20th IMEKO TC10 Conference
28-29 September 2026 | Lisbon | Portugal

20th IMEKO TC10 Conference
"Measurement for Diagnostics, Optimization and Control to Support Competitiveness and Innovation"
Lisbon, Portugal, September 28–29, 2026

When Multimodal Fusion Fails: A Study of Pose-Vibration Fusion for Robot Fault Diagnosis

Zeqi Wei[1], Zhibin Zhao[2], Ruqiang Yan[3]

[1]*Xi'an Jiaotong University, Xi'an, China, weizeqi@stu.xjtu.edu.cn*
[2]*Xi'an Jiaotong University, Xi'an, China, zhaozhibin@xjtu.edu.cn*
[3]*Xi'an Jiaotong University, Xi'an, China, yanruqiang@xjtu.edu.cn*

***Abstract* – Industrial robots inevitably suffer from performance degradation and mechanical faults during long-term operation, making accurate fault diagnosis essential for ensuring system reliability and reducing maintenance costs. Recently, multimodal fault diagnosis methods have attracted increasing attention because multiple sensor modalities are generally assumed to provide complementary fault information. However, the effectiveness of multimodal fusion strongly depends on the fault relevance of each modality, which remains insufficiently investigated in industrial applications. This paper presents an empirical study on pose-vibration multimodal fusion for industrial robot fault diagnosis, with particular emphasis on the phenomenon of multimodal fusion failure. A ResNet1D-based vibration encoder and a lightweight pose encoder are developed to extract modality-specific representations, while concatenation-based fusion and channel-attention fusion strategies are further investigated. Experimental results show that vibration signals alone achieve the best diagnostic performance with 99.23% test accuracy, whereas pose signals only achieve 67.88%, indicating limited fault-discriminative capability in the pose modality. More importantly, introducing pose information consistently degrades diagnostic performance: concatenation fusion reduces the accuracy to 94.42%, while attention fusion achieves 96.34%, both remaining inferior to the vibration-only baseline. Attention weight analysis further reveals that the learned fusion model consistently assigns higher importance to vibration features while suppressing pose-related representations. The results suggest that weakly fault-related modalities may introduce negative transfer and cross-modal interference rather than performance gains. This study provides practical insights into modality selection and reliability evaluation for multimodal industrial fault diagnosis systems.**

Keywords – Fault diagnosis; multimodal fusion; modal negative transfer; Industry, Innovation and Infrastructure

I. INTRODUCTION

Industrial robots have become indispensable in intelligent manufacturing systems due to their high precision, flexibility, and automation capability [1,2]. During long-term operation, however, robot components such as bearings, reducers, joints, and transmission systems inevitably undergo performance degradation and mechanical wear, which may eventually lead to faults or unexpected shutdowns. Accurate fault diagnosis is therefore critical for improving system reliability, reducing maintenance costs, and ensuring production safety.

Traditional fault diagnosis approaches mainly rely on model-based methods and signal processing techniques [3]. With the rapid development of deep learning, data-driven diagnosis methods based on convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Transformer architectures have demonstrated superior performance in extracting discriminative fault features from raw sensory data [4,5]. Among various sensing modalities, vibration signals remain the dominant source for mechanical fault diagnosis because they directly reflect structural dynamic variations induced by faults [6].

Recently, multimodal learning has emerged as an important research direction in intelligent fault diagnosis. Researchers attempt to combine multiple sensory modalities, such as vibration, current, acoustic, thermal, and motion signals, to improve diagnostic robustness and generalization ability [7, 8]. Existing multimodal methods are generally built upon an implicit assumption that additional modalities provide complementary information and can therefore enhance feature representation capability.

However, this assumption may not always be valid in practical industrial scenarios. Different modalities possess substantially different fault sensitivities and information densities. Some modalities may exhibit strong fault relevance, while others only weakly correlate with the underlying degradation process. Introducing weakly informative modalities may increase feature redundancy, optimization difficulty, and cross-modal conflicts during representation learning [9]. Consequently, multimodal

fusion can potentially degrade model performance rather than improve it. This phenomenon can be interpreted as a form of multimodal negative transfer caused by modality relevance imbalance. Despite this important issue, existing studies primarily focus on designing increasingly sophisticated fusion architectures, while the negative effects of multimodal integration remain insufficiently investigated.

In industrial robots, vibration signals directly characterize mechanical dynamic responses and are highly sensitive to localized faults. In contrast, end-effector pose information mainly reflects motion trajectories and operational states. Although pose data may contain indirect degradation-related patterns, its fault sensitivity is often considerably weaker than vibration signals, especially under fixed trajectory conditions. Therefore, blindly combining pose information with vibration signals may introduce irrelevant motion variations and interfere with discriminative fault feature extraction.

Motivated by this observation, this paper investigates the effectiveness of pose-vibration multimodal fusion for robot fault diagnosis and focuses on understanding when multimodal fusion fails. Specifically, we construct a vibration-based diagnosis framework using a ResNet1D encoder and introduce a pose encoder to extract motion-related representations from pose sequences. Two representative fusion strategies, including feature concatenation and channel-wise attention fusion based on attention reweighting mechanisms [10], are designed to investigate cross-modal interactions and feature contribution mechanisms.

Experimental results demonstrate that vibration signals alone already achieve excellent diagnosis performance, while pose information provides limited fault-related knowledge. More importantly, introducing pose features consistently reduces diagnostic accuracy, even when adaptive attention mechanisms are employed. Attention analysis further reveals that the learned fusion weights are dominated by vibration features, indicating that the network implicitly suppresses the weakly informative pose modality.

The main contributions of this work are summarized as follows:

(1) An empirical investigation into modality relevance imbalance in robot multimodal diagnosis is conducted; (2) A pose-vibration multimodal diagnosis framework based on ResNet1D and attention fusion is developed; (3) The mechanism of multimodal negative transfer is analyzed through accuracy comparison and attention weight interpretation; (4) Practical insights for modality selection in industrial multimodal diagnosis are provided.

II. METHODOLOGY

2.1 Overall Framework

The proposed framework consists of three major components:

- Vibration feature extraction module
- Pose feature extraction module
- Cross-modal fusion and classification module

The vibration branch extracts discriminative fault-related dynamic features from raw vibration signals using a ResNet1D backbone, while the pose branch models motion-related representations from end-effector pose sequences. The extracted features are subsequently fused and fed into a classifier for fault category prediction.

2.2 Pose Feature Encoder

The pose modality contains end-effector motion information represented by position, velocity, angular velocity, and related kinematic variables.

The pose sequence is represented as:

$$X_p \in \mathbb{R}^{L\times M} \tag{1}$$

where L denotes the sequence length of the pose signal, and M denotes the number of pose features.

Unlike vibration signals, pose data mainly describe motion states rather than direct mechanical fault responses. Therefore, a lightweight multilayer perceptron (MLP) encoder is adopted instead of deep temporal modeling. The lightweight design prevents excessive overfitting to motion patterns while preserving global trajectory characteristics.

2.3 Vibration Feature Encoder

Since vibration signals contain rich temporal and frequency-related fault characteristics, a one-dimensional residual convolutional network (ResNet1D) is employed as the backbone feature extractor.

The input vibration signal is represented as:

$$X_v \in \mathbb{R}^{T\times N} \tag{2}$$

where T denotes the sequence length of the vibration signal, and N denotes the number of vibration channels.

The encoder consists of:

- Initial large-kernel convolution
- Residual convolutional blocks
- Progressive temporal downsampling
- Adaptive global average pooling

Residual learning enables stable optimization and improves the extraction of deep discriminative temporal features. After feature extraction, the vibration encoder outputs a compact representation.

2.4 Feature Fusion Strategies

(1) Direct Concatenation Fusion

The first strategy directly concatenates vibration and pose features:

$$f_{cat} = [f_v; f_p] \tag{3}$$

A channel attention module is then applied to adaptively reweight fused features before dimensionality reduction and classification.

Although straightforward, this strategy may suffer from feature distribution inconsistency between modalities.

This section gives scope for explaining the used

methods and algorithms.

(2) Channel Attention Fusion

To alleviate modality imbalance, a channel-wise attention fusion mechanism is introduced.

The pose feature is first projected into the vibration feature dimension:

$$\hat{f}_p = W_p f_p \tag{4}$$

Fusion weights are then generated through a learnable attention network:

$$w = \mathrm{Soft\,max}(\phi([f_v; f_p])) \tag{5}$$

The final fused representation is computed as:

$$f = w_v \odot f_v + w_p \odot \hat{f}_p \tag{6}$$

where w_v and w_p denote modality-specific channel weights.

This design enables adaptive modality contribution learning and provides interpretability through attention analysis.

III. RESULTS AND DISCUSSIONS

3.1 Experimental Results

The experimental accuracy is summarized as shown in Table 1:

Table 1. Accuracy comparison

Method	Train Accuracy (%)	Test Accuracy (%)
Vibration Only	99.78	99.23
Pose Only	85.00	67.88
Concatenation Fusion	99.01	94.42
Attention Fusion	99.08	96.34

Vibration signals alone already achieve extremely high diagnostic performance, indicating that vibration features contain highly discriminative fault-related information. Pose information exhibits substantially weaker diagnostic capability. The significant gap between training and testing accuracy in the pose-only setting suggests poor generalization and weak fault correlation.

Besides, the modality weights of vibration and pose signals in both fusion methods are shown in Table 2.

Table 2. Modality weights

Method	Vibration Weight	Pose Weight
Concatenation Fusion	0.7629	0.2371
Attention Fusion	0.6892	0.3108

Integrating pose information consistently degrades diagnosis performance. Compared with the vibration-only baseline, concatenation fusion causes a noticeable performance drop from 99.23% to 94.42%. Although attention fusion partially alleviates this issue, its performance still remains inferior to the vibration-only model.

3.2 Analysis of Fusion Failure

The experimental results demonstrate a typical negative transfer phenomenon in multimodal learning.

The primary reason lies in modality relevance imbalance. Vibration signals directly capture mechanical dynamic responses induced by faults, whereas pose signals mainly describe robot motion trajectories. Under fixed operational trajectories, pose variations are weakly correlated with fault categories and thus contribute limited discriminative information.

Consequently, introducing pose information leads to several adverse effects:

- Feature redundancy
- Cross-modal distribution inconsistency
- Optimization interference
- Reduced representation discriminability

Attention weight analysis further supports this conclusion. The learned weights consistently favor vibration features, indicating that the network automatically suppresses the less informative pose modality. Even so, the residual interference introduced by pose features still negatively affects final performance.

These findings suggest that multimodal fusion should not be blindly applied in industrial diagnosis tasks. The effectiveness of fusion strongly depends on modality relevance and information complementarity. Weakly fault-related modalities may introduce negative transfer instead of performance improvement.

IV. CONCLUSIONS

This paper investigates the effectiveness of pose-vibration multimodal fusion for industrial robot fault diagnosis and presents an empirical study on multimodal fusion failure. Experimental results demonstrate that vibration signals alone already provide highly discriminative fault representations, while pose information exhibits weak fault relevance and poor generalization capability.

More importantly, integrating pose information consistently degrades diagnosis performance, even when adaptive attention fusion mechanisms are employed. Attention analysis further reveals that the model inherently prioritizes vibration features while suppressing pose contributions.

The findings challenge the common assumption that additional modalities universally improve diagnostic performance and highlight the importance of modality relevance evaluation before multimodal integration. Future work will focus on modality reliability estimation, adaptive modality selection, and uncertainty-aware multimodal diagnosis frameworks to mitigate negative transfer in industrial intelligent maintenance systems.

V. CKNOWLEDGMENTS

Express here your gratitude to those who or to other kinds of support helped for performing your research.

REFERENCES

[1] **Lee, J.; Bagheri, B.; Kao, H. A.**: A cyber-physical systems architecture for Industry 4.0-based manufacturing systems, *Manufacturing Letters*, Vol. 3., 2015, pp. 18-23.

[2] **Kusiak, A.**: Smart manufacturing, *International Journal of Production Research*, Vol. 56., No. 1-2., 2018, pp. 508-517.

[3] **Jardine, A. K. S.; Lin, D.; Banjevic, D.**: A review on machinery diagnostics and prognostics implementing condition-based maintenance, *Mechanical Systems and Signal Processing*, Vol. 20., No. 7., 2006, pp. 1483-1510.

[4] **Wen, L.; Li, X.; Gao, L.; Zhang, Y.**: A new convolutional neural network-based data-driven fault diagnosis method, *IEEE Transactions on Industrial Electronics*, Vol. 65., No. 7., 2018, pp. 5990-5998.

[5] **Li, Z.; Li, C.; Sanchez, R. V.**: Transformer-based deep learning model for machinery fault diagnosis, *Reliability Engineering and System Safety*, Vol. 222., 2022, pp. 108497.

[6] **Jin, Y.; Hou, L.; Chen, Y.**: A time series transformer based method for the rotating machinery fault diagnosis, *Neurocomputing*, Vol. 494., 2022, pp. 379-395

[7] **Randall, R. B.**: Vibration-based Condition Monitoring: Industrial, Aerospace and Automotive Applications, *Wiley*, Chichester, 2011, pp. 1-410.

[8] **Che, C.; Wang, H.; Ni, X.; et al**.: Hybrid multimodal fusion with deep learning for rolling bearing fault diagnosis, *Measurement*, Vol. 173., 2021, p. 108655.

[9] **Zhang, W.; Peng, G.; Li, C.; Chen, Y.; Zhang, Z.**: A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals, *Sensors*, Vol. 17., No. 2., 2017, pp. 425.

[10] **Wei, S.; Luo, Y.; Wang, Y.; et al.**: Robust multimodal learning via representation decoupling, *European Conference on Computer Vision*, Cham: Springer Nature Switzerland, 2024, pp. 38-54.