Jiang SQ, Zhang SX, Tao SD *et al.* Self model for embodied artificial intelligence. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY, X(X): -, XXX. 2026. DOI: 10.1007/s11390-026-6289-3, CSTR: 32374.14.s11390-026-6289-3

Self Model for Embodied Artificial Intelligence

Shu-Qiang Jiang† (蒋树强), *Senior Member, IEEE*, Si-Xian Zhang† (张思贤), Shi-Da Tao (陶士达)
Xi-Hong Zhu (朱玺宏), Tian-Liang Qi (齐天亮), and Xin-Hang Song (宋新航), *Member, IEEE*

University of Chinese Academy of Sciences, Beijing 100049, China
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

E-mail: sqjiang@ict.ac.cn; sixian.zhang@vipl.ict.ac.cn; taoshida20@mails.ucas.ac.cn; zhuxihong18@mails.ucas.ac.cn
tianliang.qi@vipl.ict.ac.cn; xinhang.song@vipl.ict.ac.cn

Received January 9, 2026; accepted February 27, 2026.

Abstract This paper presents a systematic formulation of the Self Model for embodied artificial intelligence, aiming to provide the missing internal representation that enables an agent to understand its own body, capabilities, memories, and decision processes. Unlike existing approaches that address isolated aspects such as perception, prediction, or skill adaptation, we propose a unified computational framework that integrates body schema, forward and inverse models, perceptual memory mechanisms, and agency. This framework captures how an embodied agent represents its physical structure, predicts the consequences of its actions, selects policies, and accumulates experiences to form a coherent sense of self. We further introduce a six-level hierarchy (L0–L5) that characterizes the developmental stages of self model from non-self representation to full self awareness, providing the first operational taxonomy for evaluating self-awareness in embodied artificial intelligence systems. A practical implementation is developed and validated in manipulation and navigation tasks, demonstrating improved prediction, adaptation, memory, and decision-making capabilities. Overall, this work establishes the conceptual foundation and technical pathway for building self model in embodied intelligence. It highlights their significance for achieving autonomy, robustness, long-horizon reasoning, and lifelong evolution in real-world environments.

Keywords embodied artificial intelligence, self model, self model hierarchy

1 Introduction

Embodied artificial intelligence (embodied AI) refers to machine intelligence that emerges through continuous interaction with the environment. Such interaction is mediated by an embodied agent, whose perception, decision and actions are tightly coupled with its physical body and the surrounding world. This coupling implies that effective embodied AI requires not only an understanding of the external environment, but also an internal awareness of the agent's own state, body and capabilities. Consequently, an embodied agent is expected to develop self-related capabilities, including self-perception of its body and environment, self-prediction to anticipate the outcomes of its actions, self-memory to maintain continuity of internal states over time and self-decision to select feasible and goal-directed actions. Existing studies in AI have touched on certain aspects of "self", e.g., meta-learning[1] emphasizes self-adaptation for rapid task generalization, and self-supervised learning[2] focuses on learning representations without external labels. However, these approaches are solely data-driven, while embodied tasks situate the agent within a closed perception-decision-action loop, where self-perception, self-prediction, self-memory and self-decision jointly operate. Therefore, "self" in embodied intelligence is not a single attribute, but an integrated con-

Regular Paper

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62125207, 62495084, 62272443, and U23B2012, and in part by the Beijing Natural Science Foundation under Grant No. L242020.

†Equally Contributed (Shu-Qiang Jiang proposed the idea of the self-model and contributed to its conceptualization and hierarchy, as well as project administration, supervision, and paper revision. Si-Xian Zhang contributed to model implementation, validation and paper drafting.)

struct of multiple synergistic self-related capabilities.

Modeling such a notion of "self" involves both foundational capacities, such as hardware embodiment and higher-order functions, including self-perception, self-memory, self-prediction and self-decision. Humans naturally maintain structured internal awareness models of the self, which offer mechanistic insights into how a coherent notion of the self is formed. In cognitive science, human self-awareness centers on self-identity which involves multiple processes of self-knowledge[3]. Self-identity integrates autobiographical memory for temporal continuity[4] and social identity for role and group membership[5]. Self-knowledge further includes body cognition, personality traits, and metacognition. Body cognition, such as body schema and first-person perspective, grounds sensorimotor coordination[6, 7], while personality traits and metacognition shape decision-making and enable self-monitoring and adaptation[8–10].

Inspired by these perspectives, an ideal self model for embodied intelligence aims to capture key aspects of human self-awareness. While achieving full human-level self-awareness remains a long-term objective, incremental research on computational self model appears both feasible and necessary. Such efforts may provide a foundation for embodied agents that are more autonomous, adaptive, and context-aware, supporting reasoning not only about the external world but also about the agent itself, its actions, and their consequences.

As illustrated in Fig.1, several existing works have explored partial aspects of the self in embodied AI. Multimodal self-recognition enables embodied agents to distinguish self-generated from external stimuli, grounding agency and body schema[11–14]. Morphological self-modeling supports structural inference and adaptation to damage[15–17]. Competence calibration and adaptive control emphasize self-related prediction and decision-making[18, 19], while LLM-based systems[20] and episodic memory mechanisms[21–23] extend self-knowledge, action reasoning, and long-horizon planning. Despite these advances, current research remains largely fragmented. Most studies focus on isolated components (e.g., perception, prediction, memory or decision) without converging into a unified computational framework that integrates these functions into a coherent notion of self.

In this paper, we propose the self model for embodied AI. The self model is defined as an internal representation of an embodied agent, including self-perception, self-memory, self-prediction and self-decision. It serves as a core component of the embodied system, linking hardware embodiment, such as morphology and multimodal sensing, with software modules for perception, prediction, memory, decision-making, and autonomous learning. By integrating these components, the self model supports embodied functionalities such as navigation[24] and manipulation[25–27], while enabling agents to adapt and update their internal representations through continuous interaction.

The remainder of this paper is organized as follows. In Section 2, we analyze the relationship be-

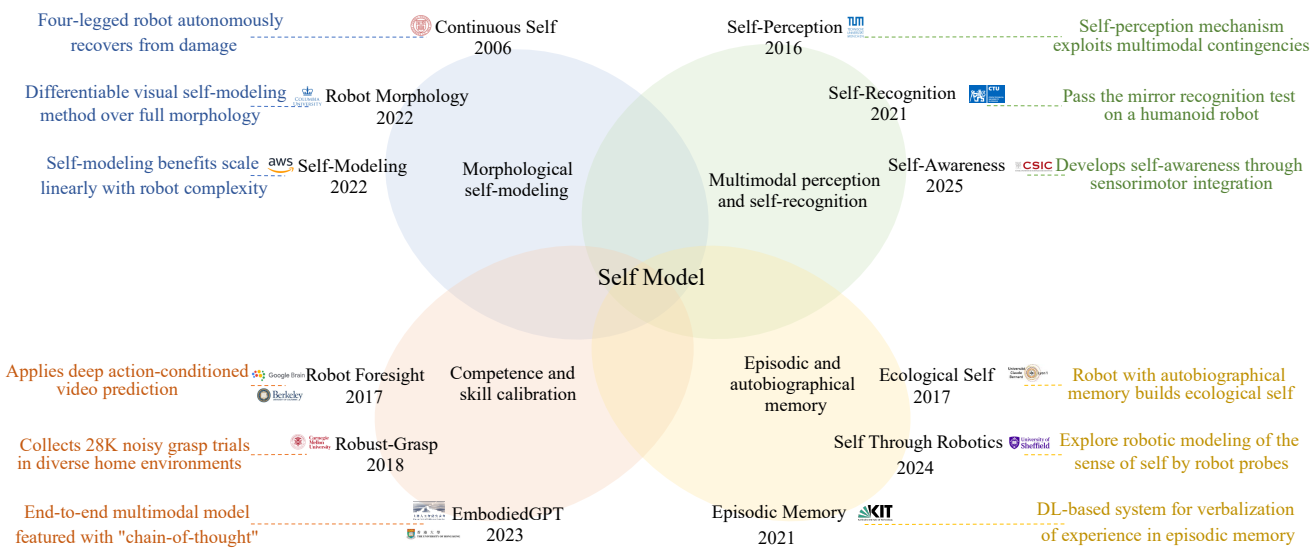


Fig.1. Related work of self model. Existing studies address isolated components of the "self", including self-perception, self-prediction and self-decision and self-memory, yet these efforts remain fragmented and do not constitute a holistic self model.

tween human self model and its computational counterpart, and present the proposed self model framework and its key functional components. In Section 3, we introduce a hierarchy (L0–L5) that characterizes different degrees of self modeling. In Section 4 and Section 5, we demonstrate an instantiation of the self model and evaluate its capabilities via real world experiments. In Sections 6 and 7, we discuss the challenges and future directions of self model.

2 Self Model

The self model denotes dynamic internal representations that an individual maintains regarding its morphology, internal state, the causal consequences of its actions, and the rules underlying its interaction with the environment. It consists of five core functional mechanisms: 1) a body schema for structural representation of self, 2) a forward model for dynamical/causal prediction, 3) an inverse model that maps desired goals to control commands, 4) mechanisms for agency that distinguish self from environment, and 5) a perceptual-memory model that tracks temporally extended self-states and interaction histories. These five mechanisms interact synergistically to form the self-awareness loop[28].

In this section, we conceptualize the self model as comprising five biological mechanisms, which are further mapped onto four implementation-oriented functional modules: perception, prediction, decision, and memory.

2.1 Self Model in Humans

For humans, the self model is an internal representation of bodily states and action intentions that enables individuals to differentiate self from the environment, predict self-generated action consequences, and maintain autonomous control[29]. In humans, these core mechanisms manifest as functional units based on neural circuits, as delineated in Fig.2.

Specifically, the human body schema functions as a representation of the body's spatial configuration, serving as the foundation for motor planning[30]. The forward model, supported by cerebellar mechanisms, predicts motor outcomes to compensate for sensory delays[31]. This prediction aids the agency mechanism in attributing actions to oneself[29] and updates the perceptual-memory model. In contrast, the inverse model is implemented by the motor cortex and basal ganglia to convert desired motion goals into specific muscle commands[32]. Furthermore, the perceptual-memory model relies on cortical regions to align multimodal inputs into a coherent self-representation[33], while the hippocampus and prefrontal circuits store and retrieve episodic memories to adapt decision-making[34].

These mechanisms form a dynamic self model in humans. Although this division aligns with neural circuits, engineering implementations require a more implementation-oriented reorganization to match the design of embodied AI system.

2.2 Self Model in Embodied AI

The human self model offers conceptual frame-

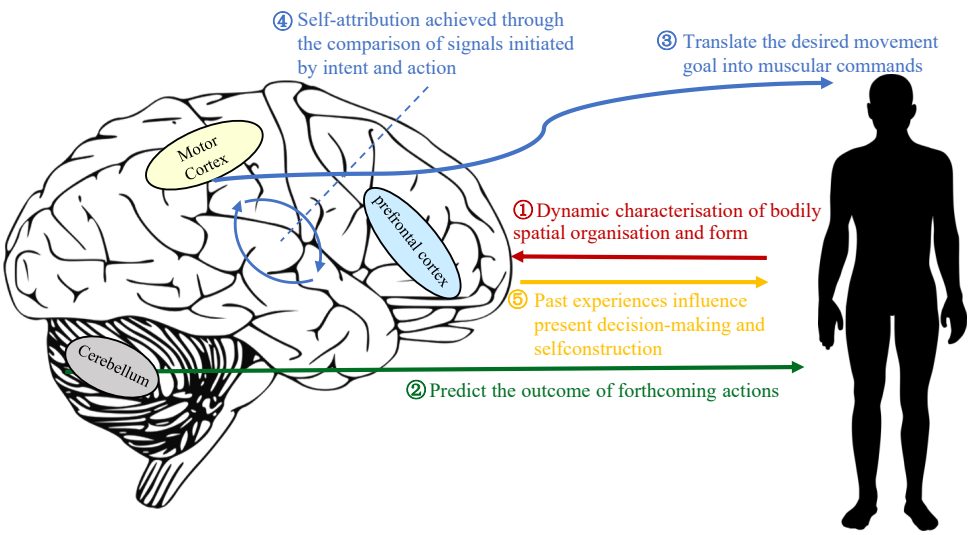


Fig.2. Internal representation of the self model in the human brain. The illustrated mechanisms offer biological inspiration for the design and construction of self model in embodied AI.

works for constructing embodied self model. In embodied AI, the self model provides an representation of all self-related aspects. Building on the human self model, the construction of embodied self model requires reorienting from the biologically tailored five-mechanism decomposition to an engineering-friendly module framework. Specifically, in the functional implementation and performance evaluation of embodied AI, we reorganize these five mechanisms into four implementation-oriented modules as shown in Fig.3: perception (body schema), prediction (forward model), memory (perceptual memory), and decision (inverse model and agency).

A key component of the perception module is a geometrically parameterized model that describes the agent's morphology and spatial configuration. In parallel, the perception module interprets sensory inputs to construct a comprehensive representation of the surrounding environment. This dual-awareness is crucial for the agent to perceive its physical presence and potential for interaction with the environment. Based on the perception module, the prediction module focuses on predicting the consequences of the agent's actions. This predictive capability facilitates the discrimination between self-generated movements and those caused by external influences, which is essential for maintaining behavioral predictability, supporting learning processes, and enabling self-monitoring. Complementary to the prediction module, the decision module is critical for purposeful action execution. It translates desired motion goals into executable control signals, with its theoretical algorithmic framework drawing on classical approaches such as inverse kinematics or optimal control[35]. The dynamic interplay between the prediction module's output and the decision module's control signals underpins adaptive motor control. Additionally, the decision module incorporates the ability to attribute actions to the self. This is a core function for developing higher-level self-awareness. This self-attribution ability can be realized through verifying the consistency between predicted and actual execution outcomes[36]. The memo-

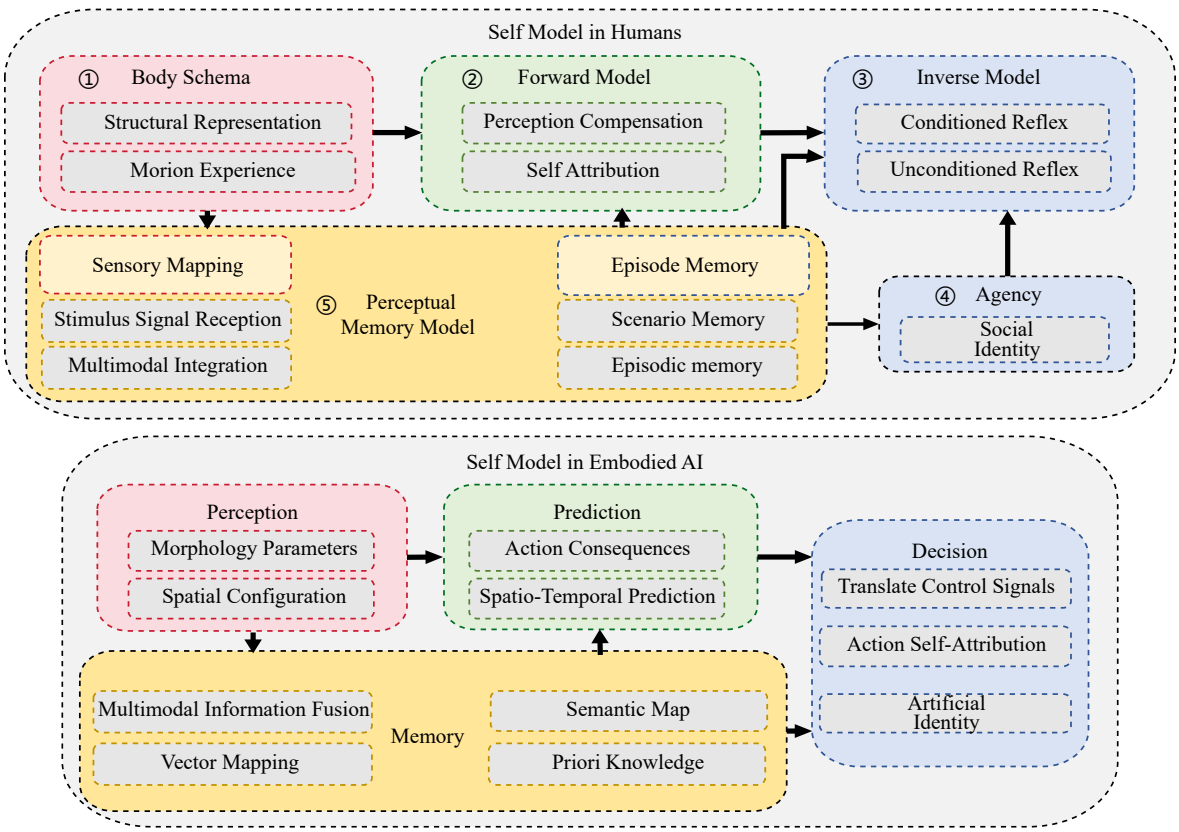


Fig.3. Our self model establishes a functional mapping between the five core mechanisms of the human self model (left: body schema, forward model, inverse model, agency, perceptual-memory model) and the four modular components of the embodied AI self model (right: perception, prediction, decision, memory). This diagram illustrates how human self model mechanisms (e.g., body schema, agency) are instantiated as domain-adapted modules in embodied AI (e.g., perception integrating geometric parameter modeling, decision incorporating artificial identity), while preserving the core cognitive capability alignment across biological and artificial systems.

ry module is essential for achieving proactive self-regulation and sustaining complex task execution. It comprises two interrelated components: sensory mapping and episodic memory. Sensory mapping integrates multimodal data to form detailed short-term state representations, updating the agent's immediate situational awareness in real time. For long-term experiential support, episodic memory stores and retrieves past interaction processes and outcomes. The awareness derived from episodic memory is significant, as it implies the agent possesses a nascent concept of self-allowing past experiences to shape current decision-making and self-construction through mechanisms such as memory networks and replaying experiences.

To realize the functionality of the four modules outlined above, the perception module continuously captures multimodal environmental and self-state information (e.g., visual, tactile) to generate real-time geometrically parameterized self-representations, which serve as the foundational input for subsequent modules. Subsequently, the prediction module fuses the real-time state from the perception module and historical experience retrieved from the memory module (including short-term sensory mapping and long-term episodic memory) to predict the sensory consequences of potential actions, while quantifying prediction uncertainty to support risk-aware decision-making. Accordingly, the decision module generates task-specific commands by integrating the prediction results and experiential constraints from the memory module. Synchronously, it initiates self-attribution verification by comparing the predicted outcomes with the actual execution feedback to distinguish self-generated actions from external perturbations. Upon execution of these commands, the agent's actual sensory outcomes are fed back bidirectionally: to the perception module for calibrating state estimation deviations, and to the memory module for updating experiential data (i.e., storing new interaction episodes into episodic memory and refreshing short-term sensory mapping). Finally, the updated perceptual state and memory data are recycled into the next iteration of prediction and decision-making, completing the closed-loop calibration.

3 Self Model Hierarchy

According to [37], the self can be systematically divided into three hierarchical levels: the implicit self, the self based on dynamic visual matching, and the self grounded in symbols, language, and artifacts. However, current embodied AI lacks a unified self model assessment framework tailored to engineering implementation. To fill this gap, we propose a hierarchy (L0–L5) that follows the natural evolution logic of self-awareness, progressing from non-self (L0) to full self-awareness (L5). This advancement is characterized by the successive acquisition of (1) basic self-awareness through a static physical self, (2) basic self-adaptation via dynamic self-environment coupling, (3) socialized self-representation, and (4) sustained self-evolution via value-oriented iteration. The self model is evaluated along four constituent dimensions (perception, memory, prediction, decision) that align with the functional modules proposed in the previous section. As depicted in Fig.4, which illustrates the overall characteristics of each level alongside the evolutionary trajectory, this hierarchy covers the complete evolution of self-awareness.

L0: Non-Self Model. At L0, no explicit representation of self is instantiated, and behavior is purely re-

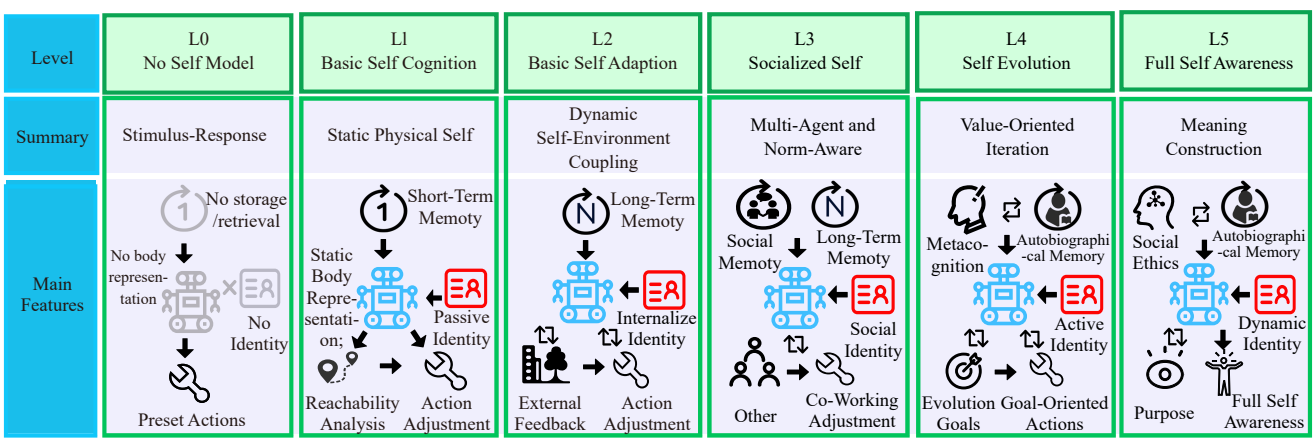


Fig.4. Overview of the six-level hierarchy of self model. The hierarchy spans from a non-self model to full self awareness, encompassing basic self cognition, basic self adaption, socialized self, and self evolution.

active without encoded self/non-self boundaries. Prediction is limited to self-state forecasting without accounting for external factors. Memory retains only prior knowledge, lacking any self-related or task-specific memory, and the capacity for self-attribution is absent. This yields a non-agentic control loop responsive only to immediate stimuli, and self-identity is not formed.

L1: Basic Self-Awareness. At L1, a static physical self is maintained. Basic self/non-self demarcation enables simple reachability and collision judgments. Short-term memory and context-bound action-outcome associations provide local competence, though they do not generalize or calibrate. Identity remains externally attributed.

L2: Basic Self-Adaptation. At L2, a dynamically physical self is maintained through closed-loop perception-action. This level enables generalized forward prediction and flexible goal-to-action mapping across various contexts. Multimodal memory consolidates self experience and stabilizes self-attribution, the process of assigning observed actions and effects to one's own agency. Identity becomes internalized, guiding action and policy adjustment.

L3: Socialized Self. At L3, the self extends to incorporate other agents and social regularities. Prediction and decision are conditioned on roles and interaction situations. Social memory and role-aware policies support cooperation and social attribution the understanding of others' actions and intentions. A distinct social identity is recognized and utilized to structure identity-based relations.

L4: Sustained Self-Evolution. At L4, long-term and counterfactual prediction capabilities support value-oriented iteration. Internally generated goals are anchored in stable preferences, such as safety and efficiency. Autobiographical memory and metacognitive monitoring enable the continual revision of strategies and the creation of explicit accounts of self-improvement, which ensure that decision-making align with an actively affirmed and evolving identity.

L5: Full Self Awareness. At L5, the computationally achievable self model supports semantic interpretation via the perception of worldviews and ethical values. This enables the formation of grand visionary plans for individuals or society. Decision is hierarchically organized by worldview-level objectives and ethical constraints. Long-term consequences of actions are foreseen using narrative-based social memory. Identity is dynamically reconfigurable, adapting and changing in response to accumulated experience, context, and behavior.

4 An Instantiation of Self Model

Rather than a universal self model, this work presents a task-specific instantiation of a self model. The proposed design represents one feasible realization of the self model.

The baseline model corresponds to an L0-level pipeline, whose components are detailed as follows. Perception (L0): The perception module takes as input a single-frame RGB-D observation I_r and a natural language instruction I_t. Object instance masks Msk_l are obtained via semantic segmentation. No body state, joint state, or collision awareness is modeled. Memory (L0): The memory module only uses visual I_r and LiDAR I_l information from the current observation without constructing a map. The observation is not accumulated across time and contains no self-related or identity-related information. Prediction (L0): Since an L0-level agent cannot predict the result and feedback of the action according to our self model hierarchy, there is no prediction module in an L0-level model. Decision (L0): The decision module selects the nearest frontier for exploration. Once the target object is detected, its relative pose is estimated from the current RGB-D observation and passed to an inverse-kinematics-based controller to execute grasping.

Building upon the L0 baseline, we present one possible instantiation of the L1 self model, which augments the L0 pipeline by introducing explicit self-related perception, memory, prediction, and identity-aware decision-making, as shown in Fig.5. Implementation details are described below.

(1) Perception (L1). The L1-level perception module SelfPerc computes the spatial regions where potential collisions may occur between the robot body and obstacles in the environment for safe motion and uses it for risk evaluation. The joint-related perception input is summarized as:

$$I_j(t) = \big(R_m(t),\, \dot{\tau}(t)\big), \tag{1}$$

where R_m is the collision risk and $\dot{\tau}$ is the temporal change rate of joint torques. The collision range R_m is defined as a joint measure of the spatial proximity and spatial extent of contact between critical parts of the robot body and surrounding obstacles. This range reflects the physical feasibility and safety of the cur-

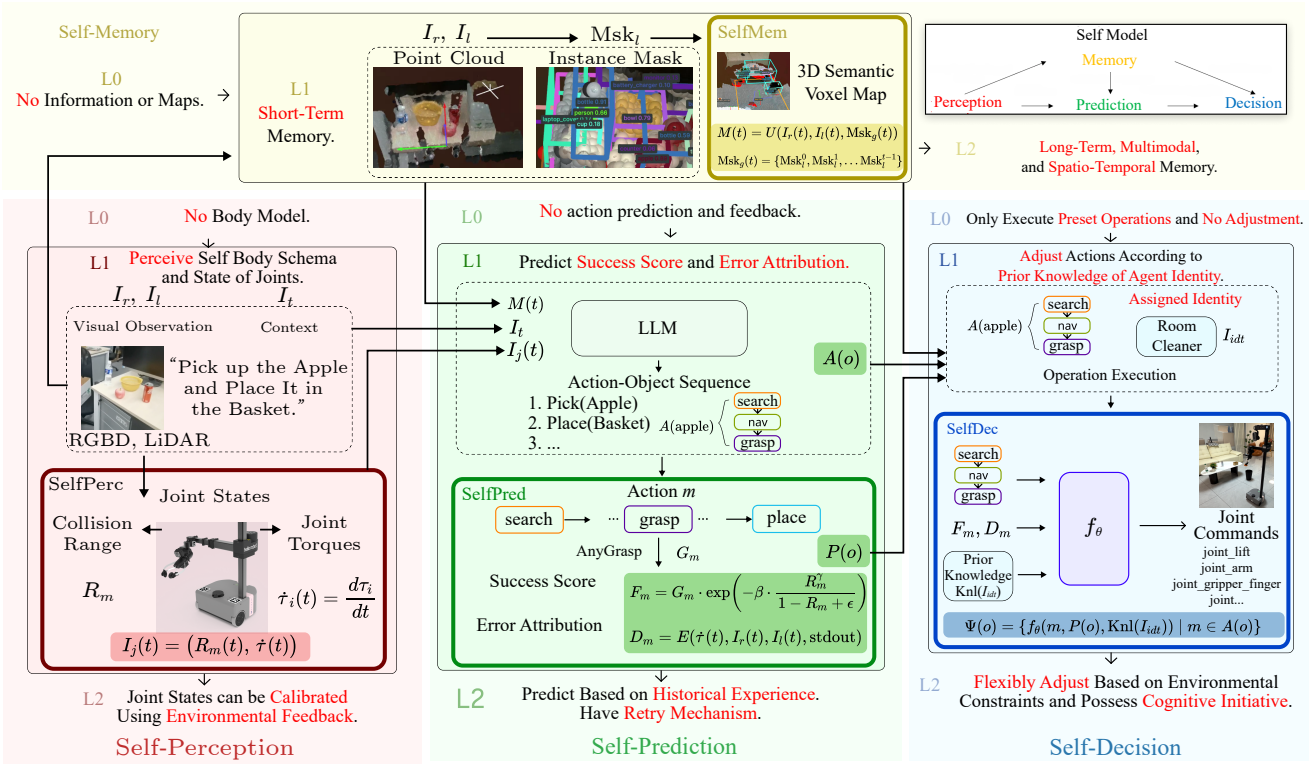


Fig.5. Framework of a self model instantiation at Level L1. For conceptual clarity, conceptual descriptions of Levels L0 and L2 are included in each component to contextualize the L1. In each module, dashed boxes denote the inputs associated with different aspects of the self. The thumbnail in the top-right corner shows alignment between the implementation and the self model definition.

rent action by jointly considering both the closeness and the distribution of potential collision points, which can be formulated as:

$$R_m(t) = \frac{|C(t)|}{|PC(t)|} \times \frac{1}{\frac{1}{|C(t)|}\sum_{p_i \in C(t)} d_{\min}(p_i, Hem(t))}, \tag{2}$$

where $Hem(t)$ denotes the collision region of the gripper, approximated by a pair of hemispherical volumes with a radius of 1.5 cm. $PC(t)$ denotes the filtered point clouds from the head and wrist cameras of the agent and $C(t)$ denotes a subset of $PC(t)$, whose points satisfy that the minimum distance to the hemispherical collision volume $Hem(t)$ is smaller than a predefined threshold. If $R(t) > 0.6$, the system classifies the state as high-risk and triggers action adjustment.

The temporal change rate of joint torques $\dot{\tau}$ is approximated by:

$$\dot{\tau}_i(t) = \frac{d\tau_i}{dt} \approx \frac{\tau_i(t) - \tau_i(t - \Delta t)}{\Delta t}, \tag{3}$$

where $\tau_i(t)$ is the joint torque. For joint i, declaring a collision at time t requires the three following conditions: the absolute change in joint torque exceeds a threshold within the time interval $\Delta t = 0.5$ s, the angular velocity is lower than 0.2 rad/s and sufficient previous data supports state comparison.

(2) Memory (L1). The L1-level memory module SelfMem incrementally maintains a real-time sparse 3D semantic voxel self-map (3DSmap), which acts as an egocentric self-map and supports downstream tasks such as path planning and active exploration. The self-map M is represented as a sparse voxel grid. Each voxel $v_i \in M$ stores a tuple $v_i = (occ_i,\ c_i,\ w_i,\ n_i)$, where $occ_i \in \{0, 1\}$ denotes occupancy status, $c_i \in \mathbb{R}^3$ is the mean RGB vector, w_i is an encoded semantic feature vector, and $n_i \in \mathbb{N}$ is the number of point-cloud observations fused into this voxel. The online self-map M at time t is denoted by a multimodal fusion operator:

$$M(t) = U(I_r(t),\ I_l(t),\ \mathrm{Msk}_g(t))\,, \tag{4}$$

where $U(\cdot)$ is the fusion operator and $I_r(t)$, $I_l(t)$ are synchronized RGB-D and LiDAR observations. $\mathrm{Msk}_g(t) = \{\mathrm{Msk}_l^0,\ \mathrm{Msk}_l^1,\ \ldots\ \mathrm{Msk}_l^{t-1}\}$ is the associated global instance-mask prediction and Msk_l^t denotes the instance segmentation mask at time t. For each input RGB image I_r, a MobileSAM-based[38] instance

segmentation model is performed to generate multiple candidate masks, each associated with a stability score. To suppress mask overlap, we apply IoU-based non-maximum suppression: if $\mathrm{IoU}_{ij} > \tau_{iou}$ (default $\tau_{iou} = 0.5$), the lower-scored mask is removed. The remaining K masks are sorted in descending score order, assigned instance IDs, and converted into the final instance segmentation map Msk_l according to per-pixel instance membership.

Given the constructed 3D semantic map, the agent is able to perform frontier explorations based on occupancy grid of the map and the agent's reachable region. The candidate frontiers extracted by temporal and task-specified heuristic algorithms are selected as next exploration targets.

(3) Prediction (L1). The L1-level prediction module SelfPred extends the L0 pipeline by incorporating a large language model, which integrates multimodal information $I_j(t)$, I_t and $M(t)$ from perception and memory modules, decomposing the task into action-object sequences $A(o)$ and utilizes its reasoning capabilities to conduct self-action feasibility inference. The prediction module also introduced two additional capabilities: manipulation pose prediction and error attribution.

For each object o, the output $P(o)$ is an ordered set whose each component corresponds to an action m in sequence $A(o)$. During manipulation, the model estimates the success probability of the action, denoted as F_m. If the action is invalid, the error attribution D_m is inferred by states of self and the environment.

For each action m, the prediction consists of two components: a success score $F_m \in (0, 1)$ and a n-dimension error attribution vector $D_m \in \{0, 1\}^n$. When multiple candidate manipulation poses are generated, F_m represents the predicted success probability. After execution, an all-zero D_m indicates predicted success, whereas a nonzero entry at index k suggests a potential failure due to the k-th failure category.

The success prediction F_m is derived by refining the manipulation likelihood output G_m by AnyGrasp[39]. Specifically, G_m is multiplied by a penalty factor determined by the collision risk R_m, yielding a physically grounded success estimate, formulated as:

$$F_m = G_m \cdot \exp\left(-\beta \cdot \frac{R_m^{\gamma}}{1 - R_m + \epsilon}\right), \tag{5}$$

where β controls risk sensitivity, γ governs nonlinearity and ϵ ensures numerical stability. The error attribution D_m is inferred by an error function:

$$D_m = E\left(\dot{\tau}(t), I_r(t), I_l(t), \mathrm{stdout}\right), \tag{6}$$

which integrates temporal change rate of joint torques $\dot{\tau}(t)$, the target object observation $I_r(t)$ and $I_l(t)$, as well as runtime feedback information from system log text outputs (denoted as stdout).

(4) Decision (L1). The L1-level decision module SelfDec extends the L0 reactive policy by incorporating inputs from memory and prediction modules. By integrating predictive feedback and identity-specific knowledge, the L1 decision module enables adaptive, multi-source decision making.

For each target object o, the decision output $\Psi(o)$ is jointly determined by the action-object sequence $A(o)$, the prediction result $P(o)$ and the self-identity representation I_{idt}, denoted as:

$$\Psi(o) = \{f_\theta\left(m,\ P(o),\ \mathrm{Knl}(I_{idt})\right) \mid m \in A(o)\}, \tag{7}$$

where m denotes actions in the predicted action sequence $A(o)$ for object o, $\mathrm{Knl}(I_{idt})$ denotes prior knowledge given identity I_{idt} and f_θ denotes an action-conditioned pose generation function, whose results are subsequently converted into joint commands via inverse-kinematics-based computation. Based on role identity, the Knl function invokes a predefined knowledge base corresponding to that identity to provide fine-grained guidance for robot behavior. During the navigation phase, it not only supplies motion constraint rules such as path priority and restricted area access but also adjusts exploration strategies using environmental prior knowledge associated with the role. Taking a cleaning robot as an example, its "Room Cleaner" role activates pre-configured garbage target detection models and common garbage distribution knowledge within Knl, enabling the robot to more rapidly and accurately identify target objects (e.g., scattered paper or bottles/cans) while moving. In the manipulation phase, Knl further offers role-specific operation templates and parameter adjustment guidelines, for instance, automatically adapting grasping force, pose, and placement methods according to different garbage types such as plastic bags and cans, thereby achieving more efficient and stable operational adjustments during task execution. This type of prior knowledge is structurally embedded into the decision module of the self model through the Knl function, equipping the system with identity-based task adaptation capability. During navigation and manipulation, if the success predic-

tion F_m and error attribution D_m derived by prediction module indicates successful execution, the policy proceeds to the next action. Otherwise, the decision module adapts the action according to the inferred failure cause and retries the operation.

Compared with L0, the L1 module maintains an explicit self-identity I_{id} that tailors behavior to task roles. For a room-cleaning agent, it activates garbage detection priors and garbage-aware manipulation adaptations, e.g., biasing pose selection toward top-down grasps for container-like trash.

5 Experiments

5.1 Experimental Setup

We evaluated the perception, memory, prediction, and decision modules of the self model through four experiments. The main goal was to compare the task performance of the L0-level self model and the L1-level self model across these modules. By analyzing the performance differences between L0 and L1, we aimed to demonstrate the necessity and sufficiency of the self model in ensuring the safe, accurate, and efficient action. Meanwhile, we also design an experiment in which the robot picks up garbage and throws it into a trash bin, as shown in the Fig.6, to compare the performance of the self model with other related methods across different stages of navigation and grasping. The experiments were conducted using the Stretch3 robot from Hello-Robot, which runs on Ubuntu 20.04 and is powered by an NVIDIA GeForce GTX 4090. Related video demonstrations can be accessed from our project page[①].

5.2 Ablation Study

5.2.1 *Ablations on self-perception*

To evaluate the effect of self-perception on grasping safety, we run repeated grasping trials on diverse objects with avoidable obstacles near each target, and report Grasp SR, collision frequency (ACR), and human takeovers.

As shown in Table 1, introducing collision range (BL+CR(L1)) yields a 9.7% improvement in Grasp SR and a 17.8% reduction in ACR over L0, indicating that basic collision range at the L1 level improves grasp stability and safety. Furthermore, adding joint force sensing reduces human takeovers by 9.5% compared to BL+CR(L1) model, suggesting that force-based collision cues enable the robot to detect impacts autonomously and intervene in a timely manner. Overall, the results show that incorporating self-

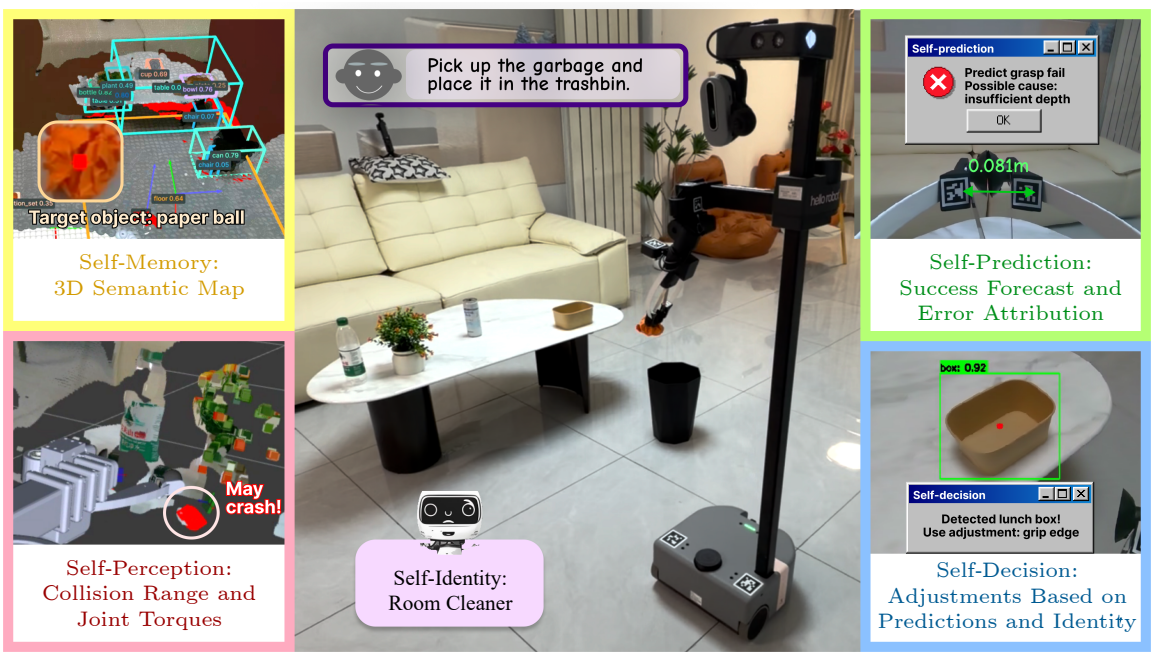


Fig.6. Overview of real-world experiments on the Stretch robot. The proposed self model is evaluated in real environments through its four components, including self-perception, self-memory, self-prediction, and self-decision. These components support embodied tasks such as navigation and manipulation, enabling the agent to perform them more effectively.

[①]https://taoshida11.github.io/Selfmodel/, Mar. 2026.

Table 1. Comparision of L0 with L1 in Perception Module

Model	Self-Perception	Grasp SR(%)	ACR (%)	AHTR (%)
BL(L0)	No joint states	72.7	27.3	18.2
+CR(L1)	Only collision range	81.0	9.5	9.5
+SelfPerc(L1)	Collision range and joint torques	81.8	9.1	1.5

Note: Perception ablation study explores the impact of collision range and joint torques on grasping success rate (Grasp SR), average collision rate (ACR), and average human takeover rate (AHTR). CR only includes collision range, while SelfPerc includes both collision range and joint torques.

perception contributes to safer grasping; if the perception module further reaches the L2 level, real-time calibration of joint states based on environmental feedback may lead to substantial gains in execution accuracy and success.

5.2.2 *Ablations on Self-Memory*

To evaluate how self-memory affect navigation, we conduct an object-goal navigation task for evaluation. Given a language instruction, the robot must locate the target object (e.g., an apple, a cup, and a screwdriver). The L0 model lacks self-memory and thus stores no map for a navigation episode, relying only on one-time observations. The L1 model has short-term memory and can store a map within a single episode, but cannot recall maps from history episodes. The L2 model can retain maps across multiple episodes and can reuse a previously built map for navigation in the same scene.

As shown in Table 2, we compare three self-memory implementations: obstacle map, implicit state map and 3D semantic map. Among them, the 3D semantic map corresponds to the proposed SelfMem model. The implicit state map[40] is implemented via an end-to-end vision-language navigation model. We further contrast L1 and L2 by configuring the L2 model to reuse a retained 3D semantic map from previous episodes. Experimental results show that employing richer memory (from obstacle maps to 3D maps) and extending memory (from single-episode maps to history maps) improves navigation performance. These results demonstrate that self-memory enables the agent to retain its own spatial relationship with the environment, facilitating more effective performance of embodied tasks.

5.2.3 *Ablations on Self-Prediction*

The prediction model is evaluated on 20 objects, each grasped multiple times. For each trial, a single object is placed and the robot attempts one grasp. We report grasp pose prediction success, error attribution accuracy, and overall grasping success. AP measures the average success rate of the top-10 confidence-ranked grasp poses, while E-AP denotes the accuracy of predicting the dominant error cause.

As shown in Table 3, we compare models with different self-prediction capabilities. The L0 baseline uses a fixed grasp pose and does not estimate grasp suc-

Table 2. Comparision of L0, L1 and L2 in Memory Module

Model	Self-Memory	Nav SR(%)	Time (s)
BL(L0)	No map	33.4	143.2
+Omap(L1)	Obstacle map(current episode)	42.1	135.4
+HS[40](L1)	Hidden state(current episode)	56.0	96.3
+SelfMem(L1)	3D semantic map(current episode)	63.6	68.2
+SelfMem(L2)	3D semantic map(history episode)	93.2	30.5

Note: Memory ablation study investigates the effects of different memory formats and memory durations of maps on navigation success rate (Nav SR) and time.

Table 3. Comparison of L0 with L1 in Prediction Module

Model	Self-Prediction		AP (%)	E-AP (%)	Grasp SR (%)
	GP Pred.	Error Attr.			
BL (L0)	×	×	–	–	23.5
+ AnyGrasp[39](L0)	✓	×	63.3	–	76.5
+ ErrorAnalysis (L0)	×	✓	–	80.0	64.7
+ SelfPred (L1)	✓	✓	64.2	83.3	82.3

Note: Prediction ablation study investigates the impact of using only AnyGrap for grasp prediction, using only error attribution, and using the SelfPred model on both the prediction outcomes and grasping success rate (Grasp SR). The prediction outcomes include the average precision (AP), error average precision (E-AP).

cess. The proposed SelfPred module leverages a large language model to jointly reason about grasp quality and failure causes. L1 baselines include AnyGrasp[39], which predicts grasp poses without error reasoning, and ErrorAnalysis, which performs error attribution with the predictive components removed. The superior performance of SelfPred demonstrates that modeling expected action outcomes and failure causes is central to the effectiveness of self model.

5.2.4 Ablations on Self-Decision

The experiment over decision model evaluates long-horizon navigation-grasp-place tasks using success rate and total execution time. In each trial, the robot is required to identify a target trash item among distractors, grasp it, and place it into a trash can. Five trash categories are considered, along with ten types of interfering objects. Each scene contains one target and five distractors, and performance is measured over the full task sequence.

We compare models with different levels of identities on self-decision capability. The proposed SelfDec model incorporates target-specific self-identity by pretraining recognition models for the five trash categories and by adapting grasping and placement strategies accordingly. The L0 baseline does not adjust its behavior based on target identity. Two ablations are included: L0-Search, which uses target-specific recognition but generic grasping and placement actions, and L0-Grasp, which adapts grasping and placement strategies while relying on a generic open-vocabulary recognizer. As shown in Table 4, SelfDec achieves higher SR and lower execution times, indicating that jointly modeling target identity and action adaptation benefits self-decision.

5.3 Comparisons with Related Work

To investigate the impact of adding the four modules of the self model on navigation and manipulation, we compare the self model with other related methods by analyzing stage-wise success rates in navigation manipulation tasks, as shown in Table 5. Baseline is our implemented L0 self model.

The experiment is conducted on moblie manipulation tasks, including navigating to the target object, grasping the object, navigating to the container, and placing the object. Performance is measured using the overall task success rate as well as the success rate at each stage. The task requires the robot to correctly identify and grasp a piece of trash and subsequently place it into a trash can. All experiments are conducted in real-world environments. Interference items include ten categories, such as apples, bananas, potted plants, scarves, and thermos cups. In each trial, one target trash item and two interference items are placed in the scene. We record the success rates and total execution time for the navigation, grasping, and placement stages. We conduct a total of 130 trials across five domestic and office environments. For navigation-only related works, such as Heuristic Baseline[47], HOZ++[24], T-diff[42], and VLN-nav[43], we only record the success rate of navigation to the target object. For manipulation-only methods, such as RL Baseline[44], AnyGrasp[39], PC-Attention[45], and GtG[46], we record solely their grasping success rate for trash items placed within the robot's predefined grasping range. For methods that involve both target-oriented navigation and manipulation, such as ManipGen[25], OVMM[47], OK-robot[48], the Baseline L0 self model, and our proposed L1 self model, we record their stage-wise success rates for target navigation, grasping, and placement into a trash bin for each piece of garbage. Additionally, we measure the performance contributions of the four modules (SelfPerc, SelfMem, SelfPred and SelfDec) to the L0-level self model. Experimental results demonstrate that each stage of the proposed self model consistently outperforms prior work, achieving the best overall performance. Video demonstrations are provided in the supplementary material.

Table 4. Comparision of L0 with L1 in Decision

Model	Self-Decision		Nav SR (%)	Grasp SR (%)	Place SR (%)	Total Time (s)
	Ident. (Nav)	Ident. (Grasp)				
BL(L0)	×	×	41.0	24.2	28.3	466.1
+Knl.nav(L0)	✓	×	86.5	25.5	47.8	389.4
+Knl.grp(L0)	×	✓	45.1	77.2	31.3	405.8
+SelfDec(L1)	✓	✓	84.2	62.1	60.0	368.9

Note: Decision ablation study investigates, from the perspective of self identity, the effects of a model that uses only navigation-related identity prior knowledge (Knl.nav), a model that uses only grasping-related identity prior knowledge (Knl.grp), and the SelfDec model on stage-wise success rates for navigation, grasping, and placement, as well as on overall execution time.

Table 5. Comparison of Navigation and Manipulation Performance Between Self Model and Related Methods in Garbage Cleaning Tasks

Method	Self Model				Partial Success Rate			Overall SuccRate	Partial SuccMetric
	Memory	Perception	Prediction	Decision	FindObj	Pick	FindRec		
Heuristic Baseline[41]	✓	×	×	×	32.7	–	–	–	–
HOZ++[24]	✓	×	✓	×	47.5	–	–	–	–
T-diff[42]	✓	×	✓	×	50.5	–	–	–	–
VLN-nav[43]	✓	×	×	×	43.8	–	–	–	–
RL Baseline[44]	×	×	✓	×	–	44.9	–	–	–
AnyGrasp[39]	×	×	✓	×	–	57.8	–	–	–
PC-Attention[45]	×	×	✓	×	–	64.2	–	–	–
GtG[46]	×	×	✓	×	–	54.6	–	–	–
ManipGen[25]	✓	✓	✓	×	31.2	16.0	9.6	6.7	53.2
OVMM[47]	✓	×	✓	×	32.8	18.4	10.1	5.1	48.5
OK-Robot[48]	✓	×	✓	×	35.6	22.6	13.6	9.2	56.8
BL (L0 self model)	×	×	×	×	20.5	5.4	2.7	0.8	31.7
BL + SelfPerc	×	✓	×	×	21.0	7.2	3.6	1.4	36.5
BL + SelfMem	✓	×	×	×	35.4	8.9	5.3	1.6	37.6
BL + SelfPred	×	×	✓	×	22.5	11.7	5.9	3.5	46.1
BL + SelfDec	×	×	×	✓	28.6	9.2	4.6	1.6	36.4
L1 self model (Ours)	✓	✓	✓	✓	42.7	26.5	17.2	12.9	61.2

Note: The Baseline method corresponds to our proposed L0 self model, while the Ours method corresponds to proposed L1 self model. BL+SelfPerc, BL+SelfMem, BL+SelfPred, and BL+SelfDec are extensions of the L0 self model with the addition of perception, memory, prediction, and decision modules, respectively. In addition to comparing navigation-manipulation approaches such as ManipGen, OVMM, and OK-Robot, we also include navigation methods such as the Heuristic Baseline, HOZ++, T-diff, and VLN-nav, as well as manipulation methods such as the RL Baseline, AnyGrasp, PC-Attention, and GtG.

6 Future Directions

Research on the self model for embodied AI is still at an early stage. Existing efforts focus on individual components and remain fragmented rather than organized into an integrated, testable framework. Against this backdrop, the hierarchy and the prototype system presented in this paper are an attempt oriented toward feasibility and standardization. Moreover, the method of self modeling is not a single fixed pathway. Depending on the platform and application scenario, future systems may adopt various forms that bridge symbolic and subsymbolic representations. Our method can be served as a reference paradigm that emphasizes modular decomposition and hierarchical capability criteria, rather than a rigid solution.

From the perspective of disciplinary development, systematic research on the self model has become increasingly critical in the current landscape of embodied AI. As noted earlier, existing studies have explored isolated aspects of self-related functions but lack integration of perception, memory, prediction, and decision into a unified framework. With the rapid emergence of multimodal perception, large-model-driven embodied learning, and long-horizon autonomous task scenarios, reliance solely on external world modeling is increasingly insufficient to support advanced autonomous behavior. As a result, explicit modeling of the self is no longer a distant vision, but an urgent practical requirement.

From the standpoint of technological ecosystem integration, the self model is not isolated from existing embodied AI approaches, but instead exhibits a high degree of compatibility and composability. On the one hand, it can be naturally embedded into current perception-planning-control pipelines, serving as an internal simulation and evaluation substrate. On the other hand, the self model interfaces effectively with reinforcement learning, meta-learning, imitation learning, and LLM-driven decision frameworks. By acting as self-related priors and competence boundary evaluators, it can provide a unified representation of self reference for multi-task, multi-scenario, and multi-agent.

Future research on the embodied self model may advance along four key directions.

1) Benchmarks and Metrics Aligned with L0–L5: establish operational benchmark tasks and evaluation metrics grounded in the proposed hierarchy, enabling cross-platform and comparable assessment of self-related capabilities (e.g., self-prediction, self-memory, etc.) under shared protocols.

2) Computable Self-Awareness Inspired by Humans: investigate human self-awareness mechanisms

(e.g., body schema, episodic memory, narrative self) and translate them into computable modules and training paradigms. A key goal is to facilitate a transition from an implicit self (emergent behavioral regularities) into an interpretable self (explicit representations with diagnostic value).

3) Co-Design with World Models and Planning Systems: develop tighter coupling between the self model and large-scale world models, planners, and simulators, so that embodied agents can answer within a unified framework: who I am, where I am, what I can do, and why I act in this way. This direction emphasizes joint reasoning over self state, world state, action feasibility, and long-horizon objectives.

4) *Ethics, Safety, and Social Integration:* incorporate ethical, safety, and social dimensions into self model, motivating research on controllability, responsibility boundaries, and human-robot relationships, especially under long-horizon autonomy and multi-agent interactions.

In summary, the self model supports embodied AI in moving from task execution to agentic understanding, and from tool-based operation to autonomous decision-making. It offers a basis for continued theoretical and applied investigation.

7 Conclusions

This paper presented the self model for embodied AI, covering its conceptual mechanisms, functional modules, hierarchical organization, and preliminary empirical validation. On the one hand, it provides a unified self-referential framework for embodied agents that enables body schema, forward prediction, inverse control, agency, and memory to operate coherently within a single system. On the other hand, the self model complements world models by offering an internal core that supports the transition from agents that merely perceive the environment to agents that can understand both the world and themselves. Research on self model remains in its early stages, and a substantial gap remains toward realizing a high-level embodied self. The hierarchy and instantiation of self model in this work serve as feasibility demonstrations and explorations, rather than complete and mature solutions. It should also be emphasized that multiple parallel technical pathways may emerge in future research. Our work offers a reference perspective, while implementation strategies, module boundaries, and evaluation metrics remain extensible and require further validation and iterative refinement.

Conflict of Interest Shu-Qiang Jiang is an editorial board member for Journal of Computer Science and Technology and was not involved in the editorial review of this article. The author declares that there are no other competing interests.

References

[1] Hospedales T, Antoniou A, Micaelli P, Storkey A. Meta-learning in neural networks: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2022, 44(9): 5149–5169. DOI: 10.1109/TPAMI.2021.3079209.

[2] Jing L, Tian Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2021, 43(11): 4037–4058. DOI: 10.1109/TPAMI.2020.2992393.

[3] Klein S B, Gangi C E. The multiplicity of self: Neuropsychological evidence and its implications for the self as a construct in psychological research. *Annals of the New York Academy of Sciences*, 2010, 1191(1): 1–15. DOI: 10.1111/j.1749-6632.2010.05441.x.

[4] Conway M A, Pleydell-Pearce C W. The construction of autobiographical memories in the self-memory system. *Psychological Review*, 2000, 107(2): 261–288. DOI: 10.1037/0033-295X.107.2.261.

[5] Tajfel H, Turner J C. An integrative theory of intergroup conflict. In *The Social Psychology of Intergroup Relations*, Austin W G, Worchel S (eds.), Brooks/Cole, 1979, pp.33–47.

[6] Gallagher S. Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences*, 2000, 4(1): 14–21. DOI: 10.1016/S1364-6613(99)01417-5.

[7] Blanke O, Slater M, Serino A. Behavioral, neural, and computational principles of bodily self-consciousness. *Neuron*, 2015, 88(1): 145–166. DOI: 10.1016/j.neuron.2015.09.029.

[8] Markus H, Wurf E. The dynamic self-concept: A social psychological perspective. *Annual Review of Psychology*, 1987, 38: 299–337. DOI: 10.1146/annurev.ps.38.020187.001503.

[9] D'Argembeau A, Ruby P, Collette F, Degueldre C, Balteau E, Luxen A, Maquet P, Salmon E. Distinct regions of the medial prefrontal cortex are associated with self-referential processing and perspective taking. *Journal of Cognitive Neuroscience*, 2007, 19(6): 935–944. DOI: 10.1162/jocn.2007.19.6.935.

[10] Merkebu J, Veen M, Hosseini S, Varpio L. The case for metacognitive reflection: A theory integrative review with implications for medical education. *Advances in Health Sciences Education*, 2024, 29(4): 1481–1500. DOI: 10.1007/s10459-023-10310-2.

[11] Lanillos P, Pages J, Cheng G. Robot self/other distinction: Active inference meets neural networks learning in a mirror. In *Proc. the 24th European Conference on Artificial Intelligence and 10th Conference on Prestigious Applications of Artificial Intelligence*, Aug. 29 -Sept. 8, 2020, pp.2410–2416. DOI: 10.3233/FAIA200372.

[12] Lanillos P, Dean-Leon E, Cheng G. Yielding self-perception in robots through sensorimotor contingencies. *IEEE Trans. Cognitive and Developmental Systems*, 2017, 9(2): 100–112. DOI: 10.1109/tcds.2016.2627820.

[13] Hoffmann M, Wang S, Outrata V, Alzueta E, Lanillos P. Robot in the mirror: Toward an embodied computational model of mirror self-recognition. *KI-Künstliche Intelligenz*, 2021, 35(1): 37–51. DOI: 10.1007/s13218-020-00701-7.

[14] Varela I D, Romero-Sorozabal P, Torricelli D, Delgado-Oleas G, Serrano J I, del Castillo Sobrino M D, Rocon E, Cebrian M. Sensorimotor features of self-awareness in multimodal large language models. arXiv: 2505.19237, 2025. https://arxiv.org/abs/2505.19237, Mar. 2026.

[15] Bongard J, Zykov V, Lipson H. Resilient machines through continuous self-modeling. *Science*, 2006, 314(5802): 1118–1121. DOI: 10.1126/science.1133687.

[16] Chen B, Kwiatkowski R, Vondrick C, Lipson H. Fully body visual self-modeling of robot morphologies. *Science Robotics*, 2022, 7(68): eabn1944. DOI: 10.1126/scirobotics.abn1944.

[17] Kwiatkowski R, Hu Y, Chen B, Lipson H. On the origins of self-modeling. arXiv: 2209.02010, 2022. https://arxiv.org/abs/2209.02010, Mar. 2026.

[18] Finn C, Levine S. Deep visual foresight for planning robot motion. In *Proc. the 2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 29-Jun. 3, 2017, pp.2786–2793. DOI: 10.1109/ICRA.2017.7989324.

[19] Gupta A, Murali A, Gandhi D, Pinto L. Robot learning in homes: Improving generalization and reducing dataset bias. In *Proc. the 32nd International Conference on Neural Information Processing Systems*, Dec. 2018, pp.9112–9122. DOI: 10.5555/3327546.3327584.

[20] Mu Y, Zhang Q, Hu M, Wang W, Ding M, Jin J, Wang B, Dai J, Qiao Y, Luo P. EmbodiedGPT: Vision-language pre-training via embodied chain of thought. In *Proc. the 37th International Conference on Neural Information Processing Systems*, Dec. 2023, pp.25081–25094. DOI: 10.5555/3666122.3667212.

[21] Pointeau G, Dominey P F. The role of autobiographical memory in the development of a robot self. *Frontiers in Neurorobotics*, 2017, 11: 27. DOI: 10.3389/fnbot.2017.00027.

[22] Bärmann L, Peller-Konrad F, Constantin S, Asfour T, Waibel A. Deep episodic memory for verbalization of robot experience. *IEEE Robotics and Automation Letters*, 2021, 6(3): 5808–5815. DOI: 10.1109/LRA.2021.3085166.

[23] Prescott T J, Vogeley K, Wykowska A. Understanding the sense of self through robotics. *Science Robotics*, 2024, 9(95): eadn2733. DOI: 10.1126/scirobotics.adn2733.

[24] Zhang S, Song X, Yu X, Bai Y, Guo X, Li W, Jiang S. HOZ++: Versatile hierarchical object-to-zone graph for object navigation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2025, 47(7): 5958–5975. DOI: 10.1109/TPAMI.2025.3552987.

[25] Gu J, Chaplot D S, Su H, Malik J. Multi-skill mobile manipulation for object rearrangement. In *Proc. the 11th International Conference on Learning Representations*, May 2023.

[26] Dalal M, Liu M, Talbott W, Chen C, Pathak D, Zhang J, Salakhutdinov R. Local policies enable zero-shot long-horizon manipulation. In *Proc. the 2025 IEEE International Conference on Robotics and Automation (ICRA)*, May 2025, pp.13875–13882. DOI: 10.1109/ICRA55743.2025.11128407.

[27] Shridhar M, Manuelli L, Fox D. CLIPort: What and where pathways for robotic manipulation. In *Proc. the 5th Conference on Robot Learning*, Nov. 2021, pp.894–906.

[28] Duval S, Wicklund R A. A Theory of Objective Self Awareness. Academic Press, 1972.

[29] de Boer D M L, Johnston P J, Namdar F, Kerr G, Cleeremans A. Predicting the bodily self in space and time. *Scientific Reports*, 2024, 14(1): 14813. DOI: 10.1038/s41598-024-65607-y.

[30] Hu Y, Chen B, Lipson H. Egocentric visual self-modeling for autonomous robot dynamics prediction and adaptation. *npj Robotics*, 2025, 3: 14. DOI: 10.1038/s44182-025-00031-6.

[31] Wolpert D M, Ghahramani Z, Jordan M I. An internal model for sensorimotor integration. *Science*, 1995, 269(5232): 1880–1882. DOI: 10.1126/science.7569931.

[32] Wolpert D M, Miall R C, Kawato M. Internal models in the cerebellum. *Trends in Cognitive Sciences*, 1998, 2(9): 338–347. DOI: 10.1016/S1364-6613(98)01221-2.

[33] Meyer N H, Gauthier B, Stampacchia S, Boscheron J, Babo-Rebelo M, Potheegadoo J, Herbelin B, Lance F, Alvarez V, Franc E, Esposito F, Morais Lacerda M, Blanke O. Embodiment in episodic memory through premotor-hippocampal coupling. *Communications Biology*, 2024, 7(1): 1111. DOI: 10.1038/s42003-024-06757-7.

[34] Elston T W, Wallis J D. Context-dependent decision-making in the primate hippocampal-prefrontal circuit. *Nature Neuroscience*, 2025, 28(2): 374–382. DOI: 10.1038/s41593-024-01839-5.

[35] Schulze L, Lipson H. High-degrees-of-freedom dynamic neural fields for robot self-modeling and motion planning. In *Proc. the 2024 IEEE International Conference on Robotics and Automation (ICRA)*, May 2024, pp.3064–3070. DOI: 10.1109/ICRA57147.2024.10611047.

[36] Poursiami H, Moshruba A, Cooper K W, Gobin D, Kaiser M A A, Singh A, Noor R, Shahbaba B, Jaiswal A, Fortin N J, Parsa M. A scalable reinforcement learning framework inspired by hippocampal memory mechanisms for efficient contextual and sequential decision making. *Scien-*

tific Reports, 2025, 15(1): 25221. DOI: 10.1038/s41598-025-10586-x.

[37] Mitchell R W. Mental models of mirror-self-recognition: Two theories. *New Ideas in Psychology*, 1993, 11(3): 295–325. DOI: 10.1016/0732-118X(93)90002-U.

[38] Zhang C, Han D, Qiao Y, Kim J U, Bae S H, Lee S, Hong C S. Faster segment anything: Towards lightweight SAM for mobile applications. arXiv: 2306.14289, 2023. https://arxiv.org/abs/2306.14289, Mar. 2026.

[39] Fang H S, Wang C, Fang H, Gou M, Liu J, Yan H, Liu W, Xie Y, Lu C. AnyGrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Trans. Robotics*, 2023, 39(5): 3929–3945. DOI: 10.1109/TRO.2023.3281153.

[40] Wei M, Wan C, Yu X, Wang T, Yang Y, Mao X, Zhu C, Cai W, Wang H, Chen Y, Liu X, Pang J. StreamVLN: Streaming vision-and-language navigation via SlowFast context modeling. arXiv: 2507.05240, 2025. https://arxiv.org/abs/2507.05240, Mar. 2026.

[41] Yamauchi B. A frontier-based approach for autonomous exploration. In *Proc. the 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97. 'Towards New Computational Principles for Robotics and Automation'*, Jul. 1997, pp.146–151. DOI: 10.1109/CIRA.1997.613851.

[42] Yu X, Zhang S, Song X, Qin X, Jiang S. Trajectory diffusion for ObjectGoal navigation. In *Proc. the 38th International Conference on Neural Information Processing Systems*, Dec. 2024, pp.110388–110411. DOI: 10.5555/3737916.3741420.

[43] Wang Z, Li X, Yang J, Liu Y, Jiang S. Sim-to-real transfer via 3D feature fields for vision-and-language navigation. In *Proc. the 8th Conference on Robot Learning*, Nov. 2024, pp.2982–2995.

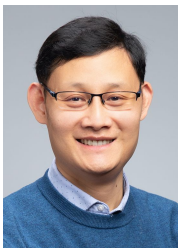
[44] Haarnoja T, Zhou A, Hartikainen K, Tucker G, Ha S, Tan J, Kumar V, Zhu H, Gupta A, Abbeel P, Levine S. Soft actor-critic algorithms and applications. arXiv: 1812.05905, 2018. https://arxiv.org/abs/1812.05905, Mar. 2026.

[45] Gui H, Pang S, He X, Wang T, Qiao S, Wang L, Yu S. High-performance grasp pose detection via point cloud serialization attention. *Pattern Recognition*, 2026, 171: 112099. DOI: 10.1016/j.patcog.2025.112099.

[46] Moghadam A R, Rastegari S, Masouleh M T, Kalhor A. Grasp the graph (GTG) 2.0: Ensemble of GNNs for high-precision grasp pose detection in clutter. arXiv: 2505.02664, 2025. https://arxiv.org/abs/2505.02664v1, Mar. 2026.

[47] Yenamandra S, Ramachandran A, Yadav K, Wang A S, Khanna M, Gervet T, Yang T Y, Jain V, Clegg A, Turner J M, Kira Z, Savva M, Chang A X, Chaplot D S, Batra D, Mottaghi R, Bisk Y, Paxton C. HomeRobot: Open-vocabulary mobile manipulation. In *Proc. the 7th Conference on Robot Learning*, Nov. 2023, pp.1975–2011.

[48] Liu P, Orru Y, Vakil J, Paxton C, Shafiullah N M M, Pinto L. OK-robot: What really matters in integrating open-knowledge models for robotics. arXiv: 2401.12202, 2024. https://arxiv.org/abs/2401.12202, Mar. 2026.



Shu-Qiang Jiang (Senior Member, IEEE) is a Professor at the University of Chinese Academy of Sciences (UCAS), Beijing, and a Professor at the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. He has authored or coauthored more than 150 papers on the related research topics. He was supported by the New-Star program of Science and Technology of Beijing Metropolis in 2008, NSFC Excellent Young Scientists Fund in 2013, Young top-notch talent of Ten Thousand Talent Program in 2014. He won the Lu Jiaxi Young Talent Award from Chinese Academy of Sciences in 2012, and the CCF Award of Science and Technology in 2012. He is the senior member of IEEE and CCF, member of ACM, Associate Editor of IEEE Multimedia, Multimedia Tools and Applications. He is the vice chair of IEEE CASS Beijing Chapter, vice chair of ACM SIGMM China chapter. He is the general chair of ICIMCS 2015, program chair of ACM Multimedia Asia2019 and PCM2017. He has also served as a TPC member for more than 20 well-known conferences, including ACM Multimedia, CVPR, ICCV, IJCAI, AAAI, ICME, ICIP, and PCM.

Sixian Zhang received his B.E. degree in automation from the University of Science and Technology Beijing, Beijing, in 2019, and his Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2025. He is currently an Assistant Researcher with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, embodied AI, and visual navigation.

Shi-Da Tao received her B.E. degree from the School of Computer Science and Technology, University of Chinese Academy of Science. She is currently a master's degree student in computer science at the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. Her research interests include computer vision, embodied AI, and visual navigation.

Xi-Hong Zhu received his B.E. degree from the School of Computer Science and Technology, University of Chinese Academy of Science. He is currently a Ph.D. student in computer science at the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His research interests include computer vision, embodied AI, and interactive navigation.

Tian-Liang Qi received his B.E. degree from the Department of Computer Science and Technology, Tsinghua University. He is currently a Ph.D. student in computer science at the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His research interests include computer vision, embodied AI, and visual rearrangement.

Xin-Hang Song received his B.E. degree in school of computer and information technology Beijing Jiaotong University, Beijing, in 2011, and his Ph.D. degree in computer science at Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2017. His research interests include image processing, large-scale image retrieval, image semantic understanding, multimedia content analysis, computer vision, and pattern recognition. He has served as PC or TPC member for well-known conferences, such as CVPR, ICCV, ECCV, ICML, and NeurIPS.