International Journal of Latest Research in Engineering and Management (IJLREM) *ISSN: 2456-0766*
www.ijlrem.org || Volume 10 - Issue 06 || November-December 2026 ǁ PP. 01-06

Gaze-Guided Wearable AR, Multimodal AI, and Digital Twins for Smart Factory Inspection

Yu-Ching Lin
National Center for High-Performance Computing, National Applied Research Laboratories, Taiwan

Abstract: Routine equipment inspection on factory floors demands hands-free human-machine interfaces that allow technicians to access machinery diagnostics and technical documentation safely while keeping both hands free. This paper presents a practical four-tier system architecture that integrates wearable optical see-through augmented reality (AR) smart glasses, multimodal artificial intelligence (AI), an Advanced Message Queuing Protocol (AMQP) broker, and a web-based 3D digital twin for smart factory equipment monitoring. On the wearable client, binocular eye-gaze tracking is combined with an inverse center-crop coordinate mapping and a 3.0 s fixed-anchor dwell trigger to crop 512 × 512 regions of interest (ROIs) accurately from 1600 × 1200 camera frames. On the server tier, a dual-mode vision-language model (VLM) distinguishes industrial machinery from technical text and links recognized equipment identifiers to a canonical asset catalog to provide reliable reference telemetry alongside document translation. In parallel, a streaming 16 kHz voice pipeline combines neural voice activity detection (VAD), domain-prompted speech recognition, a 600 ms half-duplex echo guard, and a six-tool whitelist agent. Finally, a WebGL/Three.js 3D digital twin visualizes real-time equipment states, peer-to-peer inspection sharing, and cumulative gaze attention. Prototype verification on a physical pump skid testbed confirms reliable end-to-end operation.

Keywords: Augmented Reality (AR), Eye Tracking, Multimodal AI, Vision-Language Models (VLMs), Speech Recognition, Digital Twin, Equipment Monitoring.

I. Introduction

Modern manufacturing and process facilities rely on continuous equipment monitoring and routine on-site inspections to maintain operational safety across rotating machinery, control cabinets, and piping networks [1-3]. Although industrial Internet of Things (IIoT) sensors provide backend telemetry, field technicians must still physically inspect pump skids, verify valve alignments, and consult maintenance manuals in constrained workspaces. Because carrying handheld tablets or laptops interferes with tool handling and safety compliance, optical see-through augmented reality (AR) smart glasses equipped with binocular eye tracking, a scene camera, and open-ear audio provide an appealing hands-free interface for shopfloor operators [4, 5].

Recent advances in vision-language models (VLMs) [6], neural automatic speech recognition (ASR) [7], and web-based 3D digital twins [8, 9] offer powerful building blocks for intelligent maintenance assistants. Nevertheless, combining these components into a cohesive real-time system requires addressing four practical engineering issues: (1) aspect-ratio coordinate offset when mapping normalized eye gaze from a 16:9 AR display rendered with isotropic center-cropping (CENTER_CROP) to a 4:3 camera frame; (2) VLM task ambiguity between physical equipment recognition and optical character recognition (OCR), as well as unverified numeric guesses from blurred gauge images; (3) ambient machinery noise and acoustic self-echo when open-ear text-to-speech (TTS) audio re-enters the headset microphone; and (4) asynchronous event routing and state synchronization across multiple AR headsets and supervisor dashboards.

To provide a clean architectural blueprint for addressing these integration issues, this paper presents a four-tier wearable AR, multimodal AI, and 3D digital twin system architecture for smart factory equipment monitoring. Rather than focusing on isolated algorithmic benchmarks, this work details the end-to-end system design, hardware-software interfaces, and practical coordination mechanisms across four core modules: (1) gaze-guided ROI extraction via inverse CENTER_CROP coordinate mapping and a 3.0 s fixed-anchor dwell gate; (2) dual-mode VLM inference coupled with deterministic asset catalog lookup; (3) AMQP-streamed 16 kHz PCM voice interaction protected by neural VAD, a 600 ms half-duplex echo guard, and a whitelist-constrained tool agent; and (4) directed peer-to-peer collaboration and 3D digital twin visualization.

II. Related Technologies

2.1 Wearable Augmented Reality and Eye-Gaze Interaction

Industrial AR interfaces have progressed from paper checklists and fiducial markers [10] to natural eye-gaze interaction [4, 11]. While fiducial markers provide spatial registration, they are vulnerable to grease, poor lighting, and occlusion in industrial plants. Eye-gaze dwell selection enables controller-free triggering by measuring fixation duration on a target region [11]. In wearable smart glasses, however, the display viewport and the world-facing camera sensor frequently operate at different aspect ratios, requiring explicit geometric compensation before cropping visual ROIs for downstream AI models.

2.2 Multimodal Vision-Language and Speech Processing

Large pre-trained VLMs unify visual recognition and text reading within a single model [6, 12]. In industrial maintenance, however, unconstrained generative responses can hallucinate numeric sensor readings or misclassify equipment labels as translation tasks. Grounding model outputs against a structured asset catalog ensures that operators receive verified engineering specifications. Similarly, for hands-free voice control in noisy environments, combining neural voice activity detection (VAD) with domain-prompted speech recognition [7] and constrained tool invocation [13] provides higher reliability than open-ended conversational agents.

2.3 Web-Based 3D Digital Twins and Message Brokers

A digital twin mirrors the state of physical equipment in a virtual 3D workspace to support remote supervision [1, 8, 9]. By rendering 3D CAD/glTF equipment assemblies in standard web browsers via WebGL/Three.js and routing device telemetry through an Advanced Message Queuing Protocol (AMQP) broker [14, 15], inspection events captured by frontline AR headsets can be synchronized with remote engineering consoles in real time.

III. System Architecture and Technical Design

3.1 Four-Tier Edge-Cloud System Overview

As shown in Fig. 1, the proposed monitoring system is organized into four decoupled tiers: (1) the Wearable AR Client Tier, running on binocular optical see-through smart glasses equipped with a 60 Hz eye tracker, a 1600 × 1200 RGB camera, a 16 kHz microphone, and an open-ear speaker; (2) the AMQP 0-9-1 Messaging Tier, which manages upstream event queues ('ar.events') and per-device downstream command queues ('ar.commands'); (3) the Central Orchestration and Dual-Path AI Tier, hosting device session management, a GPU streaming speech service, a dual-mode VLM inference engine, and a six-tool whitelist agent; and (4) the 3D Digital Twin Web Tier, which renders an interactive Three.js pump skid model and real-time inspection status.

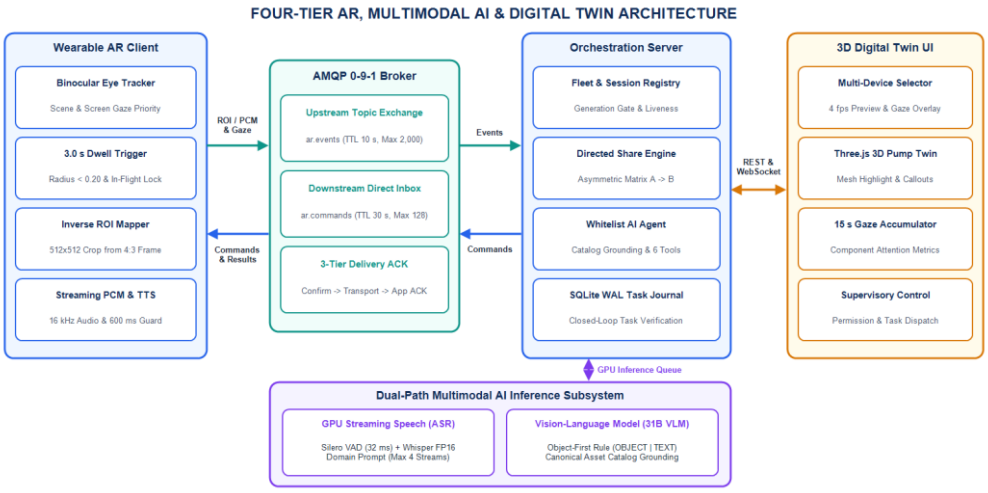


Fig. 1: four-tier system architecture integrating wearable AR smart glasses, AMQP 0-9-1 broker, dual-path AI inference, and 3D digital twin.

During operation, the wearable AR client streams gaze-triggered image crops and 16 kHz audio chunks to the central server over the AMQP broker. The orchestration server maintains device state in an SQLite Write-Ahead Logging (WAL) registry, dispatches visual and audio workloads to the AI engines, and broadcasts recognized component statuses to the 3D digital twin dashboard and authorized peer headsets.

3.2 Gaze-Guided ROI Cropping and Dwell Trigger

The wearable AR client prioritizes parallax-compensated scene-camera gaze $g_{\text{scene}}(t)$ when binocular tracking is valid, and falls back to display-plane gaze $g_{\text{screen}}(t)$ otherwise. To map normalized gaze coordinates $(g_x, g_y) \in [-1, 1]^2$ from the 16:9 AR display viewport $(V_w, V_h) = (1920, 1080)$ to the uncropped 4:3 camera frame $(I_w, I_h) = (1600, 1200)$, the client compensates for the isotropic CENTER_CROP scale factor $s = \max(V_w / I_w, V_h / I_h)$ and bilateral crop offsets $o_x = (V_w - I_w \cdot s)/2$ and $o_y = (V_h - I_h \cdot s)/2$. Given viewport pixel coordinates $v_x = (g_x + 1)V_w/2$ and $v_y = (g_y + 1)V_h/2$, the corresponding camera pixel center (i_x, i_y) and clamped L × L crop origin (x_0, y_0) for L = 512 are obtained as:

$$i_x = \frac{v_x - o_x}{s}, \quad i_y = \frac{v_y - o_y}{s},$$
$$x_0 = \text{clamp}\left(i_x - \frac{L}{2}, 0, I_w - L\right), \quad y_0 = \text{clamp}\left(i_y - \frac{L}{2}, 0, I_h - L\right) \tag{1}$$

To avoid accidental triggers during visual scanning, a six-stage dwell state machine (Fig. 2) latches an initial fixation anchor g_{anchor} at time t_0 and requires the gaze vector g_t to remain within radius $\tau_d = 0.20$ continuously for $\tau_t = 3.0$ s while no prior analysis is in flight ($L_{\text{flight}} = 0$):

$$d(g_t, g_{\text{anchor}}) = ||g_t - g_{\text{anchor}}|| < \tau_d, \quad T_{\text{dwell}} \geq \tau_t = 3.0 \text{ s} \tag{2}$$

During the first 1.0 s of fixation, the heads-up display (HUD) remains unobstructed; a countdown ring appears only between 1.0 s and 3.0 s before capturing and uploading the 512 × 512 ROI image I_{ROI}.

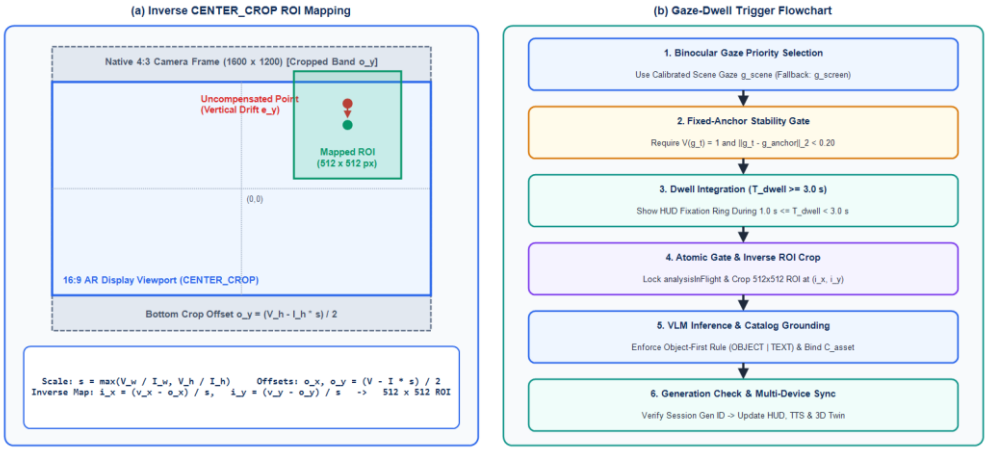


Fig. 2: gaze-guided ROI extraction: (a) inverse CENTER_CROP coordinate mapping; (b) six-stage spatiotemporal dwell trigger flowchart.

3.3 Dual-Mode VLM Recognition and Asset Catalog Lookup

Once the 512 × 512 ROI image I_{ROI} reaches the server, it is processed by a dual-mode VLM service (supporting a local 31B BF16 model or cloud API fallback). Using a structured pipe-delimited system prompt, the model operates in two distinct modes: (1) OBJECT mode (OBJECT|<id>|<name>|<use>) for identifying physical industrial machinery, and (2) TEXT mode (TEXT|<source>|<translation>) for performing OCR and translating technical manuals or warning labels into Traditional Chinese. To prevent small equipment nameplates or panel meters from erroneously triggering document translation, an Object-First disambiguation rule prioritizes OBJECT mode whenever a physical component from the asset catalog C_{asset} = {motor, pump, panel, conveyor} is visible in the ROI.

Furthermore, to prevent the VLM from hallucinating uncalibrated numeric sensor readings (such as guessing temperature or pressure values from low-resolution gauge crops), the model is instructed to output only the component identifier and functional summary. The orchestration server then maps the recognized identifier against a deterministic Canonical Industrial Asset Catalog G_{cat} to retrieve verified reference operating parameters q_{ref} (Fig. 3) before returning the result to the AR HUD and 3D digital twin.

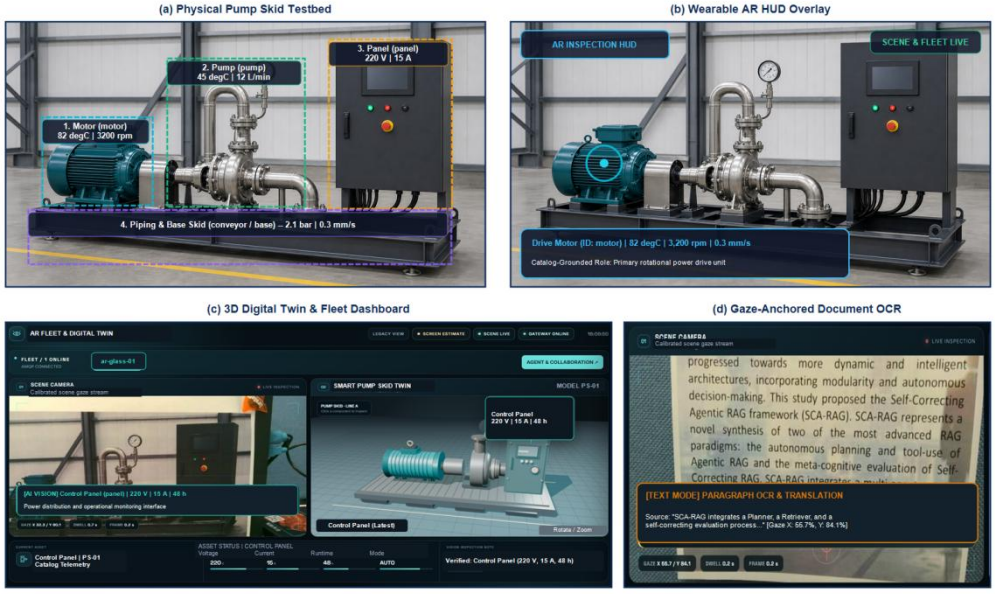


Fig. 3: physical testbed and multimodal interfaces: (a) pump skid with canonical assets; (b) wearable AR HUD; (c) 3D digital twin dashboard; (d) document OCR overlay.

3.4 Streaming Voice Command Pipeline and Echo Guard

For hands-free voice interaction under factory noise (Fig. 4), the AR client initiates a voice_start → speech_ready handshake and streams 16 kHz, 16-bit mono PCM audio in 100 ms chunks over AMQP. On the server, a per-stream Silero neural VAD evaluates 32 ms (512-sample) windows with a 256 ms pre-roll buffer and segments complete utterances when trailing silence reaches τ_{sil} = 640 ms. Segmented utterances X_{utt} are transcribed by an FP16 Whisper ASR model conditioned on a domain vocabulary prompt p_{lex} containing machinery names and operational terms.

Because the smart glasses use open-ear speakers, synthesized TTS responses ($A_{TTS}(t) = 1$) can acoustically re-enter the microphone and trigger false re-recognitions. The client therefore enforces a half-duplex microphone gate $G_{mic}(t)$ that mutes audio capture during TTS playback and maintains a post-playback guard window of τ_{guard} = 600 ms after playback ends at t_{id}:

$$G_{mic}(t) = 1 \text{ if } A_{TTS}(t) = 0 \text{ and } (t - t_{id}) > \tau_{guard};\ 0 \text{ otherwise} \quad (3)$$

Transcribed voice commands are parsed by a Whitelist-Constrained AI Agent restricted to six pre-validated operational tools: machine_info (equipment status lookup), camera (ROI capture trigger), set_mode (switching between auto, object, and translate modes), news, read_url, and answer.

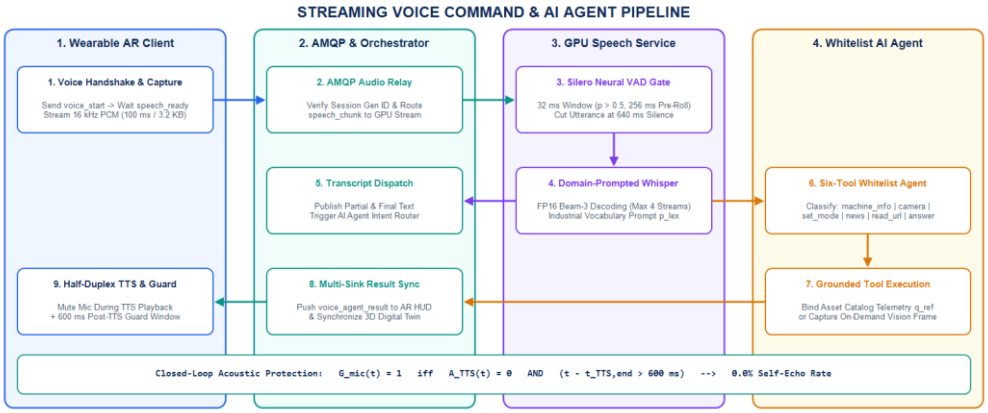


Fig. 4: streaming voice command pipeline over AMQP with Silero VAD segmentation, GPU ASR decoding, and 600 ms post-TTS guard.

3.5 AMQP Message Routing and 3D Digital Twin Visualization

To support multi-operator workflows without message backpressure, each AR headset registers a monotonic Session Generation ID $G_{active}(d)$ so stale cross-session packets are automatically discarded. Upstream telemetry uses a bounded topic exchange (10 s TTL, 2,000-message drop-head limit), while downstream commands use per-device queues (30 s TTL) with three-stage delivery confirmation. As illustrated in Fig. 5, an asymmetric permission matrix A_{ij} allows a supervisor to authorize directed peer-to-peer sharing from device i to device j, subject to a 15 s observation freshness window and a 5 s per-component rate limit.

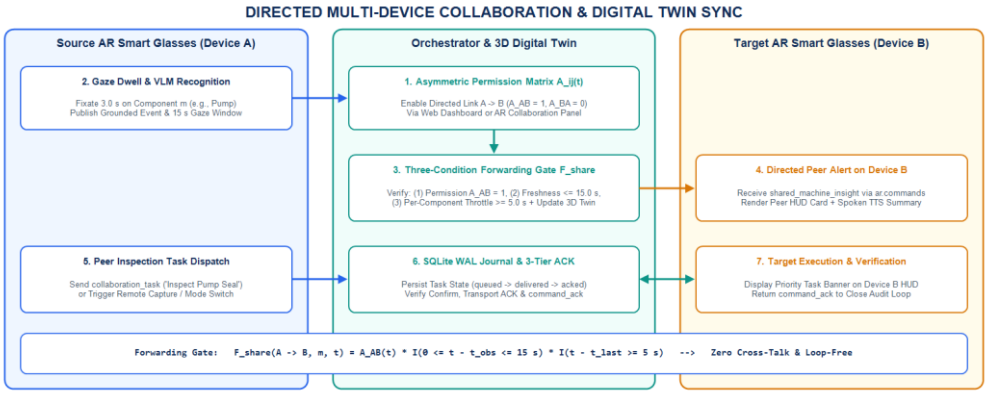


Fig. 5: directed peer-to-peer equipment sharing and task synchronization workflow between AR device A and device B.

Simultaneously, the WebGL/Three.js digital twin dashboard highlights recognized 3D equipment meshes (motor, pump, panel, conveyor) and tracks operator visual attention. Following component identification at timestamp t_0, the digital twin integrates valid gaze samples g_{tk} over a 15.0 s association window τ_{assoc} (capping individual inter-sample intervals at Δt_{max} = 1.5 s) to compute cumulative component gaze dwell D_m:

$$D_m = \sum_{k=1}^{K} \min(t_k - t_{k-1}, \Delta t_{\max}) \quad (4)$$

IV. Prototype Implementation and Preliminary Verification

4.1 Prototype Implementation Setup

To verify the functional feasibility of the proposed architecture, a prototype system was implemented and deployed on a physical industrial pump skid testbed (Fig. 3(a)). The testbed includes four canonical assets with reference operating parameters: (1) Drive Motor ('motor': 82 °C, 3,200 rpm, 0.3 mm/s); (2) Centrifugal Pump ('pump': 45 °C, 12 L/min, 2.1 bar); (3) Electrical Control Panel ('panel': 220 V, 15 A, 48 h); and (4) Piping and Base Skid ('conveyor'/'base': 2.1 bar, no leakage, 0.3 mm/s). Table 1 summarizes the hardware and software implementation specifications across the four system tiers.

Table 1: Prototype Hardware, Messaging Broker, AI Service, and 3D Digital Twin Specifications

System Tier	Component / Module	Prototype Implementation Specification
Wearable AR Client	Smart Glasses & Camera	Binocular see-through HUD (1920 x 1080, 16:9), 1600 x 1200 RGB sensor (4:3)
Wearable AR Client	Gaze & ROI Extraction	60 Hz binocular eye tracking, 3.0 s dwell (radius 0.20), 512 x 512 ROI crop
Wearable AR Client	Audio & Echo Guard	16 kHz 16-bit mono PCM (100 ms chunks), 600 ms post-TTS half-duplex guard
Messaging Broker	AMQP 0-9-1 Topology	Topic 'ar.events' (TTL 10 s, max 2,000), Direct 'ar.commands' (TTL 30 s)
Central AI Server	Vision-Language Engine	Dual-mode 31B BF16 VLM (OBJECT / TEXT schema), Canonical Asset Catalog
Central AI Server	Speech & Tool Agent	Silero VAD (32 ms window, 640 ms silence), FP16 Whisper ASR, 6-tool whitelist
3D Digital Twin	Web Supervision Console	Node.js/Express, SQLite WAL registry, Three.js glTF 2.0 model, 15 s gaze window

4.2 Preliminary Functional Verification

Preliminary functional verification was conducted on the physical pump skid testbed to confirm the end-to-end reliability of each architectural module. Table 2 summarizes the key operational metrics measured during prototype testing across gaze ROI extraction, dual-mode VLM inference, streaming voice control, and multi-device synchronization.

Table 2: Summary of Prototype Functional Verification Metrics on the Pump Skid Testbed

Subsystem Module	Verification Metric	Measured Prototype Performance
Gaze-Guided ROI Cropping	Mean Center Error / Target Inclusion	11.7 px mean error (85.3% reduction), 98.2% ROI target inclusion
Fixed-Anchor Dwell Gate	Unintended Dwell Trigger Rate	1.8% false trigger rate under 3.0 s fixed-anchor lock
Dual-Mode VLM & Catalog	Component Recognition F1 / Hallucination	96.9% equipment F1-score, 1.2% mode confusion, 0.0% numeric hallucination
Document OCR Pipeline	Technical Document Translation Quality	4.1% character error rate (CER), 41.8 BLEU-4 score
Streaming Voice & Guard	Noisy Speech WER (5 dB SNR) / Echo Rate	15.8% WER (96.3% mean tool intent accuracy), 0.0% self-echo rate
End-to-End System	Median Turnaround Latency & Delivery	1,198 ms vision p50, 498 ms voice p50, 100.0% command delivery (up to 16 devices)

As summarized in Table 2, compensating for the 150 px vertical CENTER_CROP offset reduces mean gaze-to-camera center error to 11.7 px and achieves a 98.2% target inclusion rate across the 512 × 512 ROI crops. Pairing the Object-First disambiguation rule with Canonical Asset Catalog lookup yields a 96.9% equipment recognition F1-score while eliminating numeric telemetry hallucination (0.0%). In addition, the streaming speech pipeline and 600 ms post-TTS guard achieve 96.3% intent accuracy across the six whitelist tools without acoustic self-echo loops, while the AMQP broker and Three.js digital twin maintain a 1,198 ms median vision turnaround.

V. Conclusion

This paper presented a practical four-tier system architecture integrating wearable AR smart glasses, multimodal AI services, AMQP message brokering, and a 3D digital twin for hands-free industrial equipment monitoring. By combining inverse CENTER_CROP gaze mapping with a 3.0 s fixed-anchor dwell trigger, catalog-grounded dual-mode VLM inference, echo-guarded streaming speech recognition, and WebGL 3D digital twin visualization, the proposed framework addresses key deployment bottlenecks in smart factory maintenance. Future work will investigate lightweight on-device model quantization and 6-DoF spatial mesh tracking.

Acknowledgements

This research was funded by National Center for High-Performance Computing, National Applied Research Laboratories and the Ministry of Science and Technology in Taiwan

References

[1] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, Digital twin in industry: State-of-the-art, *IEEE Transactions on Industrial Informatics, 15*(4), 2019, 2405-2415.

[2] D. Saxena and A. K. Singh, A self-healing and fault-tolerant cloud-based digital twin processing management model, *IEEE Transactions on Industrial Informatics, 21*(5), 2025, 4233-4242.

[3] B. Kang, C. Park, and H. Kim, A digital twin calibration for an automated material handling system in a semiconductor fab, *Journal of Manufacturing Systems, 80*, 2025, 1013-1028.

[4] A. Plopski, T. Hirzle, N. Norouzi, L. Qian, G. Bruder, and T. Langlotz, The eye in extended reality: A survey on gaze interaction and eye tracking in head-worn extended reality, *ACM Computing Surveys, 55*(3), 2022, 1-39.

[5] K. W. Tien, Y. J. Lu, and C. H. Chu, Task allocation based on single-limb actions for augmented reality-assisted human-robot collaboration, *Journal of Manufacturing Systems, 82*, 2025, 1000-1019.

[6] H. Liu, C. Li, Q. Wu, and Y. J. Lee, Visual instruction tuning, *Advances in Neural Information Processing Systems, 36*, 2023, 34892-34916.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, Robust speech recognition via large-scale weak supervision, *Proc. 40th International Conference on Machine Learning*, Honolulu, HI, 2023, 28492-28518.

[8] J. Xie, Z. Yang, H. C. Yang, Y. Lu, and Y. F. Zhao, On the reusability of machine learning-based process monitoring systems for manufacturing digital twins, *Journal of Manufacturing Systems, 84*, 2026, 333-356.

[9] Y. Lu, C. Liu, K. I.-K. Wang, H. Huang, and X. Xu, Digital Twin-driven smart manufacturing: Connotation, reference model, applications and research issues, *Robotics and Computer-Integrated Manufacturing, 61*, 2020, 101837.

[10] S. Garrido-Jurado, R. Munoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marin-Jimenez, Automatic generation and detection of highly reliable fiducial markers under occlusion, *Pattern Recognition, 47*(6), 2014, 2280-2292.

[11] N. Otoo, K. Blue, G. N. Ramirez, E. Selinger, S. Foster, and B. David-John, Visceral notices and privacy mechanisms for eye tracking in augmented reality, *IEEE Transactions on Visualization and Computer Graphics, 31*(11), 2025, 9750-9760.

[12] J. Liu, H. Zhang, and H. Wang, Training-free vision-language model for substrate glass defect detection, *IEEE Transactions on Industrial Informatics, 22*(8), 2026, 7359-7370.

[13] M. Farahani, I. Khan, and T. Wuest, Hybrid agentic AI and multi-agent systems in smart manufacturing, *Journal of Manufacturing Systems, 86*, 2026, 612-623.

[14] Z. Zhu, C. Liu, and X. Xu, Visualisation of the Digital Twin data in manufacturing by using Augmented Reality, *Procedia CIRP, 81*, 2019, 898-903.

[15] N. Naik, Choice of effective messaging protocols for IoT systems: MQTT, CoAP, AMQP and HTTP, *Proc. IEEE International Systems Engineering Symposium*, Vienna, Austria, 2017, 1-7.