Towards Memory-Driven Agentic AI for Human Activity Recognition

*Mohamadreza Shahabian Alashti[1], *Khashayar Ghamati[1], Hooman Samani[2], and Abolfazl Zaraki[1]

[1] School of Physics, Engineering and Computer Science (SPECS), and Robotics Research Group, University of Hertfordshire, Hatfield, AL10 9AB, United Kingdom
[2] Creative Computing Institute, University of the Arts London, United Kingdom

Abstract. This paper proposes a novel, scalable agentic AI architecture designed to enhance human activity recognition across data modalities by embedding memory-driven reasoning and context awareness. The architecture integrates multimodal sensing, deliberative reasoning through supervised learning and context-aware language models, and memory mechanisms, including short-term memory for tracking immediate activity transitions and long-term memory for embedding experiential knowledge. The evaluation of the proposed model using two major datasets namely RHM (6.7K video clips of 14 known activities) and Toyota Smart Home (16K video clips of 31 unknown activities) demonstrates significant improvements, achieving 60% accuracy when combining contextual information with supervised model output, compared to 40% accuracy with context alone and 35% with supervised models on unseen data. By overcoming the limitations of traditional HAR approaches, this research advances the development of responsive and intelligent robotic systems, facilitating more natural and effective human-robot collaboration.

1 Introduction

The advent of agentic AI represents a pivotal evolution in artificial intelligence, empowering systems to autonomously perceive, reason, and act within dynamic environments [8]. Unlike conventional AI, which operates on fixed rules or static data, agentic AI leverages experiential learning to adapt and improve over time [23]. This adaptability makes it ideal for diverse applications, such as autonomous vehicles, healthcare diagnostics, and personalised assistants, where systems must respond to unpredictable conditions. By continuously learning from interactions, agentic AI enhances its decision-making capabilities, offering robust solutions that evolve with real-world demands.

In human-robot interaction (HRI), agentic AI introduces transformative advantages by enabling robots to interpret and respond to human behaviours in a context-sensitive manner [10, 16, 12]. This capability allows robots to collaborate

[*] The first two authors contributed equally to this work. `{m.r.shahabian, k.ghamati}@herts.ac.uk`, `https://ghamati.com/har-agent`

seamlessly with humans, whether in social settings, such as assisting individuals, or industrial environments, where they adapt to workers' varying expertise. The experiential learning of agentic AI makes natural, trust-building interactions, as robots refine their responses over time. This adaptability not only improves interaction quality but also enhances user acceptance, making agentic AI a cornerstone for effective HRI [26, 27].

Human Activity Recognition (HAR) is fundamental to HRI, as it equips robots with the ability to accurately interpret human actions [4, 11]. This understanding is vital for anticipating needs, preventing misunderstandings, and delivering timely support. For instance, in an assistive living scenario, a robot must differentiate between a person cooking and momentarily checking their phone to prioritise its assistance correctly. Effective HAR ensures that robots align their behaviours with human intentions, directly influencing the success of interactions and the reliability of robotic systems in human-centric environments.

Current HAR methods exhibit strengths and limitations. Skeleton-based models, such as M-LeNet [2, 22, 28], excel at capturing human biomechanics and achieve high accuracy on labelled datasets, making them valuable for structured activity recognition. However, these models falter in distinguishing activities with similar motion patterns, like lifting an object versus standing, and lack awareness of environmental context, objects, or semantic details. For example, they might confuse *drinking water* with *raising a hand* without recognising a glass nearby. Conversely, contextual models offer broader environmental insights but often sacrifice precision in activity recognition, underscoring the need for integrated approaches that combine biomechanical and contextual understanding [25]. Memory plays a critical role in overcoming these HAR limitations by enabling systems to retain and leverage past experiences. Sequential methods like Long Short-Term Memory (LSTM) network [13], and Gated Recurrent Units have attempted to capture temporal dependencies, but they are insufficient for handling complex contextual information or retrieving it in real time. These models typically focus on short-term patterns, failing to integrate long-term knowledge, such as considering sequences of activities. Incorporating both short-term and long-term memory (LTM) into HAR is essential for tracking activity transitions, distinguishing primary from secondary actions, and adapting to evolving behaviours, thereby enhancing recognition accuracy and responsiveness [20].

To address the pressing challenges in HAR within HRI, this paper proposes an experiential agentic AI architecture that unifies multimodal sensing, contextual reasoning, and memory mechanisms. The framework integrates sensory input from diverse modalities to capture human activities and environmental cues, combines this with deliberative reasoning through supervised learning models and large contextual models, and incorporates both short-term and long-term memory to manage temporal dynamics and contextual alignment. This synergy empowers the system to interpret human actions with greater precision, adapt behaviour based on situational context, and continuously refine its understanding through accumulated experience, key capabilities for enabling responsive and adaptive robotic systems. Although Agentic AI architectures are still in their early stages, our approach demonstrates promising performance on benchmark datasets such as RHM [3] and Toyota Smart Home [9], suggesting a scalable pathway for future HRI developments. The main contributions of this work are threefold. First, we introduce a generic agentic AI architecture for HAR in HRI that is inherently adaptable to various data modalities, ensuring broad applicability across interaction scenarios. Second, we fuse supervised learning, contextual reasoning, and hierarchical memory structures to improve the system's capability to recognise activity transitions and maintain coherent context over time. Finally, we offer insights into memory-driven, context-aware activity recognition as a critical step towards more intelligent and socially aware robots, laying the groundwork for future advances in human–robot collaboration.

2 State-of-the-art Developments of Agentic AI for HRI

This review supports the proposed agentic AI architecture by examining three key areas: agentic AI in HRI, memory mechanisms in HAR, and context-aware HAR approaches. These areas enhance robots' ability to understand and react to human behaviors. The subsections cover autonomous decision-making, temporal data retention for better predictions, and the use of environmental cues for improved recognition.

Agentic AI in HRI Agentic AI systems are characterised by their ability to autonomously perceive, reason, and act in dynamic environments, drawing from social cognitive theory's concept of agency [6]. In HRI, agentic AI enables robots to interpret human behaviours and adapt responses contextually, fostering natural and effective collaboration [16]. For instance, [7] developed a cognitive architecture integrating geometric reasoning and multi-modal dialogue, allowing robots to share tasks with humans seamlessly. Recent frameworks distinguish multi-agent systems, which maintain agent autonomy, from Centaurian systems, which deeply integrate human and AI capabilities. Bornet et al. [8] propose a five-level autonomy framework, positioning adaptive learning systems as critical for advanced HRI. Our architecture aligns with this paradigm, combining reactive and deliberative components to enhance HRI through context-aware HAR.

Memory Mechanisms in HAR Memory is pivotal in HAR, enabling systems to retain temporal and contextual information to distinguish primary from secondary activities. Traditional sequential models, such as RNNs, LSTM networks [13], and GRUs, capture short-term dependencies. However, these models struggle with long-term dependencies and fail to integrate rich contextual information, limiting their ability to adapt in real-time scenarios [14]. Memory-augmented neural networks, which incorporate explicit memory units, offer a promising alternative by enabling robust storage and retrieval of high-dimensional data [15]. Autobiographical memory systems further enhance experiential learning

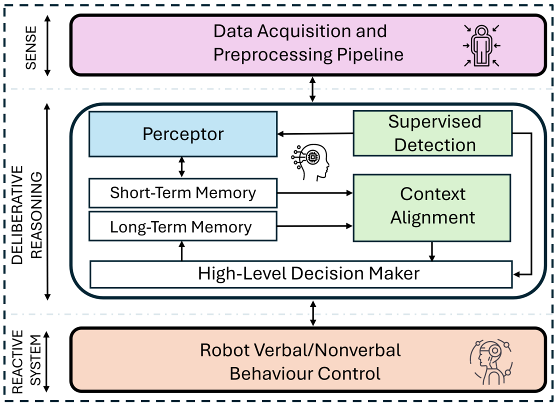


Fig. 1: Overview of the Agentic AI architecture for HRI, comprising sensing, reasoning, and reactive layers to enable memory-driven, context-aware HAR

by encoding interaction histories [20]. Our approach leverages short-term memory (STM) for tracking activity transitions and long-term memory (LTM) for contextual alignment, addressing the limitations of sequential models.

Context-Aware HAR Context-aware HAR integrates environmental and situational data to provide a holistic understanding of human activities, surpassing the limitations of motion-centric models [25]. Skeleton-based models, such as M-LeNet [4], excel at capturing biomechanical patterns but lack awareness of objects or semantic context. In contrast, context-aware approaches incorporate multimodal inputs, such as visual and sensor data, to disambiguate similar activities [18]. Neuro-symbolic methods combine data-driven learning with semantic reasoning, improving generalisation across diverse scenarios [5]. LLMs enhance HAR by providing semantic descriptions that complement skeletal data. Our architecture integrates supervised learning with LLMs to align biomechanical and contextual information, improving activity recognition accuracy in HRI.

3 Developing the Agentic AI Model for HRI

Building on the need for robots to accurately interpret human activities in dynamic environments, as highlighted in the introduction, this section presents our agentic AI architecture. The architecture integrates multimodal sensing, deliberative reasoning, and memory mechanisms to address the limitations of current HAR methods and enhance HRI. Specifically, it focuses on distinguishing primary activities from transient ones, a challenge discussed in the literature review, by leveraging both STM and LTM to track activity transitions and align environmental contexts. The proposed experiential agentic AI model is illustrated in Figure 1, with subsequent subsections describing each component in detail.

3.1 Sense

The Sense layer acquires raw streams, preprocesses them, and dispatches each modality to the right subsystem. For training, and for seeding long-term memory, we rely on the RHM dataset [1]. RHM offers four RGB viewpoints, but we keep only the robot-mounted camera to mimic a pan-tilt head that tracks the subject; the split contains 14 daily-activity classes. Generalisation is measured on the Toyota Smart Home dataset [9], which is never used for training. Its 16115 RGB-D clips (31 classes, recorded with elderly participants) provide the "unseen, real-life" test bed, with class imbalance and overlapping motions that stress-test robustness. Preprocessing within the Sense layer is essential to prepare the raw data for subsequent analysis. The Sense layer is designed to accommodate both multi-modal and single-modal inputs, channelling specific data streams to designated components. For example, skeletal data is directed to the supervised detector, while image frames are processed by the Perceptor to extract semantic and contextual features. A single YOLOv7-pose extractor (17 joints) is run on both datasets so that the supervised HAR block (M-LeNet) always receives skeletons produced by the same pipeline. To keep the context-aware branch responsive, we forward just one RGB frame out of every ten to LLaVA; the intervening frames supply skeletons only.

3.2 Deliberative Reasoning

The Deliberative Reasoning layer is the core of the agent's autonomous decision-making process, combining supervised learning, contextual reasoning, and memory mechanisms to interpret human activities accurately. This layer consists of five key components: M-LeNet for skeleton-based HAR, a Perceptor for environmental context extraction, a Context Alignment (CA) module for integrating skeletal and contextual data, and STM and LTM for tracking activity transitions and aligning contexts. High-level decision maker processes the CA outputs to predict activities.

Supervised Detection: M-LeNet for Skeleton-Based HAR The Supervised Detection block is a modular component designed to process modality-specific behavioural data using supervised learning models. In this implementation, we employ M-LeNet, a tailored 2D CNN, for skeleton-based HAR. M-LeNet processes skeleton data transformed into a 34×34 tensor image, capturing spatial and temporal dynamics efficiently.

M-LeNet Architecture Adapted from LeNet, M-LeNet features two convolutional layers with 3×3 kernels and channel configurations of 10-20 (low-capacity) or 20-40 (high-capacity). Regularisation is achieved via two dropout layers (p=0.25 and p=0.5), and the model includes three fully connected layers to enhance learning capacity.

Performance On the SK-HAR dataset, M-LeNet achieves approximately 90% accuracy with the high-capacity configuration, trained using categorical cross-entropy and Adam optimiser (learning rate 0.001) over 50 epochs.

Adaptability The block is designed to be modality-agnostic, allowing integration of alternative models such as 3D CNNs for video frames or 1D CNNs for audio signals, ensuring flexibility across diverse behavioural modalities.

Perceptor The Perceptor module analyses the environment to understand human behavior, which is crucial for accurate HAR. It processes video frames from the data pipeline, extracting detailed information and storing each frame's description in STM to track changes over time, such as activity transitions. The process begins by sending the first frame to the LLaVA [17] model, which generates a description of the scene. For each subsequent frame, the description of the previous frame is combined with the current frame and passed back to LLaVA to produce a more enriched result. To enhance accuracy, LLaVA uses the activity predicted by the M-LeNet model as additional metadata in its prompts for each frame. Each frame's description is saved in the STM, which starts empty for the first frame and grows as more descriptions are added. Once all frames are processed, the accumulated descriptions in the STM are sent to the Qwen model [24], which extracts up to five keywords summarising the environment's main context (e.g., *Drinking, Cup, Wine, Table, Chair* for a drinking activity). These keywords are then passed to the Context Alignment module. The STM's role is vital, as it aggregates frame descriptions, enabling LLaVA to detect activity transitions and providing Qwen with a comprehensive view of the environment for effective keyword extraction.

Context Alignment At the heart of our agent lies the Context Alignment module, which employs probabilistic methods to identify relationships between the extracted words obtained from the Perceptor and the known tasks stored in Long-Term Memory. Specifically, this module computes the probability that an extracted word belongs to one of the pre-trained tasks. By doing so, it enables the agent to infer additional information regarding the primary activity currently unfolding in the environment. If the computed probability exceeds 60%, the associated keypoints are considered to relate to a known task, thereby simplifying the distinction between concurrent activities. Conversely, if the probability falls below this threshold, the context is deemed unrelated to keypoints. Put differently, since LLaVA constructs an STM-based model according to the M-LeNet's predictions, a probability above 60% indicates that the extracted words correspond to a class represented in the training dataset, thus suggesting a meaningful relation between keypoints and context. To compute this probability, we utilise conditional probability in conjunction with the Jaccard Index [21] to derive the posterior probability. This calculated posterior determines whether newly encountered words are associated with any of the tasks previously encoded within the agent's LTM. In each interaction, the agent considers every keyword as a random variable. The sample space for each interaction encompasses all potential words, including both pre-existing entries in LTM and newly observed terms. We assume a uniform distribution, employing its Probability Mass Function (PMF) to facilitate probability calculations. Following each interaction, LTM is updated if the High-Level Decision Maker module validates that the association between keypoints and context yields a meaningful contribution towards activity recognition. This update integrates newly encountered keywords and incrementally adjusts frequency counts for recurring terms, a process critical for maintaining accurate likelihood estimations. Evaluation of the CA module at each interaction is governed by Eq. 1:

$$P(T_i \mid w_1, w_2, \ldots, w_n) \propto P(T_i) \cdot \prod_{j=1}^{n} P(w_j \mid T_i) \tag{1}$$

Here, $P(T_i)$ denotes the prior probability of task T_i, while $P(w_j \mid T_i)$ represents the likelihood that word w_j belongs to task T_i. Empirical evaluation reveals that the number of words per task varies dynamically due to ongoing LTM updates. As the PMF of a uniform distribution is obtained by dividing the frequency of a word by the total number of words in a task, frequently performed tasks exhibit decreasing probabilities over time, potentially leading to unreliable predictions. To counter this, we incorporate the Jaccard Index as an alternative prior, thereby accounting for disparities in word list sizes. The revised computation, employing the Jaccard Index as the prior, is expressed in Eq. 2:

$$P(T_i \mid w_1, w_2, \ldots, w_n) \propto J(T_i, W) \cdot \left(\prod_{j=1}^{n} P(w_j \mid T_i) \right) \tag{2}$$

The Jaccard Index $J(T_i, W)$ is determined using Eq. 3, where $\text{intersect}(T_i, W)$ denotes the number of new words that are already present within T_i, and $\text{union}(T, W)$ represents the total number of unique words across both the LTM and newly encountered sets:

$$J(T_i, W) = \frac{|\text{intersect}(T_i, W)|}{|\text{union}(T, W)|} \tag{3}$$

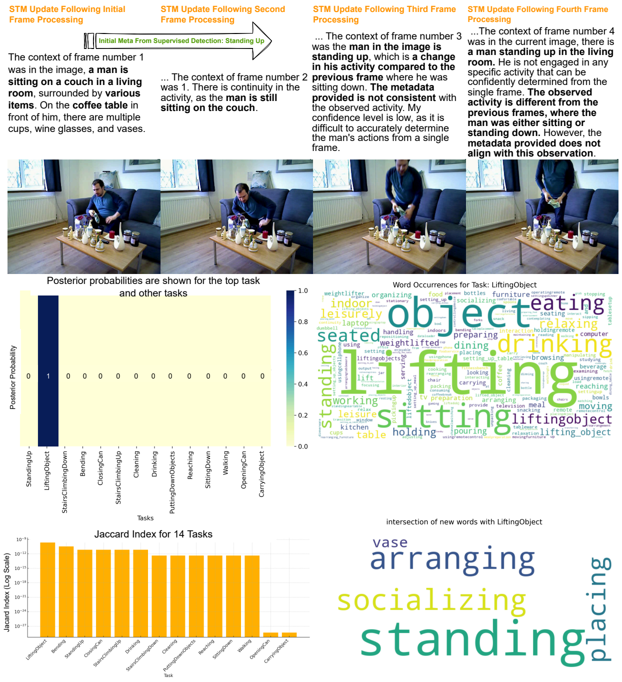
In computing the posterior probability, we first calculate the Jaccard Index (Eq. 3), then, applying Bayes' rule, transform $P(T_j \mid w_j)$ into $P(w_j \mid T_j)$ as shown in Eq. 5. The likelihood function $P(w_j \mid T_i)$ is calculated according to Eq. 4, while $P(T_i)$, being uniform across all tasks, is computed via Eq. 6:

$$P(w_j \mid T_i) = \frac{\text{Total frequency of } w_j \text{ in task } T_i}{\text{Total frequency of all words in task } T_i} \tag{4}$$

$$P(T_i \mid w_j) = \frac{P(w_j \mid T_i) \cdot P(T_i)}{\sum_j P(w_j \mid T_j) \cdot P(T_j)} \tag{5}$$

$$P(T_i) = \frac{1}{\text{Number of Tasks}} \tag{6}$$

Upon evaluating Eq. 2, the resultant posterior probability is passed to the High-Level Decision Maker module. Here, it is scrutinised to predict the principal activity occurring in the environment relative to the agent's existing LTM.



STM Update Following Initial Frame Processing

The context of frame number 1 was in the image, **a man is sitting on a couch in a living room**, surrounded by **various items**. On the **coffee table** in front of him, there are multiple cups, wine glasses, and vases.

STM Update Following Second Frame Processing

... The context of frame number 2 was 1. There is continuity in the activity, as the **man is still sitting on the couch**.

STM Update Following Third Frame Processing

... The context of frame number 3 was the **man in the image is standing up**, which is **a change in his activity compared to the previous frame** where he was sitting down. **The metadata provided is not consistent** with the observed activity. My confidence level is low, as it is difficult to accurately determine the man's actions from a single frame.

STM Update Following Fourth Frame Processing

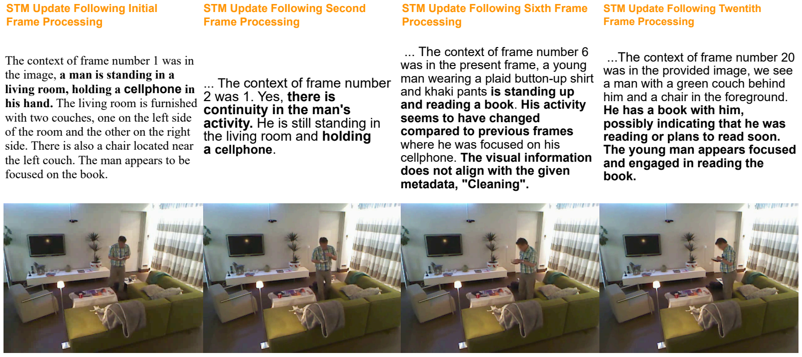
...The context of frame number 4 was in the current image, there is **a man standing up in the living room.** He is not engaged in any specific activity that can be confidently determined from the single frame. **The observed activity is different from the previous frames, where the man was either sitting or standing down.** However, the **metadata provided does not align with this observation**.

Fig. 2: Context Alignment module refines the initial *standing up* prediction to *lifting an object* by combining keypoints with contextual cues over four frames.

Should the probability fall below 60%, the combination is deemed insufficiently informative for further processing or reasoning. In this case, the results of the Supervised method should be considered. This mechanism enables the agent to utilise environmental context alongside keypoints, thereby enhancing its understanding of ongoing activities and enabling it to identify the primary activity with greater confidence and precision.

4 Proof of Concept Evaluation

To evaluate our proposed agentic AI system, we utilised two datasets. Initially, we assessed the system using the dataset on which the model had originally been trained, followed by an evaluation on a distinct dataset to investigate its adaptability to unfamiliar scenarios. Our experiments demonstrated that, upon encountering a new activity, the agent's supervised model can initially produce an activity label based on keypoint features extracted from the training data. However, the integration of contextual information significantly enhances predictive accuracy. For example, Figure 4a illustrates the agent's inference process after analysing four consecutive frames depicting a specific activity. When only



STM Update Following Initial Frame Processing

The context of frame number 1 was in the image, **a man is standing in a living room, holding a cellphone in his hand.** The living room is furnished with two couches, one on the left side of the room and the other on the right side. There is also a chair located near the left couch. The man appears to be focused on the book.

STM Update Following Second Frame Processing

... The context of frame number 2 was 1. Yes, **there is continuity in the man's activity.** He is still standing in the living room and **holding a cellphone**.

STM Update Following Sixth Frame Processing

... The context of frame number 6 was in the present frame, a young man wearing a plaid button-up shirt and khaki pants **is standing up and reading a book**. **His activity seems to have changed compared to previous frames** where he was focused on his cellphone. **The visual information does not align with the given metadata, "Cleaning".**

STM Update Following Twentith Frame Processing

...The context of frame number 20 was in the provided image, we see a man with a green couch behind him and a chair in the foreground. **He has a book with him, possibly indicating that he was reading or plans to read soon. The young man appears focused and engaged in reading the book.**

Fig. 3: Context Alignment revises the initial *Cleaning* prediction to *Carrying Object* by integrating contextual keywords from an unseen dataset.

the first frame is considered, the CA identifies the action as *sitting*, based on both the environmental context and the output of the supervised model. However, by incrementally processing all four frames, and crucially, passing the result of each frame to the next via the STM, the agent becomes aware of the activity transition (e.g., from sitting to standing), thereby constructing a more coherent and complete interpretation. Moreover, the agent is capable of inferring that the participant is lifting an object (e.g., a napkin), thereby refining its prediction to either lifting or standing up. The latter label originates from the supervised model, while the former emerges through the integration of contextual cues.

This illustrates the CA module's capacity to synthesise environmental context with the outputs of the supervised model, increasing overall confidence without an over-reliance on supervision alone. To construct the STM, the outputs from the supervised model were passed as metadata to the LLaVA model, which generated frame-level descriptions. After all four frames were processed and the STM was updated accordingly, this enriched memory was then passed to the Qwen model, which was prompted to extract the five most salient keywords from the generated descriptions. This novel mechanism enables the agent to track temporal transitions and more reliably identify the primary activity. Figure 2 demonstrates these results. The top row depicts the evolution of STM across the four frames, wherein LLaVA's outputs are concatenated with previous content and fed back into LLaVA alongside the new frame. The second row presents the original frames, where the supervised model predicted standing up. However, with the inclusion of contextual data, the CA accurately determined that the participant was in fact lifting an object (napkin), yielding a more accurate interpretation.

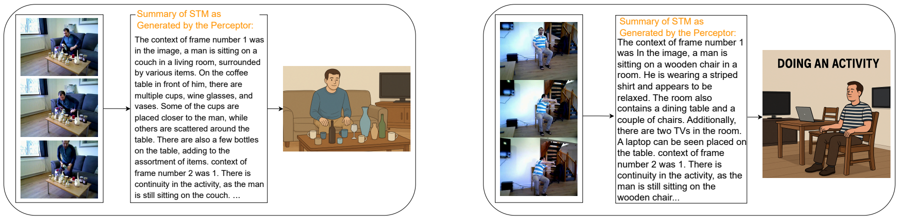
The third and fourth rows show the Jaccard index and its effect on posterior probability values. These visualisations also present all keywords associated with the Lifting Object activity during training (stored in LTM), highlighting the overlap between keywords derived from STM (via Qwen) and those previously encountered. The size of each word denotes its frequency in the training data; for example, Lifting appears larger due to its higher frequency in relevant contexts. Additionally, we tested the agent on a previously unseen dataset to evaluate its generalisability. Figure 3 shows four randomly selected frames from a video in which a participant is reading a book. After analysing all 36 frames of this sequence, the CA module inferred the activity as carrying an object, whereas the supervised model incorrectly classified it as cleaning. This example highlights how STM supports more accurate activity tracking across frames, enabling superior inference prior to final decision-making. At the same time, the LTM module contributes experiential knowledge that improves the agent's ability to generalise in unfamiliar scenarios.

To validate the CA module's output, we rely on posterior probability values. If the value exceeds 60%, the output is accepted. To visualise incorrect predictions, we are using ChatGPT [19] to generate an image based on STM content. Figure 4a shows an example from Figure 3, where ChatGPT produced an image based on the STM's detailed textual description, which preserved the core elements of the original frames. In contrast, Figure 4b depicts a case where the STM-based inference, generated by the Perceptor module, was vague and did not correspond well with the input frames and the posterior probability for this activity was below 60%. Based on this methodology, we processed videos from both datasets encompassing various activities to assess the agent's performance. We observe that combining contextual information with keypoints improved accuracy to 60%. In contrast, relying solely on context, without incorporating the supervised model, achieved only 40% accuracy. This indicates that environmental cues alone are insufficient to determine the underlying activity.

To obtain these results, we initially ran the agent without the supervised model, omitting its outputs as metadata for the Perceptor module. Accuracy was then calculated as the ratio of correctly classified videos to the total. In the second phase, we reintroduced the supervised model and passed its outputs to Perceptor to aid frame analysis via LLaVA, using the same evaluation procedure. Additionally, we observe that the supervised model alone, when applied to unfamiliar data, could at best provide a top-5 prediction set that included at least one semantically related activity. For example, when attempting to classify eating, the top labels included sitting and drinking. On this basis, we evaluated the model's standalone accuracy on the unseen Toyota dataset, which was approximately 35%. In contrast, its accuracy on the RHM dataset, on which it had been trained, was approximately 90%. We evaluated performance using 600 video clips from the Toyota dataset and 140 clips from the RHM dataset.

5 Discussion

In this study, we demonstrated the advantages of integrating memory-driven reasoning and contextual awareness into an agentic AI architecture for HAR in HRI. By combining short-term memory with contextual insights derived from LLMs such as LLaVA and Qwen, our model achieves a more nuanced and accu-



Summary of STM as Generated by the Perceptor:

The context of frame number 1 was in the image, a man is sitting on a couch in a living room, surrounded by various items. On the coffee table in front of him, there are multiple cups, wine glasses, and vases. Some of the cups are placed closer to the man, while others are scattered around the table. There are also a few bottles on the table, adding to the assortment of items. context of frame number 2 was 1. There is continuity in the activity, as the man is still sitting on the couch. ...

Summary of STM as Generated by the Perceptor:

The context of frame number 1 was In the image, a man is sitting on a wooden chair in a room. He is wearing a striped shirt and appears to be relaxed. The room also contains a dining table and a couple of chairs. Additionally, there are two TVs in the room. A laptop can be seen placed on the table. context of frame number 2 was 1. There is continuity in the activity, as the man is still sitting on the wooden chair...

(a) Correct prediction: the image generated by ChatGPT based on the STM's textual description closely matches the original input frames.

(b) Incorrect prediction: the image generated by ChatGPT from the STM's textual description fails to align with the original input frames.

Fig. 4: Comparison of ChatGPT-generated images based on STM content. (a) shows a successful case with coherent STM representation; (b) illustrates a failure case due to vague STM content.

rate understanding of human activities, effectively distinguishing primary actions from transient or secondary behaviours. This advancement addresses a critical limitation of traditional HAR methods, which often struggle to interpret complex or overlapping activities in dynamic environments. Our results underscore the pivotal role of memory and context in enhancing HAR for HRI. The integration of STM enables the model to track activity transitions across multiple frames, crucial for accurately identifying primary activities in scenarios with overlapping actions, such as using a computer while drinking tea. Single-frame analyses, in contrast, frequently misclassify secondary actions, highlighting the necessity of temporal context. Additionally, incorporating environmental context via LLMs resolves ambiguities in activities with similar skeletal movements, such as standing up versus lifting an object, where traditional skeleton-based models like M-LeNet falter. The agentic AI framework, combining supervised learning, contextual reasoning, and memory, aligns with our objective of developing a flexible, experiential model adaptable to multiple data modalities. This approach not only improves recognition accuracy but also enhances contextual responsiveness, fostering natural human-robot collaboration.

A critical aspect of our architecture is the use of prompts to guide LLMs in extracting relevant contextual information from the environment. Prompts are designed to steer models like LLaVA toward key environmental features and activity transitions, producing precise frame descriptions stored in STM. These descriptions are then processed by Qwen to identify salient keywords for the Context Alignment module. Well-crafted prompts, balancing specificity and generality, are essential to ensure focus on pertinent details without introducing bias. For example, integrating supervised model predictions into prompts as metadata enhances alignment with biomechanical cues, improving differentiation of similar activities. However, poorly designed prompts can introduce noise or irrelevant data, underscoring the need for ongoing refinement in prompt engineering. Traditional HAR approaches, whether skeleton-based or context-centric, exhibit notable limitations. Skeleton-based models excel in structured activity recognition but lack environmental and semantic awareness, leading to misclassifications in visually similar activities. Context-centric models offer broader insights but sacrifice precision in activity-specific recognition. Our architecture overcomes these shortcomings by integrating biomechanical and contextual information, achieving a comprehensive understanding of human behaviour. This synergy is evident in our experiments, where combining context with keypoints yielded 60% accuracy, surpassing context-only (40%) or supervised-only (35% on unseen data) approaches.

While our architecture significantly enhances accuracy and adaptability, it introduces challenges. The integration of sensing, reasoning, and memory components increases system complexity, potentially complicating implementation in resource-constrained environments. However, the modular design allows for component optimisation without compromising the framework. Additionally, processing multiple frames and extensive contextual data may hinder real-time performance, a critical requirement in HRI. Techniques such as CNN for efficient feature extraction could mitigate this, as demonstrated in real-time HAR studies. Notably, the primary focus of this research is to demonstrate the potential of memory-driven, context-aware agentic AI, with increased complexity as a necessary trade-off for improved performance. Future refinements will address these concerns. The ability to accurately recognise human activities contextually has profound implications for assistive robotics, healthcare, and industrial applications, enabling robots to anticipate user needs and enhance safety. The experiential learning capabilities of agentic AI foster trust and acceptance in HRI by refining responses over time. Future research should focus on optimising memory mechanisms, fine-tuning LLMs for efficiency and accuracy, and exploring hybrid approaches that balance real-time performance with robust recognition, building on the scalable foundation provided by this work.

6 Conclusion

In conclusion, our research demonstrates the substantial benefits of integrating agentic AI architectures with memory-driven reasoning for improved human activity recognition in HRI. By combining skeletal and contextual information through sophisticated memory mechanisms, our approach significantly surpasses traditional HAR methodologies, offering a robust framework capable of discerning complex human activities with greater accuracy and reliability. The inclusion of STM and LTM has proven essential for capturing both immediate activity transitions and long-term contextual associations, thereby enhancing robotic responsiveness and adaptability in dynamic environments. Future work will explore the fine-tuning of foundational models and the optimisation of memory structures to further improve computational efficiency and semantic precision. Ultimately, our architecture presents a scalable and adaptable solution, paving the way for more natural, intuitive, and effective human-robot collaborations.

References

1. Abadi, M.B., Alashti, M.R.S., Holthaus, P., Menon, C., Amirabdollahian, F.: Rhm: Robot house multi-view human activity recognition dataset. In: ACHI 2023: The Sixteenth International Conference on Advances in Computer-Human Interactions. IARIA (2023)
2. Alashti, M.R.S., Abadi, M.B., Holthaus, P., Menon, C., Amirabdollahian, F.: Lightweight human activity recognition for ambient assisted living. In: ACHI 2023: The Sixteenth International Conference on Advances in Computer-Human Interactions. IARIA (2023)
3. Alashti, M.R.S., Abadi, M.B., Holthaus, P., Menon, C., Amirabdollahian, F.: Rh-har-sk: A multi-view dataset with skeleton data for ambient assisted living research. In: ACHI 2023: The Sixteenth International Conference on Advances in Computer-Human Interactions. IARIA (2023)
4. Alashti, M.R.S., Abadi, M.H.B., Holthaus, P., Menon, C., Amirabdollahian, F.: Efficient skeleton-based human activity recognition in ambient assisted living scenarios with multi-view cnn. In: 2024 10th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob). pp. 979–984. IEEE (2024)
5. Arrotta, L., et al.: Neuro-symbolic approaches for context-aware human activity recognition. Pattern Recognition **139**, 109–123 (2023)
6. Bandura, A.: Social cognitive theory: An agentic perspective. Annual Review of Psychology **52**(1), 1–26 (2001)
7. Borghoff, U.M., Bottoni, P., Pareschi, R.: Human-artificial interaction in the age of agentic ai: a system-theoretical approach. Frontiers in Human Dynamics **7**, 1579166 (2025)
8. Bornet, P., Wirtz, J., Davenport, T.H., De Cremer, D., Evergreen, B., Fersht, P., Gohel, R., Khiyara, S., Sund, P., Mullakara, N.: Agentic Artificial Intelligence: Harnessing AI Agents to Reinvent Business, Work and Life. Irreplaceable Publishing (2025)
9. Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., Francesca, G.: Toyota smarthome: Real-world activities of daily living. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
10. Ghamati, K., Amirabdollahian, F., Resende Faria, D., Zaraki, A.: Cognitive agentic ai: Probabilistic novelty detection for continual adaptation in hri. In: 2025 34th IEEE International Conference on Robot and Human Interactive Communication (ROMAN). pp. 1–7. IEEE (2025)
11. Ghamati, K., Banitalebi Dehkordi, M., Zaraki, A.: Towards ai-powered applications: The development of a personalised llm for hri and hci. Sensors **25**(7), 2024 (2025)
12. Ghamati, K., Zaraki, A., Amirabdollahian, F.: Ari humanoid robot imitates human gaze behaviour using reinforcement learning in real-world environments. In: 2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids). pp. 653–660. IEEE (2024)
13. Graves, A., Graves, A.: Long short-term memory. Supervised sequence labelling with recurrent neural networks pp. 37–45 (2012)
14. Irfan, S., Anjum, N., Masood, N., Khattak, A.S., Ramzan, N.: A novel hybrid deep learning model for human activity recognition based on transitional activities. Sensors **21**(24), 8227 (2021)

15. Karunaratne, G., Schmuck, M., Le Gallo, M., Cherubini, G., Benini, L., Sebastian, A., Rahimi, A.: Robust high-dimensional memory-augmented neural networks. Nature communications **12**(1), 2468 (2021)
16. Lemaignan, S., Warnier, M., Sisbot, E.A., Clodic, A., Alami, R.: Artificial cognition for social human–robot interaction: An implementation. Artificial Intelligence **247**, 45–69 (2017)
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
18. Niemann, J., et al.: Context-aware human-robot collaboration in assembly tasks. Robotics and Computer-Integrated Manufacturing **72**, 102–115 (2021)
19. OpenAI: Chatgpt. https://chat.openai.com/ (2024), accessed: 2025-04-30
20. Pointeau, G., Dominey, P.F.: The role of autobiographical memory in the development of a robot self. Frontiers in neurorobotics **11**, 27 (2017)
21. Real, R.: Tables of significant values of jaccard's index of similarity. Miscel· lania Zoologica pp. 29–40 (1999)
22. Shahabian Alashti, M.R.: Human and activity detection in ambient assisted living scenarios (2024)
23. Silver, D., Sutton, R.S.: Welcome to the era of experience
24. Yang, Others: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024)
25. Yurur, O., Liu, C.H., Moreno, W.: A survey of context-aware middleware designs for human activity recognition. IEEE Communications Surveys & Tutorials **16**(3), 1406–1424 (2014)
26. Zaraki, A., Mazzei, D., Giuliani, M., De Rossi, D.: Designing and evaluating a social gaze-control system for a humanoid robot. IEEE Transactions on Human-Machine Systems **44**(2), 157–168 (2014)
27. Zaraki, A., Pieroni, M., De Rossi, D., Mazzei, D., Garofalo, R., Cominelli, L., Dehkordi, M.B.: Design and evaluation of a unique social perception system for human–robot interaction. IEEE Transactions on Cognitive and Developmental Systems **9**(4), 341–355 (2016)
28. Zaraki, A., Giuliani, M., Dehkordi, M.B., Mazzei, D., D'ursi, A., De Rossi, D.: An rgb-d based social behavior interpretation system for a humanoid social robot. In: 2014 Second RSI/ISM International Conference on Robotics and Mechatronics (ICRoM). pp. 185–190. IEEE (2014)