

Article

A Digital Twin Framework for Multimodal Operator-Centered Human–Cobot Collaboration in Assembly Tasks

David Alfaro-Viquez ¹, Mauricio Zamora-Hernandez ¹, Michael Fernandez-Vega ^{1,2}, David Ortiz-Perez ², Jose Garcia-Rodriguez ² and Jorge Azorin-Lopez ^{2,*}

¹ Department of Industrial Engineering, University of Costa Rica, San Pedro de Montes de Oca, San Jose 11501-2060, Costa Rica; david.alfaro@ucr.ac.cr (D.A.-V.); mauricio.zamora@ucr.ac.cr (M.Z.-H.); michael.fernandez@ucr.ac.cr (M.F.-V.)

² Department of Computer Science and Technology, University of Alicante, San Vicente del Raspeig, 03690 Alicante, Spain; dortiz@dtic.ua.es (D.O.-P.); jgarcia@dtic.ua.es (J.G.-R.)

* Correspondence: jazorin@ua.es

Abstract

Current digital twin frameworks focused on human–robot collaboration rarely take into account the sensory degradation of real industrial environments, nor do they integrate the operator as an active agent within the system. This research presents a multimodal digital twin framework for a dual-arm collaborative robot at an assembly station; the system was developed using ROS 2 Jazzy and CoppeliaSim as the simulator. The architecture integrates three main components: the first is a perception layer that captures voice commands using Whisper ASR and the state of the workspace using a hybrid YOLO + ViT visual pipeline, both with per-channel metadata; the second consists of a Confidence-Weighted Late Fusion engine that dynamically adjusts the weight of each modality based on real-time signal quality, so that each fusion decision can be reconstructed from the signals that generated it; and the third component is a Reference Resolver that grounds linguistic intent within the visual context of the scene and in the fusion weights, using a local instance of Llama 3.1 8B that does not transmit audio, transcripts, or images outside the system. The framework was evaluated using 210 iterations distributed across seven degradation conditions of increasing severity, comparing adaptive fusion against a baseline of fixed weights (0.5/0.5). Under clean conditions and under visual degradation of any severity, both configurations achieved 100% accuracy. Under severe auditory degradation (SNR 0 dB), adaptive fusion activated the safety gate and refrained from executing most commands (13.3% accuracy), while the fixed-weight baseline executed more commands (60% accuracy) but made three incorrect object selections; under severe dual degradation, the pattern repeated (13.3% vs. 40%, with five incorrect selections in the baseline). The adaptive system made no grounding errors in the 210 executions, compared to eight in the baseline, substituting incorrect execution with conservative abstention when no modality provided a reliable signal. The implementation, featuring a versioned degradation protocol and a fixed seed, provides a reproducible benchmark for evaluating multimodal fusion strategies in human–cobot interaction.



Academic Editors: Robert Ojstersek, Iztok Palčič and Dan Zhang

Received: 2 June 2026

Revised: 8 July 2026

Accepted: 9 July 2026

Published: 12 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: collaborative robotics; digital twin; human–robot collaboration; Industry 5.0; multimodal fusion

1. Introduction

Collaborative robots are increasingly being adopted for assembly tasks within the operator-centered manufacturing paradigm promoted by Industry 5.0 [1,2], but their inter-

action with the operator still relies heavily on rigid interfaces such as teach pendants, as well as offline programming, both of which require technical knowledge and increase the operator's cognitive load [3].

Using voice and vision offers a more natural alternative to this human–robot communication interface, but both channels can fail in different ways in a real manufacturing environment, where machinery noise degrades speech recognition, and occlusions or poor lighting reduce the reliability of visual detection. In the experiments conducted for this research, a signal-to-noise ratio of 0 dB was sufficient for Whisper to produce incoherent transcriptions for most commands (Section 6). A system that relies on a single perceptual channel has no alternative evidence to detect when that channel has deteriorated; it acts—or fails—based on the only available source of information [4].

This research addresses this vulnerability using a Confidence-Weighted Late Fusion engine, which continuously evaluates the quality of each perceptual channel and dynamically distributes the weight between the speech and visual channels based on that quality; the system's behavior in the presence of acoustic noise, visual occlusion, and simultaneous degradation of both channels is detailed in Section 3.

As a second contribution, the framework incorporates an LLM-assisted Reference Resolver that maps natural language commands to the current state of the workspace. Instead of relying on keyword matching or template-based analysis, the resolver constructs a prompt that feeds the visible scene, fusion weights, and the transcribed command into a locally hosted LLM model, thereby enabling the disambiguation of referential expressions such as color terms, spatial descriptors, and implicit references to objects. Deploying the model locally via Ollama prevents the operator's audio, transcripts, or images of the workspace from leaving the system.

The rest of the manuscript is organized as follows. Section 2 provides a review of the state of the art regarding digital twins in Industry 5.0, natural human–robot interaction, and multimodal perception for assembly tasks. Section 3 describes the proposed architecture for the system. The ROS 2 implementation, confidence-weighted fusion pipeline, and LLM-based Reference Resolver are described in Section 4. Section 5 describes the simulation environment, degradation protocols and experimental conditions. Section 6 presents quantitative and qualitative results. We discuss the interpretation of the results, the limitations and future work in Section 7. Finally, the main contributions and conclusions are presented in Section 8.

2. Related Work

2.1. The Human-Centric Paradigm: Industry 5.0 and Digital Twins

The manufacturing paradigm is transitioning from the efficiency-driven Industry 4.0 model to an Industry 5.0 (I5.0) paradigm, where human-centricity, resilience and sustainability are at the core of the process [1,2]. This new paradigm does not consider the operator as a secondary element, but a key one, whose needs, wellbeing and safety drive the design of the system [5–8]. Digital twins (DTs) have emerged as the technology of choice to bridge the physical and digital worlds as they provide the synchronization needed for this transition [1,9]. However, there is a gap in current DT architectures: the operator is considered too often a passive observer and not an active agent inside the control loop [8,10,11]. Modern frameworks need transparent interaction to comply with emerging health and safety mandates such as EU Regulation 2023/1230 [3].

Unlike static or monitoring-oriented models, an operator-centric DT must enable real-time bidirectional synchronization, in which the human's cognitive or physiological state directly affects the robotic system's behavior [10,12]. We address this by giving the operator intuitive, multimodal tools in the DT environment [7,13].

2.2. Natural Interaction and Cognitive Load in Collaborative Robotics

Human–robot collaboration (HRC) is intended to blend human adaptability with robotic accuracy, but conventional programming interfaces often contradict this objective [10,14]. Teach pendants, offline programming and Learning from Demonstration (LfD) methods have a high entry barrier as they require technical skill and rarely generalize well [7,15,16]. These rigid interfaces increase the cognitive load of the operator, which is related to mental fatigue, human error and workplace accidents [3,5,14]. These risks seemed to be reduced by “zero-programming” and natural interaction paradigms [16,17].

Augmented reality (AR) can provide visual guidance to reduce mental effort [7,13,18], and natural language processing (NLP) can allow intuitive communication through voice commands [17,19].

Large Language Models (LLMs) further add to this by providing context-sensitive, human-like reasoning that reduces training time and stress [3,20]. In this research, we use LLMs both as an interface and as a grounding mechanism that converts high-level intent into executable assembly sequences. The goal of this integration is to reduce the operator’s mental load; however, the empirical measurement of this effect is not within the experimental scope of this study.

2.3. Multimodal Perception and AI-Driven Reasoning for Assembly Tasks

The successful collaboration of the robot in assembly tasks relies on advanced “mutual cognition” (the ability to read the environment and predict the intention of humans) [7,21]. This is central to accurate visual scene understanding. Conventional RGB-D systems are still widely adopted, but recent work resorts to Vision Transformers and holistic scene understanding (VHSU) to interpret complex relationships and avoid occlusions [8,21,22]. The importance of the self-attention mechanism has been demonstrated to reach high prediction accuracy and robustness for different applications of advanced multimodal frameworks such as Dual-Extraction Modeling (DEM) and AutoAMS [23,24].

The architectures leverage multi-head attention and graph-based modeling to capture complex cross-modal interactions and long-range dependencies. Hence, they are able to handle noisy or corrupted data streams in challenging environments [23–25]. The backbone of the cognitive engine of contemporary HRC is constituted by Vision–Language Action (VLA) models and LLMs that map linguistic abstractions to physical objects [19,26,27]. These frameworks can resolve ambiguities in unstructured environments by combining visual data with linguistic intent [28,29]. Hierarchical graphs and Graph Convolutional Networks (GCNs) also allow the reasoning layer to handle complex spatial–temporal relationships in disassembly or assembly operations [30,31].

In this work, we integrated these multimodal elements in a ROS 2- and CoppeliaSim-based digital twin, where action plans can be autonomous but still follow real-time safety protocols [1,32].

Table 1 summarizes the development of HRC systems with LLM integration and multimodal fusion mechanisms. Before 2024, most of the research was based on visual modalities such as skeletons and RGB-D, and the modalities were either used separately or combined with simple heuristic rules. The literature from 2024 to 2026 marks a clear step forward towards complex neural architectures.

The current state of the art integrates vision, natural language and emerging channels (physiological signals, AR/VR environments) using more sophisticated mathematical approaches: Transformer cross-attention; convolutional graph networks with neural gating; and semantic alignment via Kullback–Leibler divergence. In turn, LLMs have become a central cognitive core (“Dispatcher” or “Planner”) that can reason about complex scenarios such as assembly and disassembly, plan sub-tasks and extract structured knowledge.

The validation of such systems has reached a good level of technological maturity. Most of them are evaluated on physical industrial robots actually performing real industrial tasks, not in pure simulation, and they prefer human-centric measures like cognitive load, fatigue reduction, and accuracy in correcting execution errors.

Table 1. Survey of the state of the art on multimodal fusion and LLM-based reasoning for human–robot collaboration. The proposed framework is included in the last row for direct comparison.

Work	Modality	Fusion Mechanism	LLM	Validation
[22]	Vision (RGB-D, skeleton)	None; single modality, coordinates merged post hoc	No	Real robot (ABB)
[6]	Vision + physiological (EDA)	Heuristic task-reassignment rules	No	Real robot; 20 trials
[27]	Vision + text	LLM-dispatched concatenation (neuro-symbolic)	Vicuna 32B	Real robot (UR5); 137 inferences
[3]	Voice + vision + GUI	Not specified (conceptual)	Generic GPT	None (conceptual)
[9]	Vision + voice	Sequential heuristics; voice overrides vision	GPT-3.5	Real robot
[29]	Vision + text	KL divergence + cross-entropy latent alignment	GPT-2-type	Datasets; few-shot
[30]	Text + relational graphs	Softmax gating + residual sum (LLM + GCN)	GPT/Gemini	Not specified
[7]	Vision + AR + text	Multi-head attention + RAG	GPT-4o, DeepSeek-V3	Real robots; 15 participants
[26]	Vision + voice + text + AR	Cross-attention + concatenation	GPT-4o/3.5	Real robot; 5 trials/task
[21]	Vision + text (VQA)	N/A (survey)	No	Literature review (138 works)
Proposed	Voice + vision	Confidence-weighted softmax + safety clipping	Llama 3.1 8B (local)	Simulation; 210 runs, 7 conditions

Actually, multimodal fusion is largely driven by learned strategies, which can typically be classified into three families, namely, encoder–decoder frameworks, attention-based architectures (with cross-attention being the most popular subclass), and graph neural networks [33].

These families are highly expressive but have real vulnerabilities in industrial deployment. In industrial settings, a common problem occurs when one modality degrades (e.g., due to machinery noise) and the learned layers tend to collapse to the modality that is still dominant, or to produce overconfident, poorly calibrated predictions. More regularization and calibration methods are needed to regain robustness and calibration [34].

The confidence-weighted fusion module proposed in this work, which we analyze in detail in later sections, takes a different, heuristic and interpretable path instead: it computes per-modality weights from explicit quality signals through a linear combination, a temperature-controlled softmax, and bounded clipping, so that every decision can be traced back to the signals that produced it.

3. Proposed Framework

The DT encompasses a spectrum ranging from a digital model, through a digital shadow—which is characterized by unidirectional synchronization between the physical asset and the digital model—to a complete DT characterized by real-time bidirectional synchronization between the physical asset and its digital counterpart. In this research, the CoppeliaSim environment implements the virtual layer and the execution layer of this architecture, but the physical counterpart has not yet been instantiated: the operator, the ABB YuMi cobot, and the workspace all reside entirely within this simulated environment. Therefore, the proposed system should be understood precisely as a high-fidelity simulation

framework that implements the perceptual, fusion, and reasoning logic of an operator-centered DT, and not as a fully deployed DT.

The architecture proposed in this work, shown in Figure 1 integrates a Confidence-Weighted Late Fusion engine with a Reference Resolver based on a Large Language Model (LLM), implemented in this work using Llama 3.1 8B and served locally via Ollama. Instead of treating all sensor inputs equally, it performs a mathematical evaluation of the reliability of the audio and visual signals before making any grounding decisions.

When one modality is degraded—for instance, by high ambient noise affecting speech recognition, or occlusions that may affect the vision channel—the system automatically compensates by increasing the weight of the unaffected modality. This dynamic compensation allows the robotic system to maintain a safe space, and accurate and seamless operations, effectively absorbing the ambiguity of human communication within chaotic industrial settings.

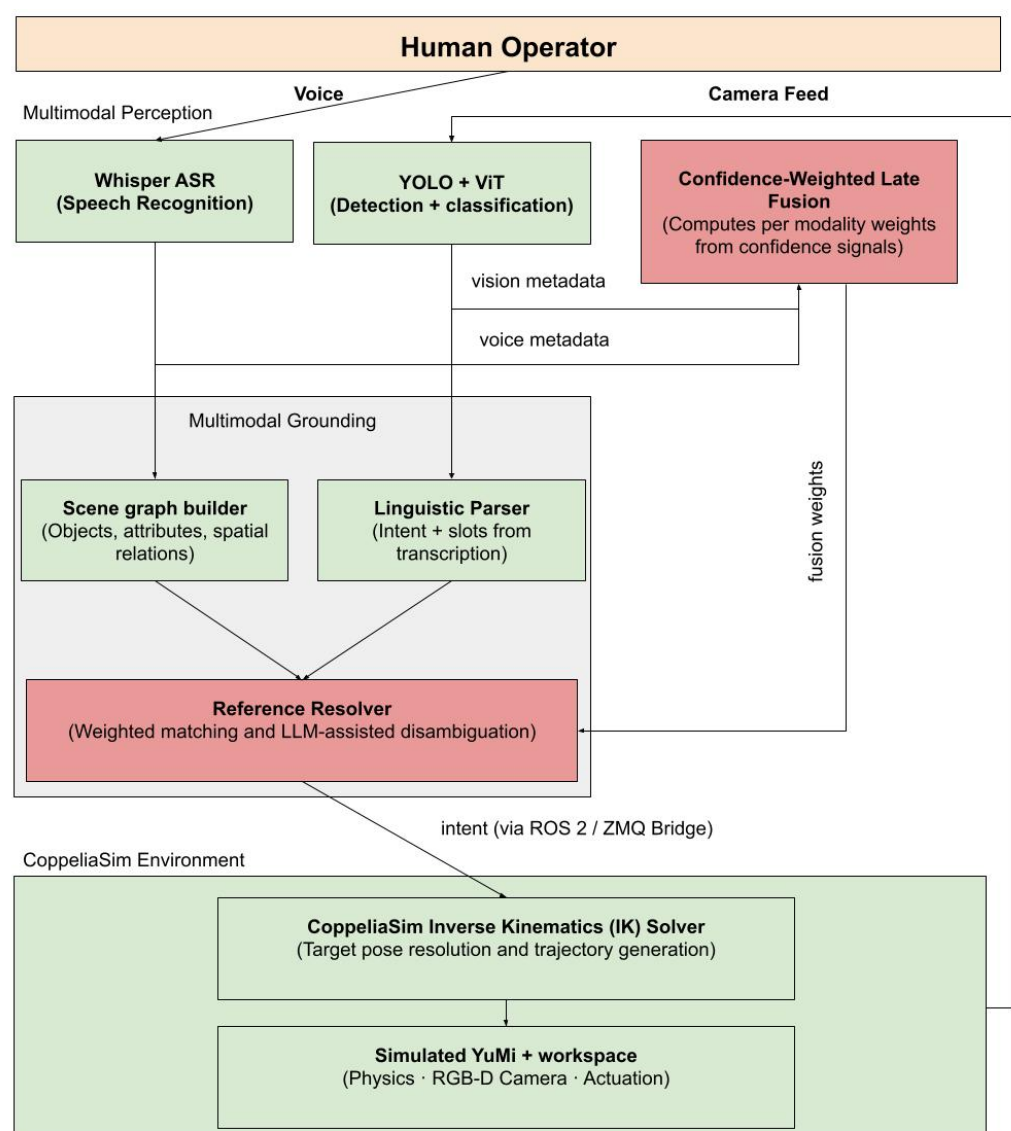


Figure 1. Proposed human-centric multimodal architecture for Industry 5.0 HRC. The boxes highlighted in red represent the contribution of this research; the “human operator” layer has been colored differently to distinguish it.

3.1. Multimodal Perception Layer

The operational cycle begins with the operator; the instructions this operator provides in natural language and within the physical workspace are captured by the multimodal perception layer. To ensure deterministic latency and data sovereignty, all processing is performed locally. This layer comprises two parallel pipelines responsible for extracting both the semantic content and the associated trust metadata.

Audio processing is performed through the integration of Whisper ASR, which, in addition to transcribing the linguistic intent, extracts signal quality metadata—such as the normalized mean logarithmic probability of the generated tokens, the signal-to-noise ratio (SNR) estimated and normalized using a sigmoid function, and the proportion of tokens decoded with low probability, which penalizes critical words decoded with hesitation.

At the same time, visual signal processing is performed using a hybrid YOLO + ViT pipeline on the data stream from an RGB-D camera; in this context, YOLO detects the geometry of objects while ViT classifies complex attributes. This pipeline generates metadata that includes the average and minimum confidence scores of YOLO's bounding boxes, ViT's normalized classification entropy—where high entropy indicates uncertainty in the classification—and the total number of detected objects, which serve as a proxy for the scene's visual complexity.

3.2. Confidence-Weighted Late Fusion Engine

The fusion process proposed in this research consists of three steps. In the first step, for each modality, a quality score is calculated using a linear combination of its positive confidence signals and its uncertainty penalties. In the second step, a softmax function with a temperature parameter converts these scores into normalized weights whose sum equals one. The temperature controls how this weight is distributed among the modalities; so, for example, a low temperature concentrates the weight almost entirely on the modality with the highest quality score, while a moderate temperature distributes the weight more gradually, thereby preserving the contribution of the weaker modality in proportion to the difference in scores. Finally, in the third step, clipping is performed to restrict the weights to the interval $[0.15, 0.85]$, thereby ensuring that $w_{vision} = 1 - w_{audio}$ and that no modality is suppressed. The complete mathematical formulation and hyperparameters are detailed in Section 4.

3.3. Multimodal Grounding and LLM-Assisted Disambiguation

The computed fusion weights are forwarded to the Reference Resolver, the cognitive core of the system powered by a local LLM.

When an operator gives an ambiguous command (for example, “pick up the screwdriver” from among three screwdrivers of different colours), the solver must select the correct object from the 3D scene graph.

The Reference Resolver computes a final candidate score ($Score_c$) for each object in the scene by combining linguistic matching (S_{ling}) and visual saliency/confidence (S_{vis}), modulated by the dynamically calculated weights:

$$Score_c = w_{audio} \cdot S_{ling} + w_{vision} \cdot S_{vis} \quad (1)$$

This architecture ensures reliable operation in a variety of industrial environments.

- **Clean Conditions:** In both modalities, the data quality is high; however, because Whisper produces consistently higher confidence scores on undegraded speech than the visual pipeline does on detection, the weights favor the audio channel

($w_{audio} = 0.841$), while the vision channel retains the minimum contribution guaranteed by the safety clipping.

- **Acoustic Degradation:** In the presence of intense noise such as that from machinery, which would be typical of an industrial plant, the SNR and Q_{audio} decrease, and the fusion engine shifts the weight toward the vision channel in proportion to the drop; under severe degradation, the weights converge to equilibrium ($w_{vision} = 0.508$). Even with a partially corrupted transcript, the resolver relies on visual context and spatial reasoning to recover the correct intent.
- **Visual Occlusion:** When the camera is partially obstructed or it is poorly lit, the ViT entropy increases sharply and Q_{vision} drops; therefore, the weights shift toward the audio ($w_{audio} = 0.85$), and the resolver decides to prioritize the audio signal over the visual signal.
- **Double Degradation:** In the event of simultaneous failure of both modalities, the overall combined confidence metric remains critically low. Rather than executing an unsafe action, the resolver detects this threshold and safely triggers a request for verbal confirmation from the operator.

3.4. Simulation and Execution Environment

Once the Reference Resolver identifies the highest-scoring candidate and deduces the precise spatial intent, the command is transmitted via a ROS 2/ZMQ middleware bridge.

This ensures asynchronous, reliable and low-latency data transfer to the physical execution layer.

The execution domain is materialized within a high-fidelity CoppeliaSim environment. To optimize computational efficiency, trajectory generation is delegated to the native CoppeliaSim Inverse Kinematics (IK) solver. The IK solver converts the target Cartesian coordinates to the exact joint configuration for the simulated ABB YuMi collaborative robot.

As the cobot executes the task, the physical state of the workspace changes. This action is immediately captured by the simulated RGB-D camera that continuously updates the multimodal perception layer, thus closing the loop of the system operation.

4. System Implementation

The proposed framework is fully implemented as a distributed system of ROS 2 Jazzy Jalisco nodes written in Python 3.12.3, deployed on a workstation running Ubuntu 24.04 LTS under the Windows Subsystem for Linux (WSL2). The simulation environment is hosted in CoppeliaSim Edu, interfaced to the ROS 2 graph via the official ZMQ Remote API. All inferences (including speech recognition and reasoning based on Large Language Models) are carried out locally, ensuring deterministic latency and full data sovereignty; neither the operators' audio, nor the transcriptions, nor the images of the workspace are transmitted to external services.

4.1. ROS 2 Communication Architecture and Custom Message Interfaces

The framework is structured as two ROS 2 Jazzy packages with separate and non-overlapping roles: an edge perception package containing the nodes for acquiring and characterizing the quality of the auditory and visual signals, and a cognition and execution package containing the fusion engine, the LLM-based interpreter, and the motion executor. This separation guarantees perception components are robot-agnostic and reusable across platforms, with robot-specific logic fully contained within a single package boundary.

To integrate task execution workflow and digital twin technology [16,35], the communication architecture is built on the basis of ROS's publish–subscribe model. The bridge to the CoppeliaSim simulator is based on segregated channels, where a high-bandwidth

pathway is used for depth camera data and a low-latency channel is used for poses and execution feedback [16,35].

Such an arrangement does not overwhelm the control pathway with visual data, which is a necessary measure to satisfy the real-time processing and safety requirements of collaborative manufacturing environments [16]. The framework introduces four ROS 2 topics that do not exist in the baseline pipeline and that represent the backbone of the communication in the contribution. Table 2 shows their message types and the nodes that publish and subscribe to them.

Table 2. Novel ROS 2 topics introduced by the proposed framework.

Topic	Type	Publisher	Subscriber	Role
/voice/metadata	VoiceMetadata	Whisper node	Fusion node	Audio quality metadata
/vision/metadata	VisionMetadata	Vision node	Fusion node	Visual quality metadata
/fusion/weights	FusionWeights	Fusion node	Interpreter node	Per-modality fusion weights
/yumi/plan	std_msgs/String	Interpreter node	Executor node	Structured JSON action plan

To enable strongly typed and self-documenting communication, a dedicated interface package defines three custom message types. The *VoiceMetadata* message encapsulates four float32 fields that characterize the quality of each utterance: *whisper_confidence*, defined as the exponentiated mean log-probability of the decoded tokens, normalized to the interval $[0, 1]$; *snr_estimate*, an estimated signal-to-noise ratio that is normalized via a sigmoid function centered at a reference SNR of 15 dB; *low_confidence_token_ratio*, the proportion of decoded tokens whose individual log-probability falls below a fixed threshold, thereby serving as a penalty for locally uncertain or hesitant decoding; and *duration_seconds*, the utterance duration in seconds, used exclusively as a diagnostic sanity-check signal.

The *VisionMetadata* message carries four fields derived from each processed camera frame: *yolo_confidence_mean* and *yolo_confidence_min*, the mean and minimum detection confidence scores across all bounding boxes; *ViT_classification_entropy*, the normalized Shannon entropy of the ViT classification distribution on the predicted top labels k , calculated as $H / \log k$ to bound the value to $[0, 1]$; and *num_objects_detected*, the total count of objects, which serves as a proxy for the complexity of the visual scene.

The *FusionWeights* message carries three float32 fields:

w_audio and *w_vision*, the per-modality fusion weights bounded within $[0.15, 0.85]$, and *confidence_combined*, a scalar in $[0, 1]$ reflecting the overall reliability of the current fusion decision.

A low confidence value *_combined* signals that both modalities are simultaneously degraded; downstream components use this field to decide whether to proceed autonomously or to request verbal confirmation from the operator.

All quality signals are normalized to $[0, 1]$ at the publishing node. The fusion node applies a secondary defensive clipping upon reception to guard against upstream normalization failures, decoupling the correctness of the fusion arithmetic from the correctness of the upstream perception nodes and improving fault isolation across the distributed system.

4.2. Confidence-Weighted Late Fusion: Internal Pipeline and Hyperparameter Configuration

The main scientific contribution of the proposed framework is the Confidence-Weighted Late Fusion module. This module continuously assesses the quality of each incoming stream of perceptual input and dynamically adjusts its relative contribution to the downstream grounding decision. The module runs as a dedicated ROS 2 node

at a fixed rate of 5 Hz, consuming quality metadata from both perception pipelines and publishing per-modality fusion weights to the Reference Resolver. Its internal processing is structured into five successive stages, each with a well-defined mathematical responsibility, as depicted in Figure 2.

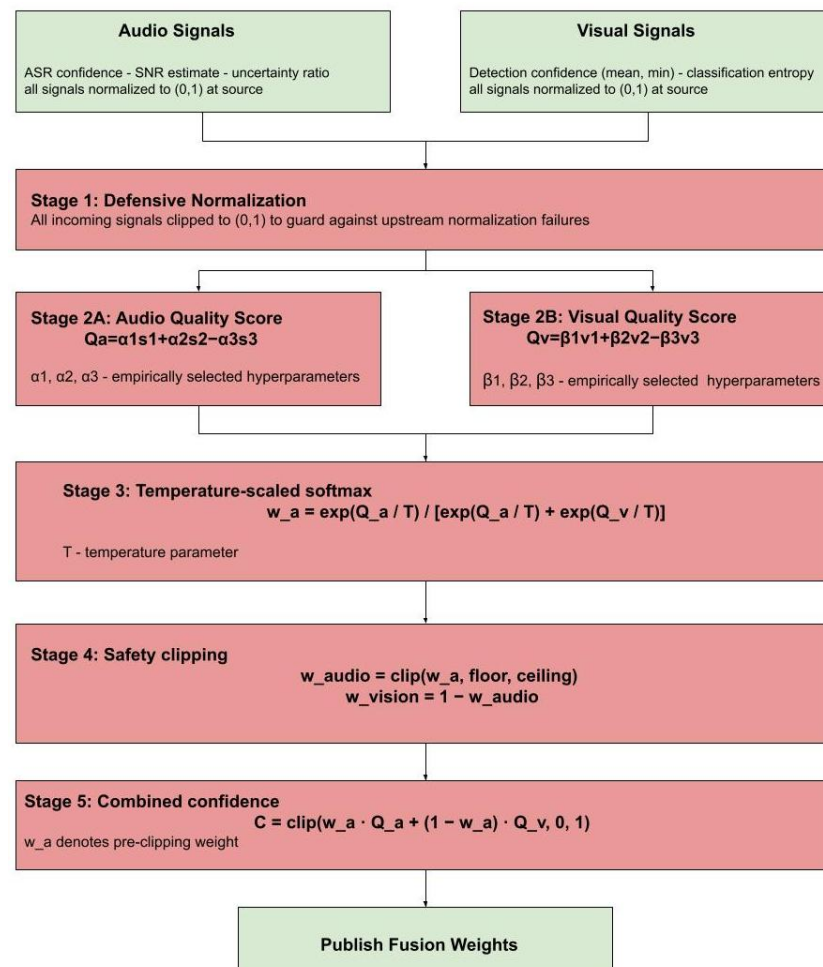


Figure 2. Internal five-stage processing pipeline of the Confidence-Weighted Late Fusion module, from per-modality quality signal ingestion to fusion weight publication. The boxes highlighted in red represent the contribution of this research

Within the broader architecture shown in Figure 1, the fusion module occupies a parallel pathway with respect to the main perception-to-action pipeline. It operates asynchronously on the quality metadata exposed by both perception nodes and publishes fusion weights that the Reference Resolver consumes at decision time. This architectural choice ensures that the computational cost of the fusion pipeline does not compound with the latency of the LLM-based grounding stage. All quality signals arrive pre-normalized to $[0, 1]$ from their respective publishing nodes, establishing a consistent numerical domain for the fusion arithmetic.

4.2.1. Stage 1: Defensive Normalization

Although all quality values are pre-normalized to the interval $[0, 1]$ prior to being emitted by their respective publishing nodes, the fusion engine applies an additional clipping operation upon receipt of each metadata message. This clipping procedure is a defensive measure designed to mitigate errors arising from faulty upstream normalization—for example, when a publishing node outputs an invalid value due to sensor malfunction or

software defects—and to decouple the correctness of the fusion algorithm employed in Stages 2–5 from potential inaccuracies introduced by upstream perception nodes.

4.2.2. Stage 2: Per-Modality Quality Scores

The normalized signals are combined into a single scalar quality score per modality, as a linear combination of positive evidence terms and negative uncertainty penalties. The audio modality quality score is as follows:

$$Q_{audio} = \alpha_1 s_1 + \alpha_2 s_2 - \alpha_3 s_3 \quad (2)$$

where s_1 is the confidence of ASR transcription, s_2 is the normalized signal-to-noise ratio estimate, and s_3 is the token ratio decoded individually with low probability. The coefficients α_1 , α_2 , α_3 are empirically chosen hyperparameters that weight the relative contribution of each signal component.

For the visual modality, the analogous quality score is as follows:

$$Q_{vision} = \beta_1 v_1 + \beta_2 v_2 - \beta_3 v_3 \quad (3)$$

where v_1 and v_2 are the mean and minimum object detection confidence scores across all bounding boxes in the current frame, and v_3 is the normalized Shannon entropy of the visual classification distribution, computed as $H / \log k$ over the labels predicted above k , where high entropy represents classification uncertainty. The coefficients β_1 , β_2 and β_3 are structurally similar to their auditory counterparts.

The additive structure of both quality scores is the same: a positive contribution for signals increasing with perceptual reliability and a negative contribution for signals increasing with uncertainty. This design ensures that Q_{audio} and Q_{vision} monotonically decrease as their respective modalities degrade, which is the property needed for the softmax stage to produce the desired weight shift.

The values of the coefficients reported in Table 3 were fixed after an ad hoc calibration process, guided during the testing phase before the formal evaluation; their fixed values were maintained for all the experimental conditions, for the sake of comparability. Instead of a black box optimization, the selection relies on three interpretable principles.

Table 3. Hyperparameter values for the confidence-weighted fusion module. All values are fixed across experimental conditions.

Parameter	Value	Role	Modality
α_1	0.5	ASR confidence weight	Audio
α_2	0.3	SNR estimate weight	Audio
α_3	0.2	Low-probability token penalty	Audio
β_1	0.4	Mean detection confidence weight	Vision
β_2	0.3	Min. detection confidence weight	Vision
β_3	0.3	Classification entropy penalty	Vision
T	0.3	Softmax temperature	Both
δ_{floor}	0.15	Minimum admissible weight	Both
δ_{ceil}	0.85	Maximum admissible weight	Both

First, within each modality, positive evidence weights were sorted according to the informational reliability of each signal. For the auditory modality, we give more weight to the ASR confidence ($\alpha_1 = 0.5$) than to the SNR estimate ($\alpha_2 = 0.3$), which directly reflects the decoder’s confidence regarding the linguistic content; the penalty for low-probability tokens ($\alpha_3 = 0.2$) is a second-order correction term.

Similarly, the average detection confidence ($\beta_1 = 0.4$) dominates the minimum confidence ($\beta_2 = 0.3$) and the classification entropy penalty ($\beta_3 = 0.3$) in the visual modality.

The softmax temperature ($T = 0.3$) was chosen to maintain sensitivity to quality differences without degenerating into a near-binary assignment; pilot tests suggested that lower values caused abrupt, unstable transitions, while higher values obscured the distinction between modalities.

Third, the limits for the clipping ($\delta_{\text{floor}} = 0.15$, $\delta_{\text{ceil}} = 0.85$) were selected so that no modality is ever fully suppressed, thus leaving residual information and avoiding discontinuities in the decision surface of the Reference Resolver. The combined confidence threshold $\tau = 0.25$ was chosen empirically based on the observed distribution of C , described in detail in Section 5.4.

4.2.3. Stage 3: Temperature-Scaled Softmax

The two scalar quality scores are converted into normalized fusion weights through a softmax function parameterized by a temperature scalar T :

$$w_a = \frac{\exp(Q_{\text{audio}}/T)}{\exp(Q_{\text{audio}}/T) + \exp(Q_{\text{vision}}/T)} \quad (4)$$

The temperature parameter modulates the decisiveness of the mapping: a low value of T produces a near-binary assignment, whereas a higher value preserves a more balanced distribution even when quality scores differ substantially. The value selected $T = 0.3$ maintains the sensitivity to quality differences while keeping complementary information from the lower modality. For numerical stability, the calculation is using a max-shift before the exponentiation, minus $Q_{\text{max}} = \max(Q_{\text{audio}}, Q_{\text{vision}})$ from either argument, which prevents floating point overflow but does not change the distribution of the output.

4.2.4. Stage 4: Safety Clipping

The raw softmax output is hard-clipped so that no modality ever gets fully suppressed, even in the face of severe degradation:

$$w_{\text{audio}} = \text{clip}(w_a, \delta_{\text{floor}}, \delta_{\text{ceil}}) \quad (5)$$

$$w_{\text{vision}} = 1 - w_{\text{audio}} \quad (6)$$

$\delta_{\text{floor}} = 0.15$ and $\delta_{\text{ceil}} = 0.85$. This constraint ensures that both modalities always have a minimum 15% contribution to the grounding decision.

This reasoning is based on two aspects: on one hand, a completely suppressed modality could still carry some residual information that would marginally improve robustness, while on the other hand, abrupt transitions between unimodal and bimodal operations would produce instabilities in the decision surface of the Reference Resolver.

When the limits of safety are reached, a warning is sent to the diagnostic stream of the system to audit offline.

4.2.5. Stage 5: Combined Confidence

The combined confidence metric $C \in [0, 1]$ is a weighted sum of the per-modality quality scores:

$$C = \text{clip}(w_a \cdot Q_{\text{audio}} + (1 - w_a) \cdot Q_{\text{vision}}, 0, 1) \quad (7)$$

One important implementation detail is that w_a in Equation (7) is the raw softmax output before clipping, not w_{audio} after clipping. This difference makes C a faithful representation of the perceptual confidence of the system irrespective of the safety correction introduced in Stage 4. When both modalities are degraded at the same time, Q_{audio} and

Q_{vision} are both low, and C is below the downstream threshold τ , which triggers the operator confirmation request discussed in Section 3. This mechanism directly provides the dual degradation safety behaviour as a function of signal quality, without the need of explicit state machines or manual thresholds per modality.

4.2.6. Output and Downstream Consumption

This pipeline outputs three quantities, w_{audio} , w_{vision} , and C , which are packaged into a `FusionWeights` message and published on `/fusion/weights` at 5 Hz. The Reference Resolver subscribes to this topic and uses the weights to modulate the relative contribution of linguistic and visual evidence in the computation of candidate object scores defined in Equation (1). The LLM-based interpreter also further uses the combined confidence C to gate command execution: commands with $C < \tau$ are rejected before reaching the language model, avoiding the cost of an inference cycle on a signal that is unlikely to yield a reliable grounding decision.

4.2.7. Design Rationale: Interpretability over Expressiveness

Inspired by the operational needs of collaborative industrial settings, we consider an explicit linear aggregation followed by a closed-form softmax function and hard clipping instead of a learned flow control network.

Each fusion decision can be completely traced to its input signals and hyperparameter values: with the published w_{audio} , w_{vision} , and C , an auditor can reconstruct the quality scores and discover which signals shifted the weights. This feature aligns with the traceability requirements of the European AI Act and the Industry 5.0 agenda on human-auditable automation.

As discussed in Section 7, an immediate future direction is to learn a gating variant trained on synthetic degradation data, with the current heuristic formulation providing a transparent and reproducible baseline to benchmark learned approaches against.

4.3. Reference Resolver Implementation: LLM-Grounded Strategist and Motion Executor

The Reference Resolver, presented in Section 3, is implemented as a cooperative two-node pipeline, as shown in Figure 3. The separation of functions is deliberate: the strategist node is responsible for all natural language understanding and linking, whilst the executor node is responsible for all motion control. Neither node encroaches on the other's responsibilities, allowing each to evolve independently.

The system receives three simultaneous inputs—the transcribed voice command from `/voz/comando`, the currently detected objects and their 3D coordinates from `/vision/detecciones`, and the real-time fusion weights from `/fusion/weights`—and processes them through five sequential stages.

First, a confidence gate filters out any command with a combined confidence C that is below a configurable threshold $\tau = 0.25$, implementing the dual degradation safety behavior described in Section 3 without calling the language model.

Commands that clear this gate are used to construct a grounded prompt that injects the visible scene, the fusion weights and the transcribed utterance into the reasoning context of the model.

This prompt is given to a locally hosted instance of Llama 3.1 8B served through Ollama [36], with the structured JSON output enforcement set and a low sampling temperature of 0.1 to constrain variability in the output.

The response is analyzed and validated with respect to the expected format. Incorrectly formatted outputs are detected and an empty plan is returned.

During validation, a whitelist is used to reject any object names not in the workspace inventory logging the fictitious references for offline auditing before publishing the validated plan to /yumi/plan.

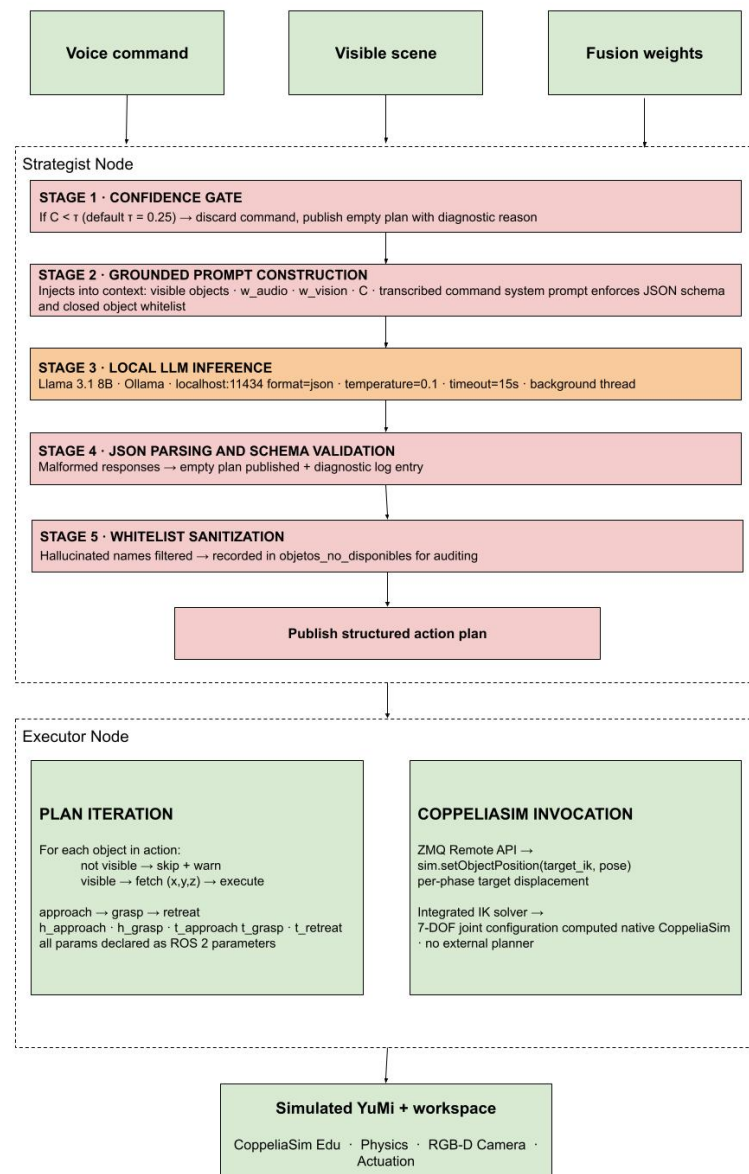


Figure 3. Internal two-node pipeline of the Reference Resolver: LLM-grounded strategist and Coppeliasim motion executor. The tables shown in red are part of the contribution of this paper

The executor node subscribes to “/yumi/plan” and iterates through the ordered list of actions. It checks whether the object is visible in the current scene for each target object. If no object is detected then the action is skipped and a warning is issued. For visible targets, the executor sends successive target pose displacements to Coppeliasim via the ZMQ Remote API to perform a three-phase approach–grasp–retreat trajectory, delegating joint configuration to the simulator’s internal Inverse Kinematics solver. All timing and height parameters are declared as ROS 2 parameters so they can be tuned between experimental conditions without recompilation.

The perception-to-action loop is closed by this pipeline, together with the fusion module in Section 4.2: the fusion weights control the confidence gate in Stage 1 and the

grounded context in Stage 2, such that the reasoning of the language model is always conditioned on the current confidence of each perceptual channel.

5. Experimental Setup

This section details the experimental setup, encompassing the simulation environment, the degradation protocols, and the experimental conditions employed to evaluate the framework under controlled sensory conditions.

5.1. Simulation Environment

The framework is fully validated in CoppeliaSim Edu, a high-fidelity robot simulator developed by Coppelia robotics (Zurich, Switzerland). CoppeliaSim was chosen for three reasons that directly satisfy the requirements of this work. It first provides a native Inverse Kinematics solver that computes 7-DOF joint configurations in real time, without the need for an external motion planning library, thus keeping the system architecture self-contained. Second, it exposes a ZMQ Remote API that enables low-latency bidirectional communication with the ROS 2 Jazzy graph, supporting the asynchronous control architecture described in Section 3. Third, its physics engine (Bullet 2.78) and integrated RGB-D camera sensor provide a sufficiently realistic visual feed for evaluating the YOLO + ViT perception pipeline under the degradation conditions introduced in this study.

Figure 4 shows the simulated workspace, which consists of a workbench where three colored target objects are placed: a red cube (cubo_rojo), a green cube (cubo_verde) and a blue cube (cubo_azul). Also present is a screwdriver as a distractor object not present in the command set. An overhead RGB-D camera observes the workspace from a fixed bird's eye view. The camera feed is streamed to the ROS 2 graph through the high-bandwidth ZMQ channel at a rate of up to 30 Hz. The dual-arm ABB YuMi collaborative robot is positioned at the edge of the workbench and all manipulation tasks are performed with its right arm driven by the CoppeliaSim IK solver. The inset in Figure 4 shows the field of view of the camera used by the visual detection pipeline in action.

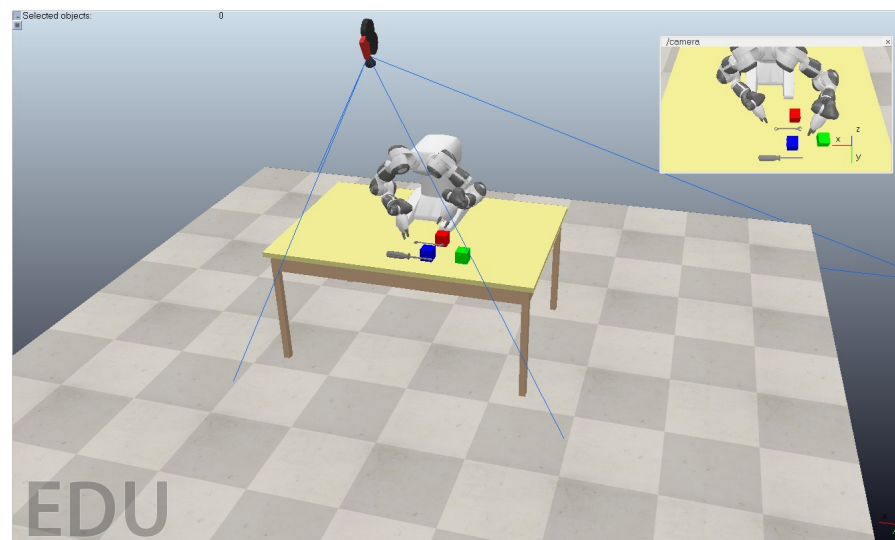


Figure 4. Simulated assembly workstation in CoppeliaSim Edu.

All experiments are performed only in this simulated environment. There is no physical robot hardware. The fully simulated implementation provides a reproducible benchmark where the scene, object positions, audio recordings, and degradation parameters are all fixed and version controlled such that every experimental condition can be exactly reproduced.

5.2. Hardware and Software Configuration

The framework runs on a workstation equipped with an Intel Core i7-12700 processor, 16 GB of RAM, and an NVIDIA RTX 4070 GPU, under Ubuntu 24.04 LTS via WSL2. Llama 3.1 8B is served locally via Ollama. All inferences are executed on-premise with no external service dependencies.

5.3. Command Set and Voice Recordings

The single operator recorded a total of 15 voice commands in a controlled acoustic environment with a directional microphone at a fixed distance of about 30 cm, sampled at 16 kHz mono PCM, the native input format of the Whisper ASR model. Five different linguistic command variants were recorded for each target object, varying verb forms and referential expressions. This lexical variation ensures that the evaluation is not biased towards a single command template. A JSON mapping file defines the ground truth object for each of the recordings. The logging system uses this to automatically determine the success of the execution.

5.4. Confidence Threshold Selection

The confidence gate in Stage 1 of the strategist node discards any command with combined confidence C below a threshold $\tau = 0.25$. Its value was chosen empirically in pilot experiments, prior to the formal evaluation. In the pilot, we observed the distribution of values C in all conditions. In clean and mildly degraded conditions, C was always above 0.40, whereas under simultaneous severe degradation of both modalities, C fell under 0.20 in the majority of frames. We thus identified the value $\tau = 0.25$ as a natural operating boundary between reliable fusion decisions and those for which no modality is providing sufficient signal quality to enable safe autonomous execution. At this threshold, the system requests verbal confirmation from the operator instead of attempting a possibly wrong manipulation. A more systematic sensitivity analysis of τ over a wider range of values is identified as an avenue for future work; the present study treats it as a fixed design parameter.

5.5. Audio Degradation Protocol

Auditory degradation was simulated by adding white Gaussian noise (AWGN), at three different SNR levels, 20, 10 and 0 dB, under controlled conditions. The noise power for each clean recording was calculated from the desired SNR and the RMS power of the clean speech signal as Gaussian noise was generated using numpy with a fixed random seed (seed = 42), to ensure full reproducibility.

$$P_{\text{noise}} = \frac{P_{\text{signal}}}{10^{\text{SNR}_{\text{dB}}/10}} \quad (8)$$

The degraded signal was clipped to $[-1, 1]$ before conversion to 16-bit PCM. The output SNR was empirically verified for all 45 degraded recordings (15 records $\times$ 3 levels) and was found to agree with the target values within ± 0.06 dB under all conditions. The speech remains intelligible at 20 dB, comparable to a noisy industrial environment at 10 dB, and at 0 dB, the noise power is equal to the signal power, corresponding to a severe degradation scenario where the ASR model produces incoherent transcriptions. In experiment mode, the degraded audio files are fed directly into the Whisper ASR node, skipping the microphone capture phase to ensure perfect reproducibility between runs.

5.6. Visual Degradation Protocol

Visual degradation was programmatically introduced at two levels applied to each camera frame before YOLO and ViT processing, selected to result in a measurable and progressive decrease in YOLO detection confidence and an increase in ViT classification entropy. During piloting, we tested a mild blur-only level (Gaussian kernel 15×15), but it was not included in the formal protocol as it produced no statistically distinguishable change in either visual quality signal relative to the clean baseline and therefore did not stress the fusion module.

For medium level, we add Gaussian blur (kernel 15×15) and reduce the brightness to 60% of the original luminance, simulating poor lighting conditions with partial occlusion. At the severe level, the medium-level transformations are augmented by a grey rectangle covering 40% of the central image area, simulating a physical obstacle partially obstructing the workspace.

5.7. Experimental Conditions and Evaluation Metric

Table 4 summarizes the seven experimental conditions evaluated in this work. Each condition is evaluated under two system configurations: *Fusion ON*, in which the Confidence-Weighted Late Fusion module operates as described in Section 3.2, and *Fusion OFF*, in which the interpreter node uses fixed weights $w_{\text{audio}} = w_{\text{vision}} = 0.5$ regardless of signal quality. This ablation study isolates the contribution of the adaptive fusion module from the rest of the pipeline.

Table 4. Experimental conditions. Each condition is evaluated under Fusion ON and Fusion OFF configurations, yielding 14 sessions of 15 commands each (210 total executions).

ID	Audio	Visual	Degradation Type
C1	Clean	None	Baseline
C2	SNR 20 dB	None	Auditory only, mild
C3	SNR 10 dB	None	Auditory only, moderate
C4	SNR 0 dB	None	Auditory only, severe
C5	Clean	Medium	Visual only, moderate
C6	Clean	Severe	Visual only, severe
C7	SNR 0 dB	Severe	Dual degradation, severe

The main evaluation metric is the command execution accuracy, which is defined as the percentage of commands where the system selected and attempted manipulation of the correct target object.

$$\text{Accuracy} = \frac{\text{correct executions}}{\text{total commands}} \quad (9)$$

A command is considered correct if the first element of the actions field in the published action plan matches the ground truth object for that recording. Commands discarded by the confidence gate are counted as failures, since, from the operator's perspective, the robot did not execute the intended action. We also track a secondary qualitative measure of grounding errors, defined as executions where the system chooses and tries to manipulate an incorrect object instead of simply abstaining. Grounding errors are particularly relevant for safety assessment, since they are active errors, not conservative abstentions.

6. Results

In this section, we report the quantitative and qualitative results obtained over the 210 experimental executions described in Section 5. The results are ordered according to the degradation axis: auditory degradation, visual degradation, and dual degradation.

We compare the Fusion ON and Fusion OFF configurations for each condition in order to remove the contribution of the adaptive fusion module.

6.1. Baseline Performance

In clean acoustic and visual conditions (C1), both system configurations achieved a perfect command execution accuracy of 100% for all 15 commands and all three target objects. The mean fusion weights for this condition were $\bar{w}_{\text{audio}} = 0.841$ and $\bar{w}_{\text{vision}} = 0.159$, which indicates the dominance of the auditory channel in the case of a clear speech signal with Whisper ASR providing high-confidence transcriptions. The overall confidence C was always above 0.50, well above the safety threshold $\tau = 0.25$, and no commands were rejected by the confidence gate. Representative execution snapshots under baseline conditions for each of the three target objects are shown in Figure 5.

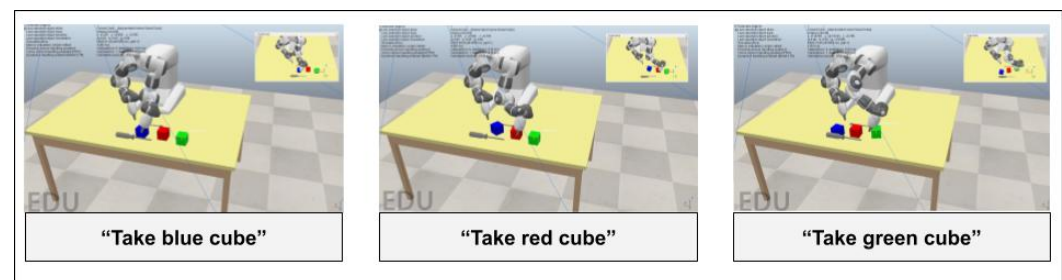


Figure 5. Representative execution snapshots under baseline condition.

6.2. Effect of Auditory Degradation

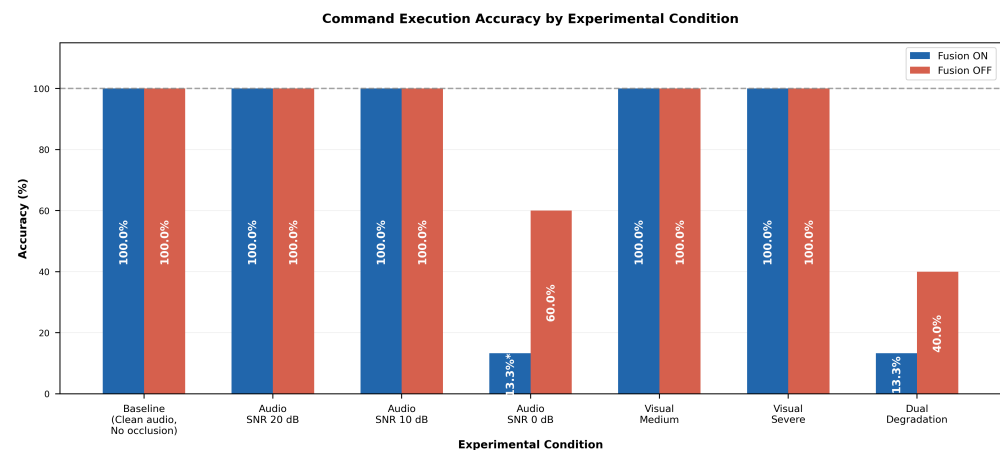
The full accuracy results over all the experimental conditions and system configurations are presented in Table 5. The results are presented as grouped bar charts in Figure 6 and the accuracy degradation curve for the auditory channel only is shown in Figure 7.

Table 5. Command execution accuracy across all experimental conditions and system configurations. GE denotes the number of grounding errors (commands where an incorrect object was selected and executed).

Condition	Audio	Visual	Config.	Success	Accuracy	GE
C1	Clean	None	Fusion ON	15/15	100.0%	0
C1	Clean	None	Fusion OFF	15/15	100.0%	0
C2	SNR 20 dB	None	Fusion ON	15/15	100.0%	0
C2	SNR 20 dB	None	Fusion OFF	15/15	100.0%	0
C3	SNR 10 dB	None	Fusion ON	15/15	100.0%	0
C3	SNR 10 dB	None	Fusion OFF	15/15	100.0%	0
C4	SNR 0 dB	None	Fusion ON	2/15	13.3%	0
C4	SNR 0 dB	None	Fusion OFF	9/15	60.0%	3
C5	Clean	Medium	Fusion ON	15/15	100.0%	0
C5	Clean	Medium	Fusion OFF	15/15	100.0%	0
C6	Clean	Severe	Fusion ON	15/15	100.0%	0
C6	Clean	Severe	Fusion OFF	15/15	100.0%	0
C7	SNR 0 dB	Severe	Fusion ON	2/15	13.3%	0
C7	SNR 0 dB	Severe	Fusion OFF	6/15	40.0%	5

Under mild auditory degradation (C2, SNR 20 dB), both Fusion ON and Fusion OFF maintained 100% accuracy. At moderate degradation (C3, SNR 10 dB), this finding suggests that, at these noise levels, Whisper ASR produces reliable enough transcriptions to allow the LLM to resolve the correct target object, regardless of whether the fusion weights are adaptive or fixed. The mean $\bar{w}_{\text{audio}}$ values decreased progressively from 0.796 at SNR 20 dB to 0.753 at SNR 10 dB, indicating that the fusion module detected the progressive

degradation and adjusted the weights accordingly even when the downstream accuracy was not affected yet.



* Confidence gate active: commands rejected due to low combined confidence ($C < 0.25$)

Figure 6. Command execution accuracy by experimental condition for Fusion ON and Fusion OFF configurations. Under clean and mildly degraded conditions, both systems perform equivalently. The performance gap emerges exclusively under severe audio degradation and dual degradation, where Fusion ON activates the confidence gate to prevent unsafe execution.

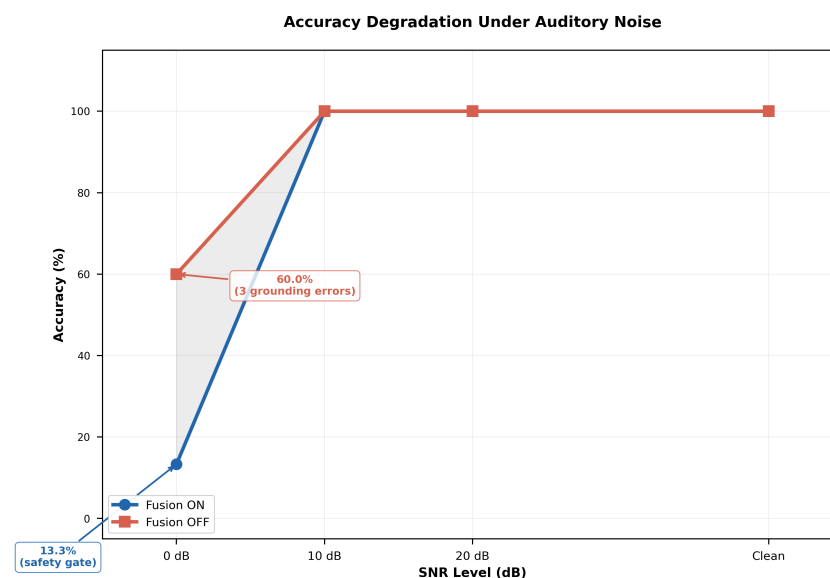


Figure 7. Accuracy degradation curve under auditory noise for isolated audio degradation conditions. Both systems maintain 100% accuracy down to SNR 10 dB. At SNR 0 dB, Fusion ON drops to 13.3% due to confidence gate activation, while Fusion OFF reaches 60.0% but incurs three grounding errors (incorrect object selections).

In the case of severe auditory degradation (C4, SNR 0 dB), the two configurations differ significantly. Fusion ON achieved 13.3% accuracy (2/15 commands), whereas Fusion OFF achieved 60.0% accuracy (9/15 commands). This counter-intuitive result is worthy of careful interpretation. With Fusion ON, the combined confidence C was below $\tau = 0.25$ for 13 of 15 commands, which were rejected by the confidence gate before reaching the language model. These rejections are considered as failures in the accuracy metric. For Fusion OFF, the fixed weights $w_{\text{audio}} = w_{\text{vision}} = 0.5$ resulted in higher C in some frames and thus more commands passing the gate. However, three of the commands that Fusion OFF attempted to run returned grounding errors, in which the system chose the wrong object to play with and played with it. Fusion ON, on the other hand, had no grounding errors

in all 210 executions. This distinction is crucial for safety assessment in an industrial environment: conservative abstention is preferable to active incorrect manipulation.

6.3. Effect of Visual Degradation

Under isolated visual degradation with clean audio (C5 and C6), Fusion ON and Fusion OFF achieved 100% accuracy regardless of the severity of the degradation. Under these conditions, the fusion module assigned $\bar{w}_{\text{audio}} = 0.849$ to 0.850 for both visual degradation levels, effectively routing the grounding decision through the clean audio channel.

The result shows that our framework can degrade gracefully when only one modality is corrupted; the fusion engine correctly recognizes the reliable channel and puts the weight on it, while the executor can retrieve the object coordinates from visual detections even with lower confidence.

6.4. Effect of Dual Degradation

Meanwhile, with both channels severely degraded (C7, SNR 0 dB + visual severe), Fusion ON obtains 13.3% accuracy and zero grounding errors, versus 40.0% and five grounding errors with Fusion OFF. The mean fusion weights under C7 were $\bar{w}_{\text{audio}} = 0.539$ and $\bar{w}_{\text{vision}} = 0.461$, indicative of a forced equilibrium where neither modality is of high enough quality to take over.

In 13 out of 15 commands, the combined confidence C fell below $\tau = 0.25$ under Fusion ON, and in all cases, this triggered the safety gate. This behavior implements the dual degradation safety scenario: the system does not take an action that has uncertain grounding and implicitly asks the operator for confirmation.

6.5. Adaptive Fusion Weight Behaviour

Figure 8 shows the mean fusion weights $\bar{w}_{\text{audio}}$ and $\bar{w}_{\text{vision}}$ per condition under the Fusion ON configuration. The figure shows three regions, with different behavior in each. For clean conditions and mild or moderate auditory degradation (C1, C2, C3), and for visual degradation of any severity (C5, C6), the audio channel dominates with $\bar{w}_{\text{audio}} \geq 0.75$. This is consistent with the high confidence of the ASR and the low token uncertainty when the audio is not severely degraded.

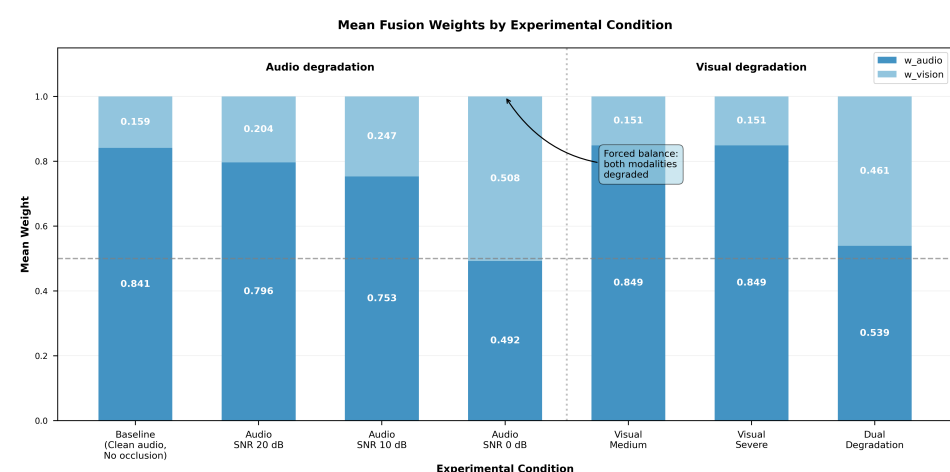


Figure 8. Mean fusion weights $\bar{w}_{\text{audio}}$ and $\bar{w}_{\text{vision}}$ per experimental condition under the Fusion ON configuration. The dashed line at 0.5 marks the balance point. Audio degradation conditions (C1–C4) show a progressive shift from audio dominance toward equilibrium. Visual and dual degradation conditions (C5–C7) show audio dominance when the audio channel is clean, collapsing toward equilibrium once both channels degrade simultaneously (C7).

For the case of severe audio degradation in isolation (C4), the weights converge to equilibrium ($\bar{w}_{\text{audio}} = 0.492$, $\bar{w}_{\text{vision}} = 0.508$) which indicates that the fusion engine correctly detects the degraded audio and shifts reliance to the visual channel.

In dual degradation (C7), the weights settle around $\bar{w}_{\text{audio}} = 0.539$, in line with the forced balance when neither modality is reliable. This progressive and bidirectional adaptation proves that the confidence-weighted fusion module responds correctly to each degradation scenario without manual intervention.

6.6. Summary

Table 5 and Figures 6–8 jointly show three important results. First, the adaptive fusion module does not cause any performance loss in favorable conditions: both configurations reach the same accuracy from C1 to C3 and from C5 to C6. Second, under severe unimodal or dual degradation, the key differentiator between the two configurations is not raw accuracy but error type: Fusion ON produces exclusively conservative abstentions (zero grounding errors across 210 executions), while Fusion OFF produces active incorrect manipulations (eight grounding errors total across C4 and C7).

Third, in Figure 8, the weight trajectories for the fusion demonstrate that the adaptive behavior is continuous and proportional to the signal quality and not a binary switch, which confirms the design rationale presented in Section 4.

7. Discussion

The experimental results in Section 6 raise three questions that need to be properly interpreted within the framework of human–cobot cooperation in Industry 5.0: the safety–accuracy dilemma that is seen in case of high degradation, the asymmetric robustness of the auditory and visual modality channels, and the relevance of an interpretable fusion method in industry.

The results should be interpreted within the context of the current study, which involved a fully simulated station, a single operator, and three differently colored target objects. The implications for generalizability are discussed in the following sections.

7.1. Safety–Accuracy Trade-off Under Severe Degradation

The most interesting result is the apparent accuracy inversion at SNR 0 dB, where Fusion ON achieves 13.3% and Fusion OFF achieves 60.0%. At first blush, one might be tempted to conclude that the fixed-weight baseline is better than the adaptive system. But this interpretation conflates two very different modes of failure.

Fusion ON results in only conservative abstentions—the confidence gate blocks commands before reaching the language model, and the robot stays in place. Fusion OFF generates active erroneous manipulations—three grounding failures in C4 and five in C7—in which the robot selects and attempts to manipulate the wrong object with high reported confidence. In an industrial assembly environment, an incorrect manipulation can damage components, disrupt production sequences or compromise the operator safety.

Considering only the subset of commands that each configuration was able to execute during the simulation (i.e., not including commands that were rejected by the confidence gate), the error rate for Fusion OFF among the attempted executions was 25% (3/12) under C4 and 45.5% (5/11) under C7, compared to a 0% error rate for Fusion ON under both conditions. These results demonstrate that, although Fusion OFF has a higher nominal accuracy, almost one in two of the active decisions under dual degradation was wrong.

We found the confidence gate threshold $\tau = 0.25$ to be an effective separator of reliable from unreliable fusion decisions. Under clean and mildly degraded conditions, C was

always greater than 0.40 and no valid commands were rejected. Under severe degradation, C fell below 0.20 in most frames and the gate fired appropriately.

We empirically show that, at $\tau = 0.25$, the separation is robust for the three-object workspace considered here. However, whether it can be extended to larger or more complex workspaces with a larger number of object classes and more perceptual ambiguity remains an open question. A systematic sensitivity analysis of τ over workspace configurations of increasing complexity is identified as a direct direction for future work.

7.2. Asymmetric Robustness Between Modalities

An interesting result is the asymmetry of degradation between audio and visual. Both setups kept 100% accuracy for any degree of isolated visual degradation (C5, C6) because the clean audio channel alone provided sufficient grounding information. In the case of isolated severe audio degradation (C4), neither configuration was able to maintain high accuracy, albeit for different reasons discussed above.

This asymmetry reflects the relative amount of information offered by each modality with respect to the current task: in a three-object workspace with color-distinct targets, a single transcribed color word (“red”, “green”, “blue”) is sufficient to uniquely identify the target object, while visual grounding alone is less discriminative under the noise levels induced in C4 as YOLO’s detection confidence degraded significantly. This suggests that the auditory channel contains a disproportionate amount of disambiguating information for color-based object selection tasks, and future work should investigate workspaces where visual context is more important for disambiguation, for example, environments with objects that are similar in color but different in geometry or spatial arrangement.

7.3. Interpretability as a Design Principle

We chose the heuristic formulation of the fusion module: linear quality scores, temperature-scaled softmax and hard clipping, in a deliberate fashion over a learned gating network. The experimental results justify this choice from an operational point of view: all the fusion decisions taken in the 210 executions are fully traceable to their contributing quality signals and hyperparameters. With a given command, an operator or a safety auditor can look at the published w_{audio} , w_{vision} and C that produced the decision, and can reconstruct exactly why the system preferred one modality or rejected the command entirely.

This interpretability is visually confirmed in Figure 8: the progressive transition from audio dominance to equilibrium as the SNR goes from clean to 0 dB is smooth, proportional, and physically interpretable without reference to internal model states. Here the gating behavior is transparent, whereas a multi-layer perceptron fusion network would present an opaque gating behavior under the same conditions.

7.4. Limitations

Some limitations of the present study should be noted. The evaluation was first performed in a fully simulated environment with a fixed three-object workspace. While the feed provided by CoppeliaSim is visually realistic enough to test the perception and fusion pipeline, the sim-to-real gap is not quantified. Further validation of the performance of the framework outside the simulation is needed with physical robot experiments with real-world acoustic noise and natural lighting variation.

In addition, the command set consisted of 15 utterances by one operator. Speaker variability, such as different accent, speech rate, and microphone placement, can affect the confidence of the Whisper transcription and the audio quality signals used by the fusion module. A multi-speaker evaluation is needed to assess generalizability to diverse operator populations.

However, the visual degradation protocol used synthetic transformations (blur, brightness reduction, central occlusion) rather than physically realistic occlusions such as moving obstacles or changing lighting conditions. These degradations are methodologically controlled and reproducible, but may not reflect the complexity of real industrial visual interference.

Finally, the hyperparameters of the fusion module (α_i , β_i , T , δ_{floor} , δ_{ceil} , τ) were chosen empirically for the current workspace. Their sensitivity to changes in workspace complexity, number of objects, and perceptual difficulty has not been systematically characterized.

7.5. Future Work

The fusion module and task complexity are the most immediate extensions. We will replace the heuristic weight formulation presented here with a learned multimodal fusion model trained on a robust degradation dataset representative of real industrial environments, moving beyond fixed hyperparameters toward a fusion policy that adapts to the statistical structure of each deployment context. At the same time, the command vocabulary will be extended from simple object selection to structured assembly instructions that involve sequential manipulation and spatial relations between components, thus stress-testing the LLM grounding stage under referential ambiguity conditions that the current three-object workspace cannot induce.

Two further extensions concern the discriminatory requirement imposed by each modality. For future evaluations, the colored objects used here will be replaced by sets of perceptually similar colors (e.g., red, purple, pink) that require the system to rely on finer visual cues and spatial reasoning rather than a single dominant color cue, as well as tools rather than objects such as cubes. The influence of the microphone type, voice timbre, and recording distance on the transcription confidence will also be systematically analyzed at the acoustic level as a complement to the assessment of multiple speakers, and will characterize how these acoustic factors propagate within the fusion module.

The longer-term roadmap addresses physical deployment and operator health. The transfer of the framework to the real YuMi platform will quantify the sim-to-real gap in the perception pipeline and the ZMQ control architecture under real-world acoustic and lighting conditions. Future work will also incorporate cognitive load and fatigue monitoring within the control loop, enabling the system to adjust its interaction modality and confirmation thresholds based on the operator's physiological state, thus closing the loop between the digital twin and the human operator that this work places at the center of the architecture, in line with the human-centric mandate of Industry 5.0.

8. Conclusions

In this paper, a multimodal digital twin approach to operator-centered collaboration between humans and cobots in virtual assembly was proposed by taking two significant steps into account, namely, a Confidence-Weighted Late Fusion module that dynamically adjusts modality dependence based on the quality of perception signal and an LLM-powered Reference Resolver module that contextualizes the ambiguous commands of natural language in the current scene and fusion scores.

The experimental evaluation on 210 executions and seven degradation conditions established three main conclusions. The adaptive fusion module does not introduce performance overhead under favorable or mildly degraded conditions: both adaptive and fixed-weight configurations achieved 100% accuracy from baseline through SNR 10 dB and across all isolated visual degradation levels. Under severe sensory degradation, the critical differentiator is not accuracy but error type: the adaptive system produced zero grounding errors across all 210 executions by activating a confidence gate that abstains from unsafe actions, while the fixed-weight baseline produced eight active incorrect ma-

nipulations. Fusion weight dynamics have verified that the adaptability of the module exhibits continuous, proportional, and physically plausible characteristics. In all 210 trials, the LLM-enhanced Reference Resolver was successful in resolving referential instructions every time it passed through the confidence gate, thus proving that grounding of natural language both visually and in terms of fusion weights works well for selecting objects in cases of uncertainty.

These results demonstrate that Confidence-Weighted Late Fusion is an effective strategy for building resilient human–cobot interfaces that degrade gracefully rather than failing catastrophically under adverse industrial conditions. The fully simulated and version-controlled implementation provides a reproducible benchmark against which future multimodal fusion approaches—including learned gating networks and cross-modal attention mechanisms—can be directly evaluated.

Author Contributions: Conceptualization, J.G.-R., J.A.-L., D.A.-V. and D.O.-P.; methodology, D.A.-V.; software, D.A.-V. and M.F.-V.; validation, D.A.-V., M.Z.-H. and M.F.-V.; formal analysis, D.A.-V. and D.O.-P.; investigation, D.A.-V.; resources, J.G.-R. and J.A.-L.; data curation, D.A.-V. and M.F.-V.; writing—original draft preparation, D.A.-V. and M.Z.-H.; writing—review and editing, D.A.-V.; visualization, M.F.-V.; supervision, J.G.-R. and J.A.-L.; project administration, J.G.-R.; funding acquisition, M.Z.-H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this study are available upon reasonable request by contacting david.alfaro@ucr.ac.cr.

Acknowledgments: The authors would like to express their gratitude to the University of Costa Rica (UCR) and University of Alicante for the institutional support provided through the approval of the research project. This support was instrumental in conducting the research on multimodal human–robot collaboration and digital twin frameworks for Industry 5.0 manufacturing environments, and contributed significantly to the development of the work presented in this article. During the preparation of this manuscript, the authors used Claude Opus 4.7, Anthropic in order to improve the English wording and translation of author-written text and to assist in the formatting of L^AT_EX tables. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AR	Augmented Reality
ASR	Automatic Speech Recognition
AWGN	Additive White Gaussian Noise
DEM	Dual-Extraction Modeling
DT	Digital Twin
GCN	Graph Convolutional Network
HRC	Human–Robot Collaboration
I5.0	Industry 5.0
IK	Inverse Kinematics
LfD	Learning from Demonstration
LLM	Large Language Model
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
PCM	Pulse-Code Modulation

RGB-D	Red–Green–Blue–Depth
ROS	Robot Operating System
SNR	Signal-to-Noise Ratio
VHSU	Vision-based Holistic Scene Understanding
ViT	Vision Transformer
VLA	Vision–Language–Action
YOLO	You Only Look Once
ZMQ	ZeroMQ (Messaging Library)

References

1. Qin, Q.; Liu, Z.; Zhong, R.; Wang, X.V.; Wang, L.; Wiktorsson, M.; Wang, W. Robot digital twin systems in manufacturing: Technologies, applications, trends and challenges. *Robot. Comput.-Integr. Manuf.* **2026**, *97*, 103103. <https://doi.org/10.1016/j.rcim.2025.103103>.
2. Hémono, P.; Nait Chabane, A.; Sahnoun, M. Leveraging digital twin and dynamic scheduling for enhanced human–robot collaboration. *Int. J. Prod. Econ.* **2026**, *293*, 109823. <https://doi.org/10.1016/j.ijpe.2025.109823>.
3. Monica, L.; Madonna, M.; Di Nardo, M.; Carra, S.; Anastasi, S. Enabling Ergonomic Human-Robot Interaction Through Large Language Modeling. *Procedia Comput. Sci.* **2026**, *277*, 2329–2338. <https://doi.org/10.1016/j.procs.2026.02.270>.
4. Yuan, G.; Liu, X.; Xiao, M.; Xiao, J.; Wang, L. Large language models in human-robot collaboration: A systematic review, trends, and challenges. *J. Manuf. Syst.* **2026**, *85*, 249–268. <https://doi.org/10.1016/j.jmsy.2026.01.011>.
5. Li, J.; Zhu, E.; Lin, W.; Yang, S.X.; Yang, S. A novel digital twins-driven mutual trust framework for human–robot collaborations. *J. Manuf. Syst.* **2025**, *80*, 948–962. <https://doi.org/10.1016/j.jmsy.2025.04.009>.
6. Chand, S.; Zheng, H.; Lu, Y. A vision-enabled fatigue-sensitive human digital twin towards human-centric human-robot collaboration. *J. Manuf. Syst.* **2024**, *77*, 432–445. <https://doi.org/10.1016/j.jmsy.2024.10.002>.
7. Ma, Y.; Tang, D.; Zhu, H.; Cai, Q.; Zhang, Z.; Wang, L.; Liu, C. Probing AR-assisted seamless HRC assembly for industry 5.0: Multi-modal mutual cognition and LLM-driven knowledge reasoning. *Robot. Comput.-Integr. Manuf.* **2026**, *97*, 103112. <https://doi.org/10.1016/j.rcim.2025.103112>.
8. Xia, G.; Ghairi, Z.; Wuest, T.; Hribernik, K.; Heuermann, A.; Liu, F.; Liu, H.; Thoben, K.D. Towards Human Modeling for Human-Robot Collaboration and Digital Twins in Industrial Environments: Research Status, Prospects, and Challenges. *Robot. Comput.-Integr. Manuf.* **2025**, *95*, 103043. <https://doi.org/10.1016/j.rcim.2025.103043>.
9. Zhang, Z.; Ji, Y.; Tang, D.; Chen, J.; Liu, C. Enabling collaborative assembly between humans and robots using a digital twin system. *Robot. Comput.-Integr. Manuf.* **2024**, *86*, 102691. <https://doi.org/10.1016/j.rcim.2023.102691>.
10. You, Y.; Cai, B.; Pham, D.T.; Liu, Y.; Ji, Z. A human digital twin approach for fatigue-aware task planning in human-robot collaborative assembly. *Comput. Ind. Eng.* **2025**, *200*, 110774. <https://doi.org/10.1016/j.cie.2024.110774>.
11. Baratta, A.; Cimino, A.; Gazzaneo, L.; Nicoletti, L.; Solina, V. Conceptual Modeling for a Simulation-based Digital Twin in Human-Robot Collaboration. *Procedia Comput. Sci.* **2025**, *253*, 3247–3256. <https://doi.org/10.1016/j.procs.2025.02.049>.
12. Cimino, A.; Longo, F.; Nicoletti, L.; Solina, V. Simulation-based Digital Twin for enhancing human-robot collaboration in assembly systems. *J. Manuf. Syst.* **2024**, *77*, 903–918. <https://doi.org/10.1016/j.jmsy.2024.10.024>.
13. Petzoldt, C.; Niermann, D.; Keiser, D.; Freitag, M. Adaptive Human-Robot Collaboration in Industrial Assembly: Augmented Reality-Supported Dynamic Task Allocation with Intuitive Process Planning. *Procedia CIRP* **2025**, *134*, 585–590. <https://doi.org/10.1016/j.procir.2025.02.173>.
14. Dhanda, M.; Rogers, B.A.; Hall, S.; Dekoninck, E.; Dhokia, V. Reviewing human-robot collaboration in manufacturing: Opportunities and challenges in the context of industry 5.0. *Robot. Comput.-Integr. Manuf.* **2025**, *93*, 102937. <https://doi.org/10.1016/j.rcim.2024.102937>.
15. Dewberry, N.K.; AlHmoud, I.; Benton, K.; Suarez, D.; Chen, Y.P.; Karkaria, V.; Tsai, Y.K.; Brock, M.; Alazzawi, N.; Chowdhury, S.; et al. A real-time VR-enabled digital twin framework for multi-user interaction in Industry 4.0. *Manuf. Lett.* **2025**, *44*, 1486–1497. <https://doi.org/10.1016/j.mfglet.2025.06.168>.
16. Mukherjee, D.; Gupta, K.; Chang, L.H.; Najjaran, H. A Survey of Robot Learning Strategies for Human-Robot Collaboration in Industrial Settings. *Robot. Comput.-Integr. Manuf.* **2022**, *73*, 102231. <https://doi.org/10.1016/j.rcim.2021.102231>.
17. Shah, R.; Doss, A.S.A.; Lakshmaiy, N. Advancements in AI-enhanced collaborative robotics: Towards safer, smarter, and human-centric industrial automation. *Results Eng.* **2025**, *27*, 105704. <https://doi.org/10.1016/j.rineng.2025.105704>.
18. Aivaliotis, S.; Papavasileiou, A.; Konstantinou, C.; Anastasiou, T.; Gkournelos, C.; Koukas, S.; Makris, S. An interactive Augmented Reality based framework assisting operators in human-robot collaborative assembly operations. *Procedia CIRP* **2024**, *126*, 170–175. <https://doi.org/10.1016/j.procir.2024.08.319>.

19. Rrapi, F.; Portelli, B.; Serra, G.; Scalera, L. From AI foundations to Large Language Models: A survey on challenges and opportunities in collaborative robotics. *Robot. Comput.-Integr. Manuf.* **2026**, *100*, 103269. <https://doi.org/10.1016/j.rcim.2026.103269>.
20. Su, B.; Cheng, F.; Lu, L.; Qing, L.; Jung, S.; Xu, X. Exploring the integration of large language models in human-robot collaboration: Effects on performance, mental stress, and trust. *Appl. Ergon.* **2026**, *134*, 104736. <https://doi.org/10.1016/j.apergo.2026.104736>.
21. Fan, J.; Zheng, P.; Li, S. Vision-based holistic scene understanding towards proactive human-robot collaboration. *Robot. Comput.-Integr. Manuf.* **2022**, *75*, 102304. <https://doi.org/10.1016/j.rcim.2021.102304>.
22. Yi, S.; Liu, S.; Xu, X.; Wang, X.V.; Yan, S.; Wang, L. A vision-based human-robot collaborative system for digital twin. *Procedia CIRP* **2022**, *107*, 552–557. <https://doi.org/10.1016/j.procir.2022.05.024>.
23. Ren, Y.; Wu, C.; Zhou, H.; Hu, X.; Miao, Z. Dual-extraction modeling: A multi-modal deep-learning architecture for phenotypic prediction and functional gene mining of complex traits. *Plant Commun.* **2024**, *5*, 101002. <https://doi.org/10.1016/j.xplc.2024.101002>.
24. Al-Sabri, R.; Gao, J.; Chen, J.; Oloulade, B.M.; Wu, Z. AutoAMS: Automated attention-based multi-modal graph learning architecture search. *Neural Netw.* **2024**, *179*, 106427. <https://doi.org/10.1016/j.neunet.2024.106427>.
25. Alzahrani, S.; Hussain, N.; Mohammad, F. Enhancing Phoneme Labeling in Dysarthric Speech with Digital Twin-Driven Multi-Modal Architecture. *Comput. Mater. Contin.* **2025**, *84*, 4825–4849. <https://doi.org/10.32604/cmc.2025.066322>.
26. Wang, X.; Dengxiong, X.; Bai, S.; Zheng, P.; Zhang, Y. VLABot: A human Vision–Language–Action models interaction framework for robotic assembly. *Robot. Comput.-Integr. Manuf.* **2026**, *100*, 103268. <https://doi.org/10.1016/j.rcim.2026.103268>.
27. Tong, X.; Zheng, H.; Wang, B.; Bao, J. Embodied Digital Twin Driven Human-Centric Collaborative Robot Behavior: Cognitive Inference of Action Strategies. *Chin. J. Mech. Eng.* **2025**, 100164, in press. <https://doi.org/10.1016/j.cjme.2025.100164>.
28. Wang, T.; Fan, J.; Zheng, P.; Yan, R.; Wang, L. Vision-Language Model-Based Human-Guided Mobile Robot Navigation in an Unstructured Environment for Human-Centric Smart Manufacturing. *Engineering* **2025**, in press. <https://doi.org/10.1016/j.eng.2025.04.028>.
29. Wu, D.; Zhao, Q.; Fan, J.; Qi, J.; Zheng, P.; Hu, J. H2R Bridge: Transferring vision-language models to few-shot intention meta-perception in human robot collaboration. *J. Manuf. Syst.* **2025**, *80*, 524–535. <https://doi.org/10.1016/j.jmsy.2025.03.016>.
30. Xiao, J.; Terzi, S. Large language model-guided graph convolution network reasoning system for complex human-robot collaboration disassembly operations. *Procedia CIRP* **2025**, *134*, 43–48. <https://doi.org/10.1016/j.procir.2025.03.007>.
31. Xiao, J.; Terzi, S. Vision-Language Model-Based Hierarchical Graph for Robotic Navigation. *Procedia CIRP* **2025**, *134*, 49–54. <https://doi.org/10.1016/j.procir.2025.03.008>.
32. Dimitropoulos, N.; Kaipis, M.; Giartzas, S.; Michalos, G. Generative AI for automated task modelling and task allocation in human robot collaborative applications. *CIRP Ann.* **2025**, *74*, 7–11. <https://doi.org/10.1016/j.cirp.2025.04.081>.
33. Han, X.; Chen, S.; Fu, Z.; Feng, Z.; Fan, L.; An, D.; Wang, C.; Guo, L.; Meng, W.; Zhang, X.; et al. Multimodal Fusion and Vision-Language Models: A Survey for Robot Vision. *arXiv* **2025**, arXiv:2504.02477. <https://doi.org/10.48550/arXiv.2504.02477>.
34. Chlon, L.; Chlon, M.; Awada, M.M. AECF: Robust Multimodal Learning via Entropy-Gated Contrastive Fusion. *arXiv* **2025**, arXiv:2505.15417. <https://doi.org/10.48550/arXiv.2505.15417>.
35. Zhang, R.; Lv, Q.; Li, J.; Bao, J.; Liu, T.; Liu, S. A reinforcement learning method for human-robot collaboration in assembly tasks. *Robot. Comput.-Integr. Manuf.* **2022**, *73*, 102227. <https://doi.org/10.1016/j.rcim.2021.102227>.
36. Ollama. Ollama: Get Up and Running with Large Language Models. 2023. Available online: <https://ollama.com> (accessed on 12 May 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.