/min, the learned decision boundaries may partly reflect both fault-type differences and speed-related distribution differences. Therefore, the term "variable speed" in this study refers to diagnosis under known speed-varying bench-test conditions, rather than generalization to completely unseen rotational speeds. A more rigorous evaluation of speed generalization requires future datasets containing complete healthy and faulty categories across multiple training and unseen testing speeds.

To further reduce these errors, several improvement strategies can be considered in future work. First, speed-normalized features can be introduced to reduce the influence of rotational speed differences on the learned representation. For example, vibration features can be normalized with respect to shaft speed or characteristic meshing frequency, so that samples with the same fault mechanism but different speeds are represented in a more comparable feature space. Second, order-tracking-based representations can be incorporated to transform time–frequency features into the order domain, where speed-dependent spectral shifts can be alleviated. Third, a speed-aware auxiliary branch or

condition-embedding module can be introduced into the network to explicitly encode rotational speed information and guide the model to distinguish fault-mechanism-related features from speed-related modulation. Finally, expanding the dataset to include more speed levels, load conditions, and healthy baseline data at different speeds would help improve the model's ability to learn more robust decision boundaries under variable operating conditions.

6.5. Noise Robustness and Practical Implications

The comparative noise robustness results demonstrate that the proposed method maintains relatively high accuracy under different AWGN levels and outperforms the selected baseline models under the same noisy test conditions. This behavior may be attributed to three aspects. First, the fractional-order differential preprocessing strengthens fault-related transient components before feature construction, which helps reduce the masking effect caused by background noise. Second, the SST-based time–frequency representation improves the concentration of fault-related energy in the visual view, making the two-dimensional input more informative under noisy conditions. Third, the collaborative temporal–visual fusion mechanism integrates the long-range temporal dependencies extracted by VibrMamba and the global time–frequency structural features extracted with VMamba, thereby providing complementary information when either single view is partially degraded by noise.

From a practical perspective, this characteristic is relevant because gearbox vibration signals under known speed variations and noisy conditions can be influenced by background vibration, external disturbances, and operating-condition fluctuations. Therefore, the observed robustness suggests that the proposed method has potential for condition-monitoring tasks involving nonstationary gearbox signals. However, further validation on field-measured data is still required before drawing stronger conclusions regarding field deployment, as summarized in the limitations and future directions in Section 7.

7. Conclusions

This study proposed a fractional-order-enhanced dual-view collaborative framework for gearbox fault classification under controlled laboratory conditions with known speed variations and noise disturbances. The proposed method integrates Grünwald–Letnikov fractional-order differential enhancement, paired one-dimensional vibration-sequence and SST time–frequency representations, and collaborative VibrMamba–VMamba state-space modeling.

Experimental results on the current gearbox bench-test dataset show that the proposed method achieves 98.90% diagnostic accuracy under clean test conditions and maintains 91.52% accuracy at -5 dB SNR. The comparative experiments, ablation study, noise robustness evaluation, t-SNE visualization, confusion matrix analysis, and fractional-order sensitivity analysis indicate that the performance improvement mainly benefits from the combined effect of fractional-order transient enhancement, concentrated SST time–frequency representation, and complementary temporal–visual feature fusion. In particular, the sensitivity analysis and frequency response comparison shows that $\alpha = 0.6$ provides a suitable balance between fault-related impulse enhancement and high-frequency noise sensitivity under the current dataset and parameter configuration.

Several limitations should still be acknowledged. First, this study is limited to controlled laboratory conditions, and the results should be interpreted as bench-level validation rather than direct evidence of field-level generalization. Second, the current dataset does not fully cover broader load levels, more diverse speed combinations, installation variations, long-term degradation processes, or field-measured industrial disturbances. In

particular, because healthy baseline data at 600 and 900 r/min are not available in the current dataset, the reported results should not be interpreted as strict evidence of cross-speed generalization. Third, the noise robustness evaluation is mainly based on AWGN, which cannot fully represent non-Gaussian impulsive noise, colored noise, or nonstationary noise distributions in practical environments. In addition, although the SST-based representation improves time–frequency concentration, it introduces the highest preprocessing cost in the full pipeline. Therefore, it should be carefully considered when evaluating future real-time deployment feasibility.

Future work will focus on four directions. First, adaptive fractional-order selection will be investigated to avoid manually fixing the fractional order under changing operating conditions. Second, field-measured gearbox data with a wider range of load, speed, and degradation conditions will be introduced to evaluate model generalization further. Third, more realistic disturbance sources, including impulsive, colored, and nonstationary noise, will be considered in future robustness evaluations, and speed-normalized features, order-tracking-based representations, and speed-aware condition embeddings will be explored to reduce speed-related misclassification under variable-speed conditions. Fourth, model compression and edge-computing deployment strategies, such as pruning, quantization, lightweight SST implementation, and parallel preprocessing, will be studied to improve the feasibility of real-time online monitoring on embedded or edge devices.

Author Contributions: Conceptualization, F.X. and K.N.; methodology, F.X. and Z.S.; validation, F.X. and K.N.; investigation, F.X., S.W. and H.C.; writing—original draft preparation, F.X. and K.N.; writing—review and editing, F.X. and Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 52265068.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, J.; Zheng, C.; Wu, Z.; Chen, H.; Wang, X.-B.; Yang, Z.-X. Review of Fault Diagnosis for Rotating Machinery: Prior Knowledge Integration in Data-Driven Methods Benefits Model Interpretability and Generalizability. *Mech. Syst. Signal Process.* **2026**, *242*, 113623. [CrossRef]
2. Neupane, D.; Bouadjenek, M.R.; Dazeley, R.; Aryal, S. Data-Driven Machinery Fault Diagnosis: A Comprehensive Review. *Neurocomputing* **2025**, *627*, 129588. [CrossRef]
3. Ahmad, H.; Cheng, W.; Xing, J.; Wang, W.; Du, S.; Li, L.; Zhang, R.; Chen, X.; Lu, J. Deep Learning-Based Fault Diagnosis of Planetary Gearbox: A Systematic Review. *J. Manuf. Syst.* **2024**, *77*, 730–745. [CrossRef]
4. Liu, W.; Zhang, Z.; Ye, Z.; He, Q. A Novel Intelligent Fault Diagnosis Method for Gearbox Based on Multi-Dimensional Attention Denoising Convolution. *Sci. Rep.* **2024**, *14*, 24688. [CrossRef] [PubMed]
5. Yuan, B.; Li, Y.; Chen, S. Efficient Gearbox Fault Diagnosis Based on Improved Multi-Scale CNN with Lightweight Convolutional Attention. *Sensors* **2025**, *25*, 2636. [CrossRef] [PubMed]
6. Nguyen, V.-T.; Diep, Q.B. Vibration-based gearbox fault diagnosis using a multi-scale convolutional neural network with depth-wise feature concatenation. *PLoS ONE* **2025**, *20*, e0324905. [CrossRef]
7. Cheng, X.; Dou, S.; Du, Y.; Wang, Z. Gearbox Fault Diagnosis Method Based on Lightweight Channel Attention Mechanism and Transfer Learning. *Sci. Rep.* **2024**, *14*, 743. [CrossRef]
8. Dou, S.; Dai, F.; Du, Y.; Liu, F.; Li, T.; Yao, D.; Bai, H. Deep Learning-Based Fault Diagnosis of Gearbox Using CBAM-DSC-ResNeXt50 Structure. *J. Vib. Eng. Technol.* **2025**, *13*, 418. [CrossRef]

9. Xie, F.; Wang, G.; Shang, J.; Sun, E.; Xie, S. Gearbox Fault Diagnosis Based on Multisensor Deep Spatiotemporal Feature Representation. *Mathematics* **2023**, *11*, 2679. [CrossRef]
10. Zhang, Y.; Ding, J.; Li, Y.; Ren, Z.; Feng, K. Multi-Modal Data Cross-Domain Fusion Network for Gearbox Fault Diagnosis under Variable Operating Conditions. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108236. [CrossRef]
11. Zhang, X.; Wang, H.; Wang, C.; Liu, M.; Xu, G. Time-Segment-Wise Feature Fusion Transformer for Multi-Modal Fault Diagnosis. *Eng. Appl. Artif. Intell.* **2024**, *138*, 109358. [CrossRef]
12. Xiao, X.; Li, C.; He, H.; Huang, J.; Yu, T. Rotating Machinery Fault Diagnosis Method Based on Multi-Level Fusion Framework of Multisensor Information. *Inf. Fusion* **2025**, *113*, 102621. [CrossRef]
13. Li, X.; Xie, F.; Pan, X.; Yu, L.; Fu, H.; Wu, Z. Intelligent Fault Diagnosis for a Wind Turbine Gearbox via Multisensor Information Fusion with Adaptive Weights. *Energy* **2025**, *335*, 138352. [CrossRef]
14. Mao, G.; Wang, J.; Sun, Y. An Explainable Intelligent Fault Diagnosis for Rotating Machinery via Multi-Source Information Fusion Under Noisy Environments and Small Sample Conditions. *Sensors* **2026**, *26*, 1713. [CrossRef]
15. Daubechies, I.; Lu, J.; Wu, H.-T. Synchrosqueezed Wavelet Transforms: An Empirical Mode Decomposition-like Tool. *Appl. Comput. Harmon. Anal.* **2011**, *30*, 243–261. [CrossRef]
16. Hazra, B.; Narasimhan, S. Gearbox Fault Detection Using Synchro-Squeezing Transform. *Procedia Eng.* **2016**, *144*, 187–194. [CrossRef]
17. Zhu, X.; Shi, W.; Huang, Z.; Shi, L. Geometry-Based Synchrosqueezing S-Transform with Shifted Instantaneous Frequency Estimator Applied to Gearbox Fault Diagnosis. *Sensors* **2025**, *25*, 540. [CrossRef]
18. Hoang, N.-P.; Nguyen, T.-D.; Nguyen, T.-H.; Pham, D.-H.; Nguyen, P.-D.; Nguyen, T.-V.-H. A Data-Driven Framework for Fault Diagnostics in Gearbox Health Monitoring Under Nonstationary Conditions. *Processes* **2025**, *13*, 1663. [CrossRef]
19. Bekaddour Benattia, M.E.; Habbouche, H.; Benkedjouh, T.; Amirat, Y.; Benbouzid, M. Robust Planetary Gearbox Fault Diagnosis through Time–Frequency Analysis and Transfer Learning. *Eng. Appl. Artif. Intell.* **2026**, *167*, 113844. [CrossRef]
20. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752. [CrossRef]
21. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. VMamba: Visual State Space Model. In *Proceedings of the Advances in Neural Information Processing Systems*; Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 103031–103063.
22. Yi, H.; Li, D.; Lu, Z.; Jin, Y.; Duan, H.; Hou, L.; Duraihem, F.Z.; Awwad, E.M.; Saeed, N.A. VibrMamba: A Lightweight Mamba Based Fault Diagnosis of Rotating Machinery Using Vibration Signal. *Measurement* **2025**, *249*, 116881. [CrossRef]
23. Xia, E.; Liu, Y.; Gong, J.; Dai, X.; Pan, T.; Wang, S. BMTM-Net: A Rotating Machinery Fault Diagnosis Network Based on 2D-1D Fusion with Bidirectional Multi-Granularity Transformer-Mamba. *Neurocomputing* **2025**, *656*, 131293. [CrossRef]
24. Xu, J.; Zhao, Y.; Bao, W.; Hao, C. Fault Diagnosis of Motor Bearing in Complex Scenarios Based on Mamba and Indicative Contrastive Learning. *Eng. Appl. Artif. Intell.* **2025**, *146*, 110216. [CrossRef]
25. Zhang, H.; Li, Y.; Zhang, X.; Feng, H. Multi-Scale Group Mamba Network with Structural Attention for Rotating Machinery Fault Diagnosis Using Multisensor Data. *Adv. Eng. Inform.* **2025**, *67*, 103521. [CrossRef]
26. Li, T.; Wu, X.; Luo, Z.; Chen, Y.; He, C.; Ding, R.; Zhang, C.; Yang, J. A Bearing Fault Diagnosis Method under Small Sample Conditions Based on the Fractional Order Siamese Deep Residual Shrinkage Network. *Fractal Fract.* **2024**, *8*, 134. [CrossRef]
27. Xie, F.; Song, C.; Wang, Y.; Song, M.; Zhou, S.; Xie, Y. Rolling Bearing Fault Diagnosis Based on Fractional Constant-Q Nonstationary Gabor Transform and VMamba-Conv. *Fractal Fract.* **2025**, *9*, 515. [CrossRef]
28. Xie, F.; Li, G.; Song, C.; Song, M. The Early Diagnosis of Rolling Bearings' Faults Using Fractional Fourier Transform Information Fusion and a Lightweight Neural Network. *Fractal Fract.* **2023**, *7*, 875. [CrossRef]
29. Wang, J.; Qiao, L.; Ye, Y.; Chen, Y. Fractional Envelope Analysis for Rolling Element Bearing Weak Fault Feature Extraction. *IEEE/CAA J. Autom. Sin.* **2017**, *4*, 353–360. [CrossRef]
30. Yang, J.; Li, X.; Mao, M. Research on Improved Transformer Fault Diagnosis Method Driven by IBKA-VMD and Hierarchical Fractional Order Attention Entropy Synergy. *Fractal Fract.* **2026**, *10*, 195. [CrossRef]
31. Jiao, R.; Li, S.; Ding, Z.; Yang, L.; Wang, G. Fault Diagnosis of Rolling Bearing Based on BP Neural Network with Fractional Order Gradient Descent. *J. Vib. Control.* **2024**, *30*, 2139–2153. [CrossRef]
32. Bohner, M.; Tunç, O.; Korkmaz, E. On the Fundamental Qualitative Properties of Integro-Delay Differential Equations. *Commun. Nonlinear Sci. Numer. Simul.* **2023**, *125*, 107320. [CrossRef]
33. Caputo, M.; Fabrizio, M. A New Definition of Fractional Derivative without Singular Kernel. *Prog. Fract. Differ. Appl.* **2015**, *1*, 73–85. [CrossRef]
34. Losada, J.; Nieto, J. Properties of a New Fractional Derivative without Singular Kernel. *Prog. Fract. Differ. Appl.* **2015**, *1*, 87–92. [CrossRef]

35. Yasin, F.; Afzal, Z.; Saleem, M.S.; Jahangir, N.; Shang, Y. Hermite–Hadamard Type Inequality for Non-Convex Functions Employing the Caputo–Fabrizio Fractional Integral. *Res. Math.* **2024**, *11*, 2366164. [CrossRef]
36. Gomolka, Z.; Zeslawska, E.; Olbrot, L. Fractional-Order Neural Network for Detecting Process Deviations in Optical Fiber Cable Manufacturing. *Sci. Rep.* **2026**, *16*, 6677. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.