Journal of Dynamics, Monitoring and Diagnostics, 2026, 5, 64-73
https://doi.org/10.37965/jdmd.2026.1107

ISTP
RESEARCH ARTICLE

A Single-Device Environment-Adaptive Mixed Reality Framework for Real-Time Industrial Fault Diagnosis

Xueyi Li,[1,2] Bo Kang,[1] Jing Tang,[1] Qi Li,[2] Tianyang Wang,[2] and Minwei Zhang[3]
[1]College of Mechanical and Electrical Engineering, Northeast Forestry University, Harbin 150040, China
[2]Department of Mechanical Engineering, Tsinghua University, Beijing 100084, China
[3]Beijing Zhongyuan Ruixun Technology Co., Ltd., Beijing 100085, China

(Received 26 December 2025; Revised 28 January 2026; Accepted 03 February 2026; Published online 04 February 2026)

Abstract: In industrial environments, monitoring and fault diagnosis of mechanical equipment face challenges such as spatial localization drift and delays in real-time data rendering, especially in complex settings with low illumination, weak textures, and strong interference. Traditional methods struggle to effectively integrate monitoring data with physical entities, increasing cognitive load and reducing diagnostic accuracy. To address these issues, we propose the Single-Device Mixed Reality (SEMR) framework, a novel solution that enhances industrial equipment monitoring and fault diagnosis. The framework integrates three key mechanisms: an environment-aware model that adjusts the confidence of Simultaneous Localization and Mapping (SLAM) to ensure precise spatial registration, a Kalman filter-based motion prediction to reduce rendering delays, and a fault-tolerant gaze interaction system for hands-free operation. Experimental results demonstrate that SEMR reduces the spatial registration error by 52.1%, from 14.2 cm to 6.8 cm, and decreases latency during dynamic inspections by 26.7%, improving diagnostic accuracy and real-time performance. The proposed method provides a cost-effective and reliable solution for enhancing industrial fault diagnosis and equipment monitoring, particularly in challenging environments.

Keywords: adaptive SLAM; fault diagnosis; industrial inspection; predictive interaction; real-time equipment status monitoring; single-device mixed reality

I. INTRODUCTION

With the rapid development of Industry 4.0 and smart manufacturing, the operational reliability and safety of large and complex mechanical systems have become increasingly important [1–3]. Equipment status monitoring and fault diagnosis, through the analysis of multi-source signals such as vibration, temperature, acoustic emission, and current, enable anomaly detection, fault localization, and health assessment [4–6]. These techniques serve as a crucial foundation for ensuring the safe operation of industrial systems and reducing unplanned downtime [7]. In recent years, the integration of digital twins and advanced mixed reality (MR)-based visualization technologies has revolutionized fault diagnosis within industrial environments, achieved through enhancing real-time data visualization capabilities and decision-making support [8–10]. However, several challenges persist, including the complexity of real-world scenarios [11], the coupling of multiple faults, the interpretability of results, and the practical implementation of the proposed solutions [12,13]. However, key bottlenecks still exist, particularly in terms of data scarcity, industrial-specific modeling, and reliable real-time performance [14,15].

In terms of methodology, fault prediction and health management for industrial scenarios are gradually shifting from purely data-driven approaches to mechanism–data fusion [16–18]. Physics-informed deep learning models, by incorporating degradation priors and consistency constraints, improve generalization and reliability in tasks like remaining useful life prediction [19–21]. However, in traditional on-site inspection and diagnostic processes [22–24], maintenance personnel still commonly face the challenge of separation between information and physical entities. Monitoring results are often presented in the form of 2D waveforms, spectrograms, or threshold alarms on remote monitoring screens or handheld devices [25]. When engineers are faced with real equipment, it is difficult to quickly correlate abstract data with specific components in space [26], which increases cognitive load and may lead to misjudgments.

MR technology [27] offers an intuitive pathway to address the above challenges. By overlaying monitoring data, fault labels, and information from 3D visualization models directly onto the physical equipment [28], MR allows inspectors to quickly localize, compare, and make decisions on-site, enhancing the efficiency of the diagnostic loop. However, industrial workshops are typically unstructured and complex, characterized by metallic specular reflections, low-texture surfaces, pronounced illumination variations, and occlusions. Most consumer-grade MR headsets rely on visual Simultaneous Localization and Mapping (SLAM) for spatial tracking [29]; under such adverse conditions, they are prone to feature loss and pose drift. In precision diagnostic scenarios, even centimeter-level registration errors may cause virtual fault annotations to be placed on adjacent healthy components, compromising diagnostic accuracy. Beyond drift, industrial monitoring also involves high-rate real-time data streams. When inspectors move around equipment or rapidly switch viewpoints to examine different measurement points, excessive

Corresponding authors: Qi Li (e-mail: liq22@tsinghua.org.cn), Tianyang Wang (e-mail: wty19850925@tsinghua.edu.cn).

end-to-end latency can introduce visual misalignment and trailing artifacts between the rendered waveforms and the physical components, degrading responsiveness and interaction smoothness. On-site operations often involve hands-busy scenarios or the use of protective gloves, rendering conventional controller or gesture-based interactions inconvenient and potentially unsafe. Gaze-only triggering is also susceptible to vibration and gaze jitter, leading to inadvertent activations and falling short of industrial-grade reliability requirements.

To address the aforementioned challenges, this paper proposes a single-device MR optimization framework for industrial equipment monitoring, named SEMR. The framework is designed to achieve high-precision enhanced diagnostics in industrial environments using consumer-grade headsets, without relying on expensive external localization devices. The main innovative contributions of this paper are as follows:

(1) Proposed a dynamic adjustment mechanism for SLAM confidence based on environmental quality perception, combined with sparse reference markers, to achieve stable anchoring of monitoring data on the equipment surface and smooth drift correction.

(2) Introduced a short-term motion prediction and semantic priority scheduling strategy based on Kalman filtering to reduce the latency and artifacts in the rendering of real-time monitoring data.

(3) A multimodal, fault-tolerant, hands-free gaze interaction mechanism was designed, which reduces false triggers through dynamic gaze threshold adjustment and micro-gesture confirmation, thereby enhancing the safety of on-site human–machine collaboration.

(4) A lightweight single-device deployment solution was implemented, reducing deployment costs and improving adaptability across different scenarios.

The remainder of this paper is organized as follows: Section II reviews related work; Section III presents a detailed description of the SEMR framework's system design and optimization algorithms; Section IV validates the effectiveness of the proposed method through experiments in simulated industrial inspection scenarios; and Section V concludes the paper and discusses future work.

II. RELATED WORKS

A. CURRENT STATUS OF MR-ASSISTED MONITORING TECHNOLOGY IN INDUSTRIAL ENVIRONMENTS

MR can overlay monitoring data and information from 3D visualization models directly onto the surfaces of physical equipment, enabling a more intuitive human–machine interface for on-site inspection [30], condition monitoring, and fault diagnosis. In recent years, monitoring and diagnostic systems designed for real industrial scenarios are expected not only to improve algorithmic accuracy but also to meet practical constraints on reliability, engineering usability, and deployment cost.

However, industrial workshops are typically unstructured and complex: specular reflections from metallic surfaces, texture scarcity, occlusions, and pronounced illumination variations can degrade visual features, thereby undermining the visual-SLAM-based [31] tracking adopted by consumer-grade MR devices and causing spatial drift and virtual–physical misalignment. Related studies on vision-based dynamic monitoring have demonstrated that the performance of sensing and measurement is highly dependent on factors such as imaging quality and the observability of key features in the environment, with degraded visual conditions leading to a significant reduction in accuracy and reliability; when visual conditions deteriorate, the reliability of monitoring and diagnosis is consequently compromised. Existing spatial registration solutions for industrial MR-assisted monitoring [32] can be broadly categorized into three types: markerless single-device approaches are easy to deploy but are sensitive to low texture and lighting changes and thus prone to drift; feature-assisted approaches improve global consistency but may suffer from marker maintenance and occlusion issues; and infrastructure-based tracking offers high accuracy but incurs high cost, requires substantial site modification, and lacks flexibility. These three categories are technically related to different components of the proposed SEMR framework. Similar to markerless single-device methods, SEMR runs entirely on a self-contained headset, but its environment-adaptive SLAM enhancement module explicitly estimates environment quality and performs drift-aware fusion between the built-in SLAM and marker-based Perspective-n-Point (PnP) pose corrections. In contrast to purely feature-assisted or infrastructure-based tracking, SEMR only uses sparse fiducial markers as an occasional correction source and does not rely on external tracking hardware, making it better suited to visually degraded and cost-sensitive industrial workshops.

B. CHALLENGES OF REAL-TIME INTERACTION AND HANDS-FREE OPERATION IN DYNAMIC DIAGNOSTICS

In addition to spatial localization stability, dynamic diagnostics in industrial environments face two key challenges: real-time visualization temporal consistency and hands-free interaction reliability [33]. On the one hand, the in situ display of high-frequency data, such as vibration waveforms, spectrograms, and alarms, requires a processing chain that includes user motion and posture update, pose calculation, rendering submission, and display refresh. When inspection personnel rapidly turn their heads or move around equipment, end-to-end delays can cause lag, ghosting, and misalignment of virtual labels relative to the physical equipment, weakening the corresponding relationship of measurement points and potentially misleading fault localization.

Related work also indicates that monitoring and diagnostic systems are shifting toward decision support and closed-loop applications, but achieving high-reliability real-time performance under resource-limited edge deployments and complex operational conditions remains a core bottleneck [34]. Traditional MR systems mainly alleviate this problem through rendering pipeline optimization, reprojection, or cloud/edge offloading, which can reduce apparent latency but often introduce additional infrastructure dependencies or bandwidth constraints. In contrast, the user-behavior-driven prediction module in SEMR targets on-device end-to-end interaction latency: it exploits characteristic head and gaze motion patterns in industrial inspection to anticipate user intent and adapt the internal scheduling of pose update and visualization, while keeping

all computation on the headset. On the other hand, on-site inspections often involve hand occupancy, wearing gloves, or maintaining a safe posture, making traditional controller interactions cumbersome and posing safety risks. While gaze-based interaction offers the advantage of being hands-free, industrial vibrations and heavy gaze load can lead to gaze jitter and false triggers. Fixed threshold triggering mechanisms struggle to balance "fast response" and "false trigger prevention." Therefore, there is a need to introduce motion prediction and task scheduling to reduce perception latency under single-device deployment, along with adaptive threshold and multimodal confirmation mechanisms to enhance the engineering reliability of hands-free interaction. Existing gaze-based selection techniques in MR typically rely on fixed dwell-time thresholds or single-modal confirmation cues, which are sensitive to gaze jitter in dynamic environments. These methods struggle to maintain accuracy, especially in industrial applications where vibrations and head movements are frequent. In contrast, SEMR's gaze interaction optimization module adjusts the dwell time dynamically based on gaze stability and incorporates angular drift tolerance. Additionally, secondary confirmation, such as micro-blinks or gestures, is used to reduce false triggers. These adaptive features are designed to ensure fast, reliable, and safe interaction, making the system well suited for industrial inspection workflows.

III. SYSTEM DESIGN AND OPTIMIZATION METHODS

A. SYSTEM ARCHITECTURE

To meet the stringent requirements of portability and rapid deployment for inspection equipment in industrial environments, the SEMR framework adopts a single-device, non-intrusive architecture design [35]. Unlike traditional industrial VR/AR systems that rely on external base stations or fiber optic connections, this system operates entirely on the edge computing unit of a head-mounted display, achieving a closed-loop operation from data acquisition, algorithm processing, to holographic rendering.

The overall system architecture is shown in Fig. 1 and consists of three logical layers: the multidimensional perception layer, the core algorithm layer, and the interaction rendering layer. The multidimensional perception layer is responsible for capturing raw data from the physical environment and the operator in real time. The system utilizes a consumer-grade MR headset, specifically the Meta Quest 3, which integrates an RGB-D camera array and an inertial measurement unit to capture high-resolution color video streams, depth point cloud data, and 6-DoF headset pose. These sensor data serve as the foundation for subsequent environmental adaptability analysis. The core algorithm layer, as the "computing engine" of the system, includes three serially coupled optimization modules that address the three core challenges in industrial on-site environments. The interaction rendering layer is built using the Unity engine and the OpenXR standard, responsible for overlaying the processed monitoring data onto the real physical environment. The system is ultimately packaged into an independent Android application and deployed on the headset, eliminating the need for an external PC and allowing the system to independently execute inspection tasks in industrial environments, significantly reducing deployment costs and maintenance complexity. In all experiments reported in this paper, sensing, computation, and rendering were performed entirely on the headset without relying on any external edge device or remote network transmission.

B. ENVIRONMENT-ADAPTIVE SLAM ENHANCEMENT MODULE

Unlike traditional SLAM methods that reset the pose through marker points, the approach proposed in this paper integrates multiple environmental quality indicators, such as illumination and texture density, and uses dynamic interpolation to smoothly combine the results of visual SLAM and a marker-based PnP camera pose estimation algorithm [36]. This enhances the stability and robustness of the system in industrial scenarios. The framework of the proposed SLAM enhancement module is shown in Fig. 2.

The environment-adaptive SLAM enhancement module is initialized using two types of input data: the estimated camera poses $T^{(t)}_{SLAM\leftarrow cam}$, which are provided by the built-in SLAM system of the Meta Quest 3 headset, and additional environmental information used to optimize the system's performance in dynamic conditions, and the current RGB frame. In this paper, the camera pose is represented as a

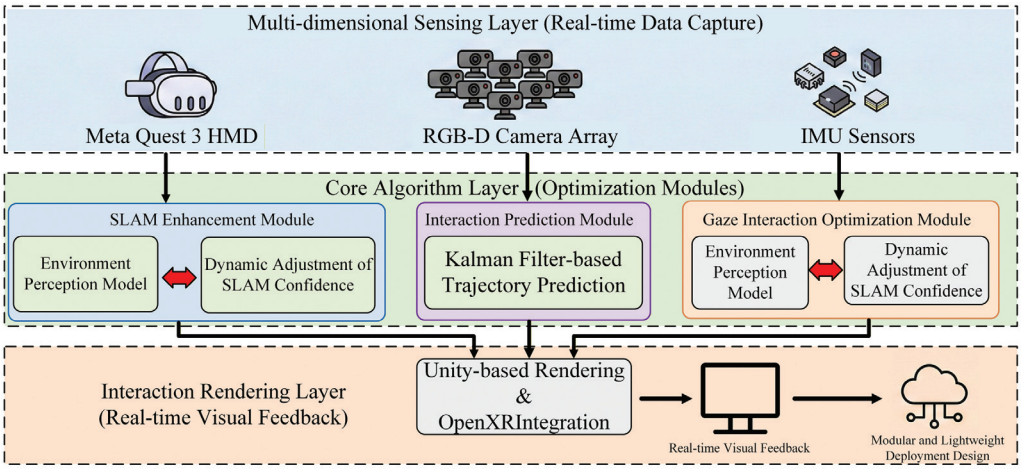

Fig. 1. SEMR system architecture diagram.

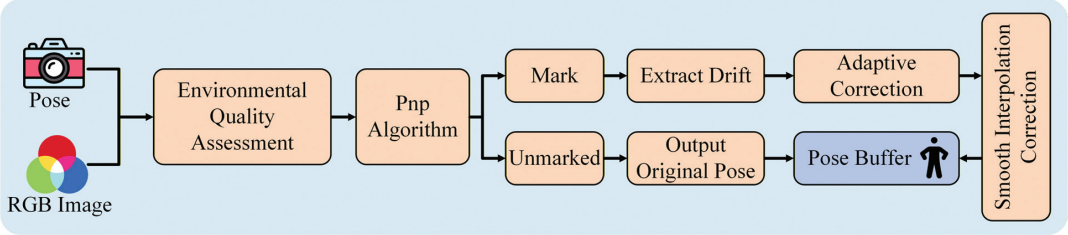

Fig. 2. Environment-adaptive SLAM enhancement module.

6-DoF rigid-body transformation between the world and camera coordinate systems. First, the environment-adaptivity score Q_t is computed as:

$$Q_t = w_L L_t + w_D D_t + w_C C_t \tag{1}$$

where $L_t \in [0,1]$ denotes the normalized grayscale value of the image, reflecting the uniformity and adequacy of environmental illumination, $D_t \in [0,1]$ denotes the normalized feature-point density per unit region (texture richness), characterizing the abundance of scene texture, and $C_t \in [0,1]$ denotes the tracking confidence provided by the SLAM system.

Acquisition of SLAM confidence C_t. The confidence value C_t is produced by the headset's native SLAM subsystem (e.g., Meta Quest 3) and reflects the reliability of the current pose estimate. In practice, C_t is typically computed from factors such as the number (and spatial distribution) of tracked features, the magnitude of the reprojection error, and the confidence state exposed by the SLAM API. When the SLAM system can stably track sufficient features, C_t is high; under low-texture scenes or rapid motion, C_t becomes low:

$$C_t = clip\left(\lambda_1 \frac{N_{inlier}}{N_{feat}} + \lambda_2 \left(1 - \frac{\bar{e}_{repr}}{e_{max}}\right), 0, 1\right) \tag{2}$$

where N_{feat} denotes the number of features in the current frame, N_{inlier} is the inlier count estimated by PnP, representing the number of valid feature matches in the current frame, $\bar{e}_{repr}$ denotes the mean reprojection error, quantifying the projection error of the frame's feature points, e_{max} is the maximum reprojection-error threshold used for comparison and normalization, and λ_1 and λ_2 are weighting coefficients that balance the two terms. Where w_L, w_D, and w_D are the weighting coefficients for illumination, texture, and tracking confidence, respectively, and they satisfy

$$w_L + w_D + w_C = 1 \tag{3}$$

The environment quality of an image is evaluated by illumination intensity L_t, texture density D_t, and SLAM tracking confidence C_t. These three indicators are combined into the environment score Q_t, which is used to assess the ability of the scene. This score serves as the basis for determining the interpolation smoothness in subsequent pose correction and is thus the key reference for dynamic reweighting.

Subsequently, a 2D planar fiducial marker and a PnP pose estimation algorithm are employed to estimate the camera pose transformation $T^{(t)}_{cam \leftarrow marker}$. If no marker is detected or if the confidence of the detected marker is insufficient, the system skips this frame's correction and directly outputs the original SLAM pose. Otherwise, the incremental transformation is computed as:

$$\Delta T_t = T_{world \leftarrow marker} (T^{(t)}_{cam \leftarrow marker})^{-1} T^{(t)}_{SLAM \leftarrow cam} \tag{4}$$

where $T_{world \leftarrow marker}$ denotes the predefined marker pose in the world coordinate system, $T^{(t)}_{cam \leftarrow marker}$ denotes the marker-to-camera transformation estimated from the detected marker using the PnP algorithm, and ΔT_t denotes the incremental correction between the current SLAM pose and the marker-based pose.

Through the decomposition of translation and rotation from ΔT_t, the translational drift and rotational drift are defined as:

$$d_{trans,t} = \|t_\Delta\| \tag{5}$$

$$\theta_{rot,t} = arccos\left(\frac{tr\left(R_\Delta\right) - 1}{2}\right) \tag{6}$$

where t_Δ, R_Δ, and ΔT_t denote the translational component, rotational component, and incremental transformation, respectively. $d_{trans,t}$ represents the translational error in three-dimensional space (meters). $\theta_{rot,t}$ denotes the angular error of the rotation axis (radians). By extracting the translational drift $d_{trans,t}$ and rotational drift $\theta_{rot,t}$, and combining them with the environment score Q_t, the interpolation weight α_t is dynamically computed as:

$$\alpha_t = min\left(\alpha_{max}, k_t \frac{d_{trans,t}}{T_{trans}} + k_r \frac{\theta_{rot,t}}{T_{rot}}\right) \times Q_t \tag{7}$$

T_{trans} and T_{rot} denote drift thresholds for translation and rotation, respectively; k_t and k_r denote proportional coefficients for translational and rotational weighting, used to adjust the sensitivity to different error types; α_{max} denotes maximum interpolation weight, preventing over-smoothing; and $\alpha_t \in [0, \alpha_{max}]$ denotes dynamic interpolation weight for pose correction. The previous alignment matrix $T^{(t-1)}_{align}$ is updated toward the target pose $\Delta T_t T^{(t-1)}_{align}$ through interpolation as:

$$T^{(t)}_{align} = Interp(T^{(t-1)}_{align}, \Delta T_t T^{(t-1)}_{align}, \alpha_t) \tag{8}$$

$$\alpha_t = min\left(\frac{max(\epsilon^{trans}_t, \epsilon^{rot}_t)}{\tau}, \alpha_{max}\right) \tag{9}$$

where $\alpha_t \in [0,1]$ with τ denotes the activation threshold and α_{max} represents the maximum interpolation ratio. Here, $T^{(t-1)}_{align}$ is the alignment matrix after the previous correction. The interpolation function *Interp* adopts linear

Table I. Experimental environmental condition settings

Condition ID	Environment description	Illuminance (lux)
E_1	Standard environment	300–500
E_2	Low-illumination environment	< 50
E_3	High-intensity illumination	> 1000
E_4	Dynamic illumination	100–800
E_5	Low-texture surface	280–320

interpolation for translation and spherical linear interpolation for rotation, ensuring smoothness and continuity during the correction process. The resulting alignment matrix is denoted as $T_{align}^{(t)}$. The pipeline combines continuous tracking from visual SLAM with environment sensing and lightweight fiducial-based corrections to suppress drift, ensuring stable, high-accuracy alignment across diverse scenes. By preventing abrupt pose resets that lead to perceptual discontinuities and implementing adaptive drift detection, the proposed method significantly enhances alignment stability, particularly under varying illumination and texture conditions, thereby ensuring smoother and more reliable user interactions.

Parameter selection for the environment-adaptive SLAM module: In practice, illumination L_t, texture density D_t, and SLAM tracking confidence C_t are first normalized to a comparable numeric range so that their typical values lie in [0,1]. The environment weights are then set such that no single indicator dominates the environment-quality score Q_t, and the score varies smoothly across the five visual conditions defined in Table I. For the drift-aware interpolation, the translational and rotational drift thresholds and the proportional coefficients are chosen according to the observed distribution of drift on pilot sequences so that pose correction is activated only when the accumulated drift exceeds a perceptually noticeable level, while small fluctuations are left to the underlying SLAM system. The activation threshold in the interpolation function and the upper bound on the interpolation ratio are selected to cap the maximum correction step and to avoid abrupt changes of the alignment matrix. The concrete parameter values used in all experiments are summarized in Section IV.A.

C. USER BEHAVIOR-DRIVEN PREDICTION MODULE

To mitigate the inevitable end-to-end latency caused by the on-device rendering pipeline and internal data processing in consumer-grade MR systems, we develop a user-behavior-driven interaction prediction module that proactively estimates forthcoming interactions and decouples system response preparation from the delayed sensing–rendering loop, thereby enhancing responsiveness under real-time constraints. This module anticipates user actions based on real-time behavior, allowing the system to reduce perceived latency and improve interaction responsiveness. By combining short-horizon motion prediction with an optimized interaction-trigger strategy, the module improves temporal consistency and interaction smoothness.

The core of this module is a Kalman-filter-based short-horizon predictor that fuses the user's historical position and velocity to forecast the next few positions. This approach effectively alleviates interaction misalignment caused by rendering latency, especially during rapid movements or head turns. The Kalman filter performs linear state estimation by combining historical observations with a motion model. We use a state vector comprising position and velocity to predict future locations, and we tailor the filter to characteristic user-motion patterns to better adapt to the dynamics of MR interactions. The state transition is modeled linearly over this state vector, without presenting explicit equations in the text:

$$x_k = F_{k-1}x_{k-1} + B_{k-1}u_{k-1} + w_{k-1} \tag{10}$$

where x_k is the state vector at time k, F_{k-1} is the state-transition matrix, B_{k-1} is the control-input matrix with control input u_{k-1}, and w_{k-1} denotes the process noise.

In interactive scenarios, user operations often involve the parallel processing of multiple tasks. To enhance the overall interaction experience, this study proposes a semantic priority scheduling mechanism that dynamically adjusts the execution order of tasks based on their urgency and the significance of the interaction. By prioritizing tasks according to their relevance to the user's immediate needs, this approach ensures more efficient and context-aware interaction, improving the system's responsiveness and user satisfaction. Specifically, when the user performs multitask operations, the system prioritizes tasks closely related to the user's current actions, ensuring that critical interactions are promptly responded to. For instance, when the user rapidly turns their head, the system prioritizes gaze-related interaction tasks, while when the user performs hand gestures, the system prioritizes content related to hand movements. This priority scheduling significantly improves the response speed of interaction tasks and reduces the impact of delays on the user experience. The semantic priority scheduling is calculated using the following formula:

$$Priority(t) = \sum_{i=1}^{n} w_i \cdot Importance_i(t) \tag{11}$$

where $Priority(t)$ denotes the task priority at time t, $Importance_i(t)$ denotes the importance of task i at time t, and w_i is the weight of task i. Task priorities are dynamically adjusted according to the user's interaction context and real-time requirements. By continually retuning the process- and measurement-noise covariance matrices in the Kalman filter, the system adapts to different user-motion patterns and interaction needs. This dynamic adjustment improves prediction accuracy in complex environments and reduces interaction issues caused by latency.

In our rotating machinery fault diagnosis testbed, multiple monitoring streams coexist on the same equipment, including vibration signals from different bearings, bearing-housing temperatures, rotational speed, and derived health indicators such as RMS vibration and kurtosis. Within the proposed SEMR framework, the semantic priority scheduling module assigns the highest priority to safety-critical and fault-related information, such as over-limit vibration alarms and over-temperature warnings for the bearing under inspection. Medium priority is allocated to diagnostic trend indicators, for example, a steadily increasing RMS vibration level suggesting incipient degradation, while routine operating parameters on non-critical components are treated as low-priority background data. During rapid viewpoint transitions or when multiple alarms and trend changes occur within a short time window, the scheduler first guarantees timely updates and stable

rendering of high-priority alarm labels and their associated measurement points in the user's field of view. Medium-priority trend visualizations are refreshed at a slightly relaxed rate, whereas low-priority elements are down-sampled or deferred until critical information is updated. This mechanism keeps safety-relevant and fault-localization-critical information visually salient and up to date under limited headset resources without overloading the user.

D. GAZE INTERACTION OPTIMIZATION MODULE

To improve the stability of interaction on consumer-grade MR headsets in challenging environments, such as those with complex backgrounds, non-uniform illumination, and sparse textures, we propose an optimization module for gaze interaction, designed to address these specific conditions and enhance overall system performance. The module employs adaptive tuning and multimodal data fusion to effectively suppress false activations triggered by subtle head micromotions or environmental disturbances. By intelligently adjusting to dynamic user and environmental conditions, the module enhances the quality of interaction between the user and the virtual environment, ensuring more accurate and reliable responses.

To enhance the stability of interactions on consumer-grade MR headsets, particularly in environments characterized by complex backgrounds, non-uniform lighting, and sparse textures, this paper presents a gaze interaction optimization module designed to address these challenges and improve overall system performance in such conditions. This module effectively reduces false triggers caused by slight head movements or environmental factors and optimizes the interaction experience between the user and the virtual environment through adaptive adjustment and multimodal data fusion strategies. In response to varying interference levels in different scenarios, the module dynamically adjusts the gaze confirmation threshold T_{fix} and the confirmation window ΔT_{conf}^{min} and ΔT_{conf}^{max}, maintaining a low false trigger rate while minimizing user wait time:

$$T_{fix} = T_{fix}^{0} + \lambda \cdot \sigma_{gaze} \tag{12}$$

where T_{fix}^{0} denotes the baseline fixed threshold, typically set to 0.8 s, and σ_{gaze} denotes the variance of gaze drift, calculated within the temporal window to evaluate fixation stability. λ denotes a scaling factor empirically determined to ensure that unstable gaze conditions lead to longer fixation confirmation times, thereby reducing false activations. In practice, T_{fix}^{0} is chosen based on preliminary tests under the standard condition to provide a comfortable confirmation duration for most users and is kept fixed across all visual environments, while λ is adjusted so that larger σ_{gaze} leads to a noticeably longer effective confirmation time; the final values of T_{fix}^{0} and λ used in the experiments are summarized in Section IV.A. For angular tolerance, a gaze drift of up to 3° is allowed while still considering it as gazing at the same target. This value is based on the typical gaze tolerance angle used in eye-tracking devices and was calibrated in the multi-user experiments of this study to balance switch sensitivity and false trigger rates. When a gaze ray cast hit is detected, the final interaction is triggered only if, within a brief verification window, a micro-blink or subtle gesture is observed. This mechanism helps to significantly reduce false triggers by ensuring that the interaction is confirmed through an additional layer of validation.

Confidence from multiple modalities is then combined using a weighted fusion scheme, with modality-specific weights balancing reliability and responsiveness [37].

$$C_{final} = w_1 C_{gaze} + w_2 C_{blink} + w_3 C_{hand} \tag{13}$$

where C_{gaze}, C_{blink}, and C_{hand} denote the detection confidence of gaze, blink, and gesture inputs, respectively, and wi are adaptive weights determined by the environmental context. To prevent repeated triggering when the user briefly moves away and returns their gaze, the module introduces a short-term history cache. Once the threshold for interrupting the confirmation process is reached, the process becomes non-recoverable, effectively preventing repeated command executions that could otherwise be triggered by minor head movements or gaze re-scanning. This mechanism ensures that only intentional interactions are recognized, enhancing the accuracy and reliability of the system. This approach significantly enhances the smoothness of interaction. The module continuously monitors the gaze hit status using coroutines in Unity, enabling real-time tracking of the user's gaze to ensure precise interaction. Once the predefined conditions for gaze-based interaction are met, the system activates the corresponding interaction mechanisms, ensuring seamless and responsive user engagement. The system activates multimodal listening, enabling it to respond seamlessly and promptly, thereby ensuring a fluid and responsive user experience.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP AND EVALUATION METHODS

To validate the effectiveness of the SEMR framework in industrial monitoring and diagnostic tasks, a rotating machinery fault diagnosis testbed was developed, and a 3D visualization model was constructed in Unity, which aligns geometrically with the physical structure and key components. This model is used for in situ overlay display and interaction validation in the MR scenario. The testbed simulates typical industrial rotating equipment for inspection and diagnostic workflows, as shown in Fig. 3. It provides a reproducible object for observation under stable operating conditions and generates typical disturbances, such as mechanical vibrations and metallic reflections, during the experiment. These features serve as the basis for evaluating the spatial registration stability and real-time interaction performance of the MR system.

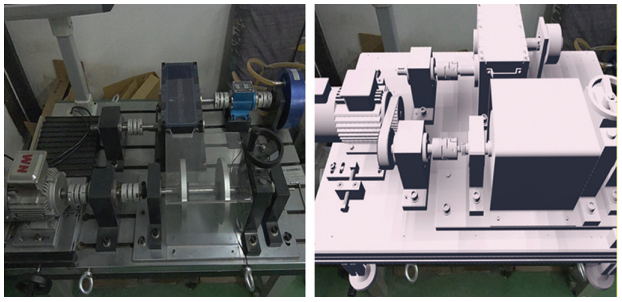

Fig. 3. Rotating machinery fault diagnosis testbed and its 3D visualization model.

To evaluate the algorithm's adaptability to visually degraded industrial conditions, five representative lighting and texture scenarios $E_1 - E_5$ were designed, corresponding to standard laboratory illumination, low illumination, strong specular reflections, dynamic illumination variations, and low-texture surfaces, respectively. In this study, the term "baseline" denotes a reference configuration in which the headset operates with its native tracking and interaction pipeline, and none of the proposed SEMR components or optimizations are enabled. Specifically, the baseline condition uses only the standard SLAM algorithm integrated into the Meta Quest 3 headset to provide pose estimation and scene tracking, and it adopts a conventional gaze-based ray-casting interface in which user selections are triggered by a fixed dwell time. To ensure clarity and reproducibility, Table I systematically summarizes all parameter settings and experimental conditions, enabling a direct comparison between the two systems.

Illuminance was measured using a lux meter at approximately 0.5 m from the equipment surface, and the illuminance range during each task was recorded. Twenty-four participants aged 23–31 years were recruited via public advertisement. A within-subject design was adopted, in which each participant completed the same set of tasks under two system configurations. Each participant performed three standardized tasks, namely stationary spot observation, slow inspection, and rapid fault troubleshooting, once under each of the five environmental conditions and for both system configurations, resulting in 30 experimental trials per participant. To mitigate learning and fatigue effects, the order of system configurations was counterbalanced across participants using a Latin-square scheme, and the task order was randomized. Prior to the formal experiment, participants completed device familiarization and eye-tracking calibration, followed by a brief practice session to learn the operation procedure; prior MR or VR experience was also recorded.

Each participant performed three standardized tasks around the testbed: stationary spot observation, slow inspection, and rapid fault troubleshooting. Each task was conducted under the five environmental conditions $E_1 - E_5$. During the experiment, the system recorded data at 60 Hz and logged the start or end time of each trial as well as invalid segments. To quantify the benefits of SEMR, two system configurations were evaluated. The baseline configuration relied on the headset's built-in tracking and used gaze ray-casting interaction with a fixed dwell time, without enabling SEMR. In contrast, the enhanced configuration fully enabled the SEMR framework, incorporating its adaptive SLAM optimization and motion prediction features.

To ensure the reproducibility of the proposed SEMR framework, the key parameters in the algorithm were determined before the formal experiment in a pre-experimental calibration phase involving five users under standard laboratory lighting conditions. Illumination L_t, texture density D_t, and SLAM tracking confidence C_t were first normalized to the range [0, 1]. The environment weights were then set to $w_L = 0.3$, $w_D = 0.3$, and $w_C = 0.4$, assigning a slightly higher weight to SLAM confidence so that the internal tracking state is prioritized, while illumination and texture serve as complementary indicators. For the drift-aware pose correction, the translational threshold was set to $T_{trans} = 0.15$ m and the rotational threshold to $T_{rot} = 0.2$ rad, corresponding to the upper bound of visually acceptable misalignment before correction is triggered. The proportional coefficients were set to $k_t = 0.8$ and $k_r = 0.8$ so that the interpolation between SLAM and PnP poses remains smooth. The activation threshold in the interpolation function was set to $\tau = 0.1$, and the interpolation ratio was limited by $\alpha_{max} = 0.5$ to avoid abrupt pose jumps. In the gaze interaction module, the baseline dwell-time threshold was set to $T^0_{fix} = 0.8$s, which provided a comfortable confirmation duration in pilot trials, and the scaling factor was set to $\lambda = 0.5$ so that unstable gaze patterns with larger variance σ_{gaze} lead to noticeably longer effective confirmation times, while stable gaze still allows fast triggering. Once determined, these parameter values were kept constant for all participants and all environmental conditions in the subsequent experiments.

B. SYSTEM PERFORMANCE VALIDATION AND RESULT ANALYSIS

Spatial registration stability analysis: Under the standard condition E_1, 24 participants performed a 60 s continuous inspection task. The temporal evolution of the spatial registration error (SRE), defined as the Euclidean distance between ground-truth and estimated positions in centimeters, is shown in Fig. 4. The solid lines represent the group-mean trajectories, the error bars indicate the inter-subject standard deviation (SD), and the shaded areas denote the 95% confidence intervals of the mean. As observed, the baseline system exhibits error accumulation and noticeable drift over time during inspection, whereas SEMR substantially suppresses long-term drift via the environment-adaptive correction mechanism, maintaining the error at a consistently lower level.

To further quantify the differences, Table II summarizes the SRE statistics under the two system configurations. The baseline system yields a mean SRE of 14.2 cm, while the SEMR framework reduces the mean SRE to 6.8 cm. This corresponds to a significant reduction in average error of approximately 52.1%, demonstrating the effectiveness of the proposed system in improving spatial registration accuracy. Moreover, the maximum drift decreases from 18.6 cm to 8.2 cm, indicating that SEMR substantially improves the stability and consistency of spatial registration under the standard condition.

Furthermore, to evaluate the algorithm's adaptability under visually degraded conditions, additional comparative tests were conducted across the five environmental scenarios defined in Table I. Fig. 5 reports the SRE for the baseline

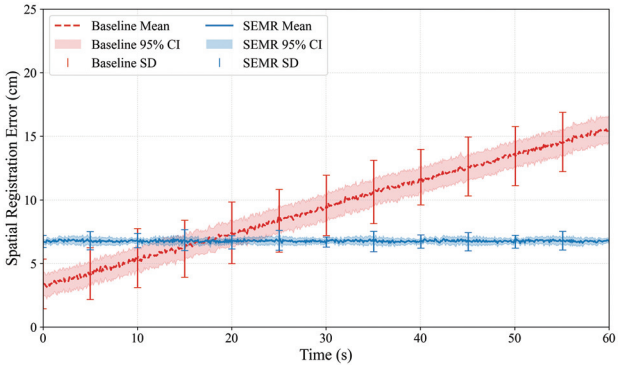

Fig. 4. Temporal trajectories of SRE for the baseline and SEMR configurations under the standard condition.

Table II. Comparison of SRE for baseline and SEMR

System configuration	Mean error (cm)	SD (cm)	Median (cm)	95% CI (cm)	Max drift (cm)
Baseline	14.2	4.5	13.8	[12.5,15.9]	18.6
SEMR (ours)	6.8	1.2	6.5	[6.1,7.5]	8.2

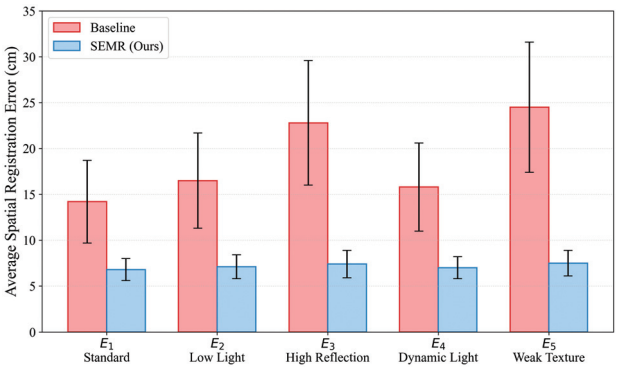

Fig. 5. SRE of the baseline and SEMR systems across the five environmental conditions defined in Table I.

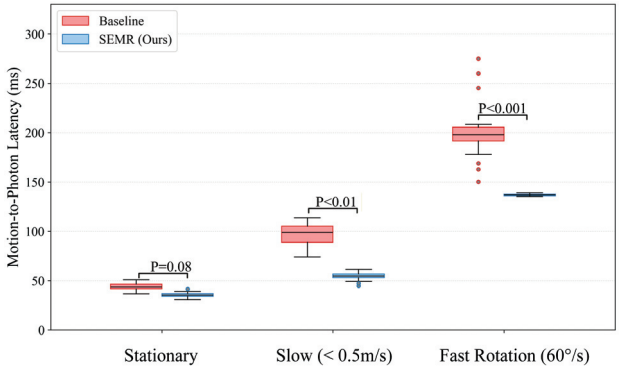

Fig. 6. End-to-end latency of the baseline and SEMR systems under three motion conditions.

and SEMR configurations under each condition. The baseline exhibits more pronounced error accumulation in scenarios such as low illumination and strong specular reflections; in some trials, the maximum instantaneous drift exceeds 25 cm, leading to substantial misalignment between virtual labels and physical components. In contrast, SEMR achieves more stable spatial registration across all scenarios by utilizing environment-quality awareness, which dynamically adjusts the system's parameters based on environmental conditions, and by incorporating lightweight reference-assisted stabilization. This dual approach ensures consistent and reliable tracking performance by dynamically adjusting to varying environmental conditions, even in complex and unpredictable scenarios. By leveraging both environment-quality awareness and reference-assisted stabilization, the system maintains stable performance, effectively addressing challenges posed by dynamic and diverse environments. These results suggest that SEMR is more robust to typical industrial visual degradations, including illumination changes, specular-reflection interference, and texture scarcity, thereby providing a reliable spatial foundation for subsequent in situ visualization and interaction.

To evaluate real-time responsiveness under different operator motion intensities, Fig. 6 presents box plots of end-to-end latency for 24 participants under three motion conditions: stationary observation (no movement), slow inspection (walking speed <0.5 m/s), and rapid viewpoint switching (head angular velocity >60°/s). A within-subject paired design was used to compare latency between the baseline and SEMR for each participant under the same motion condition. Given that the latency distribution may deviate from normality, the Wilcoxon signed-rank test was employed, and Holm–Bonferroni correction was applied for multiple comparisons across the three conditions [38], with the significance level set to $\alpha = 0.05$. The results show no significant difference under the stationary condition ($p = 0.08$). When participants started moving, the baseline exhibited increased latency dispersion, whereas SEMR maintained lower and more stable latency through motion prediction and scheduling, yielding a significant difference under the slow condition ($p < 0.01$). Under rapid viewpoint switching, the baseline showed a higher median latency and a broader distribution, while SEMR markedly compressed the latency distribution and reduced outliers; the difference was highly significant ($p < 0.001$). These findings suggest that the proposed prediction mechanism effectively compensates for rendering lag in highly dynamic scenarios, thereby mitigating delays that could otherwise disrupt the user experience. Furthermore, it enhances both the timeliness and consistency of interactions, ensuring smoother and more reliable real-time performance in dynamic environments.

To quantitatively assess the system's subjective performance in terms of ergonomics, a questionnaire evaluation was conducted using a five-point Likert scale and the System Usability Scale (SUS) [39]. All 24 participants who took part in the experimental tasks also completed the subjective questionnaire once at the end of the session. To ensure the clarity, consistency, and reproducibility of the evaluation dimensions, five key performance metrics were carefully defined and tailored for the industrial inspection scenario. These metrics were designed to capture the most relevant aspects of system performance, providing a comprehensive basis for assessing the effectiveness of the proposed approach in real-world applications. The results were then summarized by calculating the mean scores across the 24 participants, providing a comprehensive overview of the system's performance based on subjective user feedback. Fig. 7 presents a radar chart comparing the two system configurations across these dimensions.

The results show that SEMR achieves higher scores in dimensions such as visual stability and robustness, indicating a more stable in situ visualization experience under illumination variations and specular-reflection interference. However, the baseline attains a slightly higher *usability* score (4.2) than SEMR (3.8). Interview feedback suggests that the native bare-hand interaction is more intuitive and

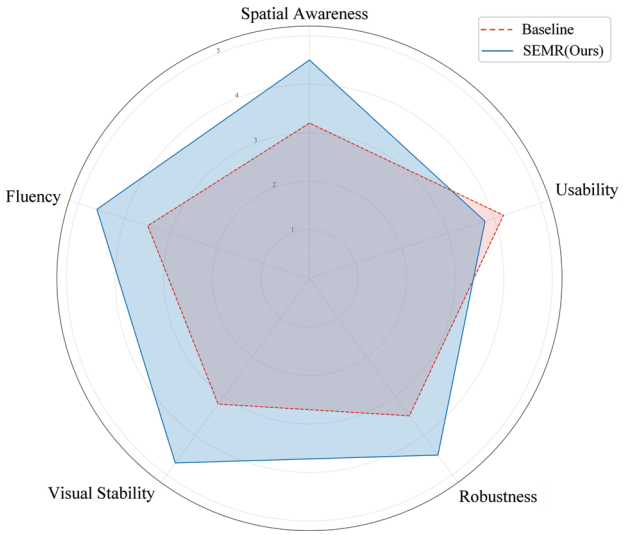

Fig. 7. Radar chart of subjective user experience scores for the baseline and SEMR systems across five evaluation dimensions.

requires less learning, whereas SEMR's "gaze confirmation + blink triggering" mechanism improves the reliability of hands-free operation but introduces a modest learning curve for first-time users. Overall, SEMR maintains good usability while substantially enhancing stability and safe interaction in challenging conditions, making it better suited for on-site industrial inspection scenarios.

V. CONCLUSIONS

This study targets two practical bottlenecks in in situ industrial condition monitoring and fault diagnosis: pose drift induced by visually degraded environments and latency-related temporal misalignment in the visualization of high-rate monitoring data during dynamic inspection. To address these issues, we propose SEMR, a single-device, environment-adaptive mixed-reality enhancement framework. SEMR suppresses long-term drift via environment-quality-aware SLAM adaptation coupled with lightweight reference-assisted stabilization, reduces interaction latency through Kalman-filter-based short-term motion prediction and rendering scheduling, and improves operational reliability in hands-busy scenarios by incorporating a fault-tolerant hands-free gaze interaction mechanism. Experimental validation on a rotating machinery fault diagnosis testbed shows that SEMR reduces the SRE to 6.8 cm under the standard condition and decreases end-to-end latency to 137 ms under fast viewpoint changes. These results indicate that the proposed approach improves both spatial consistency between monitoring information and physical components and the stability of real-time in situ visualization, providing a low-cost and reliable solution for enhanced industrial inspection. Nevertheless, the current evaluation is primarily conducted on a controlled testbed with representative visual-degradation conditions and does not yet cover more complex real production factors. Future work will extend validation to real factory deployments and explore lightweight integration of intelligent diagnostic capabilities on MR edge devices to further improve inspection efficiency and strengthen the "monitor–diagnose–act" closed-loop support.

ACKNOWLEDGMENTS

This work is supported in part by the National Key Research and Development Program of China (2024YFB3310702), in part by the National Natural Science Foundation of China (52505091), and in part by the China Postdoctoral Science Foundation.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

[1] C. Hu *et al.*, "A state of the art in digital twin for intelligent fault diagnosis," *Adv. Eng. Inf.*, vol. 63, p. 102963, 2025.
[2] X. Li *et al.*, "The bearing multi-sensor fault diagnosis method based on a multi-branch parallel perception network and feature fusion strategy," *Reliab. Eng. Syst. Saf.*, vol. 261, p. 111122, 2025.
[3] X. Li *et al.*, "A fault diagnosis method for bearings based on multiscale graph convolutional network under non-stationary speed conditions," *Mech. Mach. Theory*, vol. 217, p. 106267, 2025.
[4] N. Li *et al.*, "Six-dimensional digital twin modeling and software platform design for complex industrial systems," *J. Intell. Manuf.*, pp. 1–22, 2025.
[5] X. Li *et al.*, "A fault diagnosis data augmentation method integrating multimodal non-Gaussian denoising diffusion generative adversarial network," *Adv. Eng. Inf.*, vol. 68, p. 103776, 2025.
[6] S. Zhi *et al.*, "An unsupervised transfer learning bearing fault diagnosis method based on multi-channel calibrated Transformer with shiftable window," *Struct. Health Monit.*, p. 14759217251324671, 2025.
[7] X. Hu *et al.*, "A novel integrated fault diagnosis method based on digital twins," *Signals*, vol. 6, no. 2, p. 18, 2025.
[8] U. Asad *et al.*, "Human-centric digital twins in industry: A comprehensive review of enabling technologies and implementation strategies," *Sensors*, vol. 23, no. 8, p. 3938, 2023.
[9] J. Qi *et al.*, "Attention-guided graph isomorphism learning: A multi-task framework for fault diagnosis and remaining useful life prediction," *Reliab. Eng. Syst. Saf.*, p. 111209, 2025.
[10] H. Minghui *et al.*, "Digital twin model of gas turbine and its application in warning of performance fault," *Chin. J. Aeronaut.*, vol. 36, no. 3, pp. 449–470, 2023.
[11] Y. Xiao, H. Shao, and B. Liu, "Evaluating calibration of deep fault diagnostic models under distribution shift," *Comput. Ind.*, vol. 171, p. 104334, 2025.
[12] R. Huang *et al.*, "Compound fault diagnosis for rotating machinery: State-of-the-art, challenges, and opportunities," *J. Dyn. Monit. Diagn.*, pp. 13–29, 2023.
[13] S. Zhi *et al.*, "Local entropy selection scaling-extracting Chirplet transform for enhanced time-frequency analysis and precise state estimation in reliability-focused fault diagnosis of non-stationary signals," *Eksploatacja i Niezawodność–Maintenance Reliab.*, vol. 28, no. 1, 2025.
[14] Y. Hou *et al.*, "Cognitive load classification of mixed reality human computer interaction tasks based on multimodal sensor signals," *Sci. Rep.*, vol. 15, no. 1, p. 13732, 2025.

[15] X. Chen *et al.*, "Large models for machine monitoring and fault diagnostics: Opportunities, challenges, and future direction," *J. Dyn. Monit. Diagn.*, vol. 4, no. 2, pp. 76–90, 2025.
[16] S. Wang *et al.*, "Bearing prognostics and health management based on hybrid physical mechanism and data models: A systematic review," *Meas. Sci. Technol.*, vol. 36, no. 5, p. 052002, 2025.
[17] X. Li *et al.*, "Knowledge extraction and retrieval-augmented generation for intelligent maintenance of wind power equipment based on graph attention networks," *Chin. J. Mech. Eng.*, p. 100141, 2025.
[18] Q. Li *et al.*, "Transparent operator network: A fully interpretable network incorporating learnable wavelet operator for intelligent fault diagnosis," *IEEE Trans. Ind. Inf.*, vol. 20, no. 6, pp. 8628–8638, 2024.
[19] X. Chen *et al.*, "Physics-informed deep neural network for bearing prognosis with multisensory signals," *J. Dyn. Monit. Diagn.*, pp. 200–207, 2022.
[20] J. Zhuang *et al.*, "A neuro-fuzzy approach with hypergraph convolution for fault diagnosis in industrial devices," *J. Reliab. Sci. Eng.*, vol. 1, no. 3, p. 035301, 2025.
[21] L. Ge *et al.*, "High-dimensional optimized extraction chirplet transform: Algorithm and applications," *ISA Trans.*, 2026.
[22] H. Lin *et al.*, "Matching pursuit network: An interpretable sparse time–frequency representation method toward mechanical fault diagnosis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 7, p. 12377–12388, 2024.
[23] Y. Peng *et al.*, "Dual-stage interpretable domain generalization fault diagnosis: Integrating prior knowledge and gradient-weighted class activation mapping," *Eng. Appl. Artif. Intell.*, vol. 166, p. 113655, 2026.
[24] H. Shao *et al.*, "Small sample gearbox fault diagnosis based on improved deep forest in noisy environments," *Nondestruct. Test. Eval.*, vol. 40, no. 8, pp. 3935–3956, 2025.
[25] S. H. Lee *et al.*, "Fully portable continuous real-time auscultation with a soft wearable stethoscope designed for automated disease diagnosis," *Sci. Adv.*, vol. 8, no. 21, p. eabo5867, 2022.
[26] W. Jia, W. Wang, and Z. Zhang, "From simple digital twin to complex digital twin Part I: A novel modeling method for multi-scale and multi-scenario digital twin," *Adv. Eng. Inf.*, vol. 53, p. 101706, 2022.
[27] H. Wang *et al.*, "A survey on the metaverse: The state-of-the-art, technologies, applications, and challenges," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14671–14688, 2023.
[28] A. Sadhu *et al.*, "A review of data management and visualization techniques for structural health monitoring using BIM and virtual or augmented reality," *J. Struct. Eng.*, vol. 149, no. 1, p. 03122006, 2023.
[29] I. A. Kazerouni *et al.*, "A survey of state-of-the-art on visual SLAM," *Expert Syst. Appl.*, vol. 205, p. 117734, 2022.
[30] J. Mertes *et al.*, "Evaluation of 5G-capable framework for highly mobile, scalable human-machine interfaces in cyber-physical production systems," *J. Manuf. Syst.*, vol. 64, pp. 578–593, 2022.
[31] L. Qin *et al.*, "RSO-SLAM: A robust semantic visual SLAM with optical flow in complex dynamic environments," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 14669–14684, 2024.
[32] C. Li *et al.*, "Unleashing mixed-reality capability in deep reinforcement learning-based robot motion generation towards safe human–robot collaboration," *J. Manuf. Syst.*, vol. 74, pp. 411–421, 2024.
[33] C. Menezes, B. França, and Y. Lopes, "A survey on extended reality, digital twins, and metaverse applications in power systems," *IEEE Internet Things J.*, vol. 11, no. 21, pp. 34953–34977, 2024.
[34] Z. Guo *et al.*, "A survey on applications of large language model-driven digital twins for intelligent network optimization," *IEEE Commun. Surv. Tutor.*, 2025.
[35] T. Park, S. Mondal, and W. Cai, "Interfacing nanophotonics with deep neural networks: AI for photonic design and photonic implementation of AI," *Laser Photonics Rev.*, vol. 19, no. 8, p. 2401520, 2025.
[36] R. Han *et al.*, "Neupan: Direct point robot navigation with end-to-end model-based learning," *IEEE Trans. Rob.*, 2025.
[37] L. Liu *et al.*, "Cross-modal object tracking via modality-aware fusion network and a large-scale dataset," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 4, pp. 6981–6994, 2024.
[38] H. Rezk, M. Aly, and A. Fathy, "A novel strategy based on recent equilibrium optimizer to enhance the performance of PEM fuel cell system through optimized fuzzy logic MPPT," *Energy*, vol. 234, p. 121267, 2021.
[39] P. Vlachogianni and N. Tselios, "Perceived usability evaluation of educational technology using the system usability scale (SUS): A systematic review," *J. Res Technol. Educ.*, vol. 54, no. 3, pp. 392–409, 2022.