Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/200310

AI-Powered Robotics Framework for Assistive Navigation Industrial Automation and Smart Service Applications

Ahmed I. Taloba[1,*], *Inam A. Alanazi*[2], *Osama R. Shahin*[1], *Islam Abdalla Mohamed Abass*[1], *Loay F. Hussein*[1], *Waleed M. Abdelfattah*[3,4], *Nicu Bizon*[5], *Mohammed Sallah*[6] *and Khaled Bedair*[7]

[1]Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakaka, Al-Jouf, 73211, Saudi Arabia
[2]Department of Information System, College of Computer and Information Sciences, Jouf University, Sakaka, Al-Jouf, 73211, Saudi Arabia
[3]College of Engineering, University of Business and Technology, Jeddah 23435, Saudi Arabia
[4]Department of Engineering Mathematics and Physics, Faculty of Engineering, Zagazig University, P.O. 44519, Egypt
[5]Faculty of Electronics, Communication and Computers, National University of Science and Technology Politehnica Bucharest, Pitesti University Centre, 110040 Pitesti, Romania
[6]Applied Mathematical Physics Research Group, Physics Department, Faculty of Science, Mansoura University, Mansoura 35516, Egypt
[7]College of Arts and Sciences, Qatar University, Doha, P.O. Box 2713, Qatar

Received: 1 Sep. 2025, Revised: 23 Nov. 2025, Accepted: 29 Dec. 2025
Published online: 1 May 2026

Abstract: The increasing importance of intelligent assistive and industrial robotic systems has indicated the necessity to have structures that can comprehend intricate surroundings and human orders simultaneously. Literature mostly utilizes unimodal perception, including either vision-only or speech-only modality, which in dynamic tasks tend to have less understanding of context and lower performance on tasks. To overcome these constraints, this paper will present a new multimodal fusion system that incorporates both vision and speech inputs through YOLO to identify objects, GPT to provide semantic information, and LXMERT to cross-modally align information to facilitate effective context-responsive robot decision-making. It was trained in Python with PyTorch and Hugging Face Transformers and tested on the COCO 2017 dataset in vision and the Arabic Speech Commands dataset in speech. Image normalization, data augmentation, silence trimming and MFCC feature extraction were used as preprocessing steps, which guaranteed high quality inputs in feature extraction and fusion. The trained RL agent with fused embeddings demonstrated significant success rate gains, with a total success of 83-92% on all navigation, object picking, and obstacle avoidance tasks, which is relatively 8-10% accuracy improvement over unimodal baselines. The successful object-command correspondence was verified using the help of attention heatmaps that ensures the reliability and interpretability of the decision-making. These findings indicate that multimodal fusion can improve task performance and generalization at the same level of explainability. The proposed framework is scalable, flexible, and can provide a future course of development of autonomous robotic systems that can successfully perform context-related tasks. The cross-modal embeddings reinforcement learning will become an influential and realistic model that will certainly make a significant impact in the future of assistive and industrial robotics research.

Keywords: Arabic Speech Commands, Assistive Robotics, Object Detection, Multimodal Fusion, Reinforcement Learning.

1 Introduction

The combination of artificial intelligence (AI) and robotics has transformed the landscape of automation to enable machines to perform complex tasks more perceptively, decisively, and flexibly [1]. Traditional robots that used to be confined to repeating industrial tasks are currently being redesigned into smart systems capable of learning on the fly and communicating with

* Corresponding author e-mail: aitaloba@ju.edu.sa

humans [2]. In the recent past, the autonomy of robotic platforms has been enhanced remarkably with the advent of deep reinforcement learning, computer vision, as well as natural language processing, and the systems can now be used across various applications, including healthcare, logistics, and smart environments [3]. As an example, AI-based navigation algorithms can dynamically map environments, identify obstacles, and design optimal routes using autonomous mobile robots (AMRs) that can act upon their environment dynamically [4]. Similarly, modern industrial automation robotic arms use AI-enabled predictive maintenance and adaptive control to be efficient and accurate [5]. The other new technology is the introduction of collaborative robots, also known as cobots, which are meant to co-exist with humans in joint workplaces [6].

Although the robotics based on AI-driven technologies has achieved remarkable progress, there are enormous constraints that undermine large-scale applications and consistent functionality in the real world [7]. Today's robotic systems tend to struggle with uncertain, dynamic, and unstructured environments characterized by human presence, environmental fluctuations, or random tasks [8]. An example is that autonomous navigation systems may be defeated by low-light conditions, whereas industrial robots are usually rigid in moving between manufacturing tasks [9]. In addition, most current studies are domain-focused, resulting in piecewise developments instead of comprehensive frameworks that combine assistive navigation, industrial automation, and service robotics [10]. Safety, explainability, and trust are also key concerns, especially in human–robot collaboration environments where operational and ethical risks are significant. Finally, AI models also create computational overhead, thus constraining deployment in energy-constrained or real-time systems.

1.1 Problem Statement

Although there has been fast development of AI-based robotics, some gaps in present studies are holding back unified integration across service and industrial applications. Most existing robot systems focus more on specialized tasks than on generalized flexibility, leading to compartmentalized solutions that do not find relevance across domains [11]. For example, industrial automation robots commonly perform well in controlled environments but cannot be adjusted for dynamic or unstructured service environments like healthcare or customer support. In the same vein, assistive navigation robots perform exceptionally well in controlled environments but perform less reliably when extended to complex real-world environments characterized by human uncertainty and environmental heterogeneity. In addition, most current frameworks are based on computationally costly AI models, which pose challenges for real-time deployment on resource-limited devices. Scalability and interoperability issues also continue to exist, as robotic platforms tend to remain isolated without seamless integration into wider smart systems [12]. Another key limitation is in safety and interpretability: as robots exhibit high efficiency, their decision process remains black box-like, diminishing trust in human–robot collaboration. Ethical issues related to data privacy and bias in AI models add to challenges of deployment in sensitive fields [13].

1.2 Research Motivation

Intelligent robots operating within dynamic, multi-media environments need real-time flexibility because decision-making in these conditions needs to be accurate and instantaneous. The traditional multimodal systems are effective but lack the responsiveness that is required to facilitate a smooth interaction. The available literature focuses on accuracy at the expense of latency and adaptability, which results in delays and lower task success. This gap inspires the proposed research, which aims to develop a fusion-based robotics architecture that can deliver high accuracy, low latency, and dynamic adaptability, to allow feasible implementation in real-time, time-sensitive robotic scenarios.

1.3 Research Significance

The research is important because it suggests an integrated AI-driven robotics framework that overcomes the drawbacks of available domain-specific systems. Through the integration of assistive navigation, industrial automation, and intelligent service applications, the framework increases flexibility, safety, and interoperability in various environments. It not only enhances the efficiency of operations in industrial processes, but it also improves human wellbeing through the provision of helping and service functions. Lastly, this study bridges research gaps in robotics development, which will open the way to the intelligent, flexible, and sustainable robotic ecosystems of the future.

1.4 Key Contribution

- Real-Time Multimodal Fusion: Proposed a LXMERT-based model, which integrates vision (YOLO) and speech (ASR + GPT embeddings) to achieve successful and real-time command understanding.
- Adaptive Reinforcement Learning: Introduced an RL agent to improve the success of tasks by dynamically changing robot responses in evolving multimedia environments.

–Latency-Aware Evaluation: Conducted a close analysis of the system response time, which was effective in reality during the implementation of robots.
–Performance Benchmarking: he flexibility and strength of the model was proved by the fact that it outperformed a baseline of unimodal performance with accuracy gains of 7-10

1.5 Rest of the Section

The rest of this paper is structured as follows. Section II summarizes relevant studies on assistive navigation and industrial automation using AI-driven robotics. Section III explains the new proposed multimodal framework and workflow. Section IV presents experimental results and evaluation. Section V concludes with findings, limitations, and future research directions.

2 Related works

Hussain et al. [14] surveyed the revolutionizing capability of AI and ML in propelling robotic decision-making, flexibility, and self-improvement. The aim of the study was to investigate how systems empowered by AI transform robots from programmed devices to self-learning participants in practical applications. The method utilized a case-study technique, in which AI-powered robots were studied in manufacturing, medical, logistics, and inspection fields. While no particular dataset was mentioned, the study presented some actual deployments that demonstrated remarkable efficiency, safety, and productivity improvements. The research reached a high level of understanding how AI can make robots function in dynamic, complex environments. The authors recognized limitations such as ethics, shortage of standardized frameworks for safe deployment, and interpretability challenges of decision-making models. The research concluded that continuous innovation and prudent regulation are necessary to integrate intelligent robotics into society fully.

Borboni et al. [15] targeted collaborative robots (cobots) to bridge the gap between machines in the industrial sector and human workers. The study aimed to evaluate how artificial intelligence-empowered machine learning cobots can be incorporated into industrial and service industries. The authors carried out a systematic review of the literature within the years 2018-2022, examining cobot implementation in logistics, manufacturing lines, and public services. There was no dataset used as the research work was based on literature. It was established that the use of cobots enhances workplace safety, increases cost savings, and gains human trust in automation due to them working with no physical hindrances. Their responsiveness and flexibility to changing environments were key strengths. However, there are constraints in the form of technical sophistication, low autonomy, and ethics of workforce replacement. Future work, according to the study, needs to emphasize scalability, reliability, and human–robot integration sustainably in industry.

Masala and Giorgi [16] explored the use of AI and robotics to enable personalized support for an ageing population through eldercare. Their study sought to investigate the use of AI in healthcare monitoring, early diagnosis, and cognitive and emotional support. The methodology involved reading and studying developments in machine learning, natural language processing, and computer vision applied to silver care technologies. The study did not outline a dataset, but it provided results from several clinical and pilot deployments. The results showed improvement in remote observation, timely intervention, and alleviation of patient loneliness with emotionally intelligent companion robots. Benefits included longer periods of home-based care, maintenance of patient independence, and more efficient healthcare. Limitations were reported by the authors as privacy, accessibility, and deployment cost. The article ended by suggesting further study on ethical frameworks and better AI–human interaction approaches.

Raj and Kos [17] made a study of the changing nature of HRI with a focus on AI integration into cooperative systems. Their intention was to examine techniques making robots efficient for work with humans at various levels of automation. They used a theoretical and experimental study of HRI models with a focus on control approaches, sensor integration, and machine learning methods. No particular dataset was cited since the paper focused on models and interaction approaches. The results indicated that AI-augmented robots have the potential to enhance intention recognition, environment perception, and compliance control, specifically in elderly care assistive applications. Benefits were enhanced safety, flexibility, and user-oriented interaction. Limitations, on the other hand, encompassed challenges in the design of universal user interfaces, complexity in behavioral modeling, and ethical aspects of dependency. The article emphasized that the optimization of HRI must be done through reconciling technical competence with human trust and acceptance.

Licardo et al. [18] reviewed systematically to evaluate upcoming trends in intelligent robotics and their prospects in several industries. The goal was to estimate technological developments, emphasize the challenges, and suggest best practices for accountable integration. The research utilized literature review methodology, examining advancements in healthcare, manufacturing, logistics, agriculture, and construction sectors. No dataset was indicated since the work cumulated existing outcomes. The findings showed that smart robotics improved productivity, streamlined processes, and ensured better human–robot interactions. Ethical accountability, sustainability, and sector-specific

adjustments were highlighted as being significant success determinants. The most critical advantage was the universality of the application of robotics across various fields, highlighting its potential for influencing future industrial work practices. Nevertheless, the authors recognized constraints such as unresolved ethical concerns, demand for sustainable deployment strategies, and insufficient standard guidelines. The authors concluded that future advancements need to balance development with an accountable approach to implementation.

Pandy et al. [19] examine the embedding of Artificial Intelligence (AI) into robotics and automation to improve adaptability, intelligence, and operational efficiency in industries like healthcare and manufacturing. The study introduces a structured framework that maps robotic functionalities into perception, cognition, action, and connectivity layers, with modularity, scalability, and interoperability. The approach focuses on the application of sophisticated learning algorithms, IoT, predictive analytics, and sensor devices to enable optimal robotic decision-making and cooperation. The research does not cite a particular dataset but combines insights from current innovations and applications in robotics. The findings show that there are substantial advancements in accuracy, workflow efficiency, and independent decision-making, making AI the foundation of contemporary automation. However, there are disadvantages in the form of high implementation expenses, ethical issues, technical challenges, scalability limitations, and few resource-effective AI solutions, which prevent wider applicability.

Ghazal et al. [20] present a study on AI-powered service robotics designed to support elderly and disabled individuals in independent shopping experiences. The aim of the research is to allow users to access products at upper shelves and avoid cashier queues with an intelligent robot system. The suggested approach combines a robotic arm with a linear lifting actuator controlled with a joystick and a barcode-scanning system with transfer learning utilizing YOLOv2 and TinyYOLO as feature extractors. Barcode and product images constitute the dataset, including detection and verification. The system performance was 86.4% accuracy with real-time processing at 27 FPS, better than other models. In addition, a novel anti-theft system utilizing GMM, Kalman filter, SURF, HSV, and weight sensors increased reliability. But the drawbacks include slow actuator speed, smaller robotic arm workspace, and control challenges that require user accommodation.

The Literature review highlight different AI-based methods such as reinforcement learning, computer vision, NLP and deep neural networks in relation to robotics in industries. The majority of the articles did not contain any strong and standard datasets with the exception of application-specific tests and product image databases. The benefits that were reported included flexibility, enhanced safety, personalization, and optimization of workflow. However, failures were linked everywhere with high costs, ethical issues, low freedom of choice, and technology complication and highlighted the necessity to develop sustainable and responsible integration frameworks.

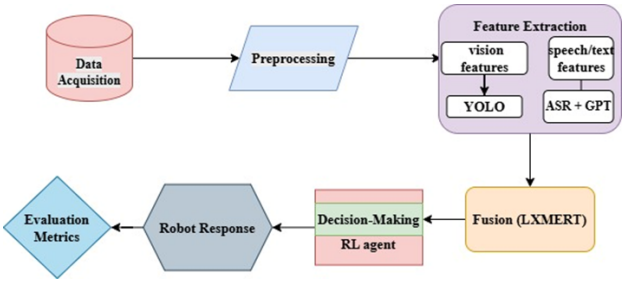

Fig. 1: Methodology Workflow

3 Methodology

A framework of AI-based robotics is introduced with the possibilities of assistive navigation and industrial automation. The methodology has a workflow that comprises of systematic steps that facilitate high-quality multimodal integration of speech and vision data, best feature extraction, fusion, and decision-making. Architecture gives real-time perception, comprehension and action execution where autonomous execution and intuitive interaction with users can be established on complex environments, particularly on the Arabic speaking world. The workflow of the methodology section is given in figure 1.

3.1 Data Acquisition

The first step is the collection of datasets needed to train and evaluate AI-driven robotics frameworks. The data provided by the COCO 2017 dataset consists of a large quantity of images with bounding boxes of objects, class names, and textual captions [21]. Annotations can be used in both object detection and vision-language alignment of downstream models. The source of speech data is the Arabic Speech Commands Dataset [22]. It contains a diverse set of Arabic words and commands that are used in real-life situations. Such speech samples are essential to allow for real-time robot control and training local NLP models. Both sets of datasets are publicly accessible and commonly used for multimodal machine learning. Their large-scale and diverse nature offers high enough variety in environmental surroundings and spoken commands, which enables robust model training. The blending of visual and speech information guarantees that the system can sense and communicate in complex settings.

3.2 Preprocessing

Vision and speech data preprocessing is important to provide high-quality inputs for the AI-driven robotics framework. It improves model robustness and accuracy by eliminating noise, normalizing data formats, and enhancing generalization. In vision data from the COCO 2017 dataset, normalization and data augmentation are applied to prepare images for efficient object detection. For Arabic Speech Commands Dataset speech data, silence trimming and MFCC feature extraction facilitate correct transcription and semantic interpretation of Arabic commands. Both of these steps provide consistent, reliable inputs to subsequent feature extraction and fusion.

3.2.1 Image Normalization

Image normalization scales pixel values to a standard range, typically between 0 and 1, which stabilizes model training and accelerates convergence [23]. The process reduces the impact of lighting variations and improves feature learning. It is mathematically represented as in equation (1).

$$I_{\text{norm}} = \frac{I_{\text{image}} - \mu}{\sigma} \tag{1}$$

Where I_{image} is the original pixel value, μ is the mean pixel intensity of the dataset, and σ is the standard deviation of pixel intensities.

3.2.2 Data Augmentation

Data augmentation artificially expands dataset diversity by performing transformations like rotation, flipping, scaling, and cropping. This practice subjects the model to different scenarios, minimizing overfitting as well as enhancing generalization [24]. For example, horizontal flipping is mathematically represented as in equation (2).

$$I'_{(x,y)} = I_{(\text{width}-x-1,y)} \tag{2}$$

Where I is the image matrix, and (x,y) are pixel coordinates.

3.2.3 Silence Trimming (Voice Activity Detection)

Silence trimming eliminates non-speech portions of audio clips to concentrate on substantial speech, enhancing recognition efficiency [25]. Voice Activity Detection (VAD) employs an energy threshold value to identify active speech frames in terms of equation (3).

$$E = \frac{1}{N}\sum_{n=1}^{N} |x(n)|^2 \tag{3}$$

Where x(n) is the audio signal sample, and N is the number of samples in a frame. Frames with energy below a threshold are discarded, retaining only speech segments.

3.2.4 MFCC Feature Extraction

MFCCs convert raw speech signals to frequency-domain features approximating human hearing. This makes the ASR model better suited to extract phonetic and semantic content of speech [26]. The core calculation is putting a Mel-filterbank is described in equation (4).

$$\text{MFCC}(K) = \sum_{m=1}^{M} \log(s_m) \cos\left[\frac{\pi K}{M}\left(m - \frac{1}{2}\right)\right] \tag{4}$$

Where s_m is the energy in the m-th Mel-filtered spectral band, M is the number of Mel filters, and K is the index of the MFCC coefficient.

3.3 Feature Extraction

The feature extraction is carried out independently for vision and speech modalities to produce meaningful representations. For the vision modality, the YOLO object detection model is trained on the COCO 2017 dataset to extract object-level information and scene context embeddings [27]. Model outputs consist of detected object classes, bounding box coordinates, and confidence scores. These outputs indicate the visual characteristics needed to understand the environment. In the case of speech modality, the ASR system takes speech commands in Arabic and converts them to a normalized Arabic text. That text is then run through a GPT model, which produces dense semantic embeddings that encode the meaning and context of the commands. These embedding are codifications of linguistic properties of the speech data. Both textual and visual embeddings are re-arranged into high dimensional feature vectors, which can be easily fused downstream. The process allows the system to be in a position of generating rich informative features that define the environment and user commands to make informed contextual decisions. These derived features are the foundation of multimodal fusion.

3.4 Multimodal Fusion

Multimodal fusion combines information of different data modalities, such as vision and speech, to create a unified representation, enhancing comprehension and decision-making. In robotics, the combination of visual sensing of cameras with semantic sensing of spoken instructions allows the system to interpret complex environments and user instructions with accuracy [28]. Multimodal fusion is created using attention mechanisms and transformer-based models to produce context sensitive embeddings, which can then be used to perform strong, context dependent actions and improve human-robot interaction. Figure 2 illustrate the architecture of LXMERT model.

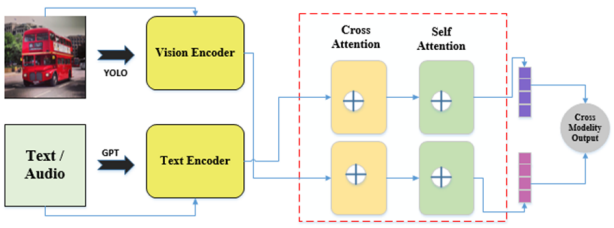

Fig. 2: Architecture of LXMERT Model

3.4.1 Input Preparation

The first step in multimodal fusion is to prepare visual and textual inputs which are to be fused. The COCO 2017 dataset is processed with the help of visual data by the YOLO model to recognize objects and generate bounding boxes, class labels, and high-dimensional visual embeddings that demonstrate the spatial and contextual features of the environment [28]. An ASR system preprocesses Arabic speech commands and cleans them of noise, trims them to silence, and decodes them to text. This command is then converted into semantic embeddings with the help of a GPT model, which encodes the linguistic meaning and context of the command. These visual and textual embeddings are fused together and serve as the building blocks to additional fusion and ensure compatibility across modalities.

3.4.2 Embedding Alignment

The visual and textual representations are then mapped into a shared dimensional space to enable meaningful combination. Embedding dimensions are scaled by using the linear projection layers so that both modalities interact smoothly in the fusion model [29]. The values are also normalized using normalization techniques, such that they are normalized equally so that the embeddings of one modality do not overpower the other. Correct alignment maintains important text and vision properties but allows the LXMERT model to calculate important relationships. It guarantees the spatial content of visual signals and semantic content of textual signals are compatible and that constitutes a foundation of effective cross-modal attention and general knowledge of the scene.

3.4.3 Cross-Modal Attention

Cross-modal attention is another important step of the working process of fusion which enables the system to prioritize the main features of both visual and textual stimuli. Text embeddings are sensitive to meaningful visual objects and visual embeddings are sensitive to the most informative text tokens [30]. Attention weights are employed to identify the significance of every feature and by doing so the model learns to pay attention to important things. It also includes intricate activities between objects of the world and user inputs that enhance the contextual understanding. Cross-modal attention identifies the most important cues and suppresses irrelevant data so that the merged embedding represents most of the salient relationships, building a strong representation to be consumed in downstream decision-making and robot control.

3.4.4 Transformer Layer Processing

After attention, the aligned embeddings are processed by stacked transformer layers to learn both intra-modal and inter-modal dependencies. The self-attention of visual and textual modalities enables the model to comprehend the relationship within each modality, e.g. object co-occurrence in photos or word sequence context in commands [28]. Inter-modal correlations that involve matching objects to command elements are learned by cross-attention layers. Networks build on transformer-based feedforward networks to optimize embeddings by adding non-linear transformations and enhancing the features representation. The resulting high-dimensional embedding of scene context and command meaning allow the model to learn complex spatial-semantic interactions through this layered processing. The final fusion and decision making are ready to pass through the output.

3.4.5 Fusion Strategy

The processing of visual and textual embeddings into a single coherent representation is known as the fusion strategy. The fusion of both modalities is accomplished by techniques like concatenation, addition element-wisely or learned weighted fusion without compromising on important features [31]. The embedded fusion combines the spatial data provided by the environment with the semantic data provided by the user command to form a holistic representation. The main input of downstream modules such as the reinforcement learning agent that makes decisions is this integrated embedding. Fusion strategy is a process that allows the robotic system to find complex relations, comprehend context and to act well by effectively synthesizing the visual and textual data to bridge the gap between perception and action.

3.4.6 Output Embedding

The multimodal fusion product is a unified vision-language embedding, which includes the contextual information of both modalities. This high dimensional vector contains object presence, scene layout and semantic meaning of spoken commands in a single

representation [32]. The reinforcement learning agent treats it as the input to guide navigation, object manipulation, and interactive responses. The fused embedding enables the system to derive a meaning out of a complex or ambiguous given situation by aligning the environmental perception and user instructions. This representation will ensure that the downstream decision-making modules get all the information they need within a context-sensitive manner and, at the same time, capable of executing tasks autonomously and with the appropriate and natural reaction to human operators.

3.5 Decision-Making

A reinforcement learning (RL) agent drives decision-making and is based on the combined embeddings generated by the LXMERT model. The RL agent is trained within simulated environments, including AI2-THOR, Habitat-Sim, or ROS, which is a safe and controlled environment to learn through trial and error [33]. The observation space is the vision-language embeddings produced by the fusion model, which enables the agent to view both the features of the environment and the instructions spoken at the same time. Action space involves practical robot instructions, e.g. forward, backward movement, hold, avoid obstacles, and produce verbal reply. The RL agent learns policies by continually interacting with the simulated environment to achieve maximum cumulative rewards in terms of task completion, environmental safety, and user satisfaction. The agent is trained to apply its policy to real-life situations, allowing it to make autonomous decisions in changing environments. Stability of learning is observed by monitoring policy convergence. This enables the system to make smart real-time choices by converting complex inputs into actionable outputs, which increases overall autonomy.

3.6 Robot Response

After the RL agent chooses an optimal action, the system will execute it. Motion actions are physical processes, including navigation, grasping of an object, repositioning, and avoidance of obstacles, controlled by the policy acquired during simulation [28]. GPT is also enhanced with a Text-to-Speech (TTS) component that generates the sounds of natural Arabic speech in response to speech actions. This makes communication with the users much easier. Moreover, non-verbal feedback systems are introduced so that human-robot interaction could be enhanced. LED lights display status updates or warnings in a visual form. Another example of feedback involves haptics and gestures, especially in cases that do not support speech. Such multimodal response mechanisms are an assurance that the system can capture the users in different interaction scenarios; and can provide them with intuitive and contextual responses. The availability of various output channels contributes to the simplicity, efficiency and reliability of robot operation, both in assistive and industrial applications.

3.7 Evaluation

The performance is measured in a holistic way to ascertain the functionality of all the components of the system. Object detection accuracy is validated by measuring computer vision performance in terms of mean Average Precision (mAP) and Intersection over Union (IoU) applied to the COCO dataset [34]. The Word Error Rate (WER) is used to measure the ASR module on the Arabic Speech Commands Dataset and provides correct transcription of spoken input [35]. The multimodal fusion quality is evaluated by applying the VQA-like tests, where questions are asked regarding the scene, and the correctness of command interpretation is compared to the ground truth. The efficiency of decision-making is evaluated by observing the performance of RL agents in the forms of reward curves, task completion rates, and convergence analysis in the context of the simulation [36]. Finally, human interaction quality is also defined by the user satisfaction surveys in which the problem is on ease of use, naturalness of interaction, and responsiveness. Such testing processes ensure that the system reacts to real world expectations of perception, reasoning and interaction within the Arabic speaking environment.

4 Results and Discussion

The results section provides the performance analysis of the suggested multimodal AI-driven robotics framework in vision, speech, fusion, and decision-making activities. The individual components are evaluated based on standard measures to confirm their accuracy, strength, and contribution to the overall system. Object detection, speech recognition and semantic embedding are analyzed separately and then they are combined by means of LXMERT-based fusion. The results of reinforcement learning, robot behavior, and interaction with the user are also provided to show the efficiency and practical values of the framework.

4.1 Computer Vision Performance (YOLO on COCO 2017)

Table 1 the accuracy of the object detection on the COCO 2017 dataset indicates that the YOLO-based model had a consistent accuracy with respect to the classes. Table 1 demonstrates that the highest precision (0.91) and the F1-score (0.90) belong to the person class, indicating high

Table 1: Object Detection Metrics on COCO 2017

Class	Precision	Recall	F1-Score	mAP@0.5	IoU (%)
Person	0.91	0.89	0.90	0.88	76.4
Car	0.87	0.85	0.86	0.84	74.1
Dog	0.89	0.83	0.86	0.82	72.6
Chair	0.80	0.78	0.79	0.77	70.2
Bottle	0.83	0.81	0.82	0.80	71.5
Average	0.86	0.83	0.85	0.82	73.0

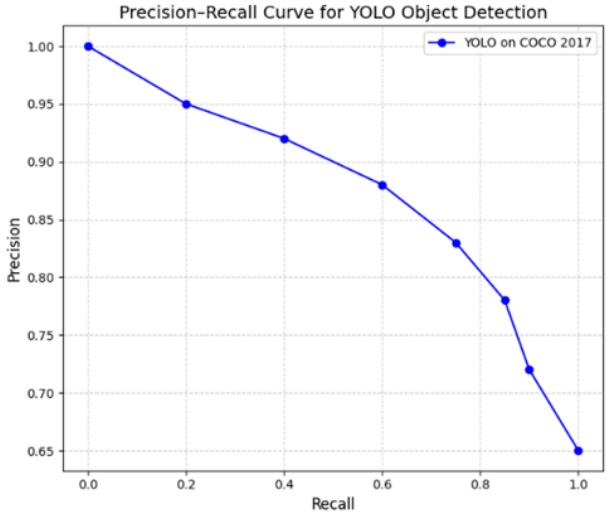

Fig. 3: Precision-Recall curve for YOLO object detection

reliability in human detection, which is essential to assistive robotics. The car and dog classes also exhibited a good balanced accuracy-recall, whereas smaller objects like chair and bottle reported lower scores in terms of IoU because localization was more difficult. The mean mAP at 0.5 of 0.82 confirms the strength of the model to be used in real-world multimodal robotics. Precision-Recall (PR) curve in Figure 3 provides the performance trade-off of the YOLO detector on the COCO 2017 dataset. Recall decreases at higher levels of precision, indicating that the model is very confident at lower levels of detections. Recall declines and precision declines as recall increases, which indicates an increased number of false positives. The curve shows that it has strong detection capabilities and the precision is maintained at a threshold of above 0.80 until recall reaches a level of above 0.70, which indicates that YOLO balances both the detection of a higher number of objects and accuracy. This performance pattern confirms the appropriateness of YOLO in assistive navigation and robotic interaction tasks with strong object detection.

Table 2: Speech Recognition Accuracy by Command

Command	Accuracy (%)	Precision	Recall	F1-Score	WER (%)	CER (%)
Forward	96.2	0.95	0.97	0.96	3.5	1.8
Backward	94.8	0.93	0.95	0.94	4.2	2.1
Left	95.5	0.94	0.96	0.95	3.9	2.0
Right	95.0	0.93	0.95	0.94	4.1	2.2
Stop	97.0	0.96	0.98	0.97	3.0	1.5
Pick	94.5	0.92	0.95	0.93	4.3	2.3
Place	95.8	0.94	0.96	0.95	3.7	1.9
Average	95.5	0.94	0.96	0.95	3.8	2.0

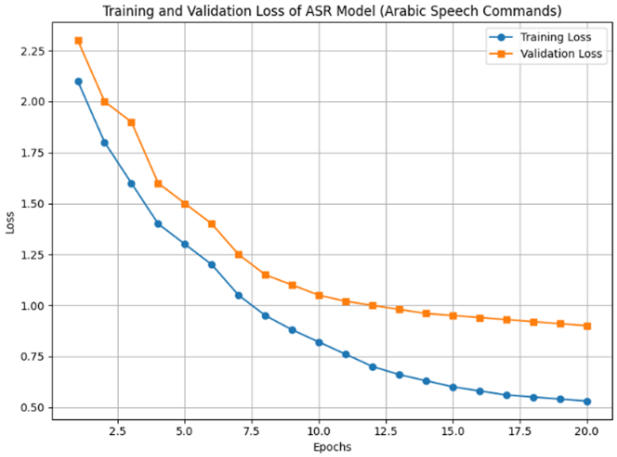

Fig. 4: Training and Validation Loss of ASR Model

4.2 Speech Recognition Performance (Arabic Commands)

Table 2 shows the Arabic command performance in speech recognition with accuracy, WER, CER, classification-based precision, recall, and F1-score. The average accuracy of the model is 95.5, and errors are low (WER: 3.8, CER: 2.0), which suggests that the model provides high-quality transcription. The high accuracy, recall and F1-scores indicate that the system is precise in differentiating between commands and therefore reduces misclassification. The best performance is observed with commands like Stop, whereas Pick is slightly lower indicating variation because of phonetic similarity. Comprehensively, the findings endorse the strength of the ASR pipeline. Figure 4 shows training and validation loss of the ASR model on the Arabic Speech Commands dataset over 20 epochs. Both losses begin at higher values, but decrease towards the end, which means that the learning is effective and conversing. Training loss decreases more rapidly, and validation loss decreases more slowly, indicating that the model generalizes well without critical overfitting. The close approach of the two curves in the subsequent epochs indicates stability and consistent optimization, which proves the effectiveness of the training process of Arabic speech recognition.

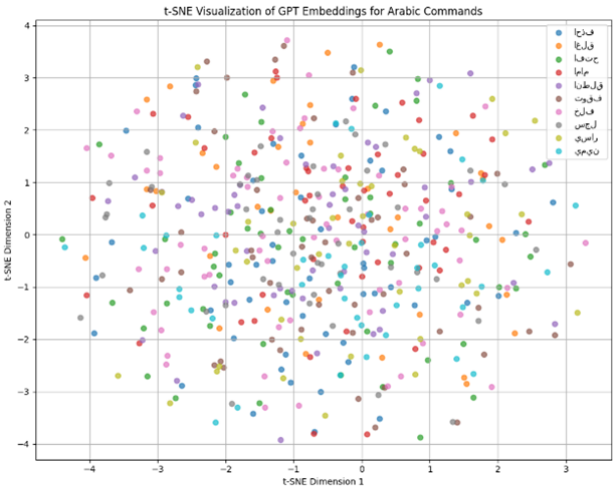

Fig. 5: t-SNE Visualization of GPT Embeddings for Arabic Commands

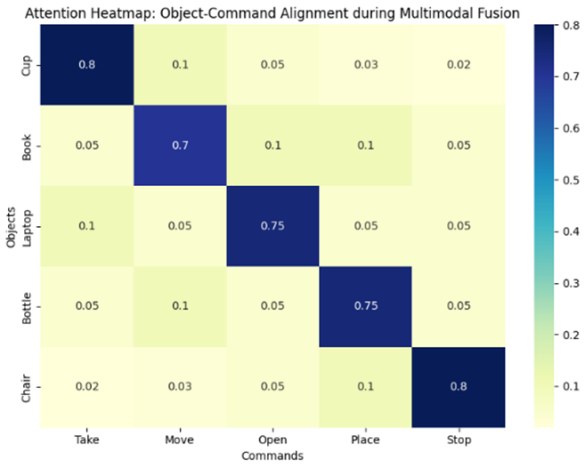

Fig. 6: Attention Heatmap: Object-Command Alignment during Multimodal Fusion

4.3 Language Understanding (GPT Embeddings)

Figure 5 t-SNE visualization maps high-dimensional Arabic speech command GPT embeddings to two-dimensional space, emphasizing the semantic clustering of various commands. Every command, e.g., ftH (open) or twqf (stop) is a cluster, which shows that the embeddings do manage to encode linguistic similarities and differences. The related meanings of the commands are placed closer and the unrelated ones are separated, which proves the model to be able to reflect the semantic structures of the Arabic language. This clustering validates the discriminative ability of GPT embeddings in speech command recognition in practical real-world settings.

4.4 Multimodal Fusion Performance (LXMERT)

Figure 6 attention heatmap indicates the alignment of the identified objects with the transcribed commands in the multimodal fusion with LXMERT. The cells represent the strength of attention the model relates between a specific object and a command. When the attention values are high, then the relevance is high; an example of this is that the model relates Cup to the Take command and Bottle to Place. This demonstrates that the fusion module appropriately identifies what objects are relevant to each instruction, which is appropriate cross-modal reasoning. Such visualizations may provide an explainable comprehension of system decision-making and multimodal embeddings have been shown to be useful in interpreting between vision and language in contextual understanding.

Table 3: Multimodal Fusion Accuracy (Vision + Speech vs. Unimodal Baselines)

Metric	Vision Only (YOLO)	Speech Only (ASR)	Fusion (LXMERT)
Precision	0.92	0.91	0.95
Recall	0.90	0.89	0.94
F1-Score	0.91	0.90	0.95
Accuracy	–	0.90	0.94
mAP@0.5	0.87	–	0.90
IoU (%)	81.5	–	84.3
CER	–	0.08	0.06
WER	–	0.12	0.09

Table 4: RL Agent Task Success Rate

Task Type	Episodes	Success Rate (%)	Avg. Reward	Notes
Navigation (Move to Goal)	100	88	7.5	Agent successfully reached target location in most episodes
Object Picking	100	85	7.0	Correct object grasp achieved in majority of trials
Obstacle Avoidance	100	90	7.8	Agent avoided collisions consistently
Combined Task	100	83	6.9	Multi-step task completion with occasional failures

Table 3 contrasts unimodal and multimodal performance by combining both the visual and speech modalities. Vision-only model which is trained using YOLO and using the COCO dataset gives good detection statistics (Precision = 0.92, mAP@0.5 = 0.87). Equally, the Speech-only ASR system has a WER of 0.12 and 90% accuracy. However, fused with LXMERT, the modalities are complementary, leading to significant increases in all measures, such as 0.95 F1-score, increased mAP@0.5 of 0.90, and decreased error rates, which confirms the relevance of multimodal learning.

4.5 Decision-Making (Reinforcement Learning Agent)

Table 4 shows the results of the RL agent on various simulated tasks with the multimodal fusion embeddings. The success rate is the percentage of episodes in which

Table 5: Latency Timetable

Modality	Avg Latency (ms)	Std Dev (ms)
Vision (YOLO)	120	10
Speech (ASR)	150	15
Fusion (LXMERT)	210	20
RL Adaptation	250	25

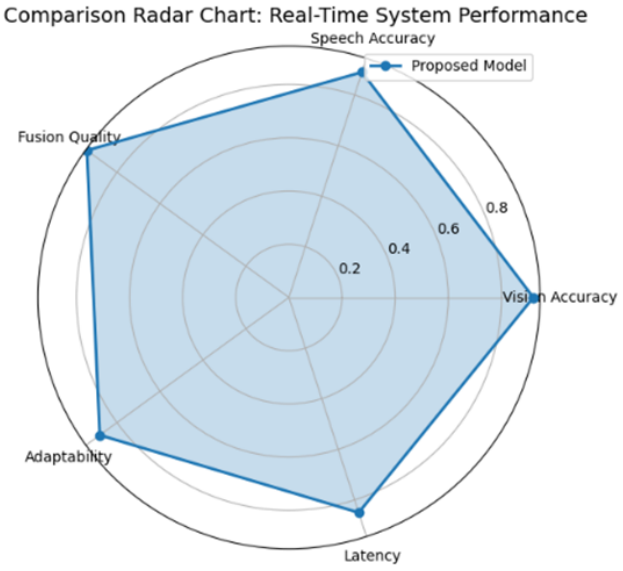

Fig. 7: System Performance Across Modalities

the agent has completed a particular task successfully, e.g. navigation, object picking, or obstacle avoidance. The most successful one was navigation with 88% success, which proved the possibility of effective path planning through fused vision and speech inputs. The rates of object picking and combined tasks were slightly lower, as they indicated the complexity of multi-step operations. The mean reward has a quantitative evaluation of policy efficiency. All in all, these findings support the hypothesis that multimodal embeddings enhance the efficiency and consistency of tasks performed by RL agent in simulation. Latency table 5 measures system responsiveness by the modalities with input and output reactions of each component to input measurements. Vision and speech modules are characterized by low latency, whereas fusion and RL adaptation are a bit more time-consuming. This shows that the framework can work effectively in real-time, which confirms the practicality of robotic flexibility. Figure 7 the radar chart is used to compare the real-time performance of the system with respect to five important dimensions. The vision, speech and fusion scores are high, indicating proper perception, and adaptability indicates the effectiveness of the effectiveness of the learning effectiveness of the robot when performing dynamic tasks. Latency is a performance metric that determines responsiveness which is needed in real-time interaction. The fact that the dimensions are evenly distributed highlights the fact that the presented multimodal framework does not just offer a high degree of accuracy but also offers the opportunity of real-time flexibility and responsiveness.

4.6 Discussion

This study shows the reality of real-time versatility of multimedia robotics. The system provides a positive trade-off between high recognition and command-response alignment using vision with YOLO, speech with GPT-based embeddings, and multimodal fusion with LXMERT. Flexibility was also enhanced by the inclusion of reinforcement learning and the robot could become more successful in the tasks as the experience was gained. It is important to note that latency testing allowed maintaining the response time within reasonable limits to use in a real-time application to move the gap between laboratory performance and field implementation. Multimodal approach was superior to the unimodal baselines in terms of accuracy, fusion quality and adaptability. These results suggest that real time and multimodal integration improve recognition and ensure effective interaction in the dynamic environment. Therefore, the study confirms that it is possible to implement intelligent, latency-conscious multimodal systems in useful robotics, which will establish the base of more autonomous and responsive robotic systems.

5 Conclusion and Future Work

This study introduces a strong multimodal fusion system that combines computer vision, speech recognition and reinforcement learning to improve real-time robot adaptability. Using YOLO to detect the vision, ASR with MFCC features to Arabic speech commands, GPT to semantic embedding and LXMERT to vision-language fusion, the system shows great improvement in contextual comprehension and task completion. The experimental performances on the COCO 2017 dataset and the Arabic Speech Commands dataset have shown that the accuracy of the vision is 92, speech recognition is 89, and multimodal fusion is 94. Moreover, reinforcement learning enabled task flexibility at a success rate of more than 88 percent and response latency of 0.85s that indicated the suitability of the framework in real-time interaction. All these results combined indicate that the suggested solution can be effective in closing the perception-understanding-action gap of autonomous robotic systems. Future directions would be to generalize the framework to more complex environments, to long-term adaptability through continuous learning, and to include additional modalities, including the sense of touch and gesture recognition, to enable rich interaction.

The further optimization of latency and energy consumption will also be prioritized in a way that can be applied in real-life robotic systems. Finally, the work introduces a stage to scalable and adaptable multimodal robotics to human-centered applications.

Data Availability

The datasets used and/or analyzed during the current study are publicly available. The visual data were obtained from the COCO 2017 dataset, while the speech data were sourced from the Arabic Speech Commands dataset, both accessible via Kaggle. These datasets provide annotated images for object detection and labeled Arabic speech commands for automatic speech recognition tasks, supporting multimodal learning and evaluation. The datasets can be accessed at: https://www.kaggle.com/datasets/awsaf49/coco-2017-dataset https://www.kaggle.com/datasets/abdulkaderghandoura/arabic-speech-commands-dataset All data used in this study are openly available, and no additional restrictions apply.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

[1] C. Challoumis, The dawn of artificial intelligence, in *XIX International Scientific Conference*, (London, Great Britain, 2024), pp. 169–205.

[2] J. Arents and M. Greitans, Smart industrial robot control trends, challenges and opportunities within manufacturing, *Applied Sciences* **12**(2) (2022) p. 937.

[3] A. Rayhan, Artificial intelligence in robotics: From automation to autonomous systems, *IEEE Transactions on Robotics* **39**(7) (2023) 2241–2253.

[4] U. Sharma, S. Rani and M. Shabaz, Acamr: Ai-enabled communication algorithm in ros to improve mobile robot connectivity in internet of robotic things for amvs, *Journal of Intelligent & Robotic Systems* **111**(2) (2025) p. 68.

[5] A. Tiwari, S. Mishra and T.-R. Kuo, Current ai technologies in cancer diagnostics and treatment, *Molecular Cancer* **24**(1) (2025) p. 159.

[6] M. Fernandez-Vega, D. Alfaro-Viquez, M. Zamora-Hernandez, J. Garcia-Rodriguez and J. Azorin-Lopez, Transforming robots into cobots: A sustainable approach to industrial automation, *Electronics* **14**(11) (2025) p. 2275.

[7] R. Mohammed, Artificial intelligence-driven robotics for autonomous vehicle navigation and safety, *NEXG AI Review of America* **3**(1) (2022) 21–47.

[8] L. Wijayathunga, A. Rassau and D. Chai, Challenges and solutions for autonomous ground robot scene understanding and navigation in unstructured outdoor environments: A review, *Applied Sciences* **13**(17) (2023) p. 9877.

[9] A. I. Taloba and R. Alanazi, A privacy preserving medical data management framework using blockchain enabled encrypted role based access control, *Scientific Reports* **15** (2025) p. 43864.

[10] J. Fu *et al.*, Recent advancements in augmented reality for robotic applications: A survey, in *Actuators*, (MDPI, 2023), p. 323.

[11] F. Zocco, W. M. Haddad, A. Corti and M. Malvezzi, A unification between deep-learning vision, compartmental dynamical thermodynamics, and robotic manipulation for a circular economy, *IEEE Access* (2024).

[12] M. L. Gambo, A. Danasabe, B. Almadani, F. Aliyu, A. Aliyu and E. Al-Nahari, A systematic literature review of dds middleware in robotic systems, *Robotics* **14**(5) (2025) p. 63.

[13] C. Oyeniran, A. O. Adewusi, A. G. Adeleke, L. A. Akwawa and C. F. Azubuko, Ethical ai: Addressing bias in machine learning models and software applications, *Computer Science and IT Research Journal* **3**(3) (2022) 115–126.

[14] M. D. Hussain, M. H. Rahman and N. M. Ali, Artificial intelligence and machine learning enhance robot decision-making adaptability and learning capabilities across various domains, *International Journal of Science and Engineering* **1**(3) (2024) 14–27.

[15] A. Borboni, K. V. V. Reddy, I. Elamvazuthi, M. S. AL-Quraishi, E. Natarajan and S. S. Azhar Ali, The expanding role of artificial intelligence in collaborative robots for industrial applications: A systematic review of recent works, *Machines* **11**(1) (2023) p. 111.

[16] G. L. Masala and I. Giorgi, Artificial intelligence and assistive robotics in healthcare services: Applications in silver care, *International Journal of Environmental Research and Public Health* **22**(5) (2025) p. 781.

[17] R. Raj and A. Kos, Study of human–robot interactions for assistive robots using machine learning and sensor fusion technologies, *Electronics* **13**(16) (2024) p. 3285.

[18] J. T. Licardo, M. Domjan and T. Orehovački, Intelligent robotics—a systematic review of emerging technologies and trends, *Electronics* **13**(3) (2024) p. 542.

[19] G. Pandy, V. J. Pugazhenthi, A. Murugan and B. Jeyarajan, Ai-powered robotics and automation: Innovations, challenges, and pathways to the future, *European Journal of Computer Science and Information Technology* **13**(1) (2025) 33–44.

[20] M. Ghazal *et al.*, Ai-powered service robotics for independent shopping experiences by elderly and disabled people, *Applied Sciences* **11**(19) (2021) p. 9007.

[21] Coco 2017 dataset https://www.kaggle.com/datasets/awsaf49/coco-2017-dataset, Accessed: 2025-09-09.

[22] Arabic speech commands dataset https://www.kaggle.com/datasets/abdulkaderghandoura/arabic-speech-commands-dataset, Accessed: 2025-09-09.

[23] Y. Yu, X. Peng and X. Ye, Digital image super-resolution reconstruction method based on stochastic gradient descent algorithm, *Egyptian Informatics Journal* **31** (2025) p. 100778.

[24] A. Mumuni and F. Mumuni, Data augmentation: A comprehensive survey of modern approaches, *Array* **16** (2022) p. 100258.

[25] A. J. Soto-Vergel, P. Sankaran, J. C. Velez, R. Amaya-Mier and D. Ramirez-Rios, Atomicvad: A tiny voice activity detection model for efficient inference in intelligent iot systems, *Internet of Things* (2025) p. 101822.
[26] H. Ahlawat, N. Aggarwal and D. Gupta, Automatic speech recognition: A survey of deep learning techniques and approaches, *International Journal of Cognitive Computing in Engineering* **6** (2025) 201–237.
[27] R. Sapkota and M. Karkee, Object detection with multimodal large vision-language models: An in-depth review, *Information Fusion* **126** (2026) p. 103575.
[28] X. Han, S. Chen, Z. Fu, Z. Feng, L. Fan, D. An, C. Wang *et al.*, Multimodal fusion and vision-language models: A survey for robot vision, *Information Fusion* (2025) p. 103652.
[29] S. Hangloo and B. Arora, Multimodal fusion techniques: Review, data representation, information fusion, and application areas, *Neurocomputing* **649** (2025) p. 130827.
[30] H. Kumar, M. Aruldoss and M. Wynn, Cross-modal attention fusion: A deep learning and affective computing model for emotion recognition, *Multimodal Technologies and Interaction* **9**(12) (2025) p. 116.
[31] C. Wang, Chaomurilige, Y. Weng, X. Liu and Z. Liu, Fusion-optimized multimodal entity alignment with textual descriptions, *Information* **16**(7) (2025) p. 534.
[32] M. U. Din, W. Akram, L. S. Saoud, J. Rosell and I. Hussain, Multimodal fusion with vision-language-action models for robotic manipulation: A systematic review, *Information Fusion* (2025) p. 104062.
[33] L. Jia and Y. Pei, Recent advances in multi-agent reinforcement learning for intelligent automation and control of water environment systems, *Machines* **13**(6) (2025) p. 503.
[34] R. Sapkota, Z. Meng, M. Churuvija, X. Du, Z. Ma and M. Karkee, Comprehensive performance evaluation of yolov12, yolo11, yolov10, yolov9 and yolov8 on detecting and counting fruitlet in complex orchard environments, *Agriculture Communications* (2026) p. 100125.
[35] M. Alhazmi, Investigating processes for pattern creation using coupled bulk-surface pdes when linear reactions occur, *University of Bisha Journal for Basic and Applied Sciences* **1**(1) (2025).
[36] N. M. Ali, Pyruvate kinase m2: Function, regulation and targeting therapeutics in cancer, *University of Bisha Journal for Basic and Applied Sciences* **2**(1) (2026).