Multi-Scenario Reasoning: Unlocking Cognitive Autonomy in Humanoid Robots for Multimodal Understanding

Libo Wang

Abstract— **To improve the cognitive autonomy of humanoid robots, this research proposes a multi-scenario reasoning architecture to solve the technical shortcomings of multi-modal understanding in this field. It draws on simulation based experimental design that adopts multi-modal synthesis (visual, auditory, tactile) and builds a simulator "Mahā" to perform the experiment. The findings demonstrate the feasibility of this architecture in multimodal data.**

I.INTRODUCTION

In tandem with the support of cutting-edge GPUs, artificial intelligence technologies such as model predictive control, multi-sensor fusion, visual SLAM, and sim-to-real continue to stimulate the development potential of humanoid robots (Makoviychuk et al. , 2021; Chen et al., 2024; Dao et al., 2024). Given the previous literature on the development of cognitive architecture as a solution to human-level autonomy, cognitive autonomy is still a major problem in the thinking, planning and action selection modules of humanoid robots (Ogunsina et al. al., 2024). For example, Burghart et al (2005) have demonstrated a three-level cognitive architecture that from environmental input to task execution (Fig. 1).
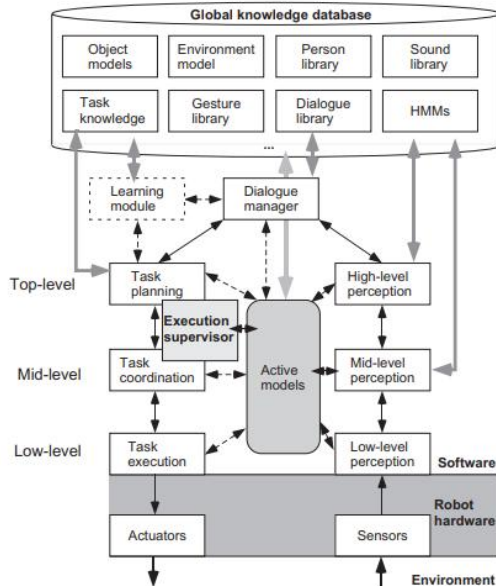

Fig. 1   Cognitive architecture of the Karlsruhe Humanoid Robot (Adapted from Burghart et al., 2005)

However, current technology supports humanoid robots to imitate humans to perform repetitive operations, but it does not mean that the simulation reaches human-level cognitive autonomy (Katiyar & Katiyar, 2021; Ogunsina et al., 2024). The perception and execution capabilities cannot make up for the significant challenges in human-level autonomy that are reflected in cognitive abilities such as thinking, planning, and decision-making (Ogunsina et al., 2024). Target perception, tracking, and repetitive movement planning still rely on visual processing (Guo et al., 2020). Humanoid robots are not only based on pre-trained architectural models of static data, but also need to process multi-modal data (visual, auditory, tactile) and flexible and coherent reasoning (Fisher et al., 2021; Xiao et al., 2023).

The technical shortcomings of multimodal understanding are the key gaps that make it difficult for humanoid robots to achieve cognitive autonomy. Multimodal understanding requires extracting relevant features from different sensory inputs and integrating them into a consistent semantic representation through cross-modal alignment techniques (Duan et al., 2022; Navarro-Guerrero et al., 2023; Tong et al., 2024). However, in practice, it is difficult to efficiently integrate and process multi-modal data such as visual, auditory and tactile data, resulting in ambiguous semantics and incoherent responses.

As mentioned before, architectures that rely on static pre-training data to complete tasks lack the ability to integrate cross-modal data (Ye et al., 2023). This directly causes semantic ambiguity when humanoid robots process feedback with hearing or touch or response incoherence (Pramanick & Rossi, 2024). Although some research attempts multi-modal fusion technology, progress is still limited and is not enough to provide humanoid robots with the same adaptive capabilities as humans (Yuan et al., 2024). To address this gap, this research proposes a multi-scenario reasoning architecture as an innovative solution. It aims to leverage multi-scenario reasoning to optimize the key challenges of cognitive autonomy in humanoid robots' cross-modal understanding of visual, auditory and tactile data based on current technology shortcomings.

II.MULTI-SCENARIO REASONING

A. Theoretical Foundation

The principle of multi-scenario reasoning proposed in this research and applied to humanoid robots is inspired by situated cognition theory. This theory emphasizes that the environment is inextricably linked to the behavior of knowledge generation and application, and that it embodies

meaning through individual interactions in real situations (Jenlink & Austin, 2013).

Specifically, reasoning ability is affected by the real-time integration of multi-modal sensory information from the five senses of vision, hearing, smell, taste, and touch and the behavior of combining situation selection (Wang et al., 2003; Thagard, 2010). The human brain's response to situations The construction and integration process of semantics is carried out in the prefrontal cortex, temporal lobe, and parietal lobe (Jouen et al., 2015). The memory retrieval and contextual association of the hippocampus play an important role in processing environmental information for cross-modal reasoning in multiple scenes (Morici et al., 2022). This process involves the dynamic allocation of attention resources to ensure the human brain's immediate understanding of complex situations and selected optimal actions, which profoundly affects the operation process of the multi-scenario reasoning architecture designed in this research (Nicolini et al., 2024).

As mentioned before, the principle of multi-scenario reasoning is inspired by situated cognition theory, thereby extending its application to the cognitive autonomy design of humanoid robots. Different from traditional technologies that only focus on static scenes or single modal data, it imitates the principle of the human brain to instantly integrate multi-modal data in complex situations to achieve dynamic processing of visual, auditory and tactile information of humanoid robots across scenarios.

B. Conceptual Principles

As a concept proposed by this research to improve the cognitive autonomy of robots, multi-scenario reasoning has been applied to the thinking, planning and decision-making fields of humanoid robots for the first time. It focuses on semantic integration and synchronization of visual, auditory and tactile data from external sensors to make optimal action choices in different dynamic scenarios. It builds a multi-scenario reasoning architecture to simulate real-time processing of multi-modal data and improves the dynamic adaptability of adaptive systems in cross-scenario learning.

Based on the principle of situated cognition theory, multi-scenario reasoning simulates the semantic integration and decision-making capabilities of the human brain's reasoning process in an uncertain environment. In the field of humanoid robots, the application of multi-scenario reasoning focuses on global situation modeling based on multi-modal data, and uses this to perform continuous reasoning and dynamic adjustment. Multi-scenario reasoning promotes cognitive autonomy in complex environments through semantic alignment, synchronized processing and scenarios of multi-modal data during thinking, planning and decision-making.

The core principle of multi-scene reasoning lies in dynamic scene modeling and semantic integration of multi-modal data, thereby solving the key shortcomings of existing multi-modal understanding technology. Based on the principle of situated cognition theory, this concept simulates the semantic integration and decision-making of the human brain in an uncertain environment. From a technical practice perspective, it is based on semantic alignment and uses situational analysis to uniformly represent visual, auditory and tactile data, which optimizes the robot's cross-modal understanding capabilities.Sparse attention effectively optimizes and dynamically adjusts the weight of each modality to highlight key information, thereby improving the accuracy and efficiency of multi-modal data processing (Song et al., 2024). The memory-augmented module provides the ability to trace back data in long-term and dynamic scenarios, ensuring that robots can perform efficient reasoning based on historical context in real-time scenarios (Muthirayan et al., 2020). The scene reasoning results are entity mapped through the sim2real module to promote the seamless connection of the actual operation of the robot (Zhao et al., 2023).

III. PROPOSED ARCHITECTURE & ALGORITHMS

This section is closely connected with the previous theories and concepts to explain in detail the core modules and algorithm display of the multi-scenario reasoning architecture. Fig. 2 explains modules including data input, scenario processing, attention-based prioritization, memory-augmented reasoning, action-decision modeling, sim2real and selected optimal action.
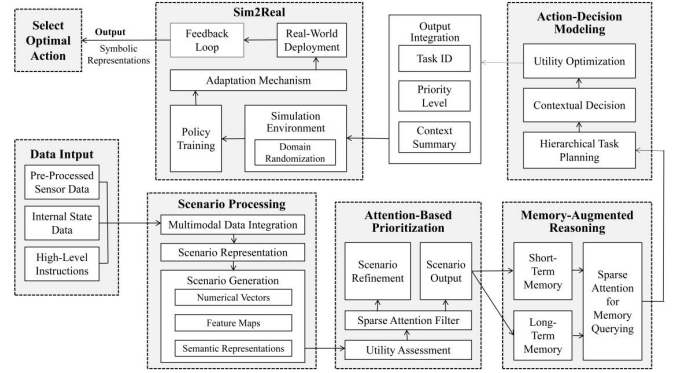
A. Abbreviations and Acronyms

This module is responsible for preprocessing and structuring data from multi-modal perceptrons, and providing high-quality data sources to subsequent modules. The Data input contains three subcomponents. Among them, pre-processed sensor data is used to integrate visual, auditory and tactile data; pnternal state data focuses on processing the internal status information of humanoid robots; high-level instructions are used to parse high-level instructions and generate preliminary semantic data.

The following algorithm is divided into four parts: data reprocessing, multimodal normalization, feature extraction, integrated multimodal output. D is the raw data set; $T(d)$ is a trustworthiness function; τ is the threshold. Normalizes data $D(m)$ for each modality using mean $\mu(m)$ and standard deviation $\sigma(m)$. F_m is extracted features. $m \in \{$visual, auditory, tactile$\}$. Combines extracted features from all modalities into a unified output O.

$$D_{filtered} = \{d \in D \mid T(d) > \tau\}$$
$$D_{norm}(m) = [D(m)-\mu(m)] / \sigma(m),$$
$$F_m = \text{Extract}(D_{norm}(m))$$
$$O = \bigcup_m F_m$$

B. Scenario Processing

This module performs semantic modeling and scene representation for multi-modal data to ensure data consistency and adaptability in dynamic environments. In components, multimodal data integration for semantic integration of cross-modal data; scenario representation converts data into structured representations for robot processing; scenario generation builds scenarios based on feature maps and semantic labels, and generates semantic consistency through semantic representations data output.

The algorithm of this module supports functions such as multimodal data integration, feature mapping, scenario generation and evaluation, and sparse attention-based scenario selection. Notably, R_k is a semantic feature, U_k is data point.

● Input:

S, I, H, weights $\alpha_s, \alpha_i, \alpha_h$.
Unified data vector $U = [U_1, U_2, ..., U_n]$
Feature map M
Scenarios $S = \{S_1, S_2,...,S_m\}$, utilities $U(S_j)$, selection size k.

● Multimodal Data Integration

$$U_k = \alpha_s s_k + \alpha_i i_k + \alpha_h h_k, \quad \forall k \in \{1, 2, ..., n\}$$

● Feature Mapping

$$R_k = f(U_k), \; f(U_k) = \text{round}(U_k, 2)$$
$$M_k = g(U_k), \; g(U_k) = e^{-U_i}$$

● Scenario Generation

$$S_{j,k} = M_k + \Delta(M_k), \; \Delta(M_k) \sim U(-0.1, 0.1)$$

Compute scenario utility:

$$U(S_j) = \Sigma_{k=1}^n S_{j,k}$$

● Sparse Attention-Based Scenario Selection

Select top-k scenarios:

$$S_k = \text{Top - k}(U(S_i)), \quad \forall_j \in \{1, 2, ..., m\}$$

● Output: Selected scenarios S_k.

C. Scenario Processing

The attention-based prioritization module is designed to dynamically filter and weight multi-modal data to optimize inference accuracy and efficiency. Sparse attention filter technology is used to filter out key data; Scenario refinement ensures semantic clarity during the optimization of scene output; Utility assessment evaluates data priority based on scene semantics; output scenario is responsible for generating these outputs.

In algorithms, $r(S_j)$ is the relevance score for S_j. Top-k($\cdot$) identifies the k scenarios with the highest scores; S'_i is the refined scenario. $f(\cdot)$ is a function incorporating memory-based adjustments. a'_j is adjusted attribute of the scenario; $g(\cdot)$: Adjustment function based on memory m_j. $\mu(m)$ represents mean, $\sigma(m)$ represents standard deviation.

● Input

Raw data D, consisting of multimodal input $\{d_1, d_2,..., d_n\}$.
Trustworthiness function $T(d)$ for evaluating data quality.
Threshold τ for data filtering.
Modalities $m \in \{\text{visual, auditory, tactile}\}$.

● Sparse Attention Filtering

$$r(S_i) = \exp(U(S_i)) / \Sigma_{k=1}^m \exp(U(S_k))$$

$$S_k = \text{Top - k}(r(S_j))$$

● Scenario Refinement

$$a'_j = g(a_j, m_j)$$

● Output

Refined top k scenarios $S'_k = \{S'_1, S'_2, ..., S'_k\}$

D. Memory-Augmented Reasoning

This module introduces memory-augmented network to improve the accuracy and coherence of reasoning. Drawing on short-term memory to record the data of the current scene to support instant reasoning, long-term memory saves historical data for cross-scenario correlation analysis. Sparse attention for memory querying retrieves key data through attention screening technology. It simulates dynamic interaction with attention-based prioritization and outputs the reasoning results to the action-decision modeling module for strategy generation (Lou et al., 2024).

These algorithm explains that STM(t) is spdated short-term memory at time t; ΔScenario is new contextual information from current input; LTM$_{\text{relevant}}$ is retrieved memory entry most similar to the query; Sim($\cdot$) is similarity function; LTM$_j$ is historical memory entries. A$_i$ is attention weight for memory entry i; Query is input query vector; Memory$_i$ includes STM or LTM; Output is weighted sum of relevant memory entries.

● Input

Query vector: Query.
Memory entries: Memory = {STM, LTM}.
Historical memory entries LTM$_i$.

● Operations

$$\text{STM}(t) = \text{STM}(t - 1) + \Delta\text{Scenario}$$

$$\text{LTM}_{\text{relevant}} = \arg\max(\text{Sim}(\text{Query}, \text{LTM}_i))$$

$$\alpha_i = \exp(\text{Score}(\text{Query},\text{Memory}_i)) / \Sigma_j \exp(\text{Score}(\text{Query}, \text{Memory}_j))$$

$$\text{Score}(\text{Query}, \text{Memory}_i) = (\text{Query} \cdot \text{Memory}_i) / (\|\text{Query}\| \|\text{Memory}_i\|)$$

● Output = $\Sigma_i \alpha_i \cdot \text{Memory}_i$

E. Action-Decision Modeling

The action-decision modeling module develops optimal action strategies based on the results of memory reasoning. Among them, utility optimization is responsible for calculating action utility; Contextual decision adjusts dynamics based on contextual information; action strategy is decomposed into specific executable steps by hierarchical task planning, and output integration is performed through task ID, priority level, and context summary.

In the following algorithm, h$_i$ is a subtask derived from task t; U is utility based on constraints C; w$_i$ is weights for context factors C$_i$; λ adjusts the impact of previous outcomes. D is decisions

● Input

High-level task T.
Environmental context E.
Context factors C$_i(D)$.
Historical feedback data.

- Operations

$$H(t) = \{h_1, h_2, ..., h_n\}$$
$$P(h_i) = U(h_i|C)$$
$$D_{opt} = \arg \max \text{Utility}(D|E)$$
$$\text{Utility}(D|E) = \Sigma^n_{i=1} w_i \cdot C_i(D)$$
$$U(D) = \text{Predicted Outcome}(D) + \lambda \cdot \text{Historical Feedback}(D)$$

F. Sim2Real

Sim2real is responsible for mapping the planning plans generated by action-decision modeling in the simulation environment to real scenarios to achieve from reasoning to selected optimal action. The simulation environment in this module is used to simulate dynamic scenarios; policy training aims to strengthen the adaptability of behavioral strategies; adaptation mechanism is dedicated to mapping simulated behaviors to real environments; domain randomization enhances generalization capabilities; Real-world deployment implements it in practice.

These components draw on algorithms by Clavera et al (2018), Luo et al. (2018), Pateria et al. (2021), Moerland et al. (2023) and Zhang et al (2024). Specifically, ε is the randomized environment; E is base environment; μ and σ define distribution of variations, P is the probability distribution governing randomization; π is the policy, τ represents trajectories in the simulation; $R(s_t, a_t)$ is reward for state s_t and action a_t, γ is the discount factor; φ is the feature encoder; D is the discriminator distinguishing between simulation and real-world data; L_{adv} is the adversarial loss, L_{task} isthe task-specific loss; R_{real} represents real-world; R_{sim} represents simulated rewards; δ is the reward discrepancy; α is adjustment factor.

- Input

Base environment E, including environmental parameters.
Simulation trajectories τ with states s_t and actions a_t.
Real-world feedback: $R_{real}(s, a)$.
Simulated rewards: $R_{sim}(s, a)$.
Hyperparameters: $\mu, \sigma, P, \gamma, \alpha, \lambda$.

- Operations

$$\varepsilon = \text{Randomize}(E; \mu, \sigma, P)$$
$$\pi = \arg \max E_{\tau \sim \pi}[\Sigma^T_{t=0} \gamma^t R(s_t, a_t)]$$
$$min_\varphi max_D L_{adv}(\varphi, D) + \lambda L_{task}(\varphi)$$
$$\delta = R_{real}(s, a) - R_{sim}(s, a)$$
$$\pi^+ = \arg \max_\pi [\Sigma^T_{t=0} \gamma^t (R_{real}(s_t, a_t) + \alpha\delta)]$$

G. Selected Optimal Action

Serving as the output module of this architecture, selected optimal action integrates the results generated by sim2real to execute optimal action instructions. Through dynamic analysis and decision-making optimization, this module ensures the accuracy of action selection performed by the humanoid robot and optimizes scene processing with feedback data.

IV. EXPERIMENTATION

This research draws on the principle of simulation based experimental design in the category of quantitative experimental research methods to test the effectiveness of the multi-scenario reasoning architecture in solving the shortcomings of multi-modal understanding technology (Ekren et al., 2010; Saglam & Papelis, 2024). Given that many development institutions such as Tesla, Boston Dynamics and NVIDIA currently do not open source technical details in the field of thinking, planning and decision-making of humanoid robots, the objectivity of real experiments is challenged. To this end, the researcher adopted a single-group design to focus on the detection and evaluation of architecture performance.

A. Experimentation Setup

The researcher used code (uploaded to Github) and prompt engineering to train an experimental tool called Mahā based on a custom GPTs model that is used to simulate humanoid robots to perform multi-scenario reasoning. Developed by Meta, Sapiens-2B is primarily intended for high-resolution tasks centered around human vision, reasoning that the applicability of synthetic data is not just visual. In contrast, Mahā is more suitable as a preferred research tool that serves multi-scenario reasoning architectures. Another advantage of Mahā based on LLMs is that there is no need to consider hardware configuration, and the data synthesis method is simple and practical.

B. Dataset

Based on the objectivity of simulation experiments, the design is selected to generate synthetic data to satisfy the sparse attention filter and memory-augmented reasoning based on scene decision-making and prioritization. Given the above agency restrictions on publicly available training data, employing synthetic data increases legal and ethical freedoms. In addtion, synthetic data is not unique to this research, it has been integrated into research in related fields such as training and testing robots (Kim et al., 2024).

Motivated by the need for more advanced multimodal LLMs to integrate text, images, and sensor outputs, the researcher used Gemini 2.0 Experimental Advanced to generate visual, auditory, and execution multimodal synthetic data through prompt engineering.

C. Implementation

To ensure that it can be understood and executed by Mahā, the multi-modal synthetic data created through the sample code generation loop are all based on precise computer language. The researcher executed the code generated by Gemini 2.0 Experimental Advanced in Python 3.13 IDLE. This code obtained the JSON document as the data set for this research after iterating 10,000 times. The researcher used consistent prompts in the design of the running process in each modality to facilitate objective recording. To ensure the objectivity of multi-scenario reasoning ability assessment, the same piece of synthetic data can only be executed once in Mahā.

Since the current OpenAI knowledge base is updated to October 2023, Mahā may still be based on GPT 4-Turbo and has limited computing power. The researcher separately executes and calculates each step of the multi-modal data in the architecture to enhance the objectivity and accuracy of the experimental results. During the experiment, it was found that Mahā, which has limited computing power based on GPTs, frequently made errors in calculations during data analysis. Mahā only performs data in the experiment and records TP

(true positives), TN (true negatives), FP (false positives), FN (false negatives) according to the confusion matrix principle. And the researcher used prompts to continuously revise Mahā's code errors during the calculation process, which also led to multiple experiments.

V. RESULTS

Due to computer language conversion, each modality is input in the same way in Mahā, and the same evaluation indicators can be used.It provides objective evidence for evaluating the indicators of precision, recall, F1-score, specificity and accuracy. Taking into account ensuring accuracy, the researcher used the formula function of Microsoft Excel to calculate the data results of each indicator and display them in Table 1-3.

Specifically, Table 1 shows the results of running the Mahā simulation architecture to perform scene reasoning on visual synthetic data. The researcher recorded the seven modules represented by "Step 1" to "Step 7" respectively, and calculated the indicators based on the recorded TP, TN, FP, and FN.

Table 1: Visual performance metrics

Mahā	Precision	Recall	F1-score	Specificity	Accuracy
Step 1	0.932	0.918	0.925	0.900	0.911
Step 2	0.925	0.919	0.922	0.888	0.907
Step 3	0.931	0.921	0.926	0.898	0.912
Step 4	0.930	0.931	0.931	0.897	0.917
Step 5	0.921	0.922	0.922	0.882	0.906
Step 6	0.923	0.929	0.926	0.885	0.911
Step 7	0.924	0.928	0.926	0.888	0.912

Table 2 presents the results of Mahā's computational metrics after execution on auditory synthesis data converted to computational language.

Table 2: Auditory performance metrics

Mahā	Precision	Recall	F1-score	Specificity	Accuracy
Step 1	0.914	0.892	0.903	0.874	0.885
Step 2	0.911	0.894	0.902	0.870	0.884
Step 3	0.909	0.900	0.905	0.876	0.890
Step 4	0.902	0.901	0.901	0.869	0.887
Step 5	0.909	0.900	0.904	0.868	0.887
Step 6	0.907	0.906	0.906	0.859	0.887
Step 7	0.907	0.903	0.905	0.867	0.888

The following data analysis results in Table 3 reflect Mahā's implementation of tactile synthetic data converted into computational language.

Table 3: Tactile performance metrics

Mahā	Precision	Recall	F1-score	Specificity	Accuracy
Step 1	0.892	0.879	0.885	0.858	0.870
Step 2	0.889	0.891	0.890	0.851	0.874
Step 3	0.893	0.888	0.890	0.863	0.877
Step 4	0.895	0.886	0.891	0.863	0.876
Step 5	0.890	0.876	0.883	0.856	0.868
Step 6	0.891	0.886	0.888	0.855	0.873
Step 7	0.892	0.879	0.886	0.856	0.869

VI. DISCUSSION

As evidenced by findings, Mahā fully demonstrated rationality and feasibility in testing multi-scenario reasoning, although it failed to compare with data from companies such as Tesla, Nvidia, or Boston Dynamics. Table 1-3 shows that in the three modes of vision, hearing and touch, the five indicators of precision, recall, F1-score, specificity and accuracy of each module in the architecture are maintained at qualified and stable levels. Among them, the F1-score and accuracy of the attention-based prioritization and memory-augmented reasoning modules are particularly outstanding, which reflects their core role in multi-scenario reasoning. It proves the effectiveness of this architecture in solving the shortcomings of multi-modal understanding technology and thereby improving the autonomous cognitive capabilities of humanoid robots.

VII. LIMITATIONS

This research focuses on the thinking, planning and action selection aspects of humanoid robots, which are abstract and logical in nature. This makes it difficult for the data synthesized by Gemini 2.0 Experimental Advanced to simulate physical dynamic interactions. The execution simulator Mahā may still be based on GPT4-Turbo in terms of computing power. Difficulty-differentiated prompts may affect the experimental results during execution. Synthetic data is created through generative code loops, which may be of variable quality when converted into computer language. In addition, the synthetic data used for simulation experiments does not take into account interfering factors in the physical scene such as reflection, noise, vibration, etc. The above situation means that the performance of the multi-scene reasoning architecture in real physical scene practice may be The results are slightly worse than those of the simulation experiment. Mahā has the potential to grow as OpenAI updates its ability to customize GPTs.

VIII. CONCLUSION

Inspired by the principles of situated cognition theory, this research proposes a multi-scenario reasoning architecture in the field of humanoid robot thinking, planning and decision-making as an innovative solution to improve cognitive autonomy. By building and simulating experiments based on an architecture based on sparse attention and memory-augmented network, it proposes a new concept of "multi-scene reasoning" for the field to achieve multi-modal understanding of visual, auditory and tactile data. Experimental results show that the precision, recall, F1-score, specificity and accuracy of the above three modes displayed by the experimental simulator Mahā all remain at a stable level above 0.85. In particular, the attention-based prioritization and memory-augmented reasoning modules perform outstandingly. Despite the limitations of synthetic data and simulated environments, this finding confirms the effectiveness of multi-scenario reasoning in promoting cognitive autonomy in humanoid robots. It not only makes up for the technical shortcomings of multi-modal understanding in this field, but also provides feasible solutions and new development directions for future humanoid robots in thinking, planning and decision-making.

REFERENCES

[1] C. Burghart et al., "A cognitive architecture for a humanoid robot: A first approach," in *Proc. 5th IEEE-RAS International Conference on Humanoid Robots*, 2005, pp. 357-362.

[2] K. Chen et al., "Accurate Perception and Association of Objects for Humanoid Robots Under Dynamic Visual SLAM," *Int. J. Humanoid Robot.*, vol. 21, no. 3, 2024.

[3] I. Clavera et al., "Model-based reinforcement learning via meta-policy optimization," in *Proc. Conference on Robot Learning*, 2018, pp. 617-629.

[4] C. Condevaux and S. Harispe, "LSG Attention: Extrapolation of Pretrained Transformers to Long Sequences," in *Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Cham, Switzerland: Springer Nature, 2023, pp. 443-454.

[5] J. Dao, H. Duan, and A. Fern, "Sim-to-real learning for humanoid box loco-manipulation," in *Proc. 2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 16930-16936.

[6] E. Datteri and V. Schiaffonati, "Computer simulations and surrogative reasoning for the design of new robots," *Synthese*, vol. 202, no. 1, p. 5, 2023.

[7] S. Duan, Q. Shi, and J. Wu, "Multimodal Sensors and ML-Based Data Fusion for Advanced Robots," *Adv. Intell. Syst.*, vol. 4, no. 12, p. 2200213, 2022.

[8] B. Y. Ekren, S. S. Heragu, A. Krishnamurthy, and C. J. Malmborg, "Simulation based experimental design to identify factors affecting performance of AVS/RS," *Comput. Ind. Eng.*, vol. 58, no. 1, pp. 175-185, 2010.

[9] M. Fisher et al., "An overview of verification and validation challenges for inspection robots," *Robotics*, vol. 10, no. 2, p. 67, 2021.

[10] D. Guo et al., "Repetitive motion planning of robotic manipulators with guaranteed precision," *IEEE Trans. Ind. Inform.*, vol. 17, no. 1, pp. 356-366, 2020.

[11] P. M. Jenlink and F. S. Austin, "Situated cognition theory," in *The Handbook of Educational Theories*, 2013, pp. 185-198.

[12] A. L. Jouen et al., "Beyond the word and image: characteristics of a common meaning system for language and vision revealed by functional and structural imaging," *NeuroImage*, vol. 106, pp. 72-85, 2015.

[13] S. Katiyar and K. Katiyar, "Recent trends towards cognitive science: from robots to humanoids," in *Cognitive Computing for Human-Robot Interaction*, Academic Press, 2021, pp. 19-49.

[14] D. Kim, M. Choi, and J. Um, "Digital twin for autonomous collaborative robot by using synthetic data and reinforcement learning," *Robot. Comput.-Integr. Manuf.*, vol. 85, p. 102632, 2024.

[15] C. Lou et al., "Sparser is faster and less is more: Efficient sparse attention for long-range transformers," *arXiv:2406.16747*, 2024.

[16] Y. Luo et al., "Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees," *arXiv:1807.03858*, 2018.

[17] V. Makoviychuk et al., "Isaac gym: High performance GPU-based physics simulation for robot learning," *arXiv:2108.10470*, 2021.

[18] T. M. Moerland et al., "Model-based reinforcement learning: A survey," *Found. Trends® Mach. Learn.*, vol. 16, no. 1, pp. 1-118, 2023.

[19] J. F. Morici, N. V. Weisstaub, and C. L. Zold, "Hippocampal-medial prefrontal cortex network dynamics predict performance during retrieval in a context-guided object memory task," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 119, no. 20, p. e2203024119, 2022.

[20] D. Muthirayan, S. Nivison, and P. P. Khargonekar, "Improved Attention Models for Memory Augmented Neural Network Adaptive Controllers," in *Proc. 2020 American Control Conference (ACC)*, 2020, pp. 639-646.

[21] N. Navarro-Guerrero et al., "Visuo-haptic object perception for robots: an overview," *Auton. Robot.*, vol. 47, no. 4, pp. 377-403, 2023.

[22] D. Nicolini and J. Mengis, "Toward a practice-theoretical view of the situated nature of attention," *Strateg. Organ.*, vol. 22, no. 1, pp. 211-234, 2024.

[23] M. Ogunsina et al., "Cognitive architectures for autonomous robots: Towards human-level autonomy and beyond," *International Journal of Frontline Research in Engineering and Technology*, 2024.

[24] S. Pateria et al., "Hierarchical reinforcement learning: A comprehensive survey," *ACM Comput. Surv.*, vol. 54, no. 5, pp. 1-35, 2021.

[25] M. Pawłowski, A. Wróblewska, and S. Sysko-Romańczuk, "Effective techniques for multimodal data fusion: A comparative analysis," *Sensors*, vol. 23, no. 5, p. 2381, 2023.

[26] P. Pramanick and S. Rossi, "Multimodal Coherent Explanation Generation of Robot Failures," *arXiv:2410.00659, 2024.*

[27] A. Saglam and Y. Papelis, "A Simulation-Based Approach for Evaluating Shared Control Algorithms for Mobile Robots," in *Proc. 2024 Annual Modeling and Simulation Conference (ANNSIM)*, 2024, pp. 1-13.

[28] L. Song et al., "Low-Rank Approximation for Sparse Attention in Multi-Modal LLMs," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13763-13773.

[29] T. Taniguchi et al., "World models and predictive coding for cognitive and developmental robotics: Frontiers and challenges," *Adv. Robot.*, vol. 37, no. 13, pp. 780-806, 2023.

[30] P. Thagard, "How brains make mental models," in *Model-based Reasoning in Science and Technology: Abduction, Logic, and Computational Discovery*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 447-461.

[31] Y. Tong, H. Liu, and Z. Zhang, "Advancements in humanoid robots: A comprehensive review and future prospects," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 2, pp. 301-328, 2024.

[32] Y. Wang et al., "A layered reference model of the brain," in *Proc. Second IEEE International Conference on Cognitive Informatics*, 2003, pp. 7-17.

[33] X. Wu, G. Wang, and N. Shen, "Research on obstacle avoidance optimization and path planning of autonomous vehicles based on attention mechanism combined with multimodal information decision-making thoughts of robots," *Front. Neurorobot.*, vol. 17, p. 1269447, 2023.

[34] X. Xiao et al., "Robot learning in the era of foundation models: A survey," *arXiv:2311.14379*, 2023.

[35] J. Ye et al., "Cross-modality data augmentation for end-to-end sign language translation," *arXiv:2305.11096*, 2023.

[36] Y. Yuan et al., "Robot synesthesia: In-hand manipulation with visuotactile sensing," in *Proc. 2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 6558-6565.

[37] J. Zhang, E. Dean, and K. Ramirez-Amaro, "Hierarchical Reinforcement Learning Based on Planning Operators," in *Proc. 2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, 2024, pp. 2006-2012.

[38] Y. Zhao et al., "Sim2Plan: Robot Motion Planning via Message Passing Between Simulation and Reality," in *Proc. Future Technologies Conference*, Cham, Switzerland: Springer Nature, 2023, pp. 29-42.