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

Quality-Gated Motion-Context Fusion for Industrial 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 - Motion context provides useful operating-condition information for vibration-based industrial robot fault diagnosis, but conventional multimodal fusion may become overly dependent on auxiliary inputs when they are missing or corrupted. This paper proposes quality-gated motion-context fusion (QMF), a posterior-level fusion framework that combines an independent vibration-only expert with a context-assisted expert through a learned context-quality gate. The vibration-only expert serves as the diagnostic anchor, while the gate adaptively controls the contribution of synchronized robot pose sequences and active-axis codes according to their intrinsic quality. QMF is evaluated on a nine-class RV reducer fault diagnosis task using synchronized two-channel vibration signals, robot pose sequences, and active-axis codes. Across ten random seeds, QMF achieves an accuracy of 95.53±0.92%, outperforming the strongest fusion baseline, Modality-dropout, by 5.47 percentage points. It also achieves the highest mean accuracy under all evaluated vibration-degradation conditions. When the pose input is missing or heavily corrupted by noise, the quality gate suppresses the context-assisted prediction and restores performance close to that of the vibration-only expert. Shuffled-pose evaluation further shows that a context-only quality gate cannot reliably detect plausible but incorrectly paired context, distinguishing intrinsic context quality from cross-modal consistency. These results demonstrate that QMF enables controlled use of motion context while preserving a complete vibration-based fallback under intrinsic context failures.

Keywords – Fault diagnosis; multimodal fusion; context quality gate; Industry, Innovation and Infrastructure

I. INTRODUCTION

Reliable fault diagnosis is essential for industrial robots because degradation in reducers, joints, and transmission components can reduce production reliability and increase maintenance costs [1]–[3]. Vibration signals are widely used for this task because they directly capture mechanical responses associated with localized faults [4].

Data-driven diagnostic methods, including convolutional, recurrent, and Transformer-based models, can automatically extract discriminative features from sensor measurements [5]–[7]. Multimodal learning can further improve diagnosis by incorporating complementary operating information [8]. However, auxiliary modalities should not be assumed to be consistently reliable or independently fault-informative.

For industrial robots, end-effector pose and active-axis information primarily describe the operating condition under which a vibration signal is acquired. Such information can help distinguish motion-dependent variations in the vibration response. Nevertheless, pose and active-axis inputs may be missing, noisy, delayed, or incorrectly paired with a vibration window. Conventional feature-level fusion may therefore improve diagnosis under matched conditions while introducing undesirable dependence on invalid motion context.

To address this problem, this paper proposes quality-gated motion-context fusion (QMF). QMF adopts an asymmetric architecture in which an independent vibration-only expert provides the primary diagnosis and remains available as a complete fallback. A context-assisted expert jointly uses vibration, pose, and active-axis information, while a learned context-quality gate controls the contribution of the context-assisted posterior to the final prediction.

The proposed gate targets intrinsic context failures, particularly missing and heavily noise-corrupted pose inputs. Under these conditions, the gate suppresses the context-assisted prediction and restores the vibration-only posterior. The method does not claim to detect every cross-modal inconsistency. For example, a shuffled pose sequence may remain intrinsically plausible even when it is paired with the wrong vibration window. This condition is therefore used to characterize the boundary of context-only quality estimation.

The study investigates whether motion context can improve vibration-based diagnosis without sacrificing the vibration-only fallback when the context is intrinsically invalid. QMF is compared with vibration-only diagnosis and three representative multimodal controls: Direct fusion, Modality-dropout fusion, and FiLM fusion.

The main contributions are summarized as follows:

- A quality-gated posterior-fusion architecture that preserves a complete vibration-only fallback while conditionally incorporating pose and active-axis context.
- A counterfactual training strategy that teaches the quality gate to suppress intrinsically invalid context and recover the vibration-only prediction.
- A controlled evaluation under matched conditions, vibration degradation, pose corruption, and temporal pose perturbation, including an explicit analysis of the boundary under plausible but incorrectly paired context.

II. METHODOLOGY

2.1 Overall Framework

The proposed quality-gated motion-context fusion (QMF) framework consists of a vibration-only expert, a context-assisted expert, and a context-quality gate. Each temporally aligned sample contains a vibration window x_v, a robot pose sequence x_p, and an active-axis code a, where a is represented as a one-hot vector.

The vibration-only expert performs diagnosis using only x_v, whereas the context-assisted expert jointly uses vibration, pose, and active-axis information. The context-quality gate estimates the intrinsic usability of the motion context and controls its contribution to the final prediction. This posterior-level fusion design preserves vibration as the primary source of fault evidence and provides a complete vibration-only fallback when the auxiliary context is missing or severely corrupted. The whole framework is shown in Fig. 1.

2.2 Vibration and context-assisted experts

A one-dimensional ResNet-18 encoder extracts the vibration representation

$$h_v = f_v(x_v) \tag{1}$$

where $f_v(\cdot)$ denotes the vibration encoder. The vibration-only expert produces the class-probability vector

$$p_v = \mathrm{softmax}(g_v(h_v)) \tag{2}$$

where $g_v(\cdot)$ is the vibration-only classification head.

The motion context consists of the synchronized pose sequence and active-axis code. A lightweight pose encoder $f_p(\cdot)$ extracts the pose representation, while an embedding layer $f_a(\cdot)$ encodes the active-axis information. Their outputs are concatenated to form

$$h_c = \left[f_p(x_p); f_a(a) \right] \tag{3}$$

where $[\cdot\,;\,\cdot]$ denotes feature concatenation. A projection layer then maps h_c to a compact context representation

$$\tilde{h}_c = f_{\mathrm{proj}}(h_c) \tag{4}$$

The projected context representation is concatenated with the vibration feature, and the context-assisted expert produces

$$p_{vc} = \mathrm{softmax}\left(g_{vc}([h_v; \tilde{h}_c]) \right) \tag{5}$$

Pose and active-axis information are not regarded as independent fault evidence. Instead, they describe the operating condition under which the vibration window was

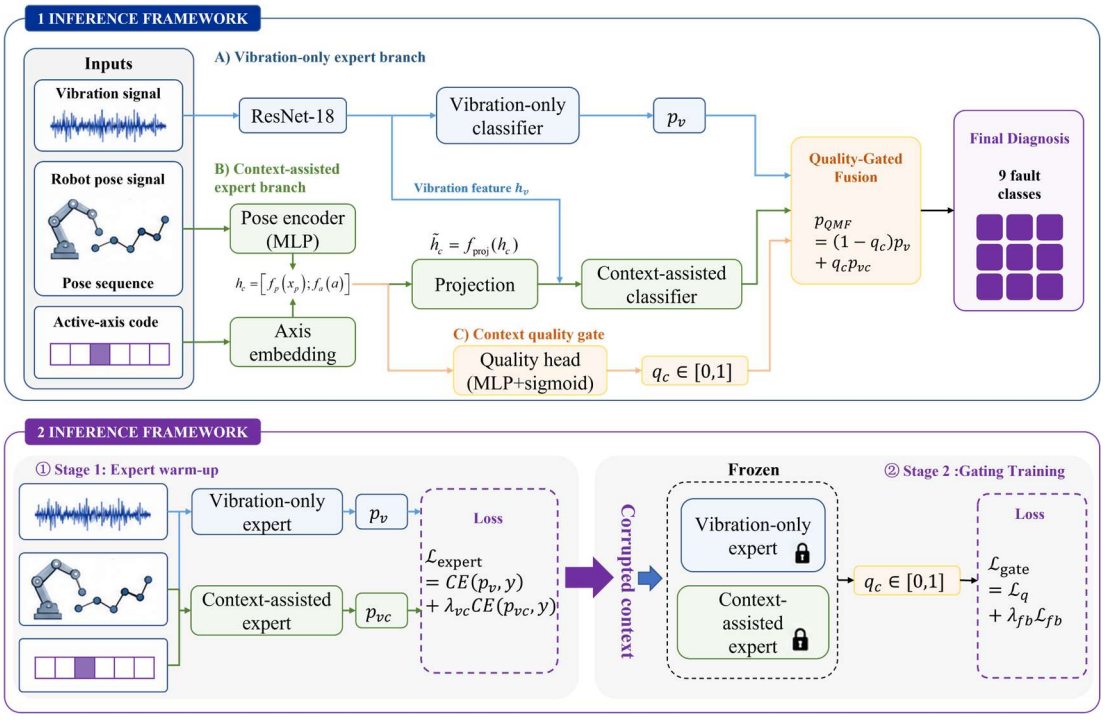

Fig. 1. The framework of the proposed QMF approach

acquired and help the context-assisted expert interpret motion-dependent variations in the vibration response.

2.3 Quality-gated motion-context fusion

The context-quality gate estimates a sample-specific scalar $q_c \in [0,1]$ from the context representation:

$$q_c = \sigma(g_q(h_c)) \tag{6}$$

where $g_q(\cdot)$ is a multilayer perceptron and $\sigma(\cdot)$ is the sigmoid function. Because the gate uses only the context representation, q_c measures the availability and intrinsic quality of the motion context rather than its cross-modal consistency with the vibration signal.

The final QMF posterior is obtained through output-level interpolation:

$$p_{\text{QMCF}} = (1-q_c)p_v + q_c p_{vc} \tag{7}$$

When the motion context is valid, q_c approaches one and the context-assisted expert contributes more strongly to the diagnosis. When the pose signal is missing or severely corrupted, q_c approaches zero and the final posterior converges to the vibration-only prediction:

$$\lim_{q_c \to 0} p_{\text{QMCF}} = p_v \tag{8}$$

Therefore, the quality gate controls whether motion context is allowed to affect the diagnostic result, while the vibration-only expert remains available as a complete fallback.

2.4 Counterfactual context-quality training

QMF is trained in two stages. In the first stage, the vibration-only and context-assisted experts are trained on the same nine-class diagnostic task using synchronized vibration, pose, and active-axis inputs. Their training objective is

$$\mathcal{L}_{\text{expert}} = \mathcal{L}_{CE}(p_v, y) + \lambda_{vc}\mathcal{L}_{CE}(p_{vc}, y) \tag{9}$$

where y is the fault label and λ_{vc} controls the contribution of the context-assisted classification loss.

In the second stage, the two diagnostic experts are frozen and only the context-quality gate is optimized. For each vibration window, a valid context input and an intrinsically corrupted counterpart are constructed while keeping the vibration input and fault label unchanged. The gate predicts an adaptive context-use coefficient $q_c \in [0,1]$ is an emergent decision variable that determines the degree to which motion context contributes to the final diagnosis. The gate is learned through diagnostic supervision, fallback consistency, and paired ranking:

$$\mathcal{L}_{\text{gate}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{fb}}\mathcal{L}_{\text{fb}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}} \tag{10}$$

where $\mathcal{L}_{\text{cls}}$ is the classification loss computed from p_{QMCF}. For intrinsically corrupted context inputs, the fallback-consistency term encourages the fused posterior to recover the frozen vibration-only decision:

$$\mathcal{L}_{\text{fb}} = D\left(p_v, p_{\text{QMCF}}\right) \tag{11}$$

where $D(\cdot,\cdot)$ denotes a posterior-divergence function. In addition, a paired ranking objective promotes a higher gate value for valid context than for its corrupted counterpart:

$$\mathcal{L}_{\text{rank}} = [m - q_c(x_v, x_c) + q_c(x_v, \tilde{x}_c)]_+ \tag{12}$$

where m is a positive ranking margin, $[\cdot]_+$ denotes the hinge operator and $\tilde{x}_c$ is corrupted context inputs. This design allows the gate to learn context acceptance from its diagnostic utility and counterfactual behaviour: valid context is retained when it supports fault discrimination, whereas intrinsically degraded context is discouraged from altering the vibration-supported prediction.

Zeroed pose and heavily noise-corrupted pose are treated as intrinsic context-quality counterfactuals because they directly alter the availability or signal quality of the auxiliary input. Shuffled and delayed pose sequences are used only as evaluation controls. Although such sequences may remain individually plausible, they can be temporally inconsistent with the corresponding vibration window. These perturbations therefore evaluate the boundary of context-only quality estimation rather than serving as intrinsic-quality supervision.

III. RESULTS AND DISCUSSION

3.1 Dataset description

The dataset was collected using a BRTIRUS1510A six-degree-of-freedom industrial robot [9]. Faults were introduced by replacing components in the RV reducers of axes 2 and 3, as shown in Fig. 2. The diagnostic task comprises nine classes, including one healthy condition and eight reducer fault conditions. For axis 2, the fault classes are sun-gear pitting, sun-gear tooth breakage, planet-gear cracking, and planet-gear tooth breakage. The same four fault types were introduced into the RV reducer of axis 3.

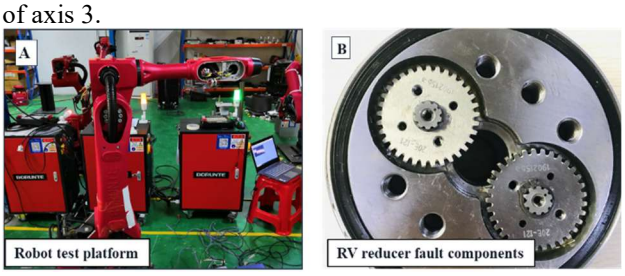

Fig. 2. Experimental platform and fault components (A) industrial robot test platform. (B) RV reducer components

Two types of signals were synchronously collected: vibration signals and robot pose signals, sampled at 100 kHz and 100 Hz, respectively, which are shown in Fig. 3. The signals were segmented into temporally aligned 0.5-s windows, with each window containing 50 consecutive pose samples. The vibration signals were downsampled by a factor of 10, retaining 5,000 samples per channel in each window. Accordingly, each vibration input had a size of 5000×2, whereas the corresponding pose input had a size of 50×9, comprising 50 time points and 9 pose channels.

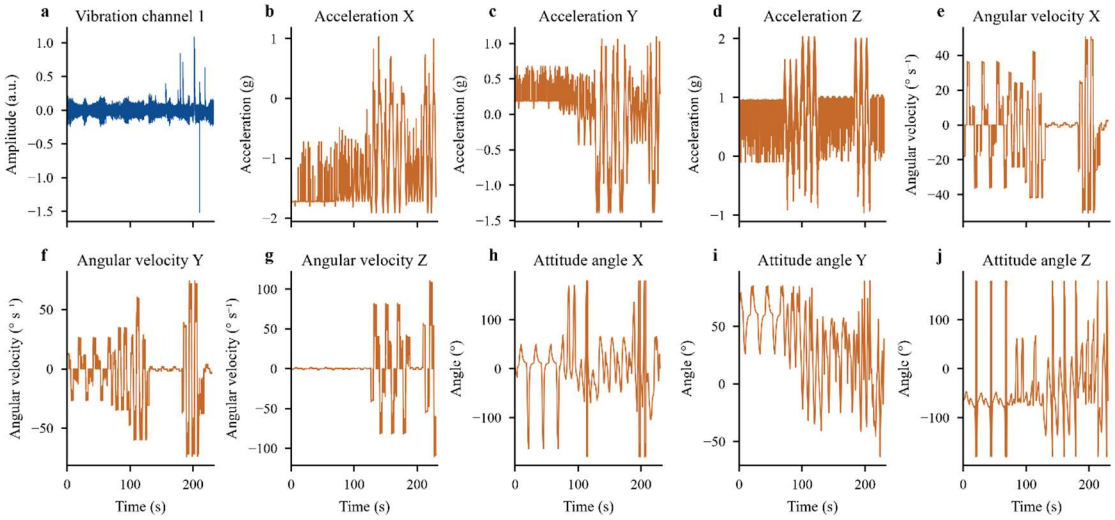

Fig. 3. Full-sequence vibration and nine-channel pose signals of the industrial robot.

3.2 Fusion baselines and comparison protocol

To determine whether the performance improvement arises from posterior-level quality gating rather than merely from incorporating motion context, QMF is compared with a vibration-only reference and three representative multimodal fusion baselines: Direct fusion, Modality-dropout fusion, and FiLM fusion [10].

The vibration-only model uses the same one-dimensional ResNet-18 encoder and classification head as the vibration expert in QMF. It predicts the fault class solely from the vibration representation h_v and provides the reference performance expected when motion context is unavailable.

Direct fusion represents conventional feature-level multimodal integration. It concatenates the vibration, pose, and active-axis representations and predicts the fault label as

$$\mathrm{softmax}\left(g_{\mathrm{Direct}}([h_v;h_p;h_a])\right). \tag{13}$$

All input modalities therefore contribute to a shared diagnostic representation, without an explicit mechanism for rejecting unreliable motion context.

Modality-dropout fusion adopts the same feature-concatenation architecture as Direct fusion but randomly masks the complete context representation during training:

$$\tilde{h}_c = m h_c, \qquad m \sim \mathrm{Bernoulli}(1-p_{\mathrm{drop}}) \tag{14}$$

where p_{drop} is the context-dropout probability. This strategy exposes the classifier to both multimodal and vibration-dominant samples. However, it does not estimate a sample-specific context-quality score or explicitly recover the posterior of an independently trained vibration-only expert.

FiLM fusion treats motion context as a conditioning variable for the vibration representation. The context representation generates feature-wise scaling and shifting parameters, and the vibration feature is modulated as

$$\gamma(h_c)\odot h_v + \beta(h_c) \tag{15}$$

where $\odot$ denotes element-wise multiplication. Although FiLM provides context-dependent interpretation of vibration features, unreliable context can directly perturb the diagnostic representation because no explicit rejection or posterior-level fallback mechanism is included.

For fair comparison, all methods use identical data partitions, input preprocessing procedures, backbone configurations, corruption settings, seed sets, and evaluation metrics. Each method retains its architecture-specific fusion module and training objective. Direct fusion, FiLM fusion, and Modality-dropout fusion are optimized using the fault-classification objective on the common training samples. For QMF, the same intrinsic context corruptions are used during the gate-training stage after the two diagnostic experts have been frozen.

3.3 Evaluation protocol

Accuracy is used as the primary evaluation metric, and all methods are evaluated on the same held-out test set. All multimodal methods are trained under the same mixed context- and vibration-degradation protocol. Table 1 reports the mean and standard deviation across ten fixed random seeds, while Tables 2–4 report the corresponding ten-seed results under the controlled evaluation conditions.

The evaluation covers three groups of evidence: matched-condition diagnosis, vibration degradation, and pose counterfactuals. The vibration-degradation conditions include 50% signal truncation, additive noise at 20 and 10 dB, and removal of vibration channel 0. Because each vibration window contains two channels, channel-0 removal sets the first channel to zero while retaining channel 1 and all motion-context inputs. It therefore represents partial channel loss rather than complete loss of the vibration modality. The pose counterfactuals include real, shuffled, delayed, zeroed, and heavily noise-corrupted pose sequences.

3.4 Main diagnostic performance

Under the final training protocol, QMF achieved the highest performance, as shown in Table. Its mean accuracy was 95.53±0.92%, compared with 84.85±5.47% for vibration-only diagnosis, 87.37±1.09% for Direct fusion, 88.74±4.67% for FiLM fusion, and 90.06±1.32% for Modality-dropout fusion. QMF therefore exceeded the strongest external fusion baseline by 5.47 percentage points. The pose-only model achieved an accuracy of 54.15±1.03%, indicating that robot motion context contains information associated with the diagnostic task. However, this result does not imply that end-effector pose directly reflects reducer fault characteristics. Instead, pose describes the operating condition and trajectory stage under which the vibration response is generated. The higher performance of QMF indicates that motion context is most useful when it conditions the interpretation of vibration evidence rather than replacing it.

Table 1. Matched-condition diagnostic performance.

Method	Vibration degradation	Context degradation	Accuracy (%)
Pose only			54.15±1.03
Vib. only			84.85±5.47
Direct	✓	✓	87.37±1.09
FiLM	✓	✓	88.74±4.67
Modality-dropout	✓	✓	90.06±1.32
QMF	✓	✓	**95.53±0.92**

3.5 Robustness to vibration degradation

Table 2 compares the multimodal methods under four vibration-degradation conditions. QMF achieved the highest mean accuracy in every condition, reaching 89.31% under 50% truncation, 90.03% at 20 dB noise, 86.39% at 10 dB noise, and 88.55% after ch0 removal. Relative to the strongest baseline in each condition, the corresponding improvements were 0.61, 1.63, 3.49, and 0.69% points. The magnitude of the improvement depended on the degradation type. The advantage of QMF was relatively small under signal truncation and single-channel removal but became more pronounced under 10 dB noise. Because ch0 removal retains the second vibration channel and all motion-context inputs, this condition is interpreted as partial vibration-channel loss rather than diagnosis without vibration evidence.

Table 2. Diagnostic accuracy under vibration degradation.

Condition	Direct	Modality-dropout	FiLM	QMF
Clean	87.37±1.09	90.06±1.32	88.74±4.67	**95.53±0.92**
Trunc. 0.5	84.95±0.88	88.70±0.73	86.69±4.66	**89.31±2.58**
Noise 20dB	86.92±0.46	86.42±1.03	88.40±3.88	**90.03±2.15**
Noise 10dB	78.04±1.72	79.11±2.34	82.90±7.36	**86.39±0.46**
No Ch0	85.13±1.15	87.86±1.45	87.07±4.97	**88.55±1.82**

3.6 Pose validity and fallback behaviour

The context-quality score responded strongly to intrinsic pose failures, as shown in Table 3. The mean q_c values were 0.828 for real pose, 0.0 for zeroed pose, and 0.007 for heavily noise-corrupted pose. Consequently, QMF returned to the vibration-only performance range under both intrinsic failures. It achieved 85.40% accuracy with zeroed pose and 85.36% with noisy pose, compared with 84.85% for vibration-only diagnosis. Under the same conditions, Modality-dropout achieved 76.72% and 75.65%, respectively. QMF therefore improved upon Modality-dropout by 8.68 and 9.71 percentage points.

Delayed pose retained a relatively high quality score of 0.784, and QMF achieved 91.54% accuracy, exceeding Modality-dropout by 7.16 percentage points. A different behavior was observed for shuffled pose. Its quality score remained high at 0.773, close to 0.828 for real pose, because the quality head observes only the motion-context representation. A shuffled sequence can remain intrinsically plausible even when it is paired with the wrong vibration window.

Under shuffled pose, QMF achieved 68.30% accuracy. Although this result remained higher than those of Direct fusion, FiLM fusion, and Modality-dropout fusion, it was below the 84.85% vibration-only reference. The shuffled-pose result therefore defines the boundary of the context-only quality gate: it can identify intrinsic corruption but cannot necessarily detect cross-modal pairing errors.

Table 3. Pose counterfactual evaluation.

Pose condition	Direct	Modality-dropout	FiLM	QMF	q_c
Real	87.37±1.09	90.06±1.32	88.74±4.67	**95.53±0.92**	0.828
Shuffled	36.63±3.09	59.01±1.33	51.04±3.82	**68.30±4.87**	0.773
Delayed	76.79±7.35	84.38±1.53	80.43±4.65	**91.54±3.30**	0.784
Zero	45.73±8.13	76.72±3.83	72.70±4.29	**85.40±5.55**	0.000
Noisy	43.79±5.35	75.65±2.76	68.75±5.73	**85.36±2.61**	0.007

3.7 Gate ablation

Table 4 evaluates the two components required for quality-controlled fallback: counterfactual quality training and sample-adaptive gating. Removing counterfactual training reduced QMF accuracy by 3.40 percentage points under truncation, 10.78 points at 20 dB noise, 9.94 points after removal of vibration channel 0, and 8.17 points under zeroed pose. In addition, the quality score under zeroed pose remained high at (q_c =0.910), indicating that ordinary diagnostic supervision alone does not teach the gate to identify intrinsic context failure.

Replacing the adaptive quality score with a fixed value of q_c =0.5 also reduced robustness. Compared with the complete QMF model, the fixed-gate variant was lower by 2.55 percentage points under truncation, 4.30 points at 20 dB noise, 3.79 points after removal of vibration channel 0, and 3.58 points under zeroed pose. Because the fixed gate continues to include the context-assisted prediction regardless of context quality, it cannot provide sample-specific fallback.

Table 4. Ablation results.

Variant	Trunc. 0.5	Noise 20dB	No Ch0	Zero pose
QMF w/o counterfactual training	85.91±2.41	79.25±2.65	78.61±2.42	77.23±7.36
QMF with q_c=0.5	86.76±3.25	85.73±1.98	84.76±2.15	81.82±6.70
QMF	**89.31±2.58**	**90.03±2.15**	**88.55±1.82**	**85.40±5.55**

These ablation results show that counterfactual quality supervision enables the gate to distinguish intrinsically valid and invalid context, while the adaptive quality score allows the final diagnosis to respond to sample-level changes in context quality.

3.8 Discussion

QMF treats robot motion context as a conditional operating cue rather than independent fault evidence. Unlike conventional feature-level fusion, it retains an independently trained vibration-only expert and introduces motion context only through posterior-level interpolation. This asymmetric design is appropriate because vibration is physically closer to reducer faults, whereas pose and active-axis information primarily describe the operating condition under which the vibration response is generated.

The matched-condition and vibration-degradation results show that motion context can improve the interpretation of motion-dependent vibration variation. More importantly, under zeroed and heavily noise-corrupted pose, the quality score approached zero and QMF recovered performance close to that of the vibration-only expert. The comparison with Modality-dropout indicates that exposure to degraded context alone does not provide the same explicit fallback behavior as posterior-level quality gating.

The shuffled-pose evaluation clarifies the scope of the proposed method. Because an incorrectly paired pose sequence may remain intrinsically plausible, a quality gate based only on the context representation cannot reliably detect every cross-modal mismatch. QMF is therefore most suitable for applications in which the main context failures involve missing values or severe signal corruption while vibration evidence remains available. Synchronization and pairing errors require a complementary cross-modal consistency mechanism.

IV. CONCLUSION

This paper proposed quality-gated motion-context fusion (QMF) for nine-class RV reducer fault diagnosis in industrial robots. QMF combines an independent vibration-only expert with a context-assisted expert through a sample-specific quality gate. By performing fusion at the posterior level, the framework preserves vibration as the primary source of diagnostic evidence while conditionally incorporating synchronized robot pose and active-axis information.

Across ten random seeds, QMF achieved an accuracy of 95.53±0.92%, outperforming Direct fusion, FiLM fusion, and Modality-dropout fusion. It also achieved the highest mean accuracy under all evaluated vibration-degradation conditions. When the pose input was zeroed or heavily corrupted by noise, the quality gate suppressed the context-assisted prediction and restored performance close to the vibration-only reference. These results demonstrate the benefit of explicitly controlling the contribution of auxiliary motion context rather than relying solely on feature-level fusion or modality masking.

The shuffled-pose evaluation further showed that intrinsic context quality and cross-modal consistency are distinct problems. Because an incorrectly paired pose sequence can remain individually plausible, the current context-only gate cannot reliably detect all pairing errors. The present evaluation is limited to one industrial robot platform, one motion protocol, and controlled signal-degradation conditions. Future work will investigate cross-modal consistency estimation for detecting synchronization and pairing failures.

V. ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Program of China (Grant No. 2024YFB4709200).

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

[5] **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.

[6] **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.

[7] **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.

[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] **Long, J.; Mou, J.; Zhang, L.; et al**.: Attitude data-based deep hybrid learning architecture for intelligent fault diagnosis of multi-joint industrial robots, *Journal of manufacturing systems*, Vol. 61., 2021, pp. 736-745.

[10] **Perez, E.; Strub, F.; De Vries, H.; et al.**: FiLM: Visual reasoning with a general conditioning layer, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32., 2018, pp. 3942-3951.