Multimodal Artificial Intelligence for Interactive Maintenance Training Using AR Glasses

Ming-Chi Luo*

*Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan; e-mail: micky920620.cs10@nycu.edu.tw

*Corresponding Author: micky920620.cs10@nycu.edu.tw

DOI: https://doi.org/10.30211/JIC.202604.015

Submitted: Aug. 29, 2026 Accepted: Oct 17, 2026

ABSTRACT

Industrial maintenance training increasingly requires technicians to interpret complex equipment states while accessing procedural knowledge under time and safety constraints. Conventional manuals and video-based instruction separate information from the physical task environment, whereas many augmented-reality systems still rely on fixed workflows and limited contextual understanding. This study proposes a multimodal artificial intelligence framework for interactive maintenance training using augmented-reality (AR) glasses. The framework integrates visual equipment perception, speech and text understanding, maintenance-document retrieval, and procedural context into a confidence-aware fusion mechanism. A context-aware reasoning layer then selects the next instructional action and presents component labels, step guidance, warnings, and conversational feedback directly in the user's field of view. The manuscript defines the system architecture, multimodal representation and fusion model, maintenance-state formulation, retrieval process, and an experimental protocol for comparing printed manuals, video tutorials, conventional AR, and the proposed approach. The resulting design provides a reproducible basis for evaluating task efficiency, operational errors, interaction quality, usability, and cognitive workload in AR-assisted maintenance training.

Keywords: Augmented reality, Multimodal artificial intelligence, AR glasses, Maintenance training, Human-Computer interaction, Computer vision, Intelligent tutoring

1. Introduction

Industrial maintenance is a knowledge-intensive activity in which technicians must identify components, interpret equipment conditions, follow safety procedures, and execute sequential operations accurately. As production systems become increasingly automated and interconnected, maintenance procedures also become more complex and difficult to learn through conventional apprenticeship alone. Printed manuals and desktop-based digital documents remain common sources of procedural knowledge, but they force operators to divide their attention between the physical workspace and external information sources. This separation can increase search time, interrupt task

flow, and create avoidable opportunities for procedural errors, particularly for novice technicians or infrequently performed maintenance tasks [3]-[6].

Augmented reality (AR) offers an alternative by aligning digital instructions with physical equipment. Foundational AR research emphasizes the combination of real and virtual content, real-time interaction, and spatial registration [1], [2]. In maintenance and repair, head-worn AR can display labels, arrows, animated cues, and procedural instructions within the operator's field of view. Controlled studies and systematic reviews have reported benefits in task localization, instructional accessibility, and maintenance performance, while also identifying persistent challenges related to tracking, content authoring, ergonomics, interaction design, and deployment complexity [4]-[11]. These challenges indicate that visual overlays alone are insufficient for flexible industrial training.

Recent advances in multimodal artificial intelligence create an opportunity to move AR maintenance systems from fixed instruction players toward context-aware interactive tutors. Modern visual encoders, object detectors, vision-language models, speech recognition systems, semantic retrieval methods, and transformer-based architectures can represent heterogeneous information in a shared computational pipeline [16]-[29]. In a maintenance scenario, the system may need to jointly interpret what the user is looking at, what the user asks, which procedural step is active, what the technical manual states, and whether a safety constraint has been satisfied. A robust system should therefore reason over multiple modalities rather than treating camera, speech, and textual knowledge as independent channels.

The central research problem addressed in this study is how to construct an AR-glasses-based maintenance training architecture that combines multimodal perception and procedural context while remaining sufficiently transparent for industrial training. Existing AR maintenance literature has extensively investigated visualization, tracking, and task guidance [4]-[11], whereas general-purpose multimodal AI research has emphasized representation learning and language-vision alignment [16]-[24]. A practical gap remains between these streams: maintenance guidance requires reliable component grounding, domain-specific knowledge retrieval, explicit task-state management, and bounded instructional actions instead of unconstrained conversational output.

Accordingly, this study proposes a multimodal AI framework for interactive maintenance training using AR glasses. The contributions are fourfold. First, a modular architecture integrates visual perception, speech/text interaction, maintenance-document retrieval, and procedural context. Second, a confidence-aware fusion formulation combines modality representations while reducing the influence of uncertain inputs. Third, a context-aware maintenance reasoner constrains guidance according to procedural state and safety rules. Fourth, a reproducible evaluation protocol is defined to compare printed manuals, video tutorials, conventional AR guidance, and the proposed multimodal AR system using objective performance measures and subjective usability/workload instruments. The framework is intended as a research and implementation blueprint that can be instantiated with specific AR hardware and domain datasets.

2. Literature Review

2.1 Augmented Reality for Industrial Maintenance and Training

AR has long been investigated as a means of reducing the spatial and cognitive separation between instructions and maintenance objects. Henderson and Feiner demonstrated that tracked head-worn AR documentation could assist professional mechanics by placing text and graphical guidance in the work environment [4]. Webel et al. subsequently developed an AR training platform for assembly and maintenance skills that explicitly considered multimodal interaction and skill transfer [5]. Palmarini et al. reviewed AR maintenance research and found promising performance benefits but also substantial fragmentation in hardware, software, tracking, and authoring approaches [6].

Industrial AR surveys further show that wearable displays and markerless tracking have become central design directions, while usability, registration robustness, environmental variability, and integration with enterprise information systems continue to limit deployment [7], [8]. Eswaran et al. emphasized the role of AR guidance in assembly, maintenance, and repair and identified tracking, rendering, and human-machine interaction as recurring technical concerns [9]. Recent work on facility-management knowledge transfer and maintenance adoption likewise argues that human factors, organizational readiness, and data integration are as important as visualization fidelity [10], [11]. These findings motivate an architecture in which AR acts as an interaction surface, while intelligent reasoning is handled by dedicated perception, knowledge, and context modules.

2.2 Visual Perception and Equipment Component Recognition

Reliable visual grounding is necessary for maintenance guidance because an instruction such as "remove the upper protective cover" must be anchored to the correct physical component. Deep convolutional models established strong visual feature representations [27], while single-stage detectors such as YOLO demonstrated efficient object localization suitable for real-time systems [24], [25]. Transformer-based detection approaches such as DETR introduced set-based object prediction [26], and promptable segmentation models such as Segment Anything expanded the ability to isolate previously unseen objects with minimal task-specific adaptation [19].

For an AR training system, the visual subsystem should not only classify objects but also provide spatial coordinates, confidence values, and temporal continuity. The proposed architecture therefore treats visual detection as a probabilistic observation rather than an unquestioned fact. Confidence is propagated into the fusion layer so that ambiguous visual states can trigger clarification, prompt a request for a different viewpoint, or fall back on procedural context. This design is particularly important in maintenance environments where visually similar components, occlusion, lighting variation, and partial disassembly can degrade recognition performance.

2.3 Multimodal Representation, Vision-Language Models, and Retrieval

Transformer architectures provide a general mechanism for learning long-range dependencies and cross-token interactions [16]. BERT demonstrated contextual language representation [17], while CLIP showed that visual and textual concepts can be aligned through large-scale contrastive learning [18]. More recent vision-language instruction models further support conversational reasoning over images [20]. These developments are relevant to maintenance because a technician's question often

refers to an object currently visible in the scene, requiring the system to ground language in visual context.

However, general-purpose vision-language generation is not sufficient for procedural maintenance. Technical guidance must remain consistent with verified manuals, task prerequisites, and safety rules. Retrieval-augmented generation addresses part of this problem by conditioning responses on externally retrieved evidence [21]. Dense retrieval and sentence-level embeddings provide practical mechanisms for matching user questions to relevant document passages [22], [29]. The proposed system adopts retrieval as a constrained knowledge-access step: the language module retrieves candidate SOP passages, while the maintenance-state reasoner determines whether the retrieved instruction is valid for the current procedural step.

2.4 Speech Interaction, Human Factors, and Research Gap

Hands-free speech is particularly attractive for AR maintenance because operators may be manipulating tools or wearing gloves. Large-scale weakly supervised speech recognition has improved robustness across languages and acoustic conditions [23], but recognition errors remain consequential when commands refer to safety-critical actions. Therefore, speech confidence and semantic consistency with the current maintenance state should influence downstream decision-making. Similarly, usability and cognitive workload must be measured rather than inferred from system accuracy alone. The System Usability Scale (SUS) [30] and NASA Task Load Index (NASA-TLX) [31] are widely used instruments for evaluating perceived usability and workload in interactive systems.

Table 1 summarizes the positioning of the proposed study. Earlier systems typically optimize one or two aspects, such as overlay, recognition, or procedural display. The proposed framework instead combines multimodal perception, knowledge retrieval, explicit procedural state, safety constraints, and closed-loop interaction. The novelty lies not in claiming a new foundation model but in defining an integrated and testable maintenance-specific architecture that constrains multimodal AI within a procedural training workflow.

Table 1. Functional comparison of maintenance training approaches

Approach	Visual grounding	Speech interaction	Manual/SOP retrieval	Explicit task state	Safety-aware guidance
Printed / PDF manual	No	No	Manual	No	Manual
Video tutorial	Limited	No	Fixed content	Implicit	Fixed
Conventional AR	Yes / marker based	Limited	Pre-authored	Partial	Rule based
General VLM assistant	Yes	Optional	Optional	Weak	Unconstrained
Proposed multimodal AR	Yes	Yes	Yes	Yes	Yes

Source: By authors.

3. Research Design

3.1 Overall System Architecture

The proposed system is organized into five cooperating layers: (1) AR-glasses sensing and interaction, (2) multimodal encoding, (3) confidence-aware fusion, (4) maintenance reasoning and knowledge retrieval, and (5) AR instruction rendering. Figure 1 illustrates the architecture. The separation between perception and reasoning is intentional. Perception modules may be replaced as hardware or models evolve, while the procedural reasoner preserves maintenance constraints and traceability.

The AR glasses provide RGB video, audio, and user interaction events. A visual encoder identifies candidate components and spatial regions; a speech recognizer converts spoken input into text; a language encoder embeds the user's query; and a context encoder represents the current maintenance step, previously completed actions, tool availability, and relevant safety conditions. The retrieval module indexes authorized maintenance documents and retrieves the most relevant passages for the current query. These signals are fused into a maintenance-state representation, after which the reasoner selects an instructional action, such as highlighting, explaining, warning, confirming, repeating, or advancing to the next step.

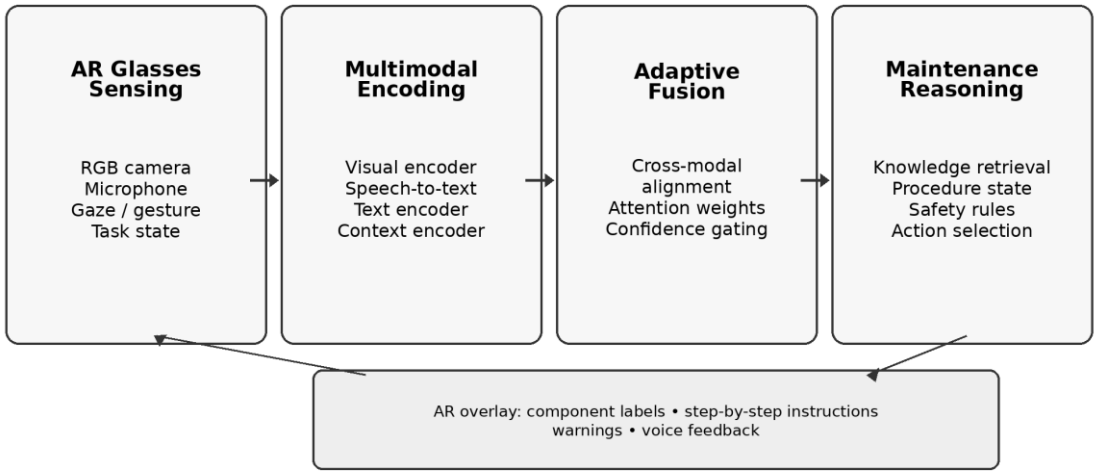


Figure 1. Overall architecture of the proposed multimodal AR maintenance training system

Source: By authors.

3.2 Visual Equipment Recognition and Spatial Grounding

Let the RGB frame captured at time t be denoted by $I_t \in R^{H\times W\times 3}$. A visual encoder $f_v(\cdot)$ transforms the frame into a feature representation v_t. An object detector produces a set of N candidate equipment components with class labels, bounding boxes, and confidence values. The representation is defined as follows:

$$v_t = f_v(I_t) \quad (1)$$

$$O_t = \{(c_i, b_i, p_i)\}_{i=1}^{N} \quad (2)$$

where c_i denotes the predicted component category, b_i denotes the image-space bounding box, and p_i ∈ [0,1] is the detector confidence. For AR rendering, image-space regions are associated with device/world coordinates through the tracking subsystem. The current implementation blueprint is model-agnostic; YOLO-style detectors [24], [25], DETR [26], or promptable segmentation [19] may be selected according to latency and data availability.

Temporal consistency is maintained by associating detections across adjacent frames. When a target component is not detected with sufficient confidence, the system does not automatically advance the procedure. Instead, it can request a change in viewpoint or ask the user to confirm the target. This behavior reduces the risk of presenting an instruction on an incorrectly grounded component.

3.3 Speech, Text, and Maintenance-Knowledge Representation

The audio stream A_t is transcribed into a textual query Q_t by a speech-recognition function f_s(·). The query is encoded by f_q(·), while maintenance-document passages D_j are embedded by f_k(·).

$$Q_t = f_s(A_t), \quad q_t = f_q(Q_t) \quad (3)$$

$$k_j = f_k(D_j) \quad (4)$$

Cosine similarity is used as a transparent retrieval score between the query representation and candidate maintenance passages:

$$s_j = (q_t^T k_j) / (\|q_t\|_2 \|k_j\|_2) \quad (5)$$

The top-K passages are then filtered by equipment identifier, maintenance procedure, and current step. This filtering is important because semantically similar instructions can be operationally invalid if they belong to another equipment model or a later disassembly stage. The retrieval subsystem therefore returns not only text but also metadata such as procedure ID, prerequisite step, component ID, hazard class, and version date.

3.4 Confidence-Aware Multimodal Fusion

A key design requirement is that unreliable modalities should not dominate the decision. The proposed fusion layer assigns each modality a reliability score derived from detector confidence, speech-recognition confidence, retrieval similarity, and context completeness. Let m ∈ {v,q,k,c} represent visual, query, knowledge, and context modalities. A normalized modality weight α_m is calculated as:

$$\alpha_m = \exp(e_m) / \Sigma_{j\in\{v,q,k,c\}} \exp(e_j) \quad (6)$$

The fused representation is obtained from the weighted concatenation of modality embeddings:

$$z_t = W_f [\alpha_v v_t ; \alpha_q q_t ; \alpha_k k_t ; \alpha_c c_t] + b_f \quad (7)$$

This formulation allows the system to down-weight a low-confidence visual detection or noisy speech transcription while relying more heavily on procedure context and retrieved knowledge. In an implementation based on cross-attention, Equation (7) may be replaced with learned token-level fusion; however, the explicit weighted form is useful for system auditing and ablation experiments.

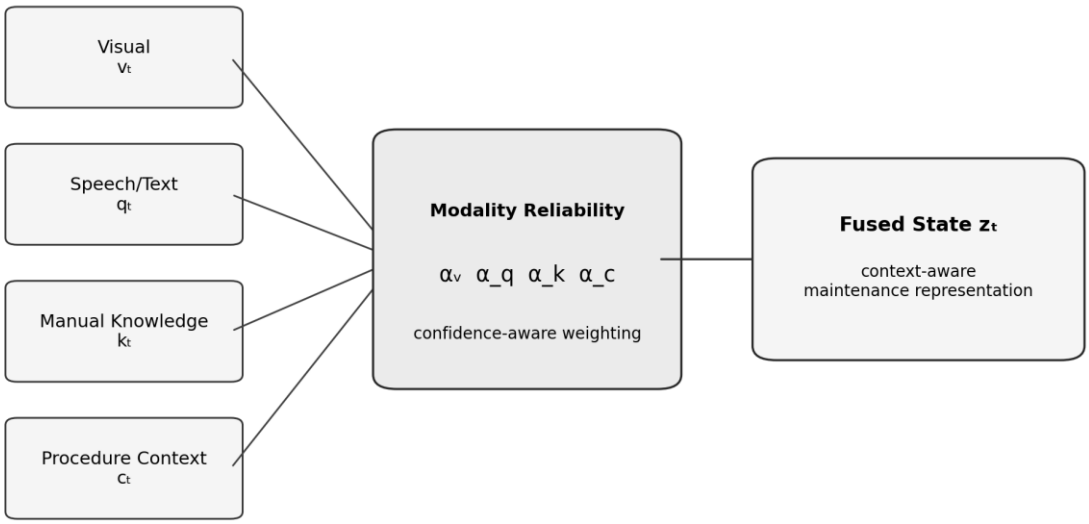


Figure 2. Confidence-aware multimodal fusion mechanism

Source: By authors.

3.5 Context-Aware Maintenance State and Action Selection

The maintenance process is represented as a finite set of procedural states. At time t, the state S_t includes detected objects O_t, the user query Q_t, the active procedure step P_t, equipment condition E_t, and safety flags H_t:

$$S_t = (O_t, Q_t, P_t, E_t, H_t) \quad (8)$$

The instructional action set A contains actions such as DISPLAY_LABEL, SHOW_ARROW, SHOW_TEXT, SPEAK, WARN, REQUEST_CONFIRMATION, REPEAT, and ADVANCE. The reasoner selects the action that maximizes the conditional action score while satisfying procedure and safety constraints Ω(S_t):

$$a_t^* = \arg\max_{\{a \in A,\ a \in \Omega(S_t)\}} P(a \mid z_t, S_t) \quad (9)$$

The constraint set Ω(S_t) prevents the system from issuing an action that violates a prerequisite or unresolved hazard. For example, a cover-removal instruction can be blocked until power isolation has been confirmed. This maintenance-state design also enables traceability because every transition can be logged together with the detected component, retrieved passage, user query, system action, and confidence values.

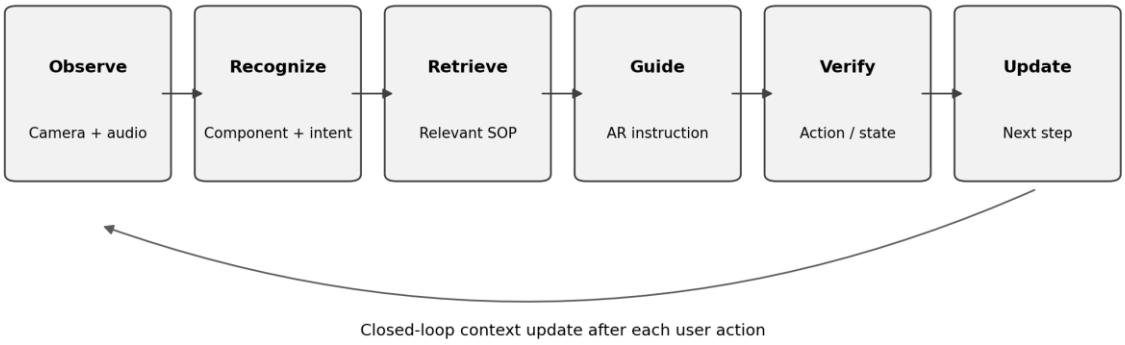


Figure 3. Closed-loop procedure state and interaction flow

Source: By authors.

3.6 AR Instruction Rendering and Interaction Design

The AR interface follows a minimal-information principle: only information needed for the current step is displayed, while additional explanation is available on demand. The primary overlay includes the target component, step number, concise action text, and safety warning when applicable. Speech can be used for clarification questions, repetition, and navigation. Figure 4 presents an illustrative interface arrangement rather than a hardware-specific screenshot.

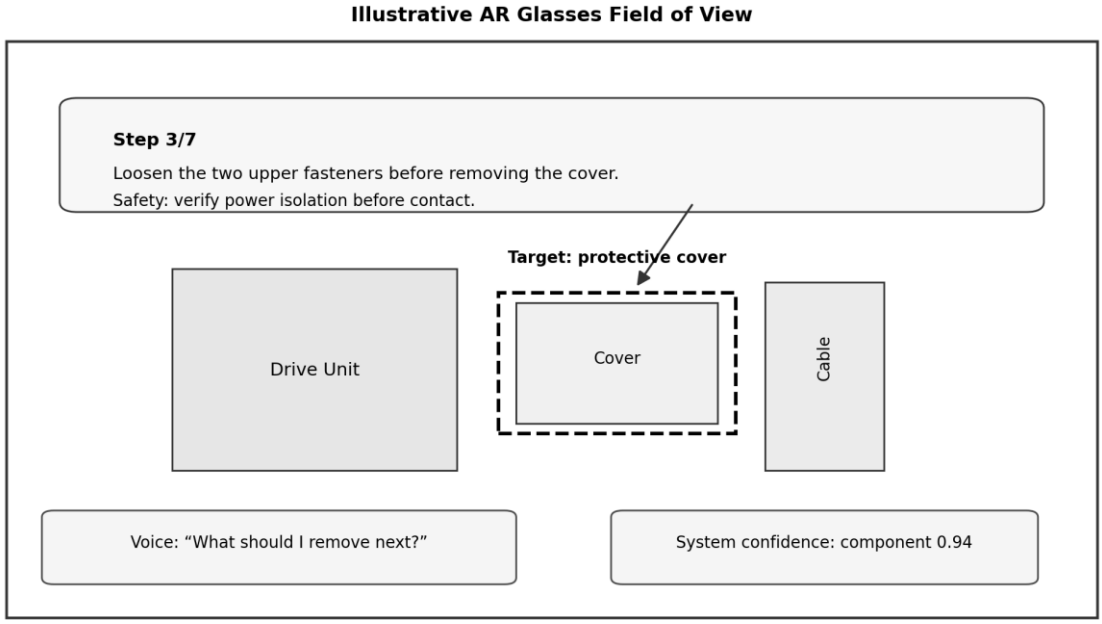


Figure 4. Illustrative AR maintenance instruction interface

Source: By authors.

3.7 Multimodal Maintenance Guidance Algorithm

Algorithm 1 summarizes the runtime loop. The procedure is deliberately conservative: advancement occurs only after the current target and completion condition are validated. If the fused confidence is below a threshold τ, the system requests clarification instead of generating an unverified instruction.

Algorithm 1. Context-aware multimodal maintenance guidance

Input: frame I_t, audio A_t, procedure state P_t, maintenance corpus D, threshold τ
1: Capture I_t and A_t from AR glasses
2: Detect equipment objects O_t and compute visual confidence
3: Transcribe A_t to Q_t and encode query representation q_t
4: Retrieve top-K maintenance passages conditioned on equipment and P_t
5: Encode procedural context c_t and safety flags H_t
6: Compute modality weights α_v, α_q, α_k, α_c
7: Fuse modalities to obtain z_t
8: if fused confidence $< \tau$ then request clarification / new viewpoint
9: else apply procedure and safety constraints $\Omega(S_t)$
10: select action a_t^* and render AR / voice instruction
11: observe user operation and verify completion condition
12: update P_t and repeat until maintenance task is completed
Output: traceable sequence of context-aware maintenance instructions

3.8 Implementation Blueprint

The architecture can be implemented using a contemporary optical see-through or video see-through AR headset with an RGB camera, microphone, inertial tracking, and network connectivity. The perception and retrieval modules may execute on-device, on an edge computer, or on a local server depending on latency and privacy requirements. For industrial deployment, the maintenance corpus should be restricted to version-controlled manuals and SOP documents. Equipment identity should be resolved before retrieval, and all generated language should be grounded in retrieved passages or predefined procedural rules.

Table 2. Functional modules of the proposed system

Module	Recommended function	Key output	Design priority
AR sensing	Camera, audio, tracking, gaze/gesture	I_t, A_t, pose, interaction events	Low latency
Visual perception	Detection / segmentation / tracking	Component ID, region, confidence	Robust grounding

Module	Recommended function	Key output	Design priority
Speech & language	ASR + query encoding	Transcript, intent, confidence	Hands-free interaction
Knowledge retrieval	Dense / hybrid retrieval	Top-K SOP passages + metadata	Traceable evidence
Fusion	Confidence-aware multimodal fusion	z_t	Reliability
Reasoner	Procedure state + safety rules	Constrained action	Safety and auditability
Renderer	AR labels, arrows, text, voice	In-view instruction	Low cognitive load

Source: By authors.

4. Results and Discussion

4.1 Experimental Protocol

A controlled evaluation should compare four instructional conditions: G1 printed/PDF manual, G2 video tutorial, G3 conventional AR with pre-authored step overlays, and G4 the proposed multimodal AR system. Participants should perform the same set of maintenance tasks under counterbalanced ordering to reduce learning effects. Suitable tasks include component identification, cover removal, connector inspection, fastener replacement, and reassembly. The equipment should be de-energized or otherwise configured for safe training use.

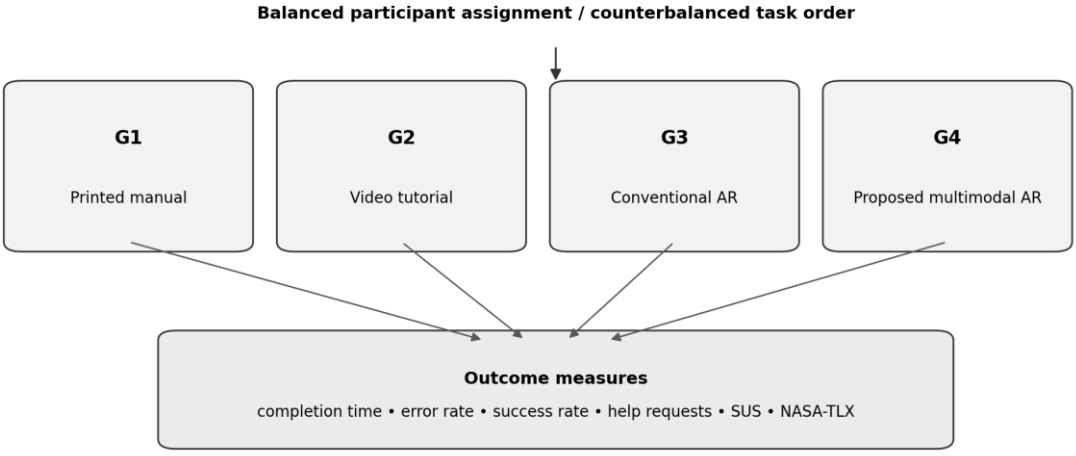


Figure 5. Proposed comparative evaluation protocol

Source: By authors.

Objective outcomes include completion time, operation errors, task success rate, number of help requests, and system response latency. Perception-specific outcomes include component precision,

recall, F1-score, and mean average precision where bounding-box ground truth is available. Subjective outcomes include SUS and NASA-TLX. The following equations define three core task measures.

$$Error\ Rate = N_error / N_operation \quad (10)$$

$$Success\ Rate = N_success / N_total \quad (11)$$

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (12)$$

Statistical analysis should first test distributional assumptions. If repeated measurements are normally distributed, repeated-measures ANOVA with appropriate post-hoc comparisons can be used. Otherwise, a Friedman test followed by corrected pairwise tests is suitable. Effect sizes and confidence intervals should be reported in addition to p-values.

4.2 Component Recognition and Grounding Results

This subsection is reserved for measured visual-recognition results. At minimum, the final manuscript should report component-level precision, recall, F1-score, and latency on a test set that is separated from training images. Performance should also be stratified by lighting condition, occlusion level, and component size because these factors directly affect AR grounding reliability. The proposed system should not be described as outperforming a baseline until actual measurements are available.

Table 3. Component recognition performance (replace TBD with measured results)

Method	Precision	Recall	F1-score	mAP@0.5	Latency (ms)
Detector baseline	TBD	TBD	TBD	TBD	TBD
Detector + temporal tracking	TBD	TBD	TBD	TBD	TBD
Detector + context validation	TBD	TBD	TBD	TBD	TBD

Source: By authors.

4.3 Maintenance Task Performance

The primary system-level hypothesis is that contextual AR guidance can reduce information-search overhead and procedural errors relative to external documentation. The appropriate evidence is measured task performance rather than user preference alone. Completion time should be measured from task start to successful completion; an operation error should be defined in advance, for example selection of an incorrect component, execution of an out-of-order action, or violation of a prescribed step. Help requests should record both user-initiated clarification and system-triggered recovery prompts.

Table 4. Maintenance task performance (replace TBD with measured results)

Condition	Completion time (s)	Error rate	Success rate	Help requests
G1 Printed/PDF manual	TBD	TBD	TBD	TBD
G2 Video tutorial	TBD	TBD	TBD	TBD
G3 Conventional AR	TBD	TBD	TBD	TBD
G4 Proposed multimodal AR	TBD	TBD	TBD	TBD

Source: By authors.

4.4 Usability, Workload, and Interaction Quality

Usability evaluation should complement task performance. SUS provides a standardized ten-item usability score [30], while NASA-TLX captures multidimensional workload [31]. Additional Likert items may assess instruction clarity, visual comfort, speech interaction quality, confidence in system guidance, and perceived usefulness. Because AR can simultaneously reduce document-search burden and introduce display or interaction load, workload should be interpreted by dimension rather than only by an aggregate score.

Table 5. Usability and workload outcomes (replace TBD with measured results)

Measure	**G1**	**G2**	**G3**	**G4**
SUS score	TBD	TBD	TBD	TBD
NASA-TLX overall	TBD	TBD	TBD	TBD
Instruction clarity (1-5)	TBD	TBD	TBD	TBD
Interaction quality (1-5)	N/A	N/A	TBD	TBD
Perceived confidence (1-5)	TBD	TBD	TBD	TBD

Source: By authors.

4.5 Ablation and Failure-Mode Analysis

Ablation experiments are important for determining whether multimodal fusion contributes beyond AR visualization. Recommended variants include visual-only, visual + speech, visual + retrieval, and full multimodal fusion. In addition, failure cases should be grouped into visual misrecognition, speech-transcription error, ambiguous query, retrieval mismatch, tracking drift, and procedure-state mismatch. Reporting recovery behavior is especially valuable: a safety-aware system should be rewarded for refusing to advance when confidence is insufficient rather than being evaluated solely on response rate.

Table 6. Recommended ablation configurations

Variant	Visual	Speech	Retrieval	Procedure context	Primary purpose
A	✓	–	–	–	Visual baseline
B	✓	✓	–	–	Effect of speech interaction
C	✓	–	✓	✓	Effect of grounded knowledge
D Full	✓	✓	✓	✓	Complete proposed system

Source: By authors.

4.6 Discussion

The proposed framework addresses a recurring limitation in AR maintenance research: the gap between spatial visualization and flexible task understanding. By representing visual observations, language, retrieved technical knowledge, and procedure context separately before fusion, the architecture supports both modular engineering and diagnostic analysis. The explicit procedure state also distinguishes the system from unconstrained visual question answering. In maintenance training, a technically fluent answer is not necessarily a valid next action; validity depends on the equipment model, current disassembly state, completed prerequisites, and safety status.

The confidence-aware design further acknowledges that multimodal inputs have different and time-varying reliability. Visual recognition may degrade under occlusion, speech may be corrupted by industrial noise, and document retrieval may return multiple semantically similar passages. A practical tutor should therefore expose uncertainty to the control logic. Clarification, re-observation, and step confirmation are legitimate actions rather than failures. This principle is consistent with the broader literature emphasizing human factors and trustworthy deployment for industrial AR [6]-[11].

Several limitations must be considered. First, the framework requires domain-specific component labels and structured maintenance documents. Second, AR tracking and display ergonomics remain hardware-dependent. Third, the proposed equations define an interpretable fusion baseline but do not by themselves guarantee optimal cross-modal representation. Fourth, the empirical benefit of the system must be established through controlled user studies using measured data. Finally, industrial deployment would require cybersecurity, privacy, change-management, and equipment-certification considerations beyond the training-focused scope of this paper.

5. Conclusions

This study presents a multimodal artificial intelligence framework for interactive maintenance training using AR glasses. The framework integrates visual component perception, speech and text interaction, maintenance-document retrieval, confidence-aware multimodal fusion, explicit procedural state, and safety-constrained action selection. Instead of treating AR as a passive display channel, the proposed design uses closed-loop perception and context updates to determine what the user is observing, what information is being requested, and which instructional action is appropriate at the current maintenance step.

The manuscript also defines an evaluation methodology that separates perception quality, task performance, usability, and workload. This separation is necessary because high recognition accuracy does not automatically translate into effective maintenance training, and favorable user preferences do not substitute for objective task outcomes. Future work should implement the framework on a selected AR platform, collect a version-controlled maintenance corpus and component dataset, complete the comparative user study, and investigate lightweight on-device multimodal models, digital-twin integration, multilingual interaction, and privacy-preserving edge deployment.

Acknowledgements

This article received no financial or funding support.

Conflicts of Interest

The authors confirm that there are no conflicts of interest.

References

[1] Azuma, R.T. A survey of augmented reality. Presence: Teleoperators and Virtual Environments, 1997, 6(4), 355-385.

[2] Milgram, P. and Kishino, F. A taxonomy of mixed reality visual displays. IEICE Transactions on Information and Systems, 1994, E77-D(12), 1321-1329.

[3] Nee, A.Y.C., Ong, S.K., Chryssolouris, G. and Mourtzis, D. Augmented reality applications in design and manufacturing. CIRP Annals, 2012, 61(2), 657-679.

[4] Henderson, S. and Feiner, S. Exploring the benefits of augmented reality documentation for maintenance and repair. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(10), 1355-1368. DOI: 10.1109/TVCG.2010.245.

[5] Webel, S., Bockholt, U., Engelke, T., Gavish, N., Olbrich, M. and Preusche, C. An augmented reality training platform for assembly and maintenance skills. Robotics and Autonomous Systems, 2013, 61(4), 398-403. DOI: 10.1016/j.robot.2012.09.013.

[6] Palmarini, R., Erkoyuncu, J.A., Roy, R. and Torabmostaedi, H. A systematic review of augmented reality applications in maintenance. Robotics and Computer-Integrated Manufacturing, 2018, 49, 215-228. DOI: 10.1016/j.rcim.2017.06.002.

[7] Egger, J. and Masood, T. Augmented reality in support of intelligent manufacturing – A systematic literature review. Computers & Industrial Engineering, 2020, 140, 106195. DOI: 10.1016/j.cie.2019.106195.

[8] Masood, T. and Egger, J. A survey of industrial augmented reality. Computers & Industrial Engineering, 2020, 139, 106159. DOI: 10.1016/j.cie.2019.106159.

[9] Eswaran, M., Gulivindala, A.K., Inkulu, A.K. and Bahubalendruni, M.V.A.R. Augmented reality-based guidance in product assembly and maintenance/repair perspective: A state of the art review on challenges and opportunities. Expert Systems with Applications, 2023, 213, 118983. DOI: 10.1016/j.eswa.2022.118983.

[10] Wang, X., Wang, S., Xiao, F. and Luo, X. Augmented reality-based knowledge transfer for facility management: A systematic review. Journal of Building Engineering, 2024, 98, 111186. DOI: 10.1016/j.jobe.2024.111186.

[11] Tayyarat, O. et al. A systematic review of different adoption strategies for augmented reality in maintenance: Bridging theory and practice. Discover Applied Sciences, 2026, 8, 18.

[12] Billinghurst, M., Clark, A. and Lee, G. A survey of augmented reality. Foundations and Trends in Human–Computer Interaction, 2015, 8(2-3), 73-272.

[13] Chatzopoulos, D., Bermejo, C., Huang, Z. and Hui, P. Mobile augmented reality survey: From where we are to where we go. IEEE Access, 2017, 5, 6917-6950.

[14] van Krevelen, D.W.F. and Poelman, R. A survey of augmented reality technologies, applications and limitations. The International Journal of Virtual Reality, 2010, 9(2), 1-20.

[15] Carmigniani, J., Furht, B., Anisetti, M., Ceravolo, P., Damiani, E. and Ivkovic, M. Augmented reality technologies, systems and applications. Multimedia Tools and Applications, 2011, 51, 341-377.

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. Attention is all you need. Advances in Neural Information Processing Systems, 2017, 30.

[17] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, 2019, 4171-4186.

[18] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, 2021, 139, 8748-8763.

[19] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P. and Girshick, R. Segment Anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, 4015-4026. DOI: 10.1109/ICCV51070.2023.00371.

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

[21] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S. and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 2020, 33, 9459-9474.

[22] Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of EMNLP-IJCNLP, 2019, 3982-3992.

[23] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I. Robust speech recognition via large-scale weak supervision. In: Proceedings of the 40th International Conference on Machine Learning, 2023, 202, 28492-28518.

[24] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 779-788.

[25] Bochkovskiy, A., Wang, C.-Y. and Liao, H.-Y.M. YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934, 2020.

[26] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S. End-to-end object detection with transformers. In: Computer Vision – ECCV 2020, Lecture Notes in Computer Science, 2020, 12346, 213-229.

[27] He, K., Zhang, X., Ren, S. and Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 770-778.

[28] Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In: Proceedings of the International Conference on Learning Representations, 2019.

[29] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W.-T. Dense passage retrieval for open-domain question answering. In: Proceedings of EMNLP, 2020, 6769-6781.

[30] Brooke, J. SUS: A ‘quick and dirty’ usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A. and McClelland, I.L., Eds., Usability Evaluation in Industry. Taylor & Francis, London, UK, 1996, 189-194.

[31] Hart, S.G. and Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Hancock, P.A. and Meshkati, N., Eds., Human Mental Workload. North-Holland, Amsterdam, Netherlands, 1988, 139-183.

Copyright© by the authors, Licensee Intelligence Technology International Press. The article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA).