

Article

Multimodal Prompt-Tuning Large Language Model for Machinery Fault Diagnosis with Sound-Vibration Signals

Yu Tian ^{1,2,*} , Juntao Wang ², Ying Zheng ³, Jiangling Wang ² and Caiming Zhong ² ¹ School of Mechanical Engineering, Tongji University, Shanghai 201814, China² College of Science & Technology, Ningbo University, Ningbo 315300, China; wjt7292000@gmail.com (J.W.); 2511170010@nbu.edu.cn (J.W.); zhongcaiming@nbu.edu.cn (C.Z.)³ Ningbo Fotile Kitchen Ware Co., Ltd., No. 1266, Binhai 6th Road, Qianwan New Area, Ningbo 315336, China; zhengying@fotile.com

* Correspondence: tianyu@nbu.edu.cn

Abstract

Vibration and sound-sensing technology has become the mainstream perception method in the field of mechanical fault diagnosis. However, most current diagnostic models still focus on single-modality data analysis, which to some extent restricts further improvement of diagnostic performance. Multimodal fault-diagnosis models are still relatively lacking in existing research. In addition, most methods usually simply concatenate or early fuse vibration and sound signals and directly input them into the model, failing to deeply explore and effectively fuse the discriminative features of each modality. Moreover, the strong generalization ability of large language models (LLMs) demonstrates their potential for application in other fields. Therefore, it is interesting to perform machinery fault diagnosis based on LLMs with sound and vibration signals. This study proposes a new multimodal DNN, namely a sound-vibration large language model (SVLLM), for multimodal feature fusion and machinery fault diagnosis. Firstly, a multimodal prompt-tuning module is proposed to fine-tune LLMs to learn domain knowledge of fault diagnosis based on vibration signals and sound signals. It can also adaptively fuse sound signals and vibration signals. Secondly, a freeze-thaw module is proposed to freeze the pre-trained key layers and thaw the layers that need to be learned to adapt the model to fault diagnosis tasks. Thirdly, the multihead cross-attention module is proposed to fully extract the characteristics of vibration signals and sound signals through modal expansion interaction. The experimental results on the motor dataset and Ottawa dataset show that SVLLM has good performance in feature extraction and machinery fault diagnosis with sound and vibration signals.

Keywords: sound–vibration; feature fusion; deep learning; machinery fault diagnosis; large language model



Academic Editor: José de Jesús Rangel Magdaleno

Received: 30 June 2026

Revised: 12 August 2026

Accepted: 23 August 2026

Published: 31 August 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Due to its excellent dielectric properties, good processability, and economy, rotating machinery has become an indispensable component of industrial systems [1]. Its compact structure, strong load-bearing capacity, and ease of standardized mass production have further promoted its widespread application in the industrial field [2]. However, under complex working conditions such as high speed, heavy load, severe temperature changes, or high dust, key components of rotating machinery (e.g., bearings, motors, gearboxes, etc.) are prone to various damages during long-term operation [3]. Once a malfunction occurs, it not only affects the normal operation of the equipment, but may also trigger a chain

shutdown accident [4]. Therefore, the ability to perform timely surveillance and intelligent analysis of rotating machinery health is crucial for enhancing equipment reliability and facilitating proactive, predictive upkeep [5].

Vibration signals often contain periodic impact components and modulation characteristics corresponding to bearing faults, making them an important source of information for bearing state recognition [6]. The fusion of signal processing and artificial intelligence has led to numerous approaches designed to automatically mine fault-relevant characteristics from vibration data, facilitating the accurate diagnosis and proactive alerting of bearing fault categories [7]. Huang et al. [8] proposed an intelligent bearing fault diagnosis method based on a local connection neural network and feature optimization, which solves the problem of limited diagnostic performance under incomplete vibration signals. Xiao et al. [9] proposed a blade fault diagnosis method based on transfer-function reconstruction and an autoencoder, which solves the problem of inaccurate diagnosis of turbine-blade faults caused by signal attenuation. However, vibration signals are collected through contact and are susceptible to interference from surrounding vibration sources [10]. This will cause a lot of noise to be mixed in the vibration signal, which will affect the effectiveness of the model diagnosis.

As a non-contact technology, sound monitoring is very suitable for fault diagnosis of rotating machinery [11]. By collecting sound signals through a microphone, remote measurement can be achieved, making it easy to quickly and comprehensively assess the health status of large equipment or components that are difficult to access in installation locations [12]. Due to the lack of physical contact during the collection process, the sound signal is less affected by mechanical vibration and environmental coupling interference, which is beneficial for obtaining purer fault characteristic information [13]. Wei et al. [14] proposed a fault detection method for high-speed train ventilation devices based on sound vibration signal fusion and time-frequency analysis. By designing anti-interference sensor packaging and embedded systems, combined with STFT feature extraction and SVM classification, the reliability of fault diagnosis is significantly improved. Su et al. [15] proposed a gear-fault diagnosis method based on Bark scale perception feature extraction, which calculates sub-band energy and selects high-energy features through a time-varying specific loudness model, and combines an SVM classifier to achieve high-precision fault recognition. Huang et al. [16] proposed a motor-rotor fault-diagnosis method based on sound signals. By using wavelet-packet denoising and MFCC feature extraction, combined with feature visualization and DeiT classifier, non-invasive high-precision diagnosis of demagnetization and eccentricity faults is achieved. The effectiveness and reliability of sound features in replacing vibration signals for motor state monitoring are verified. However, sound signals are highly susceptible to interference from surrounding noise, resulting in key information being drowned out in irrelevant noise.

Sound and vibration signals reflect mechanical states from different physical levels: vibration signals directly capture the impact and resonance transmitted by internal structures, and are sensitive to early local damage (such as peeling); the sound signal records the airborne sound field radiated by the equipment during operation, which can effectively reflect the overall operating status and aerodynamic noise characteristics [17]. The combination of the two can form a complete observation perspective from internal conduction to external radiation [18]. Most current research directly inputs sound and vibration as multichannel data into neural networks for mechanical equipment fault diagnosis. There is currently limited research on how to fully extract the connections between sound and vibration multimodality.

In recent years, multimodal fault diagnosis has emerged as a rapidly growing research direction [19]. The fusion of multimodal data has been recognized as a promising approach

to enhance diagnostic accuracy and robustness, particularly in complex industrial environments where single-modality signals may be unreliable. Among the various multimodal fusion strategies, attention-based mechanisms have gained considerable attention. For instance, Zhu et al. [20] proposed a semantic-aware and enhanced cross-attention method for multimodal bearing fault diagnosis, which employs a semantic-guided feature mapping mechanism to purify sensor signals within a unified semantic space and achieves precise cross-modal alignment through a time-event synchronization strategy. Despite these advances, most existing multimodal fusion methods still operate within randomly initialized networks trained from scratch on limited fault datasets, lacking the rich sequential priors that pre-trained models can provide.

Beyond multimodal fusion, knowledge-guided learning has emerged as another important paradigm in intelligent fault diagnosis [21]. Unlike purely data-driven approaches that rely solely on labeled samples, knowledge-guided methods incorporate human expertise, physical principles, or prior knowledge into the learning process to constrain and guide model training. Recent reviews have systematically examined the evolutionary context of fault diagnosis technologies, covering traditional machine learning, digital twins, knowledge graphs, and large language models. Among the knowledge-guided techniques, knowledge distillation has proven particularly effective for scenarios with limited or imbalanced training data. Cheng et al. [22] proposed a novel knowledge distillation framework for bearing fault diagnosis under imbalanced samples, which integrates improved activation functions and adaptive feature selection strategies to enhance diagnostic performance when fault samples are scarce. These knowledge-guided approaches highlight the importance of transferring and reusing prior knowledge in fault diagnosis, which aligns with the core motivation of our work.

More recently, large language models have opened new avenues for fault diagnosis and prognostics. Several comprehensive reviews have systematically examined the current applications, technical approaches, and development trends of LLMs in fault diagnosis. These reviews categorize LLM-based methodologies into in-context learning, fine-tuning, retrieval-augmented generation, multimodal learning, and time-series approaches. The integration of LLMs into fault diagnosis frameworks has been explored across various domains, including energy systems, power systems, and prognostics and health management. Zhou et al. [23] proposed a feature extractor-attribute learner framework for identifying electrical anomalies in co-phase power supply systems, demonstrating how advanced diagnostic frameworks can be applied to complex industrial systems beyond traditional rotating machinery. However, most existing LLM-based fault diagnosis approaches still focus on text-based or log-based analysis, and their application to raw-sensor signal processing—particularly for multimodal sound-vibration fusion—remains largely unexplored.

While existing approaches have made valuable progress, they still face notable limitations: (1) most current diagnostic models rely on single-modal data (either vibration or sound signals alone), which offers only a partial representation of equipment health. This restricts the model's ability to capture comprehensive fault characteristics, especially under variable operating conditions. (2) Even in recent multimodal studies, the fusion of sound and vibration signals is often realized through basic feature concatenation or linear weighting. Such shallow fusion fails to explore the intrinsic physical correlations and deep complementary relationships between the sound and vibration modalities, resulting in the loss of critical interactive information during the fusion process. Therefore, to achieve more robust and interpretable machinery health monitoring, it is crucial to develop an advanced fusion framework that can deeply model cross-modal interactions. This work introduces a multimodal large language model (LLM)-based diagnostic approach, aiming to leverage

the powerful pattern recognition and contextual understanding capabilities of LLMs to achieve deeper and more adaptive sound-vibration fusion.

GPT-2 is adopted not for its linguistic understanding but as a general-purpose sequential feature processor. Text and mechanical sensor signals (vibration/audio) share a fundamental nature as long-sequence data with complex temporal dependencies, hierarchical patterns, and local–global couplings. The multihead self-attention and feed-forward architectures in GPT-2 are inherently modality-agnostic. More importantly, GPT-2’s causal (masked) self-attention mechanism inherently respects the temporal causality of physical signals—the current machine state is strictly determined by past observations—rendering it conceptually more suitable for online streaming diagnosis than bidirectional models (e.g., BERT) that rely on future context. Instead of training a deep Transformer from scratch, which is prone to overfitting on small fault datasets, the rich sequential priors acquired from billion-token pre-training are leveraged. Through the proposed prompt-tuning, this pre-trained sequence model is efficiently adapted to the fault diagnosis domain with minimal parameter updates, thereby avoiding catastrophic forgetting and ensuring training stability.

The contributions of this study can be summarized as follows: (1) A novel multimodal prompt-tuning paradigm is proposed for LLM-based fault diagnosis. Unlike existing prompt-tuning methods that are primarily designed for natural language processing tasks, our multimodal prompt-tuning module is the first to generate modality-specific prompts, specifically vibration prompts and sound prompts, alongside a learnable task anchor. This design enables a pre-trained LLM to adapt to heterogeneous sensor signals without modifying its core architecture. (2) A parameter-efficient cross-modal adaptation strategy is introduced. Although selective fine-tuning has been explored in transfer learning, its integration with prompt-based LLM adaptation for sound-vibration fusion is novel. Our freeze–thaw strategy is specifically designed to preserve GPT-2’s universal sequential priors while allowing the prompt space and high-level representations to adapt to the fault diagnosis task. This combination has not been systematically investigated in the mechanical fault diagnosis literature. (3) A multihead cross-attention mechanism is developed to enable deep sound–vibration interaction. Although cross-attention has been applied in various domains, its combination with LLM-based feature refinement represents a new architectural design. In this framework, the cross-attention outputs serve as soft prompts to guide a frozen LLM, which is fundamentally different from using cross-attention merely as a fusion layer within a randomly initialized network.

Crucially, the true contribution lies in the systematic integration of these three modules into a unified framework that leverages the pre-trained sequential priors of LLMs for mechanical signal processing, achieves deep cross-modal interaction through prompt-guided attention, and enables parameter-efficient domain adaptation. The rest of this paper is organized as follows: the details of SVLLM are given in Section 2. The experiment and results analysis are presented in Section 3. Finally, Section 4 presents the conclusions and expectations.

2. Method

2.1. Network Structure

As shown in Figure 1, SVLLM consists of three key components: (1) a multimodal prompt-tuning module, (2) a freeze–thaw reuse module, and (3) a multihead cross-attention module. Firstly, a multimodal prompt-tuning module has been introduced, aimed at domain adaptation fine-tuning of pre-trained large language models based on vibration and sound signals. This module enables the model to learn professional knowledge in mechanical fault diagnosis and achieve adaptive fusion of sound and vibration signals. Secondly, the freeze–thaw reuse module selectively freezes the pre-trained key layers

and unfreezes the layers that need to be adjusted for diagnostic tasks. While retaining general semantic knowledge, it efficiently adapts to downstream fault-diagnosis tasks, improving model training efficiency and stability. Finally, the multihead cross-attention module is designed to achieve deep modal interaction between vibration signals and sound signals. This module utilizes a cross-modal attention mechanism to fully extract and fuse discriminative features from bimodal signals, enhancing the model's ability to perceive and characterize fault features. Based on the SVLLM composed of three key modules, it can effectively extract features and diagnose faults in the vibration signals of mechanical equipment.

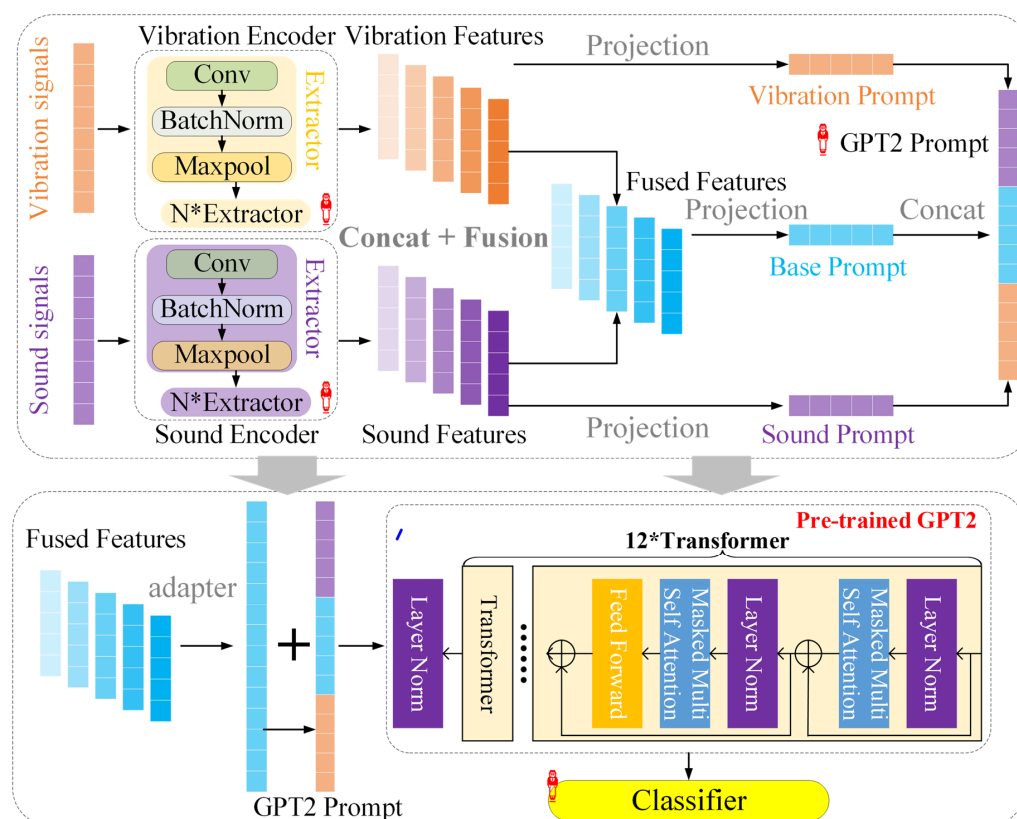


Figure 1. Sound–vibration fusion network.

2.2. Multimodal Prompt-Tuning Module

Prompt-tuning is one of the core technologies in the field of natural-language processing (NLP) in recent years. It designs prompt words to enable pre-trained models (such as GPT and BERT) to adapt more efficiently to downstream tasks, gradually replacing traditional fine-tuning methods and becoming a solution for efficient parameter tuning. Prompt-tuning only optimizes prompt embedding, with a very small number of parameters and high training efficiency. By directly modifying the input embedding layer, it is independent of the model structure and easy to implement [24].

To make the model applicable to multimodal data, a multimodal prompt-tuning module is proposed in SVLLM. By generating prompts for each modality, the model can fully learn the knowledge of each modality. In addition, basic prompts are generated based on the features of multimodal fusion. The final prompt consists of a basic prompt, a vibration prompt, and a sound prompt, allowing the large model to fully acquire domain knowledge of multimodal fault diagnosis. The multimodal prompt-tuning module is shown in Figure 2.

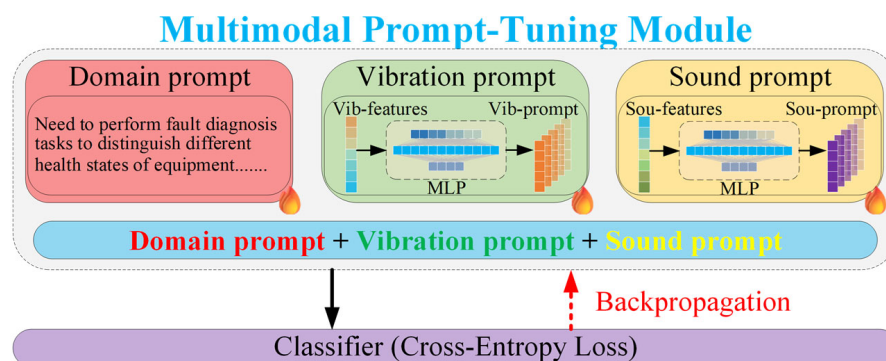


Figure 2. Multimodal prompt-tuning module.

Firstly, the vibration signal and sound signal are subjected to independent encoders for feature extraction. The formula can be expressed as:

$$v_i = f(\sum xv * k_i + b_i) \quad (1)$$

$$s_i = f(\sum xs * \bar{k}_i + \bar{b}_i) \quad (2)$$

$$v_i = \max\text{-pool}(v_i) \quad (3)$$

$$s_i = \max\text{-pool}(s_i) \quad (4)$$

where xv and xs are the vibration and sound inputs of the multiscale convolution, k is the convolution kernel, b is the bias, f is the activation function, $\max\text{-pool}$ is the maximum pooling operation, and V_i and S_i are the vibration features and sound features extracted by the encoder, respectively.

Then, the fusion features are calculated based on the extracted vibration and sound features. The calculation formula for fusion features F_{fused} is as follows:

$$F_{fused} = \text{Dropout}\left(\text{GELU}\left(\text{Norm}\left(W_f \cdot (\text{concat}(V, S)) + b_f\right)\right)\right) \quad (5)$$

where Dropout is a commonly used regularization technique, GELU is an activation function, Norm is a layer normalization operation, concat is a concatenation operation along the feature dimension, W_f and b_f are the weight matrix and bias, respectively.

After that, the prompt of SVLLM can be calculated based on the vibration features, the sound features, and the fused features. The calculation formulas for the prompts are as follows:

$$VPrompt = \text{ReLU}(W_2 \cdot \text{ReLU}(W_1 \cdot V + b_1) + b_2) \quad (6)$$

$$SPrompt = \text{ReLU}(W_2 \cdot \text{ReLU}(\bar{W}_1 \cdot S + \bar{b}_1) + \bar{b}_2) \quad (7)$$

where W_1 and W_2 , b_1 , and b_2 are the weight matrices and biases of two linear layers, respectively. ReLU is an activation function, $VPrompt$ and $SPrompt$ are the generated vibration and sound prompts, respectively.

In addition, the basic prompt $BasePrompt$ is a learnable parameter matrix designed in SVLLM. It can be formulated as:

$$BasePrompt \in R^{1 \times L \times H} \quad (8)$$

Finally, the prompt of SVLLM can be calculated by the $VPrompt$, the $SPrompt$, and the $BasePrompt$. It can be formulated as:

$$Prompt = \text{concat}(VPrompt, SPrompt, BasePrompt) \quad (9)$$

2.3. Freeze-Thaw Reuse Module

In SVLLM, GPT-2 is adopted as the backbone. Unlike conventional fine-tuning, in which all parameters are adjusted, a selective freeze–thaw strategy is employed. This design is motivated not merely by computational efficiency but by a knowledge reuse philosophy for cross-domain sequence modeling. The shallower layers of GPT-2, which are frozen, are hypothesized to capture universal sequential primitives (e.g., local oscillations, envelope fluctuations, and short-term correlations) that are transferable across text and sensor time-series. Fine-tuning is restricted to the deeper layers (the last two transformer blocks) and the output projection layers, enabling the model to recombine these primitives into high-level task-specific fault features. This strategy effectively bridges the gap between the source domain (text) and the target domain (sound-vibration signals) without modification of the fundamental sequence extraction pipeline. Meanwhile, by freezing the vast majority of parameters, the solution space is strictly constrained, which acts as a powerful regularizer against overfitting. It is a critical advantage given the limited scale of annotated fault data relative to the billions of tokens in NLP. The freeze–thaw reuse module is shown in Figure 3.

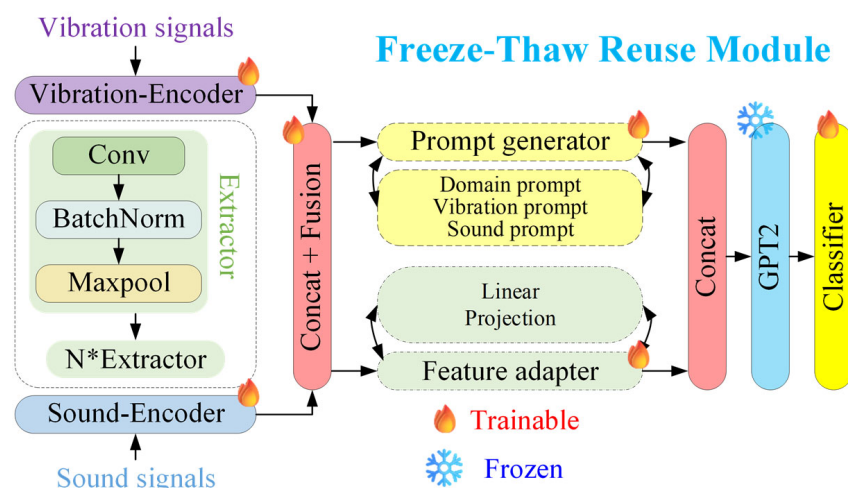


Figure 3. Freeze-thaw reuse module.

In the multimodal fault diagnosis task, we adopt a selective fine-tuning strategy, training only the last two layers of GPT-2, prompt parameters, and classifier, while freezing the first 10 layers. This approach is mainly based on the consideration of efficient parameter fine-tuning: firstly, freezing most layers can effectively avoid catastrophic forgetting and preserve the general semantic representation learned by GPT-2 in large-scale texts; secondly, significantly reducing the number of trainable parameters on small fault datasets can significantly lower the risk of overfitting. In addition, this strategy also improves computational efficiency, shortens training time, and reduces video memory usage. The prompt mechanism, as a task-adaptive “soft instruction”, can guide the entire model to adapt to fault diagnosis by adjusting only a small number of parameters. At the same time, retaining the universal feature extraction capability of the underlying layer and only fine-tuning the high-level and output-normalization layers is also beneficial for gradient stability and training convergence, thereby achieving efficient and stable few-sample cross-modal learning while ensuring performance [25].

2.4. Multihead Cross-Attention Module

Multihead attention represents an evolution of conventional attention mechanisms, designed to strengthen a model’s representational capacity and learning capability. It achieves this by dividing the input features into several subspaces, or “heads,” each of which per-

forms attention independently [26]. This mechanism has found extensive application across a range of deep learning tasks, such as machine translation, text summarization, speech recognition, and image captioning. Fundamentally, multihead attention operates by executing multiple parallel instances of scaled dot-product attention, subsequently concatenating their outputs, and projecting the combined result into the final representation [27].

In this study, the multihead cross-attention module is constructed of the multihead attention mechanism, which is used to extract cross-modal information between sound signals and vibration signals. The multihead cross-attention module formalizes the modal alignment problem as a learnable retrieval aggregation process. It provides a unified tensor interface using Q , K , and V , which enables the model to align different modalities in the semantic space through end-to-end training. The structure of the multihead cross-attention module is shown in Figure 4. The calculation procedure can be defined as follows:

$$Q = F_v, K = F_s, V = F_{sv} \quad (10)$$

$$Attention(F_v, F_s) = \frac{F_v F_s^T}{\sqrt{d_k}} \quad (11)$$

$$Scores = Softmax(Attention) \quad (12)$$

$$F'_{sv} = Scores \times F_{sv} \quad (13)$$

where F_v and F_s are the vibration features and sound features, respectively. T is the transpose operator, d_k is a scaling factor that has the same dimension with F_s , $Softmax$ is the activation function, which ensures that all output values are within the (0, 1) interval and the total sum is 1. F_{sv} are the sound–vibration features, and F'_{sv} are the attention-weighted features.

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (14)$$

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W_O \quad (15)$$

where QW_i^Q , KW_i^K , and VW_i^V are the linear transformation matrices corresponding to the i th head (obtained by dividing W_Q , W_K , W_V), and W_O is the final output linear transformation matrix.

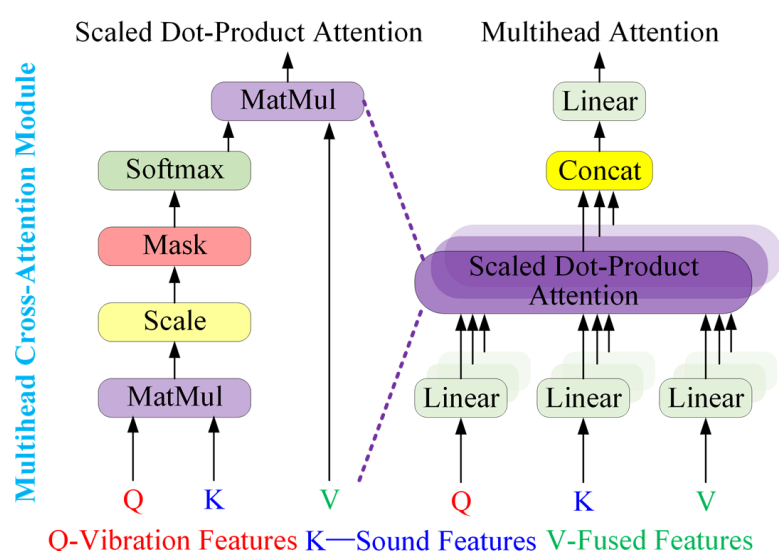


Figure 4. Multihead cross-attention module.

In this module, vibration features, sound features, and fused features are used as queries, keys, and values in the cross-attention mechanism, respectively. The cross-modal

attention module learns a data-driven alignment between the two modalities and selectively extracts modality-shared and complementary features. This design makes the feature extraction of the model more robust, which outperforms the simple summation of multimodal features.

2.5. Machinery Fault Diagnosis Flow of SVLLM

The machinery fault diagnosis flow of SVLLM is composed of four steps (Algorithm 1). Firstly, sound signals and vibration signals are collected by the sensors. Secondly, the training dataset will be input into SVLLM to train the model for fault diagnosis. Thirdly, the sound signals and vibration signals collected in real time are fed into the well-trained SVLLM to perform fault diagnosis. Finally, the machinery fault diagnosis results are output online.

The SVLLM-based machinery fault diagnosis method can be represented as follows.

Offline training:

- (1) The training data is consisting of sound signals and vibration signals.
- (2) A multimodal prompt-tuning module is proposed to integrate domain knowledge into LLM.
- (3) A freeze-thaw reuse cross-modal module is proposed for finetune the LLM.
- (4) The multihead cross-attention learning module is proposed to fuse the sound signals and vibration signals.
- (5) The trained SVLLM is obtained for online testing.

Online health monitoring:

- (1) Online signals are collected from the machines.
- (2) Feed data into SVLLM that has been trained offline.
- (3) Output the fault diagnosis results online.

Algorithm 1 Training Process of SVLLM.

Input: Training samples X_s , X_v .

Initializing the parameters of SVLLM.

If training steps < N:

Forward-propagation:

- (1) Generating the vibration feature, the sound features, and the fused features.
- (2) Generating the prompt of SVLLM based on the vibration prompt, the sound prompt, and the basic prompt.
- (3) Calculating the cross-entropy loss.

Backpropagation:

- (1) Calculating the parameters of the fully connected layer and then update.
- (2) Calculating the parameters of the prompt consist of the vibration prompt, the sound prompt, and the basic prompt, and then update.
- (3) Calculating the parameters of the vibration encoder and the sound encoder, and then update.

Update SVLLM by minimizing the cross-entropy loss.

End

Output: Trained SVLLM.

3. Experiment and Result Analysis

In this section, two datasets are constructed to validate the effectiveness of SVLLM for machinery fault diagnosis. The first dataset is constructed by setting up a motor test bench to collect sound and vibration data, forming a motor dataset. The second dataset is built using benchmark data publicly available from the University of Ottawa.

3.1. Case 1: Motor Fault Diagnosis

Figure 5 illustrates the motor test-rig configuration: a stepper motor is coupled to a magnetic-powder brake through a torque transducer, while a CT1020LC accelerometer and a CT1213 microphone capture vibration and sound signatures, respectively. A CT1020LC piezoelectric accelerometer with a nominal sensitivity of 100 mV/g captures the vibration signatures, while a CT1213 condenser microphone with a nominal sensitivity of 50 mV/Pa records the acoustic signatures. A four-channel NI-9234 cDAQ module performs simultaneous dynamic acquisition of both modalities (i.e., the accelerometer is affixed directly to the motor housing, and the microphone is positioned on an adjacent bracket). Comprehensive dataset specifications for the recorded vibration and sound signals are summarized in Table 1. A total of seven health states is considered in this experiment. The vibration and sound signals of each condition are collected at a sampling rate of 10,000 Hz, and then the samples are divided into 2048 points. A total of 1000 samples are prepared for bearings under each health state. The long-duration raw signals were segmented into non-overlapping windows of 2048 data points each by applying a fixed-length sliding window with a stride equal to the window size. Specifically, the signal was divided sequentially into consecutive 2048-point segments without any overlap between adjacent windows. The vibration and sound signals of motors in different health states are shown in Figure 6. It is worth noting that to ensure the fairness of the comparative experiments, all experiments in this paper utilized a five-fold cross-validation. In order to preserve the original mechanical information and avoid introducing artificial distortion, no preprocessing operation was performed on the original sensor signals in this study.

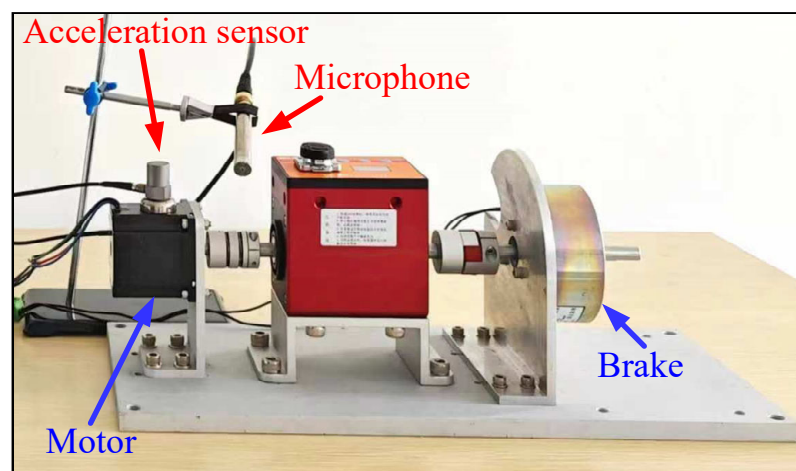


Figure 5. Motor test rig.

Table 1. Detailed information on the vibration and sound signals.

Labels	Condition	Health Types	Physical Description	Structure (Sample Size × Data Length)
C1	100 r/min	Healthy	-	1000 × 2048
C2		Bearing fault-level1	A single-point defect was introduced on the outer race of a bearing using Electrical Discharge Machining (0.2 mm diameter).	1000 × 2048
C3		Bearing fault-level2	A single-point defect was introduced on the outer race of a bearing using Electrical Discharge Machining (0.4 mm diameter).	1000 × 2048
C4		Broken bar fault-level1	One rotor bar was partially fractured by drilling a hole through the bar cross-section to simulate a crack-initiation-level fault (0.2 mm diameter).	1000 × 2048
C5		Broken bar fault-level2	One rotor bar was partially fractured by drilling a 0.2 mm diameter hole through the bar cross-section to simulate a crack-initiation-level fault (0.4 mm diameter).	1000 × 2048

Table 1. Cont.

Labels	Condition	Health Types	Physical Description	Structure (Sample Size $\times$ Data Length)
C6	100 r/min	Misalignment-level1	A controlled parallel misalignment was induced between the motor shaft and the load shaft by inserting precision-machined shim spacers at the flexible coupling (0.1 mm offset).	1000 $\times$ 2048
C7		Misalignment-level2	A controlled parallel misalignment was induced between the motor shaft and the load shaft by inserting precision-machined shim spacers at the flexible coupling (0 mm offset).	1000 $\times$ 2048

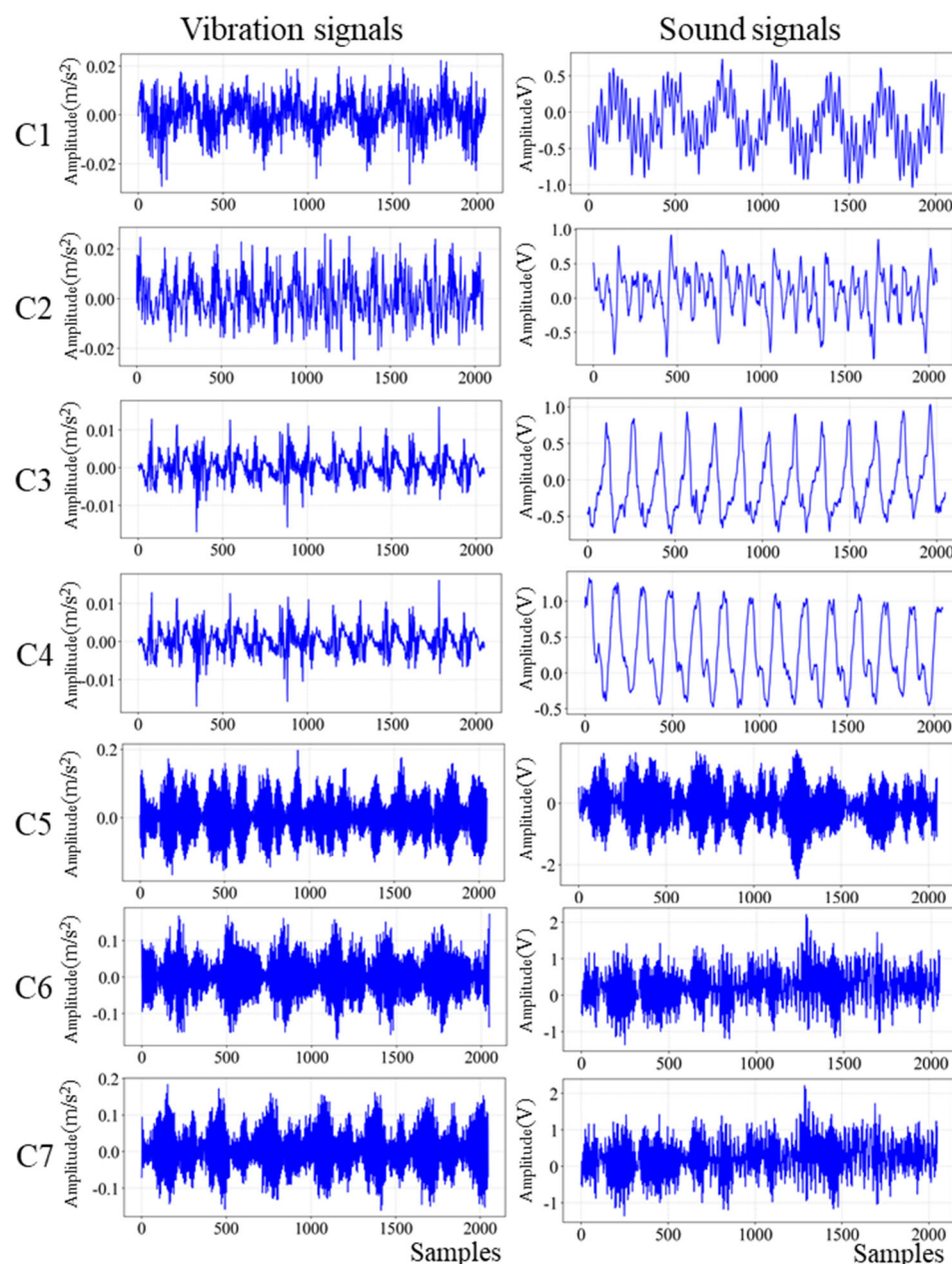


Figure 6. The vibration and sound signals of motors in different health states with the motor dataset.

(1) *Parameter Setup*: The parameter settings of SVLLM are shown in Table 2, where K is the width of the convolution kernel, S is the sliding strides, P is the padding sizes, and $[16@1024 \times 1]$ indicates that the features have 16 channels, with the size of each feature map being 1024×1 . All experiments are tested on a computer with an NVIDIA GeForce RTX 4090 GPU and Intel Xeon Gold 5218R CPU @2.10 GHz. SVLLM proposed under Torch-2.2.0+cu121 based on Python 3.8.

Table 2. Parameter setup of SVLLM.

Structure		Parameters	Outputs Sizes
	Input	Vibration signals	[1@2048 × 1]
		Sound signals	[1@2048 × 1]
Vibration-encoder	Conv	K = 3; S = 2; P = 3	[64@1024 × 1]
	BatchNorm	-	[64@1024 × 1]
	MaxPool	K = 3; S = 2; P = 1	[64@512 × 1]
	Conv	K = 3; S = 2; P = 3	[64@1024 × 1]
	BatchNorm	-	[64@1024 × 1]
	MaxPool	K = 3; S = 2; P = 1	[64@512 × 1]
Sound-encoder	Concat Feature	-	[512]
	Fusion Feature	Linear (512-256)	[256]
	Feature Adapter	Linear (256-768)	[768]
Prompt	Base Prompt	Learnable parameters	[10 × 256]
	Vibration Prompt		[10 × 256]
	Sound Prompt		[10 × 256]
	Fused Prompt		[10 × 768]
GPT2	GPT2-Input	GPT parameters	[11(10 + 1)@768]
	GPT2-Output		[768]
Classifier		Linear (768-256)	[256]
		Linear (256-7)	[7]
Learning rate = 0.001			

(2) *Result Analysis*: In this section, state-of-the-art DNNs are considered in comparison to evaluate the machinery fault diagnosis performance of SVLLM: ResNet [28] effectively alleviates the gradient vanishing problem in deep neural networks by introducing skip connections, making it possible to train extremely deep models. DenseNet (Dense Connected Network) [29] achieves feature reuse by directly connecting each layer to all previous layers, significantly improving the parameter efficiency and performance of the model. SCNet (Self-Calibrated Convolutional Network) [30] enhances the receptive field of convolutional layers through an internal self-calibration mechanism, enabling the network to adaptively capture richer contextual information within the same layer. 1D-RCAE (One-Dimensional Residual Convolutional Autoencoder) [31] combines one-dimensional convolution, residual learning, and autoencoder structure, focusing on robust feature extraction and dimensionality reduction from one-dimensional signals such as vibration. ECANet (Efficient Channel Attention Network) [32] introduces a streamlined channel attention module designed to efficiently model interdependencies between channels. By avoiding dimensionality reduction, it preserves direct cross-channel interaction, thereby improving the network's capacity to emphasize salient features. FCTransformer [33] fuses multisensor data by combining Fourier convolution with Transformer for frequency-aware global dependency modeling. Multimodal bearing fault diagnosis (MMBFD) [20] uses semantic-aware cross-attention to align and enhance cross-modal features for accurate fault classification.

Table 3 provides a detailed explanation of the model parameters. It shows that simply increasing parameters does not guarantee better performance (e.g., some model with fewer parameters sometimes outperforms larger models on certain modalities). The proposed SVLLM adopts a GPT-2 backbone network with 124 M parameters and a total parameter count of 132.35 M. However, the trainable parameters are only 22.09 M. The trainable parameter count is kept at the same level as several other models. Therefore, the effectiveness of SVLLM should be attributed to its sophisticated fusion architecture and domain-specific adaptation, rather than to brute-force parameter expansion. Moreover, in order to conduct a completely fair evaluation, a state-of-the-art model that also uses GPT-2 as the backbone network is also used for comparison in this section.

Table 3. The detailed information of model parameters.

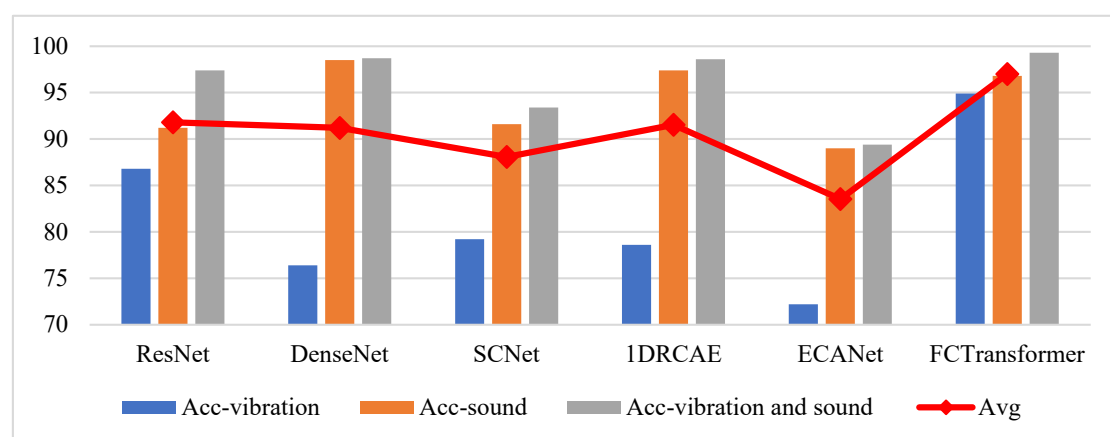
Models	Total Parameter (M)	Trainable Parameter (M)
SVLLM	132.35	22.09
ResNet	10.50	10.50
DenseNet	22.02	22.02
SCNet	10.53	10.53
1DRCAE	41.21	41.21
ECANet	5.34	5.34
FCTransformer	2.47	2.47
MMBFD	150.00	26.00

Moreover, the number of iterations is set to 1000 for all models, and the comparison results with these state-of-the-art methods are presented in Table 4. It can be easily seen that the accuracy of SVLLM is better than that of other models based on the multimodal prompt-tuning module, the freeze–thaw reuse module, and the multihead cross-attention module.

Table 4. The fault diagnosis results with the motor dataset.

Models	Acc-Vibration (%)	Acc-Sound (%)	Acc-Vibration and Sound (%)	Avg (%)
SVLLM	-	-	99.40 ± 0.24	99.40
ResNet	86.80 ± 1.07	91.20 ± 0.90	97.40 ± 0.50	91.80
DenseNet	76.40 ± 1.34	98.50 ± 0.38	98.70 ± 0.36	91.20
SCNet	79.20 ± 1.28	91.60 ± 0.88	93.40 ± 0.79	88.10
1DRCAE	78.60 ± 1.30	97.40 ± 0.50	98.60 ± 0.37	91.50
ECANet	72.20 ± 1.42	89.00 ± 0.99	89.40 ± 0.97	83.50
FCTransformer	94.90 ± 2.69	96.80 ± 0.91	99.30 ± 0.37	97.00
MMBFD	-	-	99.20 ± 0.62	99.20

To further investigate the impact of different modal data (i.e., sound signals and vibration signals) on the fault diagnosis performance of various models, we conducted comparative experiments using the most advanced models under a single modality. The results are shown in Figure 7. According to the experimental results, it can be found that DenseNet and ECANet have poor performance in feature extraction and fault diagnosis of vibration signals. SCNet has poor feature extraction and fault diagnosis for sound signals. All models perform best when using sound and vibration signals. However, it can be seen that different models may be limited by a certain mode and cannot efficiently use multimodal feature information. SVLLM has demonstrated strong capabilities in multimodal feature fusion and fault diagnosis.

**Figure 7.** Results (%) of the state-of-the-art DNNs on different modalities of data with the motor dataset.

Furthermore, the experiment also plotted the confusion matrix between SVLLM and several well-performing models to study the recognition performance of each model for different health types. The results of the confusion matrix are shown in Figure 8. It can be seen that ResNet has a low recognition accuracy for fault type seven (imbalance level two) and is prone to misidentifying it as fault type six (imbalance level one). DenseNet can easily confuse fault types three (bearing fault level two) and four (broken bar fault level one). SCNet can easily confuse fault type five (broken bar fault level two), fault type six, and fault type seven. However, SVLLM has a high accuracy in identifying each type of fault.

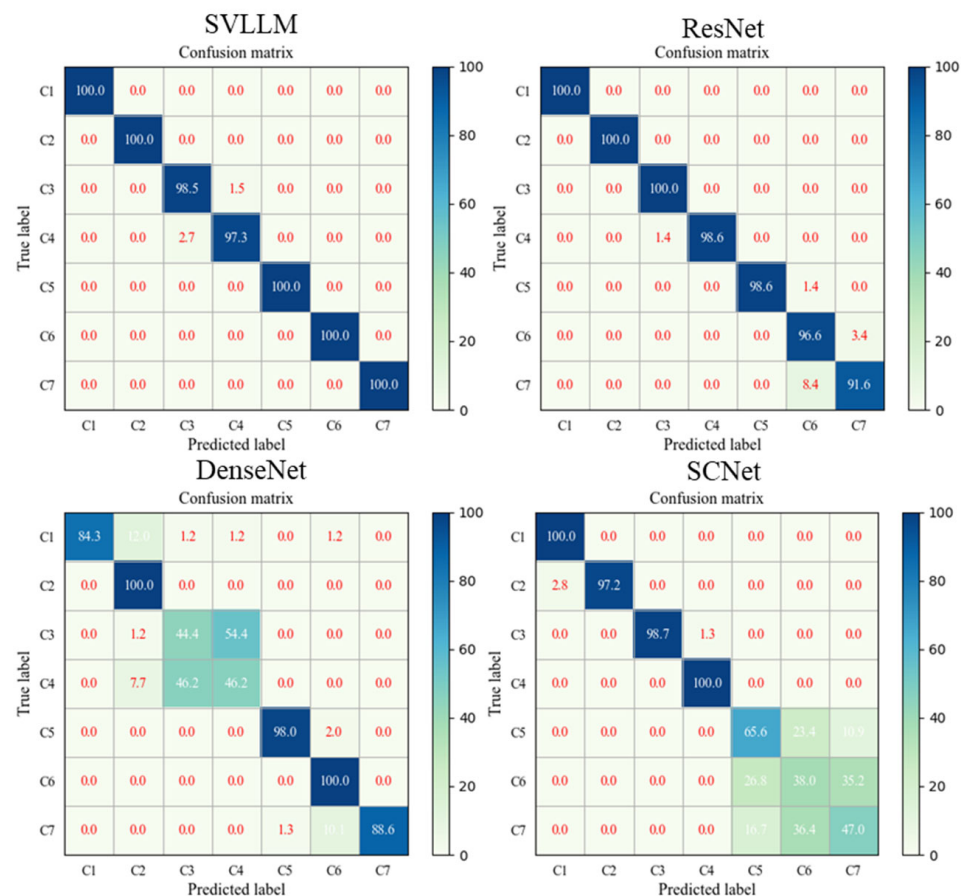


Figure 8. Confusion matrix (%) of the state-of-the-art DNNs with the motor dataset.

(3) *Visualization Analysis*: To provide an intuitive understanding of the hierarchical representations, this section employs t-distributed stochastic neighbor embedding (t-SNE) to map the high-dimensional features extracted by SVLLM onto a two-dimensional plane. Within this reduced space, the horizontal and vertical axes correspond to the first and second embedded dimensions, respectively, allowing for a visual inspection of the feature distributions learned at each layer. The result is shown in Figure 9, which presents features from the input data (a), the sound encoder layer (b), the vibration encoder layer (c), the prompt layer (d), the GPT2-output layer (e), and the fully connected (FC) layer (f). As shown in Figure 9a, features of the input data with different fault types are distributed disorderly. These features are gradually clustered after the sound encoder and the vibration encoder layers, as presented in Figure 9b,c. Figure 9d presents the results of the prompt layer, where the different health states have been divided, although there is a little confusion between the categories. As shown in Figure 9e, the sample categories are further divided after GPT2. In particular, the features of the FC layer are clustered well, and there are clear

distinctions between the different fault types, as shown in Figure 9f. It demonstrates the powerful feature learning and machinery fault diagnosis of SVLLF.

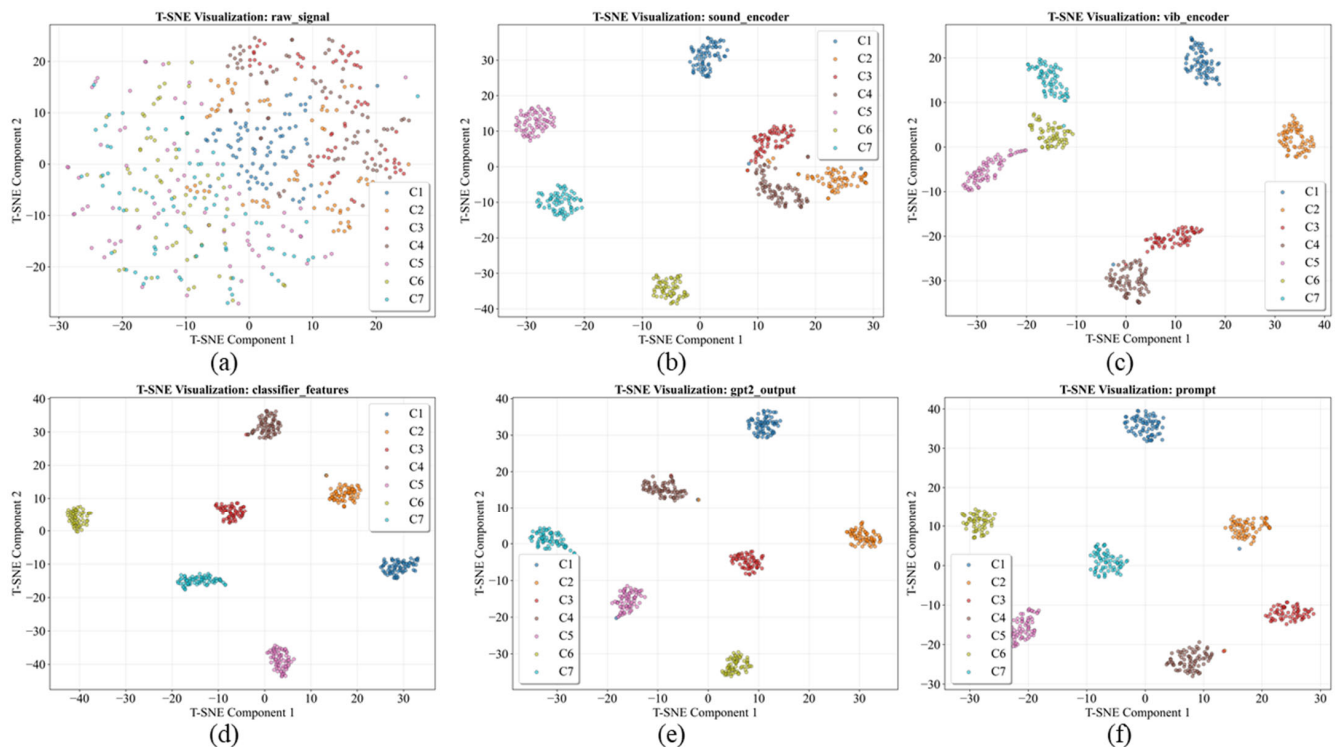


Figure 9. T-SNE-based visualization of the output features in each layer with the motor dataset. (a) Input data, (b) sound encoder layer, (c) vibration encoder layer, (d) prompt layer, (e) GPT2-output layer, (f) fully connected (FC) layer.

In order to verify the effectiveness of each key module, ablation experiments are conducted in this section. Removed Multimodal Prompt Unity Module (no multimodal prompts), Freeze–Thaw Reuse Module (freeze all large model parameters), and Multi head Cross Attention Module (directly concatenate vibration and sound features), respectively. The result is shown in Figure 10. It can be seen that after removing cross attention, the model cannot capture complementary information across modalities, resulting in a decrease in accuracy. Removing prompts for optimization leads to a lack of knowledge in the field of fault diagnosis in the model, resulting in a decrease in accuracy. Removing the freeze–thaw strategy may result in parameter shifts in the model during later iterations, leading to performance degradation.

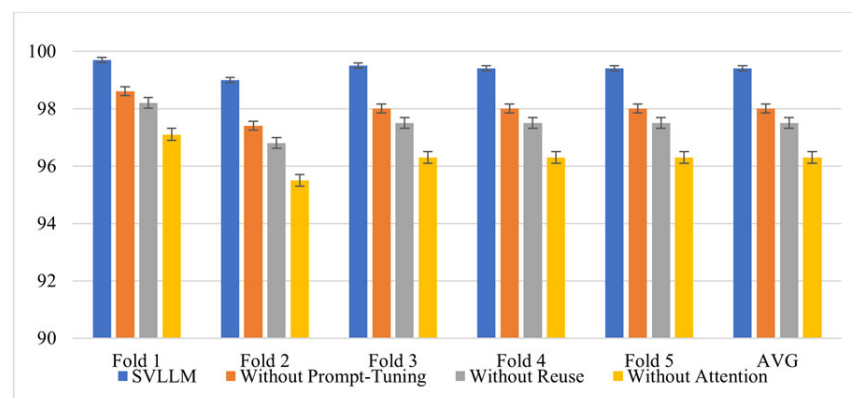


Figure 10. The ablation experiment results of Case 1.

3.2. Case 2: Ottawa Bearing Dataset Fault Diagnosis

The rolling bearing dataset used in this study is sourced from the vibration and sound fault testing experimental platform established by the University of Ottawa, which operates under constant load and speed conditions [34]. As shown in Figure 11, the experimental setup primarily consists of a single-phase motor, rolling bearing, drive screw, acceleration sensor, and sound sensor, capable of simultaneously collecting vibration and audio signals during operation. During the data acquisition phase, signals under each healthy condition are continuously collected for 10 s at a sampling frequency of 42,000 Hz, corresponding to a shaft speed of 1750 rpm. The raw signals collected are subsequently divided into multiple continuous segments, with each segment having a fixed length of 2048 sampling points.

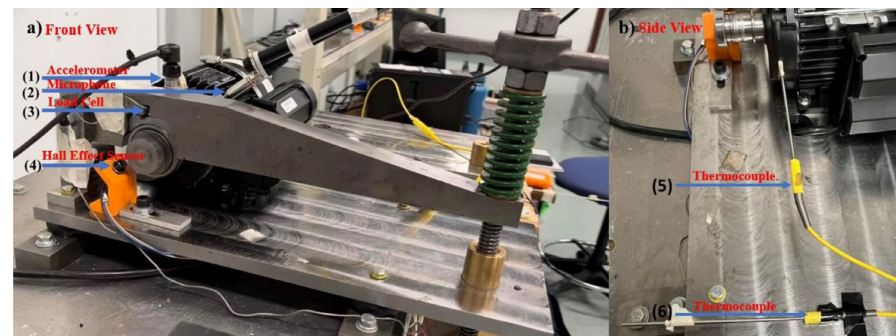


Figure 11. The Ottawa bearing test rig.

The detailed information of the Ottawa dataset for the recorded vibration and sound signals are summarized in Table 5. The experiment encompassed five distinct bearing health states, namely: normal state, inner race failure, outer race failure, ball failure, and cage failure. A five-fold cross-validation is also used in this section. For the Ottawa dataset, each health condition is represented by 1000 samples, with each sample consisting of 2048 consecutive data points. To construct these samples, the continuously recorded raw signals were partitioned into fixed-length frames of 2048 points using a non-overlapping sliding window protocol, i.e., the stride was set equal to the window size, ensuring that adjacent frames contained no shared data points. Figure 12 illustrates the typical vibration signal and sound signal waveforms corresponding to the five health states in the dataset, visually reflecting the characteristic differences in the time domain among different failure types.

Table 5. Detailed information on the Ottawa dataset.

Labels	Condition	Health Types	Structure
C1	1750 r/min	Healthy	1000 × 2048
C2		Inner fault	1000 × 2048
C3		Outer fault	1000 × 2048
C4		Ball fault	1000 × 2048
C5		Cage fault	1000 × 2048

As with the case study, this section conducted comparative experiments between SVLLM and state-of-the-art models on the Ottawa dataset. The experimental results are shown in Table 6. The experimental results show that the accuracy of SVLLM in fault diagnosis on the Ottawa dataset can reach 100% when using vibration and sound signals. This surpasses all other state-of-the-art models. This proves that under the influence of the multimodal prompt-tuning module, the freeze-thaw reuse module, and the multihead cross-attention module, SVLLM has strong capabilities in multimodal feature extraction and feature fusion to achieve high-quality fault diagnosis.

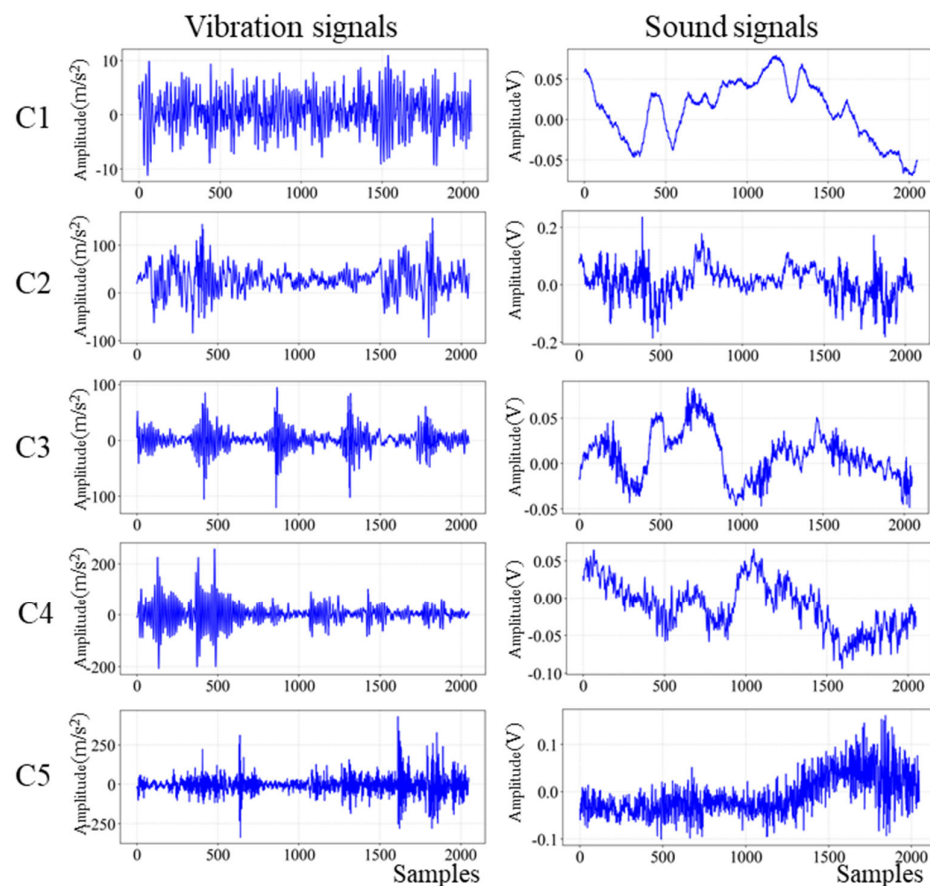


Figure 12. The vibration and sound signals of motors in different health states with the Ottawa dataset.

Table 6. The fault diagnosis results with the Ottawa dataset.

Models	Acc-Vibration (%)	Acc-Sound (%)	Acc-Vibration and Sound (%)	Avg (%)
SVLLM	-	-	100.00 ± 0.00	100.00
ResNet	96.50 ± 0.58	98.20 ± 0.42	96.30 ± 0.60	97.00
DenseNet	85.70 ± 1.11	78.70 ± 1.29	95.70 ± 0.64	86.70
SCNet	88.70 ± 1.00	75.20 ± 1.37	92.80 ± 0.82	85.60
1DRCAE	92.90 ± 0.81	89.20 ± 0.98	96.80 ± 0.56	93.00
ECANet	82.80 ± 1.19	78.20 ± 1.31	90.40 ± 0.93	83.80
FCTransformer	96.40 ± 0.67	96.80 ± 0.70	90.20 ± 0.23	97.50
MMBFD	-	-	99.40 ± 0.25	99.40

To further investigate the impact of different modal data on the fault diagnosis performance of various models, comparative experiments are performed using the most advanced models under a single modality. The results are shown in Figure 13. According to the experimental results, it can be found that DenseNet and ECANet have poor performance in feature extraction and fault diagnosis of vibration signals. SCNet has poor feature extraction and fault diagnosis for sound signals. The vast majority of scenarios perform best when using sound and vibration signals. However, the sound accuracy of ResNet is 98.2%, while the sound vibration accuracy is 96.3%. This indicates that traditional models may sometimes fail to fully utilize multimodal data. However, SVLLM maintained the best fault diagnosis performance in all cases, demonstrating its powerful multimodal feature extraction and fusion capabilities.

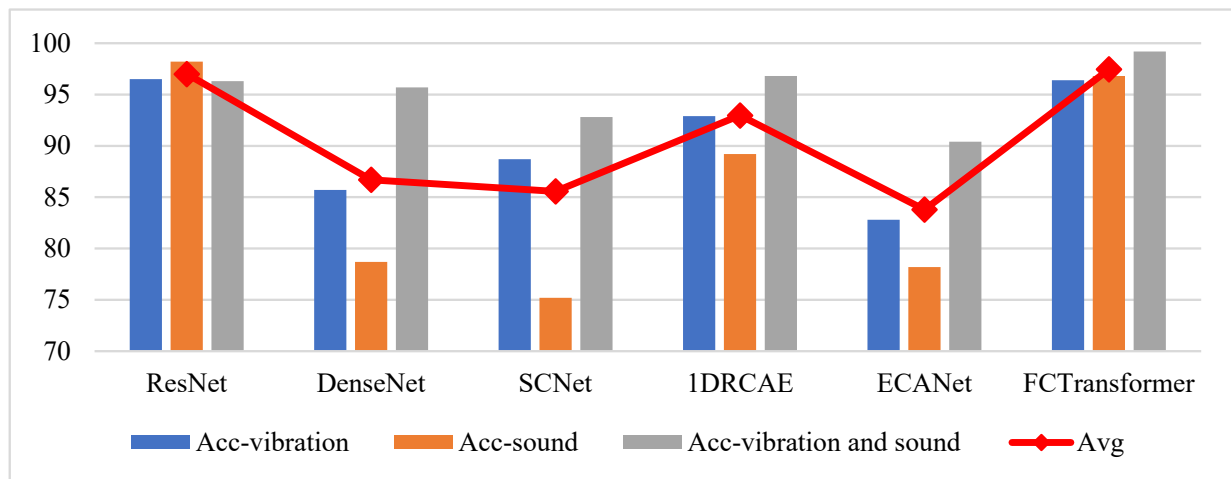


Figure 13. Results (%) of the state-of-the-art DNNs on different modalities of data with the Ottawa dataset.

Similarly, in Case 2, confusion matrices of several models were also drawn for further analysis. The results of the confusion matrix are shown in Figure 14. It can be seen that ResNet has a low recognition accuracy for fault type five (cage fault) and is prone to misidentifying it as fault type two (inner fault). DenseNet can easily confuse fault type one (health) and fault type two. SCNet can easily confuse fault type one, fault type three (outer fault), and fault type five. However, SVLLM has a high accuracy in identifying each type of fault.

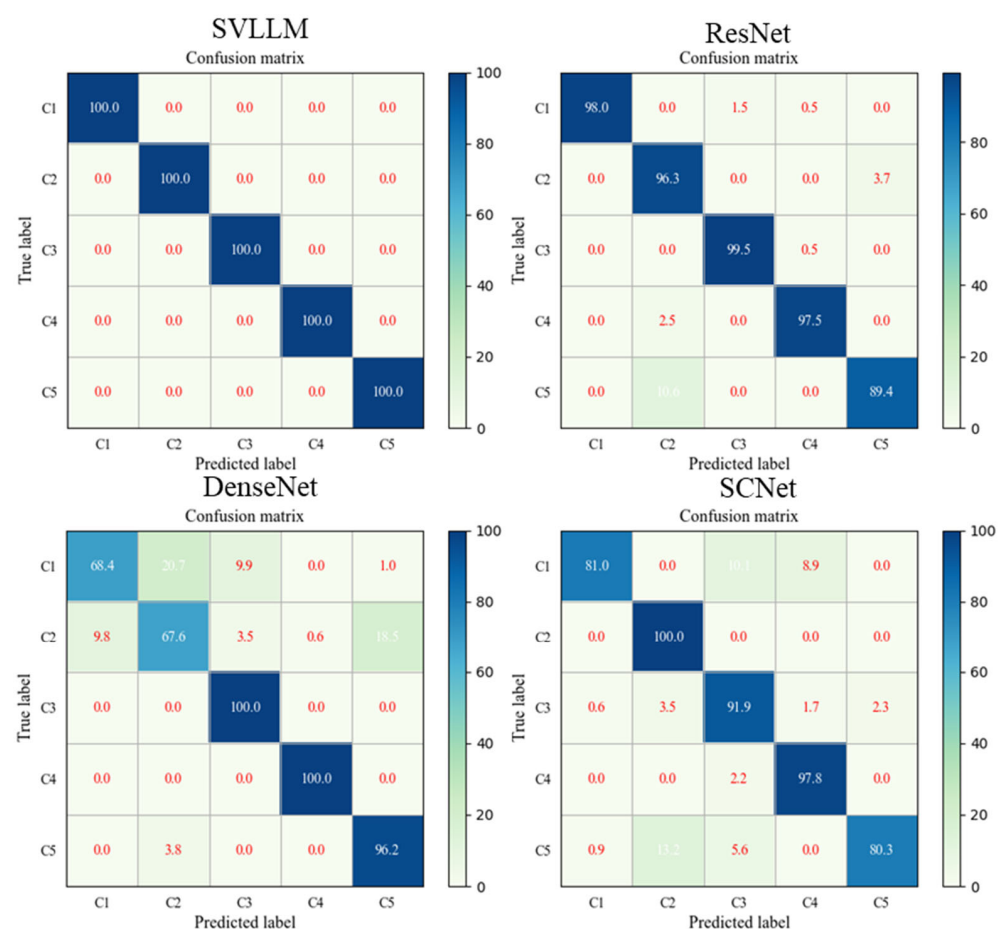


Figure 14. Confusion matrix (%) of the state-of-the-art DNNs with the Ottawa dataset.

In this section, ablation experiments are still used to verify the effectiveness of each key module. As shown in Figure 15, the results of the ablation study confirm the indispensable roles of all three proposed modules. Notably, the cross-attention module contributes the most significant performance gain, validating its effectiveness in aligning heterogeneous sensor features. The freeze–thaw strategy also exhibits a critical impact, preventing catastrophic forgetting during incremental training.

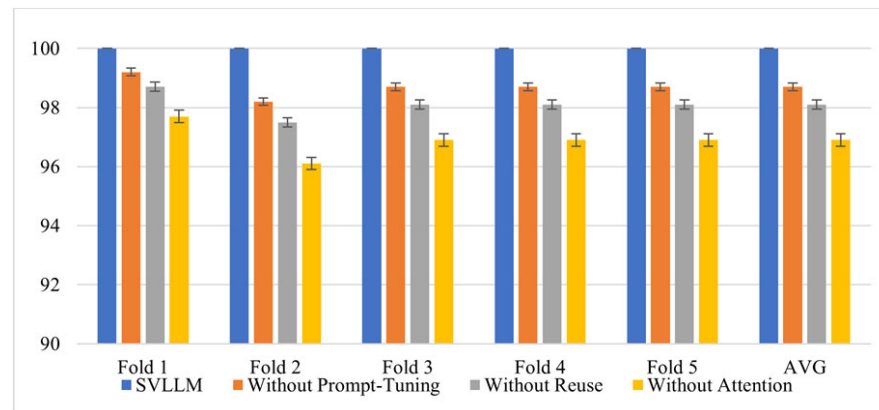


Figure 15. The ablation experiment results of Case 2.

4. Conclusions

This study proposes a multimodal large-scale language model (SVLLM) based on sound and vibration signals to address the issues of insufficient utilization of multimodal information and shallow feature fusion in mechanical fault diagnosis. Firstly, by introducing a multimodal prompt-tuning module, the model can effectively learn and adaptively integrate fault-diagnosis domain knowledge from vibration and sound signals. Secondly, a freeze–thaw reuse module was designed to efficiently adapt to diagnostic tasks while maintaining pre-trained knowledge. The third hypothesis proposes a multihead cross-attention module that fully extracts discriminative features of bimodal signals through extended interaction between modalities. Experiments on motor data sets and Ottawa data sets show that SVLLM performs well in feature extraction and fault diagnosis tasks of sound and vibration signals, which verifies its effectiveness and progressiveness in multimode mechanical fault diagnosis. Although this study has made positive progress in multimodal fault diagnosis methods, there are still several directions worth further exploration: firstly, it is possible to consider incorporating more types of sensing signals (such as temperature, current, etc.) into the model to construct a more comprehensive multimodal perception system. In addition, exploring a fusion diagnostic framework that combines physical mechanisms with data-driven approaches to enhance the interpretability and generalization ability of models will also be an important path for future research. It is also very interesting to study how to use different or multiple LLMs to guide machinery fault diagnosis.

Author Contributions: Conceptualization, Y.T.; Methodology, Y.T.; Software, J.W. (Jiangling Wang); Validation, J.W. (Juntao Wang) and C.Z.; Formal analysis, Y.T.; Investigation, Y.Z.; Resources, Y.Z. and C.Z.; Writing—Original Draft, Y.T.; Writing—Review and Editing, J.W. (Juntao Wang) and C.Z.; Visualization, J.W. (Jiangling Wang); Project administration, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 62172242), the Ningbo Science and Technology Fund (Grant Nos. 2023Z213, 2023Z228, 2024Z202, 2024Z296, 2025Z036), the Ningbo Key Laboratory of Intelligent Home Appliances, and Provincial Philosophy and Social Science Planning Project under Grant 22NDJC127YB. This work was supported

by the Ningbo ‘Innovation Yongjiang 2035’ Key R&D Programme (Grant No.: 2024Z202), 2025 Ningbo Yongjiang Talent Programme, and the Ningbo Key Laboratory of Household Appliances.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: Author Ying Zheng was employed by the Ningbo Fotile Kitchen Ware Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The Ningbo Fotile Kitchen Ware Co., Ltd. had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Zhang, R.; Chu, F.; Qiu, Y.; He, D.; Wang, F. Trend-Aware Patch Feature Fusion Transformer for Fault Diagnosis of Vibration Signals. *IEEE Sens. J.* **2025**, *25*, 24663–24674. [\[CrossRef\]](#)
2. Yao, K.C. Electric Motor Vibration Signal Classification Using Wigner–Ville Distribution for Fault Diagnosis. *Sensors* **2025**, *25*, 1196. [\[CrossRef\]](#)
3. Feng, Z.; Lin, R.; Zhang, Y.; Lin, Z. Vibration signal analytical model and signature analysis for harmonic drive flexible bearing fault diagnosis. *Mech. Syst. Signal Process.* **2025**, *237*, 112877. [\[CrossRef\]](#)
4. Habbouche, H.; Amirat, Y.; Benkedjouh, T.; Benbouzid, M. Digital twin-based gearbox fault diagnosis using variational mode decomposition and dynamic vibration modeling. *Measurement* **2025**, *246*, 116669. [\[CrossRef\]](#)
5. Xiao, C.; Chang, J.; Yu, J. Multiple Wave Shape Mode Decomposition for Bearing Fault Diagnosis. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 6507811. [\[CrossRef\]](#)
6. Han, G.; Shen, J.; Wang, Z.; Zhu, Y.; Xie, Y. A Novel Dual-Domain Adversarial Method for Vibration Signal Denoising in Bearing Fault Diagnosis. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 3522612. [\[CrossRef\]](#)
7. Hu, H.X.; Cai, Y.; Meng, Q.; Cui, H.; Hu, Q.; Zhang, Y. MFVAE: A Multiscale Fuzzy Variational Autoencoder for Big Data-Based Fault Diagnosis in Gearbox. *IEEE Trans. Fuzzy Syst.* **2025**, *33*, 180–191. [\[CrossRef\]](#)
8. Huang, D.; Zhang, W.A.; Gui, W. A General Intelligent Fault Diagnosis Framework Based on Incomplete Vibration Signals With Missing Values. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 3527712. [\[CrossRef\]](#)
9. Xiao, Y.; Feng, K.; Wang, H. Integrating Casing Vibration and Reconstructed Pressure Signal for Gas Turbine Blade Fault Diagnosis. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 3520011. [\[CrossRef\]](#)
10. You, K.; Liu, C.; Lin, Y.; Qiu, G.; Gu, Y. DTMPI-DIVR: A digital twins for multi-margin physical information via dynamic interaction of virtual and real sound-vibration signals for bearing fault diagnosis without real fault samples. *Expert Syst. Appl.* **2025**, *292*, 128592. [\[CrossRef\]](#)
11. Tian, M.; An, W.; Sun, X.; Wang, L.; Chen, C. A non-contact fault diagnosis method based on multi information fusion networks for rolling bearings. *Appl. Acoust.* **2025**, *237*, 110776. [\[CrossRef\]](#)
12. Xie, J. Fault Diagnosis of Rolling Bearings Based on Acoustic Signals in Strong Noise Environments. *Appl. Sci.* **2025**, *15*, 1389. [\[CrossRef\]](#)
13. Jia, C.; Hu, X.; Hou, D.; Chen, L.; Chen, K.; Guo, B.; Yang, L. Real-time fault diagnosis of vehicle control motors for variable operating conditions based on acoustic vibration fusion. *Meas. Sci. Technol.* **2025**, *36*, 085903. [\[CrossRef\]](#)
14. Wei, R.; Wei, P.; Pan, C.; An, Y.; Zhu, H.; Wang, M. Multisensor-Based Fault Detection Method for Air-Exchange Device in EMU Train. *IEEE Sens. Lett.* **2025**, *9*, 7002004. [\[CrossRef\]](#)
15. Su, H.; Huang, J.; Gu, J.; Xu, C.; Xu, J.; Hu, H. Non-contact Sensor-Based Gear Fault Diagnosis with Bark Domain Feature Analysis. *J. Fail. Anal. Prev.* **2025**, *25*, 3055–3066. [\[CrossRef\]](#)
16. Huang, W.; Du, B.; Li, T.; Cheng, Y.; Cui, S. Detection of PMSM Demagnetization Fault and Eccentricity Fault Based on Acoustic Images and DeiT Classifier. *IEEE Trans. Energy Convers.* **2025**, *40*, 1870–1884. [\[CrossRef\]](#)
17. Shi, H.; Li, Y.; Bai, X.; Zhang, K.; Sun, X. A two-stage sound-vibration signal fusion method for weak fault detection in rolling bearing systems. *Mech. Syst. Signal Process.* **2022**, *172*, 109012. [\[CrossRef\]](#)
18. Qiao, Y.; Wang, H.; Cao, J.; Lei, Y. Sound-vibration spectrogram fusion method for diagnosis of RV reducers in industrial robots. *Mech. Syst. Signal Process.* **2024**, *214*, 111411. [\[CrossRef\]](#)
19. Qi, Z.; Ren, J.; Gan, W.; Yu, P.S. Large Language Models for Fault Diagnosis. In *Proceedings of the 2025 IEEE International Conference on Big Data (BigData)*; IEEE: New York, NY, USA, 2025. [\[CrossRef\]](#)
20. Zhu, X.; Dong, J.; Wang, J.; Liu, G.; Zong, L.; Hu, C.; Su, S. Multimodal bearing fault diagnosis based on semantic-aware and enhanced cross-attention. *Meas. Sci. Technol.* **2026**, *37*, 106110. [\[CrossRef\]](#)

21. Nie, Q.; Geng, J.; Liu, C. A Review of Fault Diagnosis Methods: From Traditional Machine Learning to Large Language Model Fusion Paradigm. *Sensors* **2026**, *26*, 702. [\[CrossRef\]](#)
22. Cheng, Y.; Liu, L.; Liao, Z.; Chen, B.; Yan, J.; Chen, Z. A novel knowledge distillation framework for bearing fault diagnosis under imbalanced samples. *Struct. Health Monit.* **2026**. [\[CrossRef\]](#)
23. Zhou, F.; Zhu, B.; Liu, F. Identification of Electrical Anomalies in Co-Phase Power Supply System. *IEEE Trans. Transp. Electr.* **2025**, *11*, 9212–9223. [\[CrossRef\]](#)
24. Xiao, C.; Liu, X.; Wulamu, A.; Zhang, D. KG-SR-LLM: Knowledge-Guided Semantic Representation and Large Language Model Framework for Cross-Domain Bearing Fault Diagnosis. *Sensors* **2025**, *25*, 5758. [\[CrossRef\]](#)
25. Zhang, J.; Zhang, C.; Lu, J.; Zhao, Y. Domain-specific large language models for fault diagnosis of heating, ventilation, and air conditioning systems by labeled-data-supervised fine-tuning. *Appl. Energy* **2025**, *377*, 124378. [\[CrossRef\]](#)
26. Jiang, L.; Li, X.; Wu, L.; Li, Y. Bearing fault diagnosis method based on a multi-head graph attention network. *Meas. Sci. Technol.* **2022**, *33*, 075012. [\[CrossRef\]](#)
27. Wan, L.; Huang, L.; Ning, J.; Li, C.; Li, K. A novel meta-transfer learning approach via convolutional multi-head self-attention network for few-shot fault diagnosis. *Knowl.-Based Syst.* **2024**, *299*, 112113. [\[CrossRef\]](#)
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
29. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
30. Liu, J.J.; Hou, Q.; Cheng, M.M.; Wang, C.; Feng, J. Improving convolutional networks with self-calibrated convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10093–10102.
31. Yu, J.; Zhou, X. One-dimensional residual convolutional autoencoder based feature learning for gearbox fault diagnosis. *IEEE Trans. Ind. Inform.* **2020**, *16*, 6347–6358. [\[CrossRef\]](#)
32. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539.
33. Zhang, Y.; Zhao, X.; Peng, Z.; Hui, Y.; Xu, R. FCTransformer: An intelligent fault diagnosis method for multi-sensor information fusion based on Fourier convolution and transformer. *Mech. Syst. Signal Process.* **2026**, *249*, 114061. [\[CrossRef\]](#)
34. Sehri, M.; Dumond, P.; Bouchard, M. University of Ottawa constant load and speed rolling-element bearing vibration and acoustic fault signature datasets. *Data Brief* **2023**, *49*, 109327. [\[CrossRef\]](#)

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.