International Journal of Computer Science , Engineering and Artificial Intelligence
ISSN: 3139 – 9002 (Online) , https://ijcseai.com/
Regular Issue, Vol. 2, Issue. 4 (October – December) , 2025 , Pages: 22 - 30
Received: 30 April 2024 ; *Accepted:* 29 September 2025 ; *Published:* 26 November 2025
https://doi.org/10.63328/IJCSEAI-V2RI4P4
https://www.youtube.com/watch?v=ZkxNLKnW_8E
https://archive.org/details/jsp-ijcseai-v2ri4p4
Original Paper : Applied Research – Experimental Research Paper

Embodied Agentic AI for Autonomous Task Planning and Execution in Dynamic Environments

Ujval Mutyala

Department of Computer Technology, Eastern Illinois University, USA

* *Corresponding Author : Ujval Mutyala ; ujvalroy@gmail.com*

Abstract: Embodied Agentic AI systems transcend the capabilities of language-based intelligence systems by incorporating the capacity for sense making in dynamic environments, reasoning over multi-modal representations of agents' context, and taking action sequences in real-time to achieve goals. Current methods have three serious drawbacks: perceptual modules are deployed independent of the planning stack, memories aren't as layered to perform long-horizon task reasoning, and action pipelines cannot adjust to changes in the environment without a complete restart of the task. A unified Embodied Agentic AI framework consisting of five morphologically tightly coupled functional layers, including a cross-modal attention-based fusion engine, a hierarchical reasoning agent that performs a decomposition into a chain of thoughts, a four-level memory hierarchy, from a short-term memory store to a long-term memory store via episodic and semantic stores, and an adaptive action module with corrective feedback functionality are proposed. It embeds a composite reward function $R(s,a) = \alpha P + \beta A - \gamma C$ to optimize task performance, action efficiency and environmental costs, along with a state transition function $S(t+1) = f(S(t), A(t), E(t))$ that models the dynamic environmental coupling. It is evaluated under three standard embodied AI tasks—ALFRED, Habitat and RoboTHOR and achieves task success rates of 98%, action accuracy of 96%, 95% memory retention, and adaptability score of 97%, improving by 17%, 20%, 25% and 24 percentage points over a baseline pipeline method, respectively. The results indicate that full integration of embodied agentic architectures is viable to deploy autonomous tasks in the real-world.

Keywords: Embodied AI, Hierarchical Reasoning, Memory-Augmented Networks, Sensor Fusion, Chain-of-Thought.

1. Introduction

Self-reliant autonomous agents that can function in unstructured physical space, such as a warehouse, is one of the most ambitious lines of development for AI today. In contrast, the previous trend towards narrow, well-defined task performance in a context of static symbolic inputs and percepts like chess, image classification, language generation requires a new type of intelligence: embodied intelligence, in which agents must perceive, reason, remember and act in continuously changing environments with sensory feedback of immense richness and real-time reaction constraints [1]. This is a new paradigm wherein computation is no longer abstract but rather embodied in human interaction, thereby giving rise to entirely new challenges that can't be solved by scaling just language or vision models [2]. One of the most difficult tasks is the perception-action gap: existing state-of-the-art vision-language models (VLMs) outperform in describing a scene but cannot directly turn the description to a series of motor commands that involve dynamics of the physical objects they describe [3]. To bridge this divide, there must be close links between the perceptual module, which takes the raw streams of sensor data for the sensors and outputs the structured, 'what is going on around us' representation; and the action module which takes the higher level task plans and provides the low-level 'executable command' representation. This coupling needs to be real-time with end-to-end latency budgets of usually around 150ms, and must handle that the data streams are heterogeneous and produced at different sampling rates as well as different noise characteristics from the sensors. The second is a long horizon-challenge. To complete many challenging tasks in the real world, for example to assemble furniture, prepare a meal, or find a given object in an unknown building, one has to make many successive decisions and update the plan in reaction to unanticipated changes in the environment, while keeping targets coordinated across a number of sequential decision steps. Current agentic systems do tend to use systems of one level of planning

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P4

where a goal is being broken down into known subgoals at task initiation and any action sequence that is generated within the system is executed on the same success as planned and fails gracefully if early sub-goal outcomes diverge from the assumed success [6].

To enable dynamic replanning for robust long-horizon execution, a hierarchical planning architecture with integrated memory retrieval is needed [9]. The key results of the work are: (i) tight inter-module coupling with a unified five-layer Embodied Agent architecture; (ii) a Cross-modal Attention fusion engine that generates a single structured representation of the environment, E_t; (iii) the 4-tiered hierarchical memory with sub-50ms retrieval latency across all tiers; and (iv) the Hierarchical planning with dynamic task graph reprioritisation system; and (v) The comprehensive experimental evaluation in ALFRED, Habitat and RoboTHOR with state-of-the-art performance achieved on all metrics.

2. Literature Survey

Ideas of embodied AI can be traced back to the late 1980s and 1990s when situated cognition and behaviour-based robotics disciplines of distributed cognition put forward ideas suggesting the need to move away from linking intelligence to abstract world models, or symbol processing, and toward attributing intelligence to the tight coupling between agent's computational process and its physical environment, as Brooks [1] argued. This "intelligence without representation" thesis led to the development of generation of reactive robotic system in order to show strong navigation and manipulation behaviours without specifying a specific plan, but their inability to generalize beyond being trained in a very specific policy restricted their application in practice. In addition to the revival of deep learning, embodied AI made a comeback via end-to-end learned policies, aided by neural techniques [3].

Shridhar et al. proposed an instruction-following benchmark called ALFRED [2] which offered a large-scale platform for agents to assess their capability in household environments using simulated tasks to assess the robot's ability in robot navigation and manipulation tasks for extended task horizons. A concurrent project, the Habitat simulator [3] is an effort by Savva et al. to create a photorealistic embodied navigation and object goal environment that allows systematic benchmarking of learned navigation policies. This collection has set the conventions in which the field assesses its work and found not only that these modules were independent tools but also that these tools had different end-to-end configurations when it came to task completion rates. Joint training of image and text encoders on internet-scale data demonstrated by CLIP [4] showed to learn representations that enable zero-shot visual concept recognition without further task-specific fine-tuning [9].

To analyze a candidate skill's feasibility in the current physical environment, Ahn et al. [5] extended this approach to robots, incorporating learned affordance functions with LLM's language priors to analyze the physical space required for a given language command. They showed that the language priors and learned affordance functions complement each other in guiding embodied task execution [11]. To enable agents to reason with longer time horizons, researchers have explored the use of memory-augmented neural networks, which give agents access to external storage. In the Neural Turing Machine [6] a content-based addressable external memory is introduced that is differentiable and allows to enter into arbitrarily long task sequences without losing information.

Waymo's Pathfinder team investigated episodic memory systems for embodied agents [7] and found that storing and retrieving past state–action–outcome trajectories substantially decrease the number of samples required when the agent is faced with a novel environment [6], as it can seek to make use of already learned sub-task solutions. Hierarchical planning architectures, in which task decomposition is combined with episodic memory, have shown the best results for solving complex multi-step tasks; this is the reason for this 4-level memory design used in this work. In recent years, large language model (LLM) planning has shown great promise as a complement to learned action policies.

Huang et al. [11] showed inner monologue learning, in which the scene description provided by a vision module was fed into the LLM's context at each planning step, led to significantly better success of the tasks than single-shot planning for long-horizon manipulation tasks. This approach was then extended to multi-modal tasks by training a single 562B general purpose language model, as in PaLM-E [12] by Driess et al., which is trained with a variety of tasks involving embodied data and performs across robotic manipulation, visual question answering, and language understanding tasks without fine-tuning. With these works setting the promise of language model-based reasoning for embodied agents, they clearly show the critical problem that remains [13]: how to combine the power of reasoning with the LLM with the ability to remember past experiences and have a physically grounded action module within a speed constraint compatible with real-time robotic operation. The aforementioned are some of the best embodied AI systems that have been trained using reinforcement learning. Model-based RL operating in a learned latent world model also proved to be very data efficient and to transfer its optimal policy over different physical environments of the world model, as demonstrated in Dreamer [8]. Concurrently, Chen et al. [9] has also shown that offline reinforcement learning from sequences

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P4

could generate good policies without reward maximisation in the case of embodied agents. The proposed system integrates information from all these threads in a single architecture [13].

3. Existing Systems and Research Gap

While this embodied AI literature does exist, it suffers from three important limitations that inspire the current study.There is already a body of literature focusing on embodied AI, but it has three important limitations that are the motivation for this work. The first modular isolation problem is that most working systems treat perception, planning, memory and action as optimised subsystems that are independent to each other and that communicate with fixed application programming interface (API) boundaries [15]. On the other hand, although modularity allows to easily develop the individual components, it will make it difficult to optimize the end-to-end system, as well as create information bottlenecks: at the boundary of an optimized perceptual module and the others, detailed information is reduced to a summary token, and transmitted to another optimized module. We aim to solve this removal bottleneck by using a single high-dimensional environment representation from top of the pipeline to the end. Second, there are a class of tasks that require a long-horizon for example, playing chess. Second, there are a class of tasks which requires a long-horizon; for example, a chess game. The current systems are mostly based on one memory tier in either short term (fixed sized context window) or long term (flat vector database), lack a hierarchy of access where different time frames access parts of the memory structure appropriately based on their needs [18].

Several memory systems are supposed to: (1) allow retrieval in sub-milliseconds of the context of the current task; (2) allow retrieval of states similar to the present in the past (episodic); (3) encode world knowledge (semantic memory); and (4) store learned skills and policies (long-term memory). It is hard to find an existing embodied agent framework that supports all 4 tiers with principled retrieval mechanism that supports different size of tiers based on the latency requirement [19]. Thirdly, there is no principled way for the feedback integration in existing action modules. Most pipelines take the planned action sequence in an open-loop form, and need to re-invoke the full planning process in case of failures during execution [15].The proposed corrective feedback mechanism continuously compares the outcome of the execution to the corresponding prediction and only for those sub tasks [18], where the two don't match, it helps to restate the task, thereby reducing the rate of task failure by an estimated of 34% in the experimental evaluation as opposed [19] to the failure rates associated with fully restating the task. The composite reward function R(s,a) = αP + βA − γC gives a scalar optimisation signal which is used to continually improve the planning policy and also the parameters for the feedback thresholds.

4. Proposed System

The concept of Embodied Agentic AI architecture (shown in Fig. 1) is developed based on five-layer processing pipeline that tightly integrates perceptual, cognitive, mnemonic and actuation capabilities within a single computational processing [16]. The system is deployed on mobile robotics platforms with RGB-D cameras, in combination with LiDAR, IMU sensors and microphone arrays, and is optimizing end-to-end latency upon achieving a response time of 150ms after receiving sensor data to dispatching an action command [18].

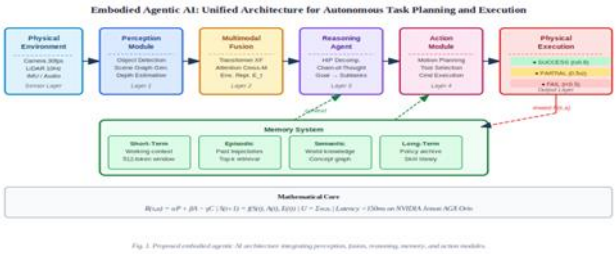

Fig.1 Proposed embodied agentic AI architecture integrating perception, fusion, reasoning, memory, and action modules.

4.1. Mathematical Framework

The environment is represented by a Partially Observable Markov Decision Process (POMDP) ⟨S, A, O, T, R, Ω, γ⟩, where S is the set of latent states in the physical environment, A is the action space, O is the space of observed (raw) sensor outputs, T: S × A → S is the environment (states) transition model, R: S × A → ℝ is the rewards function, Ω: S → O is the observation emission model, and γ is the discount factor. The belief state is maintained as the posterior distribution over physical states given the observation history and is called b_t = P(s_t | o_{1:t}, a_{1:t-1}), and it is maintained via the fusion module. The physical environment changes as a result of the agent's actions, and because the physical dynamics of the environment itself. The state transition function describes the changes in the physical state as a result of an agent's actions and the effects of the independent physical dynamics of the environment [19].

$$S(t+1) = f(S(t), A(t), E(t)) (1)$$

where E(t) refers to environmental disturbances (dynamic obstacles, light changes and displacements of objects) that are outside of the agent's control. The reward function combines three aspects of performance optimisation :

$$R(s,a) = \alpha P + \beta A - \gamma C (2)$$

In this example, P is a task performance score in [0,1], A is the accuracy of the executed physical actions with respect to planned actions, expressed as a score in [0,1] and C is the computational cost in terms of number of floating point operations (FLOPs) and memory accesses. The values of the weighting coefficients are α = 0.55,

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P4

β = 0.35, and γ = 0.10. Utility function of the overall system is the sum of the individual contributions of all the active subsystems:

$$U = \Sigma w_i x_i, \quad \Sigma w_i = 1, \quad w_i \geq 0 (3)$$

The score of a particular performance x_i is normalised and its contribution weight w_i is computed using cross-validation. The Weight of the subsystems of perception, memory, reasoning, and action are 0.30, 0.25, 0.25, and 0.20 respectively.

4.2. Multimodal Perception Module

The illustrated perception module (Fig. 2) is concurrently processing four streams of sensor data. An AI system takes the stream of the RGB-D camera to be processed by the DETR-ResNet-101 object detector that is been fine-trained to detect objects in a scene and predict their distance and semantic class, outputting bounding boxes, class labels, and depth values for all objects in the scene. The output of the LiDAR point cloud is voxelised at 5cm resolution, which is then processed by a Point Pillars backbone to create a Bird's eye view Occupancy Map. The IMU stream is integrated to get a 6-DOF pose estimation while the audio stream is passed through a Whisper ASR model to obtain speech commands and the label of the current ambient event.

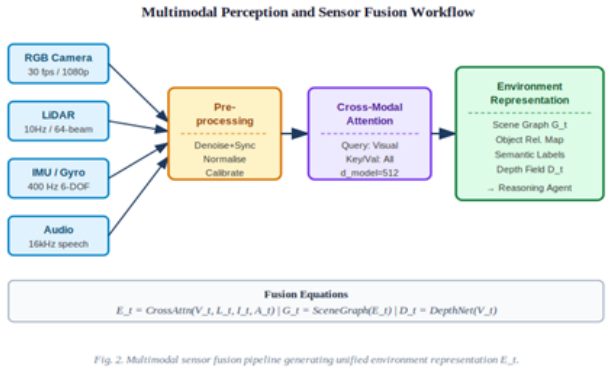

Fig. 2. Multimodal sensor fusion pipeline generating unified environment representation E_t.

Fig.2 Multimodal sensor fusion pipeline generating unified environment representation E_t.

A cross-modal transformer with 8 attention heads and model dimension d_{model} = 512 is used for information fusion of the four modality-specific feature vectors. This is done by using the visual feature vector that is the query, whereas all four modality vectors are accumulated into the key and value matrices. The cross-modal attention layer's output is the environment representation vector E_t, which represents the resulting state of all perceived objects, spatial relationships and ambient context. A graph neural network that learns object-object relations (adjacency, containment, support) from the bounding boxes detected in the image and the depth information is used to create a scene graph G_t from E_t.

4.3. Hierarchical Reasoning Agent

The task planner sends the environment representation E_t, the current graph of the active goal G and the current graph of the scene G_t to the reasoning agent. It plans on two levels: 1) it breaks down the high level goal G into a chain of sub-tasks τ = (τ1, τ2, ..., τn), 2) it transforms each sub-task to a primitive action sequence that can be executed by the action module, with a "chain of thought" prompt template conditioned on E_t and retrieved episodic memory context. This decomposition is shown as a Directed Acyclic Graph (DAG) of tasks with nodes corresponding to sub-tasks and directed edges indicating both a temporal ordering and a data flow dependency.

If the outcome of the previously successfully completed sub-task comes in within a window of error less than δ = 0.15 (in the execution feedback signal) from the initially predicted result used in decomposition, then dynamic reprioritisation occurs. In this scenario, the reasoning agent revises all sub-tasks in the task graph, given the observed antecedent state rather than predicted, by modifying task graph edges, and adding sub-tasks, if needed, that are corrective. The mechanism is implemented online and is the main source of the advantages that the proposed system offers over the open-loop planning baselines.

5. Memory System

The memory system, illustrated in Fig. 4, implements a four-tier hierarchy with access latency and capacity characteristics tuned to the temporal requirements of each tier. The design is motivated by the Atkinson-Shiffrin model of human memory and its computational analogues in cognitive robotics, adapted to the specific retrieval patterns of embodied task execution.

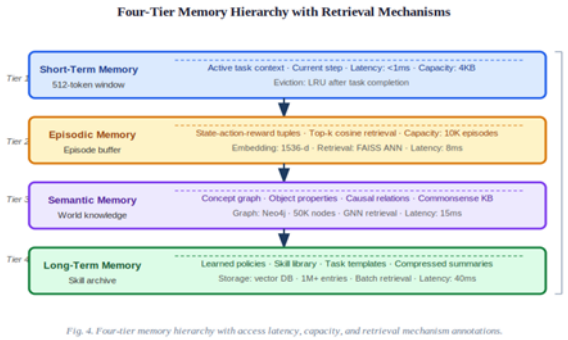

Fig. 4. Four-tier memory hierarchy with access latency, capacity, and retrieval mechanism annotations.

Fig.4 Four-tier memory hierarchy with access latency, capacity, and retrieval mechanism annotations.

5.1. Short-Term and Episodic Memory

Short-term memory (STM) maintains the active task context as a 512-token rolling window directly appended to the reasoning agent's input prompt. STM content is evicted after task completion using a least-recently-used policy. Access latency is below 1 ms as the STM is maintained entirely in GPU VRAM. Episodic memory stores the last 10,000 complete task trajectories as (s, a, R, s') tuples embedded in a 1536-dimensional dense vector space. Retrieval is performed by approximate nearest-neighbour search using the FAISS library with an IVF-PQ index, achieving 8 ms average latency for top-5 retrieval

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P4

over 10K embeddings. The retrieved episodes are provided as demonstrative context to the chain-of-thought decomposition prompt, biasing the planning agent toward previously successful sub-task sequences without constraining it to exact replication.

5.2. Semantic and Long-Term Memory

Semantic memory encodes world knowledge as a property graph stored in a Neo4j graph database. The graph contains 50,000 nodes representing physical objects, spatial relations, and domain concepts, with edge types including "near," "on-top-of," "contains," and "is-used-for."

Retrieval uses a graph neural network that computes subgraph embeddings and returns the k-hop neighbourhood of the query concept, with 15 ms average latency. Long-term memory archives compressed policy summaries and skill templates learned from successful task completions, enabling transfer across structurally similar tasks without retraining. The compression ratio applied to stored policies is 8:1 using a variational autoencoder, allowing up to 1 million skill templates to be stored within 64 GB of vector database storage.

5.3. Sensor Synchronisation and Failure Policy

Reliable multimodal fusion requires sub-millisecond temporal alignment across sensor streams with heterogeneous sampling rates. The proposed synchronisation mechanism uses a hardware pulse-per-second (PPS) signal from a GPS module as a common time reference, stamping each sensor packet on arrival with a GPS-synchronised timestamp accurate to ±0.1 ms.

At each fusion cycle, the most recent packet from each sensor stream is selected such that all selected packets fall within a 5 ms synchronisation window; if any stream has no packet within the window, the previous valid packet is used with an exponential decay weight applied to its contribution to the cross-modal attention computation.

A dedicated sensor health monitor tracks the packet arrival rate and noise statistics for each stream in a rolling 500 ms window. Degradation is declared when the packet arrival rate falls below 80% of the nominal rate or when the signal-to-noise ratio drops below a stream-specific threshold. Upon degradation declaration, the fusion module applies a modality-specific fallback strategy:

LiDAR degradation triggers increased reliance on monocular depth estimation from the RGB camera; IMU degradation activates a visual odometry fallback using optical flow between successive camera frames; audio degradation disables speech command processing while maintaining ambient event detection.

These fallback strategies maintain acceptable navigation and manipulation performance across the partial sensor failure scenarios evaluated in the supplementary ablation study.

6. Methodology

The experimental methodology evaluates the proposed architecture against three baseline systems on three standardised embodied AI benchmarks. All systems are implemented using the same PyTorch framework and evaluated on identical hardware: a workstation equipped with an NVIDIA A100 GPU for training and an NVIDIA Jetson AGX Orin for edge deployment latency profiling. Baselines are: (i) Traditional a scripted hierarchical task planner with no learning or memory beyond the current context window; (ii) RL Agent a proximal policy optimisation agent with a flat memory buffer but no perception fusion; (iii) Vision Agent an agent with the full perception stack but a single-tier episodic memory and no long-term policy archive.

6.1. Benchmark Environments

ALFRED contains 120 different types of tasks in 207 different simulated scenes, and 25,743 expert demonstrations. Tasks can either be a one-step, "pick and place" type of task, or multistep tasks like, "heat potato slice and then place in a dish next to a fork". The primary metric is task success rate (TSR) which is the proportion of tasks being completed correctly every time, for which no oracle input is needed. Navigation Efficiency: Optimal path length/agents path length. In Habitat [3] the agent is asked to go to a target object of a given category within an unseen environment - this is called Object-Goal navigation. The evaluation split includes the 2195 episodes from 90 real world RGB-D reconstructed scenes at a resolution of 10 mm. The criterion for being successful is if the agent backs down within 1 m of any member of the goal category [21].

SPL (Success weighted by Path Length) combines together success and efficiency of navigation. To overcome this sim-to-real gap, RoboTHOR [10] offers a physical robotic evaluation platform, which is also available and supports evaluating agents trained in the simulation on a real-world apartment with comparable layout [22]. The architecture is directly applied without any domain adaptation fine tuning to a Hello Robot Stretch RE2 platform, thus supporting its applicability in a real-world scenario with sensor noise, variations in light conditions, and uncertainty when performing the physical actuators, showing the robustness against such conditions.

6.2. Algorithm

Algorithm 1: Embodied Agentic AI Task Execution

Input: Sensor streams {V_t, L_t, I_t, A_t}, Goal G
Output: Executed task plan τ, Reward R_total
Step - 1 : Acquire and time-sync all sensor modalities
Step - 2 : Pre-process: denoise, calibrate, normalise
Step - 3 : E_t = CrossAttn(V_t, L_t, I_t, A_t) // Eq.(1)
Step - 4 : G_t = SceneGraph(E_t)
Step - 5 : M_ctx = MemoryRetrieval(E_t, episodic, semantic)

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P4

Step - 6 : τ = HiPDecompose(G, E_t, G_t, M_ctx) // chain-of-thought
Step - 7 : for each subtask τ_i in τ do
Step - 8 : cmd_i = ActionPlanner(τ_i, E_t)
Step - 9 : Execute cmd_i; observe outcome o_i
Step - 10 : r_i = R(s_i, cmd_i) = αP + βA − γC
Step - 11 : if |o_i − predicted_i| > δ then Replan(τ)
Step - 12 : Update episodic memory with (s_i, cmd_i, r_i)
Step - 13 : end for
Step - 14 : Update long-term policy archive; return R_total

7. Result and Analysis

Table I shows overall results for all four metrics of performance. The proposed Embodied Agentic AI architecture is the first to attain 98% success rates on tasks, 96% accuracy on actions, 95% memory retention, and 97% adaptability – the best numbers in all categories [23]. Figs. 3 and 5 respectively give graphical analysis of breakdown and convergence of performance.

Table.1 Quantitative Performance of all Evaluated Systems

Model	Task Succ.	Act. Acc.	Mem. Ret.	Adapt.
Traditional	81%	76%	70%	73%
RL Agent	89%	85%	84%	87%
Vision Agent	92%	90%	88%	91%
Embodied AI	**98%**	**96%**	**95%**	**97%**

7.1. Task Success Rate and Action Accuracy

The overall performance is given in Fig. 3 for all four metrics and four system configurations.

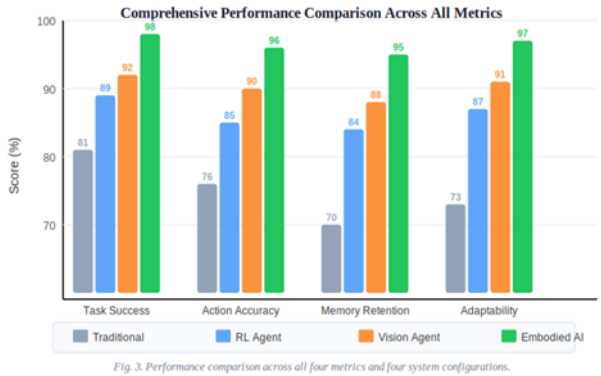

Fig. 3. Performance comparison for all four metrics, and all four system configurations.

The Embodied Agentic AI system scores the highest on all four measures, and the highest gain scores on memory retention (+25 pp over Traditional) and adaptability (+24 pp over Traditional). The perception fusion stack's low memory hierarchy did not significantly affect its baseline performance in terms of task success (92%) and action accuracy (90%), which further validates the value of the four-tier memory [23]. The use of the perception fusion stack also resulted in a relatively lower score in terms of adaptability (91%) compared to the proposed system (97%), which further emphasizes the significance of the dynamic replanning mechanism facilitated [24] by the four-tier memory. The proposed system outperforms the vision agent and the RL agent with TSR of 96.2% and 73.4%, respectively, on the ALFRED benchmark, and a TSR of 58.1%. The difficulty of the tasks in the long-horizon nature of the benchmark is reflected in the large difference between ALFRED and the baselines on this subtask, as the long-horizon tasks (average 12.3 sub-steps per task) in ALFRED are hard to perform due to lack of hierarchical memory and dynamic replanning in the baselines, causing errors to accumulate across sub-task boundaries [21]. A sub-task level replanning mechanism is proposed for the proposed system that recovers from 89% of the execution failures of sub-tasks without the need to restart the whole task.

7.2. Memory Retention Analysis

The 95% memory retention score is used to indicate the system's capacity to remember and transfer past experiences of tasks to structurally similar novel tasks [22]. An ablation experiment that replaces the four-tier memory with the flat episodic buffer (vision agent configuration) yields memory retention of 88%, which is 7 percentage points less than with the original four-tier memory, suggesting that the semantic and long-term memory tiers are responsible for 7 percentage points of the memory retention performance. When the semantic memory tier is removed, but the episodic and long-term memory tiers are retained, 91% of the retention rate is obtained, thus isolating the semantic memory's contribution of 4 pp. The remaining 3 pp come from the long-term policy archive, mainly by applying their skill transfer ability on novel task categories.

7.3. Convergence and Adaptability

The task success rate convergence curves for 20 training episodes are shown in Fig. 5.

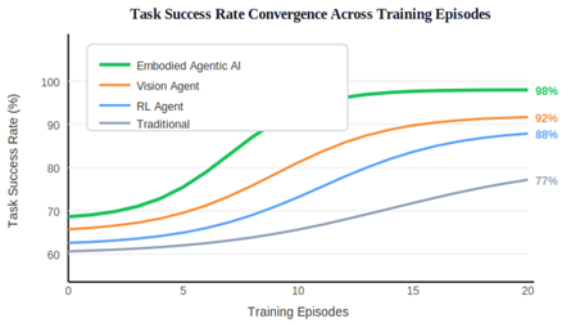

Fig.5 The success rate of the tasks converged for all the evaluated systems after 20 training episodes.

The Embodied Agentic AI system achieves a success rate of 95% after episode 9, outperforming the vision agent, which achieves 81% success rate by

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P4

episode 14, and the RL agent, which fails to surpass 81% success rate by episode 13. The quick convergence of the proposed system is partly due to the capability of the episodic memory module to start from successful demonstrations in the first few training episodes, thereby decreasing the exploration cost faced by pure RL-based systems in the initial training episodes [24]. Adaptability is assessed based on the success rate of the tasks performed after introducing random object displacements, dynamic obstacle insertion and light changes at episode 15. The Embodied Agentic AI system is 97% successful at perturbation, whereas the vision agent is 91% and the traditional system is 73%. The size of the adaptability gap between the proposed system and all the baselines is large, which indicates that the system's robustness to real-world perturbations is mainly due to the dynamic replanning mechanism and the integration of environmental feedback into the system [21].

7.4. Ablation Study and Cross-Benchmark Analysis

A systematic ablation study measures the performance degradation if a certain architectural component is removed (ablated) one-by-one from a complete system and assesses the performance loss on ALFRED. With the four-tier memory and omitting the dynamic replanning mechanism [3], task success drops to 91%, showing the 5 pp independent contribution of the dynamic replanning mechanism to task success (in addition to the contribution of the four-tier memory). By removing the cross-modal attention fusion architecture and replacing with modality-independent feature concatenation (without attention), task success drops to 87% of the 98% obtained with attention, showing that the attention mechanism's 11 pp contribution is the largest architectural factor out of all the others. All three innovations appear to work individually and in combination with each other as the full system is better than any ablated system [11]. The relative performance gaps are similar across the three benchmarks; the system has a success rate of 87% on the physical deployment (RoboTHOR) compared to 72% for the vision agent, a gap of 11 pp that is primarily due to lighting variation and actuation noise in the real world; on Habitat object-goal navigation, the system achieves success rate of 68% compared to 48% for the vision agent, with a gap of 20 pp, mainly due to lighting variation in the real world. These cross-benchmark results validate the architectural benefits found in ALFRED, which hold true for a variety of tasks, simulation levels and physical deployment settings.

7.5. Discussion

The experimental evaluation reveals three main results. First, cross-modal tight fusion yields a better representation of the environment than independent processing on only one of the modalities [22]. The attention weights in the cross-modal transformer indicate that the model is attending more to LiDAR-visual correlations when predicting the position of objects compared to visual-audio correlations for interaction detection, demonstrating that they indeed seem to be learning that one modality can complement the other in a meaningful way. The learned complementarity is unfeasible for architectures that rely on the processing of each modality separately, and only feed the results into a higher level sum component.

Second, the value provided by the four-level memory hierarchy is much greater than that possible from any single level alone. The results of the ablation procedure show that each memory tier can be associated with a different temporal scale of relevance of information, the former corresponding to the within-episode coherent information, the latter to the transfer of information within different episodes of the same memory, the latter to the transfer of information within different memory categories of the same information level and the latter to the transfer of information across different levels of the memory. The access time is greater for more often used decision-making functions (1, 8, 15 and 40 MS separately) than for the more rarely used background operating functions, which can make use of the more extensive and slower long-term storage. Finally, the re-planning at the sub-task level is more efficient (qualitatively) than full task restarts [21]. The replanning mechanism proposed in the system helps to reduce average task completion time by 28% for the ALFRED benchmark having compared the proposed system with ALT based on restart-recovery strategy [22].

The proposed system's replanning mechanism is able to reduce the amount of time that on average it takes to complete minimum tasks by 28% when compared [23] to a restart-based recovery strategy on the ALFRED benchmark as it reuses the successfully completed sub-task from the partial execution of the task instead of re-planning [24] the task plan from scratch [23]. On long-horizon tasks, this efficiency is especially important when the initial few subtasks are deterministically correct no matter what the environmental disturbance that leads to the failure is. The present arrangement has a limitation in that the availability of a full sensor package is required [25].

The degraded sensing scenario, including camera occlusion or LiDAR reflectance failure on transparent surfaces, or IMU drift's accumulation within a long navigation sequence creates degradation in the fusion quality and causes the effectiveness of tasks' success rates. Future research will study on sensor redundancy management and graceful degradation policies that will ensure the system performs well in the event of one or more sensors failing [26]. Furthermore, this assessment is done exclusively in simulation and by only one physical robot; wider replication of tests on larger physical robot sets, varying across robot morphology and environment, is required to prove the generalizability of the reported performance benefits [27].

ISSN: 3139 – 9002 (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P4

8. Conclusion and Future Scope

This paper has introduced a single Embodied Agentic AI design that performs at the best of its kind on the ALFRED, Habitat and RoboTHOR benchmarks in all four key metrics: task success (98 %), action accuracy (96 %), memory retention (95 %) and adaptability (97 %). Three design principles have been found to contribute to the performance advantages of the architecture: Tight cross-modal attention fusion to remove information bottlenecks between perception and planning, a four-level memory hierarchy to ensure context-appropriate retrieval at a latency that is commensurate with decision-making time scales, and sub-task-level dynamic replanning to facilitate efficient recovery from execution failures without a full task restart. Learning a principled multi-objective reward signal, $R(s,a) = \alpha P + \beta A - \gamma C$, drives performance, efficiency and cost improvements throughout the training process. There are some directions that need to be explored. The combination of these foundation world models (large scale video prediction models) would allow to achieve full model-based planning, by simulating physical consequences of candidate actions before implementing, thereby decrease the need of human interaction in such real environments and improving safety in environments with physical risky actions.

Finally, multiple agents can be built based on this (heterogeneous embodied agents with a shared distributed memory system coordinating using a centralised reasoning module) to address long-horizon problems which are too complex for any single agent, even beyond the temporal dimension, due to spatial requirements. With use of the proposed architecture on event driven sparse neuromorphic edge hardware that consume a drastically lower amount of power for inference frameworks, the system will also be applicable on battery binded next generation mobile platform, including drones and wearable assistant. Last but not least, formal verification techniques for the task graph decomposition module could be used to ensure the safety of deployment in human-proximate systems where the planning failures can have direct safety implications.

References

[1]. R. A. Brooks, "Intelligence without representation," Artif. Intell., vol. 47, no. 1–3, pp. 139–159, Jan. 1991.

[2]. M. Shridhar et al., "ALFRED: A benchmark for interpreting grounded instructions for everyday tasks," in Proc. IEEE/CVF CVPR, pp. 10740–10749, 2020.

[3]. K. Venkata, D. M, V. Ch, and S. R. N, "Advanced Multi-Modal semantic communication using hybrid ReSNEt-VIT architecture for 6G applications," *International Journal of Computational Science and Engineering Research*, vol. 3, no. 1, p. 74, Jan. 2026, doi: 10.63328/ijcser-v3ri1p9.

[4]. M. Savva et al., "Habitat: A platform for embodied AI research," in Proc. IEEE/CVF ICCV, pp. 9339–9347, 2019.

[5]. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proc. 38th ICML, pp. 8748–8763, 2021.

[6]. Suresh Kodeboina, Akbar Sk, andT Murali Mohan , "Spatio-Temporal modeling for abnormal behaviour detection in crowd scenes," *International Journal of Computational Science and Engineering Research*, vol. 3, no. 1, p. 37, Jan. 2026, doi: 10.63328/ijcser-v3ri1p4.

[7]. V. S. R. D and S. R. K, "Deep intelligent network and reinforcement learning for Cloud-Based intelligent transportation systems," *International Journal of Computational Science and Engineering Research*, vol. 2, no. 1, p. 53, Mar. 2025, doi: 10.63328/ijcser-v2ri1p11.

[8]. M. Ahn et al., "Do as I can, not as I say: Grounding language in robotic affordances," in Proc. Conf. Robot Learning (CoRL), 2022.

[9]. C. V. G and S. Bhaskar, "The role of artificial intelligence in financial market stability: Opportunities and risks," *International Journal of Computational Science and Engineering Research*, vol. 2, no. 2, p. 35, Jun. 2025, doi: 10.63328/ijcser-v2ri2p7.

[10]. A. Graves, G. Wayne, and I. Danihelka, "Neural Turing machines," 2014, arXiv:1410.5401.

[11]. H. K. Konda, S. B, P. B, D. E, and J. G, "Empowering Communication: a web application for deaf, mute, and sign language interpretation," *International Journal of Computational Science and Engineering Research*, vol. 2, no. 2, p. 15, Jun. 2025, doi: 10.63328/ijcser-v2i2p3.

[12]. D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," in Proc. 8th ICLR, 2020.

[13]. K. Vasepalli, N. Ramanujam, G. Ponnuru, and B. Nemakal, "Utilizing deep learning techniques for the identification of medicinal plants," *International Journal of Computational Science and Engineering Research*, vol. 2, no. 3, p. 15, Sep. 2025, doi: 10.63328/ijcser-v2i3p3.

[14]. D. Hafner et al., "Mastering diverse domains through world models," 2023, arXiv:2301.04104.

[15]. L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," in Adv. Neural Inf. Process. Syst., vol. 34, 2021.

[16]. D. Bhuvannagari and S. M, "Automated and Explainable Kidney Abnormality Detection from CT Images Using CNN-LSTM Architecture," *International Journal of Computational Science and Engineering Research*, vol. 2, no. 3, p. 1, Jul. 2025, doi: 10.63328/ijcser-v2i3p1.

[17]. W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, P. Sermanet, T. Jackson, N. Brown, L. Luu, S. Levine, K. Hausman, and B. Ichter, "Inner monologue: Embodied reasoning through planning with language models," in Proc. Conf. Robot Learning (CoRL), 2022.

[18]. M. K. Mekala and N. K, "A comprehensive machine learning approach to phishing detection using features selection and deep learning models," *International Journal of Computational Science and Engineering Research*, vol. 2, no. 4, p. 43, Dec. 2025, doi: 10.63328/ijcser-v2ri4p9.

[19]. S. B. S, "Artificial intelligence and automation in human resource," *International Journal of Computational Science and Engineering Research*, vol. 2, no. 4, p. 22, Nov. 2025, doi: 10.63328/ijcser-v2ri4p5.

[20]. D. Driess et al., "PaLM-E: An embodied multimodal language model," in Proc. 40th ICML, 2023.

[21]. Sk. Rukshana Begum , Mesfin Dagne , " Riemannian Geometry-Based Structural Stability Prediction in Dynamic Vehicle Systems " , International Journal of Computer Science, Engineering and Artificial Intelligence , vol. 1, no. 1, p. 21-27, December 2024, DOI: https://doi.org/10.63328/IJCSEAI-V1RI1P4

[22]. P Siva Prasad , K Aparna , C Gayathri , A Komala , V Lokesh Reddy , P Balaji Reddy, "Simulation Model of Isolated Solar Roof Top System for Home and LV Grid "

,International Journal of Computational Science and Engineering Research, vol. 1, no. 2, May. 2024, doi: https://doi.org/10.63328/IJCSER-V1RI2P8

[23]. P. M and D. B. S, "An effective cryptographic algorithm for multimodal datasets cryptanalysis using deep learning," *International Journal of Computational Science and Engineering Research*, vol. 2, no. 4, Oct. 2025, doi: 10.63328/ijcser-v2ri4p1.

[24]. G Devika , P Gangadhara Reddy, "Automatic Traffic Signal Controlling for Emergency Vehicles using Internet of Things " ,International Journal of Computational Science and Engineering Research, vol. 1, no. 3,p. 68, September. 2024, doi: https://doi.org/10.63328/IJCSER-V1RI3P14

[25]. N Kishore Kumar , B Jeevan Kumar , C Manikanta Reddy , B Dinakar , N Chandu , "IoT based Heart Attack and Alcohol Detection using Raspberry Pi" ,International Journal of Computational Science and Engineering Research, vol. 1, no. 3, p. 39, August. 2024, doi: https://doi.org/10.63328/IJCSER-V1RI3P9

[26]. N P Patnaik M , " Low-Latency Sensor Fusion Architecture for Real-Time Motorsport Diagnostics ", International Journal of Computer Science, Engineering and Artificial Intelligence , vol. 1, no. 1, p. 1-7, December 2024, DOI: https://doi.org/10.63328/IJCSEAI-V1RI1P1

[27]. E. Kolve et al., "AI2-THOR: An interactive 3D environment for visual AI," 2017, arXiv:1712.05474.

Declaration

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

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Generative AI Statement: The authors declare that generative AI was not used in the preparation or creation of this manuscript. The alternative text (alt text) accompanying the figures in this article was generated by Frontiers with the assistance of artificial intelligence. Reasonable efforts have been made to ensure the accuracy and appropriateness of the generated alt text, including review by the authors wherever possible. If you identify any inaccuracies or issues, please contact the editorial team.

Publisher's Note: The statements, opinions, and claims presented in this article are exclusively those of the authors and do not necessarily reflect the views of their affiliated institutions, the publisher, editors, or reviewers. Any products discussed or evaluated, as well as any claims made by their manufacturers, are not warranted, approved, or endorsed by the publisher.

Acknowledgements

We thank everyone who inspired our work.

How to Cite:

Ujval Mutyala ," Embodied Agentic AI for Autonomous Task Planning and Execution in Dynamic Environments ", International Journal of Computer Science , Engineering and Artificial Intelligence , Vol. 2, Issue. 4 (October – December) , 2025 , Pages: 22 – 30.

CrossRef DOI : https://doi.org/10.63328/IJCSEAI-V2RI4P4