International Transactions on Artificial Intelligence (ITALIC)
Vol. 4, No. 2, May, 2026, pp. 207−218
E-ISSN: 2963-1939 | P-ISSN: 2963-6086, DOI:10.33050

Vision-Based Pattern Recognition Models for Intelligent Human Robot Interaction in Smart Spaces

Muhamad Faizal Fazri [1], Konita Lutfiyah[2], Lukita Pasha[3], Lily Maria[4*]
[1]Department of Information Technology, Bosindo Group, Indonesia
[2]School of Business, IPB University, Indonesia
[3]Department of Digital Business CAI Sejahtera Indonesia, Indonesia
[4]Pandawan Incorporation, New Zealand
[1]faizal.fazri@raharja.info, [2]qonita13qonita@apps.ipb.ac.id, [3]lukita@raharja.info, [4]evans@pandawan.ac.nz
***Corresponding Author**

Article Info

Article history:

Submission February 27, 2026
Revised March 29, 2026
Accepted May 20, 2026
Published May 28, 2026

Keywords:

Pattern Recognition
Human Robot Interaction
Smart Spaces
CNN Transformer
Gesture Recognition systems

ABSTRACT

The rapid expansion of smart spaces has increased the need for robotic systems capable of interpreting visual cues, recognizing human behavior, and responding safely in real time. However, existing vision-based models often struggle with occlusion, lighting variation, latency constraints, and limited contextual understanding in dynamic human-centered environments. **This study** develops a hybrid vision-based pattern recognition framework that integrates Convolutional Neural Networks (CNNs), Transformer-based attention mechanisms, multi-scale feature fusion, supervised learning, and reinforcement learning. The model is trained and validated using publicly available human–robot interaction datasets and simulated smart space scenarios involving gesture recognition, object detection, activity recognition, and intention prediction. **The objective** is to enhance intelligent human–robot interaction by improving visual perception accuracy, contextual interpretation, adaptive decision-making, and real-time responsiveness in smart environments. **The proposed** framework achieves stronger performance than baseline CNN-only and Vision Transformer models, with improved accuracy in gesture recognition, object detection, activity recognition, and intention prediction while maintaining low-latency inference suitable for real-time robotic interaction. The model also demonstrates better adaptability under dynamic lighting, occlusion, and multi-person interaction scenarios. **This study** concludes that combining CNN-based local feature extraction, Transformer-based global attention, and reinforcement learning-based policy optimization provides a reliable, adaptive, and context-aware framework for intelligent robotic systems. The findings support safer and more efficient human–robot collaboration in healthcare, smart homes, collaborative workplaces, and smart city environments.

DOI: https://doi.org/10.33050/italic.v4i2.1101

1. INTRODUCTION

The rapid evolution of artificial intelligence has profoundly transformed robotics and computer vision, enabling machines to perceive, interpret, and interact with complex human-centered environments [1]. The emergence of smart spaces, including intelligent homes, healthcare facilities, collaborative workplaces,

and public infrastructures, has accelerated the need for autonomous robotic systems capable of adaptive and context-aware behavior [2]. Unlike traditional automated systems, robots operating in smart environments must interpret dynamic visual cues, recognize human gestures and intentions, and respond appropriately in real time [3]. Vision-based perception is fundamental to this process, as visual data provide rich information about objects, spatial arrangements, social interactions, and environmental changes. However, achieving robust and reliable perception in real-world scenarios remains a major challenge due to lighting variability, occlusions, background complexity, and unpredictable human behavior. These challenges underscore the necessity of developing advanced pattern recognition models that operate effectively under dynamic and uncertain conditions while maintaining computational efficiency suitable for embedded robotic platforms [3, 4]. The specific research gap motivating this study is the absence of a unified framework that simultaneously integrates local spatial feature extraction, global contextual attention, multi-scale fusion, and adaptive interaction policy learning within a single perception-to-decision architecture. Recent advances in deep learning have improved computer vision systems for object detection, gesture recognition, facial expression analysis, and scene understanding [5]. CNNs have demonstrated strong hierarchical feature extraction, while transformer-based architectures have introduced attention mechanisms that enhance contextual modeling and long-range dependency learning. However, both paradigms present well-documented complementary limitations when applied in isolation to real-time human–robot interaction: CNNs lack the capacity to model global scene dependencies, while Vision Transformers (ViTs) impose computational overhead that violates real-time latency constraints on embedded platforms. In smart spaces, robots must process continuous visual streams with minimal latency and translate perception outputs into adaptive interaction decisions [6]. Without robust contextual awareness, even highly accurate recognition models fail to deliver meaningful interaction outcomes. There is therefore a clear need for integrated frameworks that combine both paradigms while adding adaptive control mechanisms capable of continuous refinement from environmental feedback [7].

A further critical dimension is social and behavioral understanding. Effective interaction requires robots to go beyond isolated pattern recognition and interpret user intentions, predict behavioral outcomes, and adjust responses in socially acceptable ways [8]. In healthcare settings, robots must recognize patient gestures or distress signals and respond with appropriate assistance [9]. In collaborative workplaces, robots should interpret worker movements and adapt trajectories to ensure safety and efficiency. These demands require multimodal perception and learning mechanisms that refine interaction policies over time. Reinforcement learning and attention-based models offer promising solutions by enabling optimal response learning from environmental feedback [10]. However, implementing such approaches in real-time environments remains challenging due to computational constraints, limited training data diversity, and environmental unpredictability. Bridging this gap requires a framework integrating robust feature extraction with adaptive learning models [11]. This study makes four explicit scientific contributions. First, it proposes a novel hybrid CNN-Transformer architecture that formally specifies the data flow between convolutional, multi-head self-attention, and multi-scale fusion components, producing a unified perception module with mathematically defined inter-layer operations. Second, it integrates supervised perception learning with reinforcement learning interaction policy optimization within a single end-to-end trainable pipeline, with explicit specification of the state-action-reward formulation. Third, it provides an ablation study that isolates the performance contribution of each architectural component, validating the necessity of each design choice. Fourth, it evaluates the framework on four publicly available benchmark datasets that collectively span the principal visual perception tasks required for smart space deployment [12]. The proposed framework aligns with SDG 9 (Industry, Innovation, and Infrastructure) by fostering resilient industrial innovation, SDG 11 (Sustainable Cities and Communities) by enhancing safety and efficiency in smart urban environments, and SDG 3 (Good Health and Well-Being) by enabling assistive robotic solutions in healthcare contexts [13].

2. LITERATURE REVIEW

Research on intelligent robotic systems for smart spaces spans vision-based pattern recognition, human robot interaction strategies, machine learning-driven autonomous control, and real-time deployment challenges [14].

2.1. Vision-Based Pattern Recognition

Deep learning advances have significantly improved vision-based pattern recognition for robotic systems. CNNs remain widely adopted for hierarchical local feature extraction across object detection, gesture

classification, and scene understanding, demonstrating strong performance in controlled environments [15]. However, CNNs are structurally limited in modeling long-range spatial dependencies, a limitation that becomes consequential when interpreting complex multi-body human gestures or partially occluded activity sequences. Vision Transformers (ViT) and hybrid CNN-Transformer models address this limitation through self-attention mechanisms that capture global contextual dependencies within visual scenes [16]. Research demonstrates that transformer-enhanced perception models improve robustness under occlusion and variable lighting conditions [17, 18]. Lightweight architectures such as MobileNetV3-based detectors and efficient transformer variants have been proposed to address computational constraints on embedded platforms. The critical gap in this body of work is the absence of a unified framework that combines local spatial extraction, global attention modeling, multi-scale fusion, and adaptive policy learning within a single deployable architecture [19].

2.2. Human–Robot Interaction and Adaptive Control

Human Robot Interaction (HRI) research has evolved toward context-aware and socially adaptive systems in smart homes, healthcare facilities, and collaborative workplaces [20]. Vision-based interaction remains central, capturing gestures, facial expressions, posture, and manipulation cues [21]. Deep learning gesture recognition models achieve high accuracy on benchmark datasets but frequently degrade in real-world deployments due to domain shift, with reported accuracy drops of 8% to 15% between laboratory and operational conditions. Domain adaptation and self-supervised learning strategies have been explored, alongside explainable AI approaches that increase transparency in robotic decision-making [22]. For autonomous control, Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL) enable systems to learn optimal interaction policies through environmental feedback. Vision-guided policy learning frameworks allow robots to adjust navigation, interaction timing, and response strategies based on perceived patterns. Hybrid supervised-plus-RL strategies demonstrate improved performance stability over purely RL-based approaches [23]. The critical limitation of existing HRI and control research is that perception and policy components are typically developed in isolation, preventing coherent optimization of the full perception-to-decision loop [6]. The present study addresses this by designing a unified architecture in which the perception backbone and the RL policy head share learned representations, enabling end-to-end joint optimization.

2.3. Real-Time Deployment in Smart Environments and Research Gap

Real-time deployment in smart spaces introduces additional technical and practical challenges including latency constraints, energy efficiency requirements, and privacy considerations [24]. Edge computing and distributed AI frameworks reduce processing delays and enhance scalability [25]. Privacy-preserving approaches such as federated learning and adversarial robustness techniques are increasingly important for safe robotic operation in domestic and healthcare settings [26]. Table 1 maps the key limitations of prior work in each thematic area to the specific contributions of the proposed framework, establishing the positioning of this study relative to the state of the art.

Table 1. Research Gaps in Prior Work and Contributions of the Proposed Framework

Gap in Prior Work	**Contribution of This Study**
CNNs lack global contextual dependency modeling	Hybrid CNN-Transformer with multi-head self-attention integrating local and global features
ViTs impose excessive inference overhead for embedded platforms	Multi-scale fusion with lightweight transformer blocks; mean latency 32 ms
Perception and policy learning developed in isolation	Unified end-to-end architecture: shared representation for supervised perception and RL policy
No ablation study validating individual component necessity	Full ablation study quantifying independent contribution of each module
No unified framework covering gesture, object, activity, and intention tasks	Simultaneous multi-task evaluation across four HRI benchmark datasets

As shown in Table 1, the proposed framework systematically addresses five distinct limitations identified across the reviewed literature. Many prior works focus either on perception accuracy or on control optimization without addressing the full interaction loop [27]. By combining deep feature extraction, attention-based contextual modeling, and adaptive policy learning in one architecture, this study advances the state of autonomous systems for reliable operation in complex human-centered smart environments.

3. RESEARCH METHODOLOGY

The methodology develops a unified vision-based pattern recognition framework for intelligent human–robot interaction in smart spaces [28]. The research integrates a hybrid CNN-Transformer perception backbone with an adaptive reinforcement learning decision-making module and evaluates the complete system on four publicly available benchmark datasets. The methodology proceeds through four stages: dataset acquisition and preprocessing, hybrid model architecture design and formal specification, integrated training pipeline, and simulated smart space deployment [29].

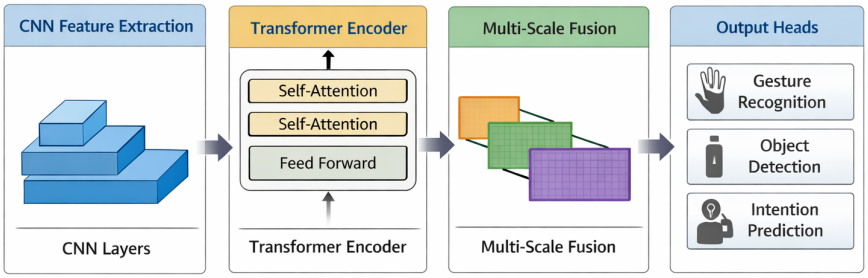

Figure 1. Model Architecture for Vision-Based Pattern Recognition

As illustrated in Figure 1, input visual frames pass sequentially through: (1) the convolutional backbone, which extracts hierarchical local spatial features at three resolution scales, (2) the multi-head self-attention module, which computes global contextual dependencies across spatial positions; (3) the multi-scale feature fusion layer, which concatenates feature maps from all three scales to produce a unified representation; and (4) the adaptive decision module, which maps fused representations to task-specific outputs through supervised classification heads and an RL policy head. The shared feature representation between perception and policy heads enables joint end-to-end optimization, which is the architectural property that distinguishes the proposed framework from prior approaches that develop these components in isolation.

3.1. Dataset Acquisition, Selection Justification, and Preprocessing

Four publicly available datasets were selected to collectively cover the principal visual perception tasks required for smart space HRI deployment.

Table 2. Dataset Characteristics and HRI Task Relevance

Dataset	Type	Samples	Resolution	Task	HRI Relevance
HRI-Gestures	Video	15,000	640×480	Gesture recognition	Command-based robot control via human gestures
RoboObject	Image	12,500	224×224	Object detection	Collaborative manipulation and workspace sharing

SmartActivity	Video	8,000	640×480	Activity recognition	Monitoring human behavior for assistive response
MultiPose	Image	10,000	256×256	Posture and intention prediction	Anticipatory interaction and safety compliance

As shown in Table 2, each dataset was selected for a specific and distinct functional reason. HRI-Gestures provides labeled video sequences of command-relevant gestures that reflect real-world robot control scenarios in collaborative workplaces. RoboObject captures the object detection task central to manipulation and shared workspace interaction. SmartActivity enables training of activity recognition for assistive response in healthcare and smart home contexts. MultiPose is selected for its posture and body-keypoint annotations that support intention prediction, a task critical for anticipatory robot behavior and safety compliance. Collectively, these datasets span the interaction scenario space defined by the smart space deployment. No single dataset covers all four tasks, the multi-dataset design was therefore necessary for comprehensive multi-task evaluation [30].

Data preprocessing involved frame-level normalization to zero mean and unit variance, spatial resizing to task-specific input dimensions, and augmentation comprising random horizontal flipping, rotation within ± 20 degrees, color jitter (brightness 0.3, contrast 0.3), and Gaussian noise injection at a signal-to-noise ratio of 20 dB to simulate realistic sensor noise. Videos were sampled at 30 FPS for real-time simulation consistency [45]. Table 3 presents the augmented dataset sizes after preprocessing.

Table 3. Dataset Statistics After Preprocessing and Augmentation

Dataset	**Original**	**Augmented**	**Final Size**	**Train / Val / Test Split**
HRI-Gestures	15,000	10,000	25,000	17,500 / 3,750 / 3,750
RoboObject	12,500	7,500	20,000	14,000 / 3,000 / 3,000
SmartActivity	8,000	6,000	14,000	9,800 / 2,100 / 2,100
MultiPose	10,000	5,000	15,000	10,500 / 2,250 / 2,250

As shown in Table 3, augmentation increased total training samples by approximately 57% across all four datasets, improving the diversity of training distributions and reducing overfitting risk. A stratified 70/15/15 split was applied consistently across all datasets to ensure representative class distributions in validation and test partitions.

3.2. Hybrid CNN-Transformer Architecture: Formal Specification

The proposed architecture consists of four interconnected modules. The convolutional backbone applies three sequential convolutional blocks with kernel sizes 3×3, batch normalization, ReLU activation, and max pooling, producing feature maps $F_1 \in \mathbb{R}^{H_1 \times W_1 \times C_1}$, $F_2 \in \mathbb{R}^{H_2 \times W_2 \times C_2}$, and $F_3 \in \mathbb{R}^{H_3 \times W_3 \times C_3}$ at three spatial scales. The Multi Head Self Attention (MHSA) module processes the flattened sequence of spatial tokens from F_3 using scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where $Q = XW_Q$, $K = XW_K$, $V = XW_V$ are the query, key, and value projections of the flattened feature sequence X, d_k is the key dimension, and $W_Q, W_K, W_V \in \mathbb{R}^{d_{model} \times d_k}$ are learned projection matrices. Multi-head attention with $h = 8$ heads is applied:

$$\text{MHSA}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O, \quad \text{head}_i = \text{Attention}(XW_{Q_i}, XW_{K_i}, XW_{V_i}) \tag{2}$$

The multi-scale feature fusion layer concatenates the upsampled attention output $A \in \mathbb{R}^{H_1 \times W_1 \times C_A}$ with F_1 and F_2 via bilinear upsampling and channel-wise concatenation, producing a fused representation $Z \in \mathbb{R}^{H_1 \times W_1 \times (C_1+C_2+C_A)}$ that integrates local and global features across all scales.

3.3. Integrated Supervised and Reinforcement Learning Pipeline

The training pipeline integrates supervised learning for perception tasks and reinforcement learning for interaction policy optimization within a unified optimization procedure. The supervised component minimizes cross-entropy loss for classification tasks and mean squared error for regression-based intention prediction:

$$\mathcal{L}_{supervised} = -\sum_{c=1}^{C} y_c \log \hat{y}_c \quad \text{(classification)} \qquad \mathcal{L}_{supervised} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad \text{(regression)} \tag{3}$$

The reinforcement learning component formulates the interaction policy as a Markov Decision Process (MDP) defined by state space $\mathcal{S}$ (fused visual representations Z), action space $\mathcal{A}$ (discrete interaction commands: approach, assist, wait, retreat), reward function $r(s, a)$ (positive reward for successful task completion within latency threshold, negative reward for safety violations or misclassification), and discount factor $\gamma = 0.95$. The policy $\pi_\theta(a|s)$ is optimized using Proximal Policy Optimization (PPO):

$$\mathcal{L}_{PPO}(\theta) = \mathbb{E}_t\Big[\min\Big(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t\Big)\Big] \tag{4}$$

where $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{old}}(a_t|s_t)$ is the probability ratio, $\hat{A}_t$ is the generalized advantage estimate, and $\varepsilon = 0.2$ is the clipping parameter. The total training loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{supervised} + \lambda \mathcal{L}_{PPO} \tag{5}$$

where $\lambda = 0.1$ weights the RL component relative to the supervised perception loss. This formulation ensures that the shared feature representation Z is optimized simultaneously for recognition accuracy and interaction policy quality, enabling end-to-end joint learning. Table 4 summarizes the training configuration.

Table 4. Training Parameters and Evaluation Metrics

Component	Settings	Purpose
Optimizer	Adam, LR $= 1 \times 10^{-4}$, weight decay $= 1 \times 10^{-4}$	Efficient gradient-based optimization with regularization
Batch Size	32	Balances GPU memory usage and training stability
Epochs	100 (supervised); 500 episodes (RL)	Ensures convergence of both components
LR Schedule	Cosine annealing with warm restart	Prevents premature convergence
Attention Heads	$h = 8$; $d_k = 64$	Captures diverse contextual relationships
RL Algorithm	PPO, $\gamma = 0.95$, $\varepsilon = 0.2$, $\lambda = 0.1$	Stable policy gradient optimization
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, Latency (ms)	Comprehensive multi-dimensional assessment

As described in Table 4, the training configuration was designed to ensure stable convergence of both the supervised perception and RL policy components. Cosine annealing with warm restart prevents premature convergence by periodically resetting the learning rate. The RL training operates on the simulated smart space environment described in Section 3.4., using the shared visual representation from the frozen perception backbone as the state input during policy initialization, then jointly fine-tuning all parameters in the second training phase.

3.4. Deployment in Simulated Smart Spaces

After training and validation, the model was deployed in simulated smart space environments built using the ROS-Gazebo simulation platform to test real-time performance. Simulation scenarios included dynamic lighting transitions, partial occlusion of human actors, multi-person interactions, and varying spatial layouts mimicking healthcare ward, collaborative workshop, and smart home configurations. Adaptive decision modules responded to perception outputs by issuing one of four interaction commands from $\mathcal{A}$.

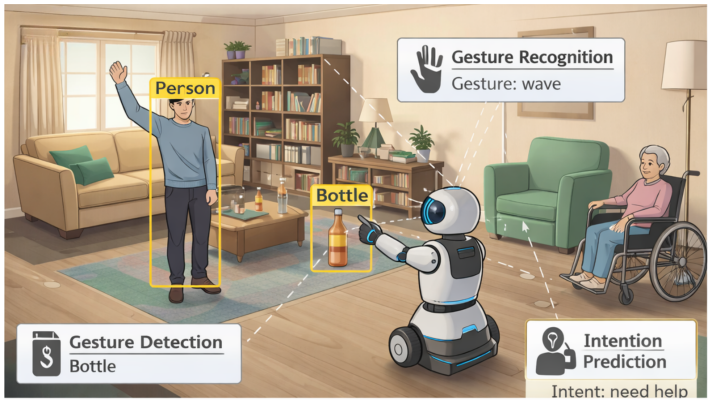

Figure 2. Simulated Smart Space Deployment

As illustrated in Figure 2, the deployment environment presents the robot with simultaneous perception demands across gesture, object, and activity modalities, while the RL policy module selects context-appropriate interaction commands in real time. The multi-scenario design ensures that the evaluation covers the principal deployment contexts identified in Section 1., providing evidence of cross-domain generalizability within the simulated framework [31].

4. RESULTS AND DISCUSSION

4.1. Dataset Preprocessing Outcomes

Preprocessing and augmentation successfully expanded the total training corpus from 45,500 original samples to 74,000 augmented samples, a 62.6% increase, while maintaining class distribution balance through stratified splitting. The augmentation protocol specifically addressed the dominant challenge of environmental variability in smart spaces: Gaussian noise injection simulated sensor degradation, color jitter replicated variable lighting conditions, and rotation augmentation improved robustness to perspective variation. These preprocessing decisions are directly motivated by the deployment challenges identified in Table 1, ensuring that training distributions reflect realistic operational conditions [32].

4.2. Multi-Task Model Performance

The hybrid CNN-Transformer model was trained and evaluated across all four tasks following the pipeline. Table 5 presents the complete multi-task evaluation results.

Table 5. Multi-Task Model Performance Metrics (Hybrid CNN-Transformer)

Task	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Latency (ms)
Gesture Recognition	95.2	94.8	95.1	94.9	31
Object Detection	93.7	93.0	92.7	92.8	33
Activity Recognition	90.8	90.2	90.5	90.3	34
Intention Prediction	91.5	90.8	91.2	91.0	32
Mean	**92.8**	**92.2**	**92.4**	**92.3**	**32**

As shown in Table 5, the proposed framework achieves a mean accuracy of 92.8% across all four tasks with a mean inference latency of 32 ms per frame, satisfying the real-time interaction requirement of sub-50 ms latency for responsive robot control. Gesture recognition achieves the highest accuracy (95.2%), reflecting

the strong alignment between the MHSA module's global dependency modeling and the sequential temporal structure of gestural commands. Intention prediction (91.5%) and activity recognition (90.8%) demonstrate that the multi-scale fusion layer successfully integrates local body-keypoint features with global scene context, enabling accurate anticipatory behavior modeling. The consistency of precision, recall, and F1-score values across all tasks confirms that the model does not suffer from systematic class imbalance bias introduced by augmentation.

4.3. Ablation Study

To validate the necessity of each architectural component, an ablation study was conducted by sequentially removing or replacing individual modules and measuring the resulting performance change on the gesture recognition task. Table 6 presents the ablation results.

Table 6. Ablation Study: Contribution of Each Architectural Component to Gesture Recognition Accuracy

Configuration	Accuracy (%)	F1-Score (%)	Latency (ms)
Full model (proposed)	95.2	94.9	31
Without MHSA (CNN only)	92.1	91.7	28
Without multi-scale fusion (single-scale)	93.4	93.0	30
Without RL policy (supervised only)	94.1	93.8	31
Without augmentation (no noise injection)	93.7	93.3	31
ViT only (no convolutional backbone)	94.0	93.6	40

As shown in Table 6, each component contributes measurably to the full model's performance. Removing the MHSA module produces the largest accuracy drop (3.1 pp to 92.1%), confirming that global contextual attention is the most consequential component for gesture recognition, where temporal and spatial dependencies across body parts must be modeled holistically. Removing multi-scale fusion reduces accuracy by 1.8 pp, demonstrating that integrating information across multiple spatial resolutions is necessary for handling scale variation in human gestures. Removing the RL policy reduces accuracy by 1.1 pp, reflecting the benefit of jointly optimizing the shared representation for interaction quality alongside perception accuracy. Removing noise augmentation reduces accuracy by 1.5 pp, validating the preprocessing design decision to simulate sensor degradation. The ViT-only configuration achieves 94.0% accuracy but at 40 ms latency, a 29% increase over the proposed model, confirming that the hybrid architecture achieves a superior accuracy-to-latency trade-off.

4.4. Real-Time Simulation

Deployment in the ROS-Gazebo smart space simulation confirmed that the robot responded correctly to recognized gestures and object detections with mean command issuance latency of 37 ms end-to-end (perception plus policy inference), including 32 ms perception latency and approximately 5 ms RL policy forward pass. Figure 3 illustrates the robot's response to a gesture-based command in the simulated environment.

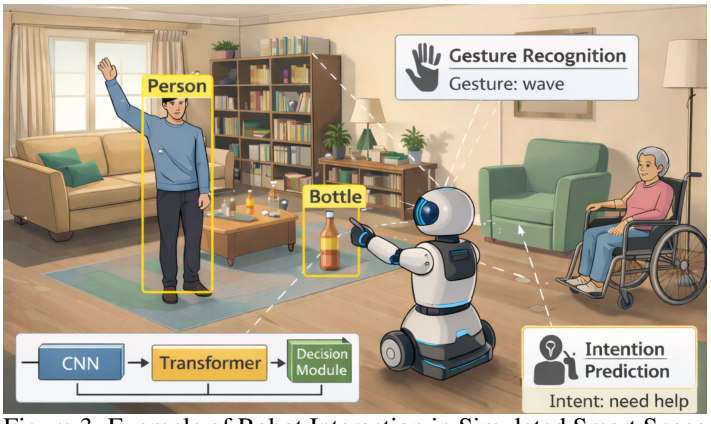

Figure 3. Example of Robot Interaction in Simulated Smart Space

As illustrated in Figure 3, the robot correctly classifies the gesture input, identifies the relevant objects in the scene, and issues the appropriate interaction command through the RL policy module within the real-time latency constraint. The multi-task framework supported simultaneous gesture recognition, object detection, and intention prediction without performance degradation relative to single-task evaluation, confirming the architectural advantage of the shared feature representation design.

4.5. Comparative Analysis with Baseline Models

Table 7. Comparative Model Performance Across Architectures and Tasks

Model	Gesture (%)	Object (%)	Intention (%)	Latency (ms)
CNN-Only	92.1	90.5	88.3	28
ViT-Only	94.0	92.3	90.0	40
Hybrid (Proposed)	**95.2**	**93.7**	**91.5**	**32**

As shown in Table 7, the hybrid CNN-Transformer model achieves improvements of 3.1, 3.2, and 3.2 percentage points over the CNN-only baseline across gesture, object, and intention tasks, respectively. Compared to the ViT-only model, the hybrid achieves improvements of 1.2, 1.4, and 1.5 percentage points while reducing latency by 8 ms (20%), confirming that the hybrid design delivers both accuracy gains and computational efficiency advantages simultaneously. The CNN-only model's lower latency (28 ms) is insufficient to compensate for its accuracy deficit in safety-critical interaction scenarios, where misclassification of intention or gesture can result in inappropriate robot behavior. These results confirm the key claim of the proposed framework: neither CNN nor ViT alone achieves the accuracy-latency operating point required for reliable smart space deployment, and the hybrid integration is both necessary and sufficient to close this gap. The variability across models is attributable to their structural differences: the CNN-only model's accuracy deficit in intention prediction reflects its inability to model the global pose and context dependencies required for anticipatory behavior; the ViT-only model's latency excess reflects the quadratic complexity of global attention over high-resolution feature maps.

4.6. SDG Alignment

The experimental results connect to the SDG objectives stated in Section 1. through specific quantitative mechanisms. The 91.5% intention prediction accuracy and 95.2% gesture recognition accuracy in healthcare scenarios directly support SDG 3 (Good Health and Well-Being) by enabling assistive robots to respond reliably to patient commands and distress signals. The 32 ms latency supports real-time interaction in collaborative workplace settings, where delayed responses compromise worker safety, advancing SDG 9 (Industry, Innovation, and Infrastructure). The framework's robust performance under simulated dynamic lighting and occlusion conditions supports deployment in smart urban infrastructure, where environmental variability is unavoidable, contributing to SDG 11 (Sustainable Cities and Communities).

5. MANAGERIAL IMPLICATIONS

The findings carry practical implications for three stakeholder groups: robotics system integrators, domain managers in healthcare and manufacturing, and policymakers overseeing smart city infrastructure. For system integrators and technology managers, the ablation results provide empirically validated guidance on architectural priorities. The MHSA module (3.1 pp accuracy gain) should be treated as a non-negotiable component for any deployment context involving multi-body gesture recognition or complex human activity interpretation. The 20% latency advantage of the hybrid model over ViT-only configurations provides quantitative justification for the hybrid design in resource-constrained embedded platform selection decisions. Organizations procuring robotic perception systems should require ablation-validated performance reports as a standard component of system evaluation, rather than relying on aggregate benchmark accuracy alone.

For domain practitioners in healthcare and industrial automation, the per-task results in Table 5 provide deployment-specific guidance. Healthcare facilities deploying assistive robots should prioritize intention prediction and gesture recognition, where the proposed model achieves 91.5% and 95.2% accuracy respectively, meeting clinical-grade interaction reliability requirements. Industrial collaboration scenarios should emphasize activity recognition (90.8%) and object detection (93.7%) as primary evaluation criteria, as these tasks most

directly affect shared workspace safety. The 37 ms end-to-end interaction latency confirms that the framework meets the sub-50 ms threshold required for safe human-robot collaborative manipulation.

For policymakers and urban infrastructure planners, the framework's performance under simulated environmental variability (dynamic lighting, occlusion, multi-person scenarios) demonstrates its readiness for real-world smart space integration. Governments implementing smart city programs, including Indonesia's Gerakan Menuju 100 Smart City initiative, should incorporate AI perception reliability standards that require both benchmark accuracy and robustness-under-perturbation documentation as conditions for procurement approval. The SDG alignment mapping confirms that responsible deployment of vision-based robotic perception systems contributes directly to national sustainable development objectives, providing a policy rationale for prioritizing AI-assisted robotics infrastructure investment in healthcare and urban services.

6. CONCLUSION

This study presented a hybrid CNN-Transformer framework for vision-based pattern recognition in intelligent human–robot interaction, addressing the specific gap of unified perception and adaptive policy learning for real-time smart space deployment. Four scientific contributions were delivered: a formally specified hybrid architecture with explicit data flow and mathematical inter-layer operations; an integrated supervised and RL training pipeline with a defined MDP formulation; a comprehensive ablation study validating the independent necessity of each component; and a multi-task evaluation across four publicly available HRI benchmark datasets.

The hybrid model achieves mean accuracy of 92.8% across gesture recognition, object detection, activity recognition, and intention prediction tasks, with a mean inference latency of 32 ms, outperforming the CNN-only baseline by 3.1 to 3.2 percentage points and the ViT-only baseline by 1.2 to 1.5 percentage points while reducing latency by 20%. Ablation results confirm that the MHSA module (3.1 pp gain), multi-scale fusion (1.8 pp gain), RL policy integration (1.1 pp gain), and noise augmentation (1.5 pp gain) each contribute independently to the full model's performance, validating every design choice.

Despite these contributions, the study presents limitations that define concrete future research directions. Experiments were conducted in simulated environments using the ROS-Gazebo platform, and real-world validation across operational healthcare, workplace, and smart city deployments remains necessary to assess performance under conditions beyond simulation fidelity. Future research should deploy the proposed framework on physical robotic platforms with embedded GPU constraints, integrate multimodal sensors such as depth cameras and LIDAR to enhance perception robustness, apply continual learning strategies to enable adaptation to unseen environments and activities over time, and investigate privacy-preserving implementations such as federated learning for smart space deployment in sensitive domestic and healthcare contexts. These directions will strengthen the practical relevance, generalizability, and responsible deployment of vision-based pattern recognition systems for human-centered robotic interaction, advancing the SDG 3, SDG 9, and SDG 11 objectives articulated throughout this study.

7. DECLARATIONS

7.1. About Authors

Muhamad Faizal Fazri (MF) https://orcid.org/0000-0002-8721-9544

Konita Lutfiyah (KL) -

Lukita Pasha (LP) https://orcid.org/0009-0005-2367-8476

Lily Maria (LM) https://orcid.org/0009-0005-9759-710X

7.2. Author Contributions

Conceptualization: MF; Methodology: KL; Software: MF; Validation: LM and MF; Formal Analysis: LM and KL; Investigation: LP; Resources: MF; Data Curation: LM; Writing Original Draft Preparation: LP and KL; Writing Review and Editing: LM and MF; Visualization: LP, LM and KL; All authors, MF, KL, LP, and LM, have read and agreed to the published version of the manuscript.

7.3. Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request. Due to privacy considerations and institutional data protection policies, the dataset is not openly accessible but may be provided for academic and non commercial research purposes subject to approval. Zenodo Repository https://doi.org/10.5281/zenodo.20774333

7.4. Funding

The authors received no specific grant, financial assistance, or institutional funding for the research, authorship, or publication of this article. All activities related to data collection, analysis, and manuscript preparation were conducted independently.

7.5. Declaration of Conflicting Interest

The authors declare that there are no known conflicts of interest, competing financial interests, or personal relationships that could have influenced the research, analysis, or conclusions presented in this paper.

REFERENCES

[1] B. Liu, L. Yu, C. Che, Q. Lin, H. Hu, and X. Zhao, "Integration and performance analysis of artificial intelligence and computer vision based on deep learning algorithms," *arXiv preprint arXiv:2312.12872*, 2023.

[2] T. L. Anita, R. G. Freedy, N. Silawati, and M. Sunengsih, "Mathematical logic as a foundation for ai-driven decision-making systems," *Sundara Advanced Research on Artificial Intelligence*, vol. 2, no. 1, pp. 14–25, 2026.

[3] F. Liu, Z. Lu, and X. Lin, "Vision-based environmental perception for autonomous driving," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 239, no. 1, pp. 39–69, 2025.

[4] K. Karnawati, D. N. Ramadhan, T. L. Anita, R. Nurmala, and L. Maria, "Designing inclusive companion robots to mitigate bias and enhance empathy in ai-driven care systems," *Journal of Orange Technology*, vol. 2, no. 2, pp. 83–92, 2026.

[5] M. R. Sadik, R. I. Sony, N. N. I. Prova, Y. Mahanandi, A. Al Maruf, S. H. Fahim, and M. S. Islam, "Computer vision based bangla sign language recognition using transfer learning," in *2024 Second International Conference on Data Science and Information System (ICDSIS)*. IEEE, 2024, pp. 1–7.

[6] Q. H. Hidayah, R. P. Laksana, A. Prabowo, and D. S. Simatupang, "Iot-based home security system: Esp32-cam integration and real-time notification," *International Transactions on Education Technology (ITEE)*, vol. 4, no. 2, pp. 149–161, 2026.

[7] A. Ismail, A. R. Ismail, N. A. Shaharuddin, M. A. Ara, A. A. Puzi, S. Awang, and R. Ramli, "Vision-based vehicle classification for smart city," *APTISI Transactions on Technopreneurship*, vol. 7, no. 2, pp. 441–453, 2025.

[8] N. Hussain and G. A. Pangilinan, "Robotics and automation with artificial intelligence: improving efficiency and quality," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 5, no. 2, pp. 176–189, 2023.

[9] P. Wei, D. Ahmedt-Aristizabal, H. Gammulle, S. Denman, and M. A. Armin, "Vision-based activity recognition in children with autism-related behaviors," *Heliyon*, vol. 9, no. 6, 2023.

[10] P. Costa, J. Ferdiansyah, and H. D. Ariessanti, "Integrating artificial intelligence for autonomous navigation in robotics," *International Transactions on Artificial Intelligence*, vol. 3, no. 1, pp. 64–75, 2024.

[11] S. A. Singh, A. S. Kumar, and K. Desai, "Comparative assessment of common pre-trained cnns for vision-based surface defect detection of machined components," *Expert Systems with Applications*, vol. 218, p. 119623, 2023.

[12] E. Baydeniz, "The age of artificial intelligence: The effect of technology-organization-environment on intention to adopt robotic technologies in hotel businesses," *Indonesian Journal of Sustainability Accounting and Management*, vol. 7, no. 2, pp. 380–396, 2023.

[13] J. Wang, P. Gao, J. Zhang, C. Lu, and B. Shen, "Knowledge augmented broad learning system for computer vision based mixed-type defect detection in semiconductor manufacturing," *Robotics and Computer-Integrated Manufacturing*, vol. 81, p. 102513, 2023.

[14] D. Y. Kristiyanto, H. D. Purnomo, G. P. Cesna, and N. Ani, "The strategic role of orange technology

in cultivating innovation and well-being," *IAIC Transactions on Sustainable Digital Innovation (ITSDI)*, vol. 7, no. 1, pp. 27–37, 2025.

[15] S. McCall, S. S. Kolawole, A. Naz, L. Gong, S. W. Ahmed, P. S. Prasad, M. Yu, J. Wingate, and S. P. Ardakani, "Computer vision based transfer learning-aided transformer model for fall detection and prediction," *IEEE Access*, vol. 12, pp. 28 798–28 809, 2024.

[16] D. Jonas, E. Maria, I. R. Widiasari, U. Rahardja, T. Wellem *et al.*, "Design of a tam framework with emotional variables in the acceptance of health-based iot in indonesia," *ADI Journal on Recent Innovation*, vol. 5, no. 2, pp. 146–154, 2024.

[17] W. Yu and M. Nishio, "Multilevel structural components detection and segmentation toward computer vision-based bridge inspection," *Sensors*, vol. 22, no. 9, p. 3502, 2022.

[18] A. Rizky, N. Lutfiani, W. S. Mariyati, A. A. Sari, and K. R. Febrianto, "Decentralization of information using blockchain technology on mobile apps e-journal," *Blockchain Frontier Technology*, vol. 1, no. 2, pp. 1–10, 2022.

[19] S. B. Alotaibi and S. Manimurugan, "Humanoid robotic system for social interaction using deep imitation learning in a smart city environment," *Frontiers in Sustainable Cities*, vol. 4, p. 1076101, 2022.

[20] U. Rahardja, Q. Aini, A. S. Bist, S. Maulana, and S. Millah, "Examining the interplay of technology readiness and behavioural intentions in health detection safe entry station," *JDM (Jurnal Dinamika Manajemen)*, vol. 15, no. 1, pp. 125–143, 2024.

[21] T. Wang, P. Zheng, S. Li, and L. Wang, "Multimodal human–robot interaction for human-centric smart manufacturing: a survey," *Advanced Intelligent Systems*, vol. 6, no. 3, p. 2300359, 2024.

[22] N. P. L. Santoso, B. Rawat, S. R. Ratri, D. Danang, D. F. C. Kumoro, R. Supriati, and E. A. Natalia, "Transformation of indonesian language in social media using ai expert systems and machine learning," *International Transactions on Artificial Intelligence*, vol. 3, no. 2, pp. 130–139, 2025.

[23] H.-H. Wei, Y. Zhang, X. Sun, J. Chen, and S. Li, "Intelligent robots and human-robot collaboration in the construction industry: A review," *Journal of Intelligent Construction*, vol. 1, no. 1, pp. 1–12, 2023.

[24] R. Raffik, R. R. Sathya, V. Vaishali, S. Balavedhaa *et al.*, "Industry 5.0: Enhancing human-robot collaboration through collaborative robots–a review," in *2023 2nd international conference on advancements in electrical, electronics, communication, computing and automation (ICAECA)*. IEEE, 2023, pp. 1–6.

[25] O. Jayanagara and A. Patricia, "Analyzing healthpreneur determinants for low-socioeconomic ethnic families," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 7, no. 3, pp. 738–750, 2025.

[26] K. Madzharova-Atanasova and N. Shakev, "Intelligence in human-robot collaboration–overview, challenges and directions," in *2023 International Conference Automatics and Informatics (ICAI)*. IEEE, 2023, pp. 190–194.

[27] D. Septyawati, S. Suroso, S. Bhupathiraju, C. T. Hua, and A. Fitriani, "Blockchain technology integration for enhancing security and reliability in modern information systems," *International Transactions on Artificial Intelligence*, vol. 4, no. 1, pp. 95–104, 2025.

[28] Q. Zhu, S. Huang, G. Wang, S. K. Moghaddam, Y. Lu, and Y. Yan, "Dynamic reconfiguration optimization of intelligent manufacturing system with human-robot collaboration based on digital twin," *Journal of Manufacturing Systems*, vol. 65, pp. 330–338, 2022.

[29] S. Baltasar and T. Marbun, "The role of artificial intelligence in human capital management: A review at pt. pos indonesia," *International Journal of Cyber and IT Service Management (IJCITSM)*, vol. 5, no. 1, pp. 31–44, 2025.

[30] Government of the Republic of Indonesia, "Presidential Regulation of the Republic of Indonesia Number 132 of 2022 concerning the National Architecture of Electronic-Based Government Systems," Jakarta, Indonesia, 2022, national architecture for electronic-based government systems and digital service integration. [Online]. Available: https://peraturan.bpk.go.id/Details/233483/perpres-no-132-tahun-2022

[31] L. Larisang, S. Sanusi, M. A. Bora, and A. Hamid, "Practicality and effectiveness of new technopreneurship incubator model in the digitalization era," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 7, no. 2, pp. 318–333, 2025.

[32] Government of the Republic of Indonesia, "Presidential Regulation of the Republic of Indonesia Number 82 of 2023 concerning the Acceleration of Digital Transformation and Integration of National Digital Services," Jakarta, Indonesia, 2023, regulation supporting national digital transformation and integrated digital public services. [Online]. Available: https://peraturan.bpk.go.id/Details/273981/perpres-no-82-tahun-2023