International Journal of Control, Automation, and Systems 23(6) (2025) 1597-1612
http://dx.doi.org/10.1007/s12555-025-0127-1
ISSN:1598-6446 eISSN:2005-4092
http://www.springer.com/12555

Reinforced Intelligence Through Active Interaction in Real World: A Survey on Embodied AI

Wookyong Kwon, Seungmin Baek, Jongchan Baek, WooSang Shin, Minseon Gwak, PooGyeon Park, and Sangmoon Lee*

Abstract: Embodied AI is a transformative field that enables intelligent systems to actively interact with and adapt to complex physical environments. This survey examines recent advancements, focusing on how embodied agents bridge the gap between cyber-physical systems and human-centered settings. Key developments include leveraging foundation models for high-level planning, affordance learning, and low-level control, alongside unifying data from internet-scale, simulation, and real-world sources. Reinforcement learning (RL), inverse RL, and imitation learning have been pivotal in advancing robotic control and skill acquisition. Additionally, the transition from transformers to deep state-space models (SSM) offers new possibilities for enhancing prediction and decision-making capabilities in embodied systems. We also discuss challenges and future directions, highlighting the importance of generalization, affordance learning, and the path toward artificial general intelligence (AGI). This survey provides a concise roadmap for researchers and practitioners shaping the future of embodied intelligence in real-world applications.

Keywords: Artificial general intelligence (AGI), deep state-space models (SSM), embodied AI, foundation models, reinforcement learning (RL).

1. INTRODUCTION

Recent advancements in artificial intelligence (AI) have accelerated its transition from cyberspace to the physical world, enabling direct interaction with and adaptation to real-world environments. Traditional AI has primarily focused on static, data-driven learning paradigms, but achieving true artificial general intelligence (AGI) necessitates the ability to learn and adapt through real-time interactions with the environment. This necessity has positioned embodied AI as a critical research paradigm, where AI systems equipped with a physical body, such as robots, can perceive, reason, and act in dynamic settings.

A key foundation enabling the realization of embodied AI is the Scaling Law, along with the development of Foundation Models. Over the past few years, AI models and datasets have grown exponentially in scale, highlighting the emergence of Emergent Capabilities unanticipated abilities that arise as models surpass critical size thresholds. This trend has driven the development of Foundation Models, which, unlike conventional AI systems, exhibit domain-agnostic generalization across diverse tasks. Particularly, prompt-based learning has demonstrated remarkable flexibility, allowing AI models to perform tasks with minimal task-specific training. One of the most transformative breakthroughs in AI research has been the advent of large language models (LLMs). Beyond traditional natural language processing (NLP), LLMs leverage in-context learning (ICL) to dynamically interpret context and employ chain-of-thought (CoT) reasoning to emulate human-like logical inference. These capabilities are particularly crucial for embodied AI, as they facilitate long-horizon manipulation and high-dimensional decision-making in complex and uncertain environments. Additionally, LLMs significantly enhance the interaction bandwidth between AI systems and humans, fostering more intuitive and adaptable communication channels.

A core enabler behind these advancements is transformer-based architecture, which plays a pivotal role in the development of embodied AI. Transformers leverage self-attention mechanisms to dynamically capture interdependencies within data, explicit position embeddings for efficient sequence encoding, and parallelization to significantly enhance computational efficiency. Furthermore, these advancements have accelerated the progress of multi-modal AI, allowing the seamless integration of heterogeneous data streams including language, vision, and sensory signals. This capability is essential for em-

Manuscript received February 21, 2025; accepted April 11, 2025. Recommended by Editor-in-Chief Hyo-Sung Ahn.

Wookyong Kwon and Jongchan Baek are with Electronics and Telecommunications Research Institute (ETRI), Korea (e-mails: wkwon@etri.re.kr, jcbaek@etri.re.kr). Seungmin Baek, Minseon Gwak, and PooGyeon Park are with the Division of Electrical Engineering, Pohang University of Sciece and Technology (POSTECH), Korea (e-mails: {preyso, minseon25, ppg}@postech.ac.kr). WooSang Shin is with Polaris3D, Korea (e-mail: we11d0ne@polaris3d.co). Sangmoon Lee is with the School of Electronic and Electrical Engineering, Kyungpook National University, Daehak-ro 80, Buk-gu, Daegu 41566, Korea (e-mail: moony@knu.ac.kr).
* Corresponding author.

bodied AI, as it enables comprehensive perception, contextual reasoning, and the execution of precise actions in dynamic environments.

Another crucial factor driving AI's evolution is the rise of Generative Models, which contribute significantly to data augmentation and simulation-based training. Autoregressive models and Diffusion models enable AI systems to generate highly realistic virtual environments, allowing embodied AI to simulate and learn from a diverse range of hypothetical scenarios without direct real-world exposure. This advancement paves the way for the development of robust world models, equipping embodied AI systems with the ability to generalize across unpredictable real-world conditions. These recent breakthroughs in AI research have collectively established a strong foundation for the progression of embodied AI. As AI methodologies converge, the feasibility of intelligent systems capable of autonomously learning, adapting, and interacting with the physical world is becoming increasingly tangible. This survey aims to provide an in-depth analysis of the current advancements in embodied AI, highlighting key technological trends and identifying critical future research directions.

Despite its long-standing research history, embodied AI remains a concept that defies a straightforward definition. This is because it extends beyond simply integrating AI with a physical body; rather, it encompasses multimodal sensory perception (visual, auditory, and tactile), language-based conceptual reasoning, and cross-modal information fusion. While many of these capabilities were once considered distant technological frontiers, rapid advancements in AI have accelerated their realization. However, there remains a lack of comprehensive references that systematically address the fundamental challenges in embodied AI and synthesize the latest research efforts. This survey aims to bridge this gap by identifying five critical challenges that hinder the advancement of embodied AI and by providing a structured review of recent developments and future research directions.

The Imperative for embodied AI-specific foundation models: Foundation Models have become a cornerstone of modern AI research, driving breakthroughs across various domains. However, most existing Foundation Models have been optimized for NLP and computer vision (CV) tasks, with limited adaptation to the distinctive requirements of embodied AI. Unlike traditional AI paradigms, embodied AI must be capable of processing egocentric, multi-modal sensor inputs, comprehending 3D spatial structures and underlying physical principles, executing human-instruction-grounded reasoning, and managing long-horizon, complex task execution. Given these challenges, there is a pressing need for Foundation Models explicitly tailored for embodied AI applications. This paper presents a comprehensive review of existing efforts to design such models and explores the emerging research trajectories in this area.

Large-scale dataset development and benchmarking for embodied AI: The efficacy of Foundation Models is inherently contingent on the availability of vast, high-quality training datasets. Conventional Foundation Models have leveraged extensive internet-scale corpora for pretraining, achieving remarkable generalization capabilities. However, existing datasets exhibit inherent limitations in viewpoint diversity, spatial dimensionality, and sensor modality representation, which are essential for training embodied AI systems. To bridge this gap, there is an urgent need for large-scale dataset curation, specifically tailored to embodied AI's unique data characteristics. Beyond dataset availability, rigorous benchmarking methodologies are equally crucial. Unlike standard AI models, embodied AI must be evaluated not only on task performance but also on safety, robustness, adaptability, and consistency across diverse, real-world environments. Thus, establishing realistic and reliable benchmarking frameworks is imperative. This survey synthesizes recent efforts in dataset curation and benchmarking methodologies, identifying gaps in current evaluation standards and outlining necessary improvements.

Advancements in control mechanisms and policy learning for embodied AI: A defining objective of embodied AI is to interact with and adapt to dynamic, real-world environments seamlessly. However, real-world interactions introduce significant stochasticity and uncertainty, posing challenges that conventional AI control strategies struggle to address. Effective high-dimensional motion planning, adaptive control strategies, and real-time optimization are crucial to ensuring robust and reliable policy learning in embodied systems. Reinforcement learning (RL) has emerged as a powerful framework for training embodied AI agents. Nevertheless, the Sim-to-Real Gap the disparity between simulated training environments and real-world deployment remains a persistent challenge. This paper reviews recent innovations in adaptive control strategies, analyzing methods that enhance policy generalization, transferability, and robustness to environmental variability.

Efficient neural architectures for real-time processing: While state-of-the-art AI models exhibit exceptional representational capacity, their computational inefficiency during inference poses a major bottleneck. Transformer-based architectures, despite their unparalleled scalability and expressive power, suffer from quadratic computational complexity concerning input sequence length, making them unsuitable for real-time decision-making and robotic control where ultra-low latency is critical. To address these challenges, next-generation neural architectures must balance computational efficiency with model expressivity. This paper explores recent advancements in hardware-aware model optimization, efficient attention mechanisms, and lightweight architectures, identify-

ing promising directions for enhancing the feasibility of real-time embodied AI applications.

The remainder of this paper is structured as follows: Section 2 explores how foundation models are leveraged for model consolidation, addressing their role in enhancing generalization and adaptability in embodied AI systems. Section 3 discusses the transition from internet-scale datasets to unified datasets, emphasizing the necessity of large-scale, high-quality data tailored for embodied AI. Sections 4 and 5 examine advances in robotic control and skill acquisition through reinforcement learning and human demonstrations, analyzing how learning-based approaches enable the development of robust and adaptive control policies. Finally, Section 6 investigates the shift from Transformer architectures to state-space models, highlighting recent efforts to develop efficient neural architectures that improve inference speed and scalability in real-time embodied AI applications.

2. CONSOLIDATING EMBODIED AI WITH FOUNDATION MODELS

2.1. The foundational backbone of embodied intelligence

Foundation models, trained on vast datasets encompassing diverse domains, have become a cornerstone of AI research. Their impact is largely driven by two key properties: emergence and homogenization [1]. Emergence citeers to the ability of these models to generalize across multiple domains, creating unexpected connections and enabling cross-disciplinary knowledge transfer. This property allows AI systems to adapt to novel problems beyond their original training scope, accelerating advancements in general-purpose intelligence. Homogenization, on the other hand, unifies AI methodologies by providing a shared framework that can be applied across various applications. This reduces the need for domain-specific architectures, enhances interoperability, and simplifies the deployment of AI solutions at scale. These characteristics collectively contribute to the rapid evolution of AI capabilities and the expansion of its applications into new areas.

In the field of embodied AI, foundation models play a pivotal role in integrating perception, reasoning, and action within interactive environments. Unlike conventional AI systems that rely on specialized models for distinct tasks, foundation models offer a unified approach, allowing autonomous agents to process multimodal data such as vision, language, and sensor inputs within a single framework. This capability is particularly beneficial in robotics and autonomous systems, where real-world environments present complex, unstructured challenges. Recent research demonstrates that foundation models significantly enhance robotic manipulation, autonomous navigation, and task generalization by leveraging their extensive prior knowledge and reasoning abilities. Moreover, the adaptability of these models allows for seamless integration of diverse sensor modalities, enabling embodied agents to operate in dynamic settings with minimal task-specific fine-tuning.

A wide range of foundation models have been developed to address different AI challenges, with notable examples spanning language understanding, vision-language interaction, image generation, and logical reasoning. In NLP, models such as BERT [2], PaLM [3], GPT [4], LLaMA [5], Qwen [6], and DeepSeek [7] have achieved state-of-the-art performance in text generation, classification, and reasoning. Vision-language models like CLIP [8] and Gemini [9] facilitate cross-modal understanding, supporting applications in image captioning, semantic retrieval, and embodied reasoning. SAM [10] and DINO, an image segmentation foundation model, generates masks based on user-provided prompts. Image generative models, including DALL-E [11] and Stable Diffusion [12], extend the capabilities of AI into creative domains, enabling high-quality content generation based on textual descriptions. More recently, foundation models designed for advanced reasoning, such as o1 [13] and DeepSeek-r1 [14], have demonstrated impressive logical and mathematical problem-solving capabilities, further pushing the boundaries of AI-driven inference.

To effectively leverage foundation models for specific tasks, various adaptation techniques have been developed. Prompting allows users to guide model responses through structured input queries, leveraging the model's pretrained knowledge without modifying its parameters. Fine-tuning enables domain-specific adaptation by updating model weights with specialized datasets, improving task performance in targeted applications. Feature-based adaptation utilizes pretrained embeddings for contrastive learning and similarity-based classification, allowing models to generalize effectively across different datasets. CoT reasoning explicitly structures intermediate reasoning steps, enhancing logical inference and decision-making capabilities. Additionally, ICL and retrieval-augmented generation (RAG) integrate external knowledge sources, supplementing model outputs with relevant, real-time information to improve response accuracy. These techniques collectively optimize foundation models for real-world deployment, ensuring that they can be efficiently adapted to diverse use cases.

2.2. Trailblazing applications of foundation models in embodied AI

Research in embodied AI is gaining significant attention. It aims to create systems that learn, adapt, and evolve while interacting with dynamic environments. To overcome Moravec's paradox, illustrated in Fig. 1, Foundation models, which have revolutionized fields like natural language processing (NLP) and computer vision (CV), are now being explored by integrating knowledge and

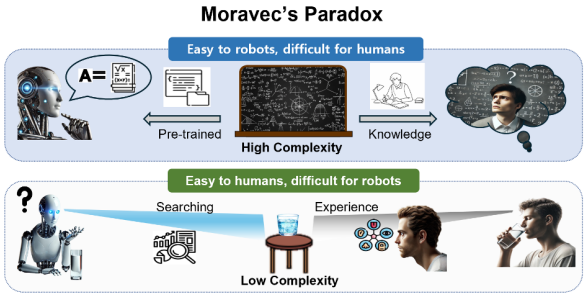


Fig. 1. Moravec's paradox.

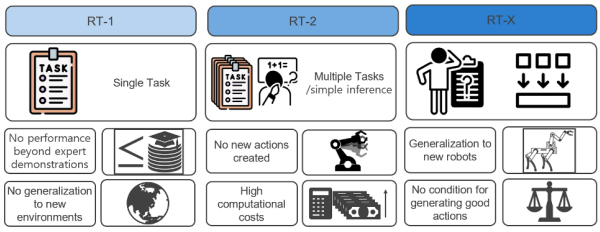


Fig. 2. Technical issues regarding robot transformers.

action-based learning. In the realm of embodied AI, foundation models have unlocked a diverse array of applications. Language-based models have become indispensable tools by enabling intuitive commands expressed in human language. This natural language interface allows for the decomposition and recombination of subtasks, facilitating complex task planning and long-horizon manipulation [15-19].

Moreover, leveraging the extensive body of human knowledge as priors has deepened the understanding of 3D spaces and physical laws. Integrating multiple modalities enriches both perception and reasoning, uncovering relationships between external knowledge elements that unimodal approaches might miss. For example, although an image might capture a person reflected in a building's glass (phenomenon stemming from light reflection) the underlying physical principles are not easily discernible from visual data alone [20-22].

Additionally, innovative research has employed code generation capabilities to program policies for robotic control [23-25], while large language models have been used to design implicit reward functions (or value functions) within reinforcement learning frameworks [26-28].

It is highly encouraging that the creative orchestration of foundation models alone paves the way for novel approaches that diverge from conventional methods.

2.3. Specialized foundation models for embodied intelligence

Thus far, discussions have focused on the innovative applications of foundation models. In this section, we turn our attention to research on developing models specialized for embodied intelligence. Unlike traditional systems, embodied intelligence must process egocentric and multimodal data, integrate perception with action within a unified framework, and employ dynamic world models that accurately simulate real-world physical laws and dynamics.

Recently, a new generation of agent models tailored for embodied AI has emerged. Fig. 2 shows technical issues regarding robot transformers. RT-1 [29] is a transformer-based model designed to generalize effectively across diverse tasks within a single learning framework. Trained on a vast and varied dataset, RT-1 enables robotic agents to operate seamlessly in multiple environments. By incorporating techniques such as TokenLearner, it reduces computational overhead and optimizes real-time performance. In systems like PaLM-SayCan [23], high-level reasoning modules are responsible for strategic planning, while RT-1 handles low-level control, allowing robots to accomplish hundreds of tasks with high success rates. RT-2 [30] further enhances the integration of vision-language understanding and execution control. By merging extensive Visual Question Answering data with robot behavior datasets, RT-2 evolves from a pure vision-language model into a comprehensive vision-language-action system. Its closed-loop control mechanism translates high-level textual commands into real-time executable actions, and the use of Chain-of-Thought prompts strengthens the reasoning process by clarifying causal relationships. Meanwhile, PaLM-E [31] expands the capabilities of large language models by incorporating raw multimodal sensor data. This approach achieves state-of-the-art performance and robust zero-shot generalization. As shown in Fig. 3, the Interactive Agent Foundation Model, which integrates various multi-task learning strategies, exemplifies the shifting paradigms in AI.

A notable example of a world foundation model is GenRL [32]. Without relying on complex, task-specific reward designs or explicit language annotations, GenRL

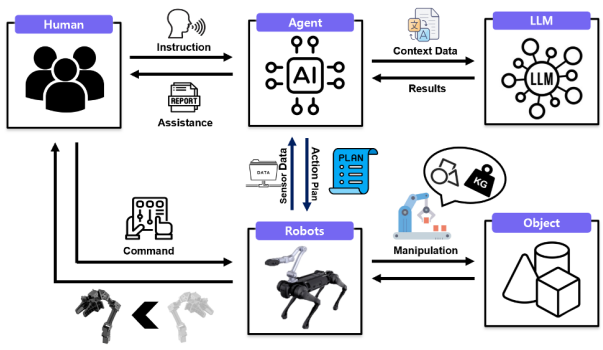


Fig. 3. Embodied AI with task-oriented interaction.

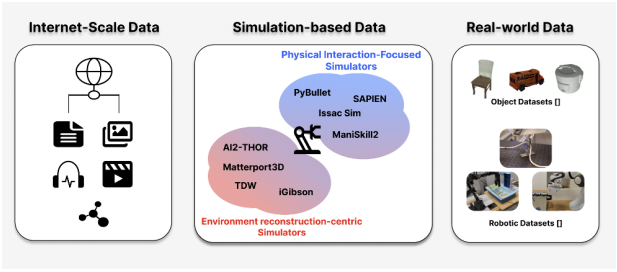


Fig. 4. Datatsets for training foundation models.

aligns vision-language models with generative world models. This alignment enables agents to learn behaviors from visual and language prompts, achieving multi-task generalization in tasks such as navigation and manipulation. GenRL effectively bridges the gap between high-level cognitive reasoning and the physical dynamics of the real world, highlighting the potential for robust policy learning. In addition, specialized language models have been developed to better capture the characteristics of multisensory data [33]. Cosmos [34] offers a generative model for promptable video synthesis, enabling users to obtain realistic world models that mimic desired environments through prompt-driven post-training.

In summary, the creative orchestration of these foundation models effectively connects high-level cognitive reasoning with precise real-world control, opening new possibilities in the field of embodied AI.

3. FROM INTERNET-SCALE DATA TO UNIFIED DATASET FOR EMDODIED AI

In the preceding section, we discussed the contributions of cognitive architectures and foundation models in advancing embodied AI toward artificial general intelligence (AGI). In this section, we explore into the critical role of data exchange between agents and their environments an indispensable component for training the inference mechanisms of foundation models and a cornerstone in enabling embodied AI to exhibit intelligence.

Embodied AI posits that true intelligence arises from the interaction between an agent and its environment, and that ensuring the quality and diversity of data obtained through such interactions is essential for its realization. Consequently, ongoing research seeks to find efficient methods for agents to organize and acquire data through their interactions and experiences within the environment. This data pipeline serves as the cornerstone of embodied AI, underpinning the development of more advanced intelligence through learning and reasoning processes. It can be derived from three primary sources: internet-based, simulator-based, and real-world environments as shown in Fig. 4.

3.1. Internet-scale data and foundation models

The first source is the vast repository of information available on the internet. While the purpose and quality of this data are often ambiguous, its sheer volume and combined with robust pre-processing techniques allow it to be transformed into a usable format. This type of data offers a significant advantage that even the datasets traditionally used in foundation model training can be continually re-processed and repurposed, enabling further refinement of intelligence concepts. Foundation models pre-trained on such massive, internet-scale datasets have exhibited remarkable capabilities in understanding, reasoning, interaction, and generation, powered by billions of parameters.

However, foundation models trained on such datasets face inherent limitations in fulfilling the requirements of embodied AI. These models are primarily designed to provide auxiliary information rather than to serve as direct decision-makers. Moreover, they face to challenges in understanding the fundamental laws of the physical world and to enable high-level intelligence and interaction based on those principles. To address these shortcomings, embodied AI requires a novel, integrative approach that leverages foundation models to comprehend environmental contexts and optimize agent behaviors.. Bridging this gap is not merely about providing information but about addressing the critical challenge of realizing actionable intelligence in the real world.

3.2. Simulators-based data for embodied AI

Another significant source of data originates from simulator-based environments, which are pivotal in advancing research on embodied AI. These simulators facilitate cost-effective experimentation, ensure safety by replicating potentially hazardous scenarios, and offer scalability by enabling testing across a wide range of diverse and complex environments. Additionally, simulators enable rapid prototyping, controlled settings for precise experimentation, data generation for training and evaluation, and the establishment of standardized benchmarks for algorithm comparison. By incorporating physical properties, object characteristics, and their interactions within realistic simulated environments, these platforms empower agents to engage meaningfully with their surroundings, fostering the development of advanced capabilities.

Key simulation platforms can be broadly classified into two major categories: Physical Interaction-Focused Simulators, which prioritize modeling complex physical interactions such as force dynamics and object manipulation, and environment Reconstruction-Focused Simulators, which concentrate on replicating realistic environments to enhance navigation and spatial understanding. Together, these categories address the multifaceted requirements of embodied AI, offering essential tools to drive forward research and development in this rapidly

evolving field.

Physical interaction-focused simulators: Simulators focusing on physical interactions specialize in modeling intricate physical interactions such as force dynamics, collision detection, and object manipulation These platforms are particularly well-suited for applications that demand high-fidelity physical modeling, including robotic arm manipulation and whole-body simulations. Prominent simulators include PyBullet [35], a lightweight physics engine that supports diverse physical interactions; SAPIEN [36], known for its precise control over object interactions; and ManiSkill2 [37], which specializes in deformable object manipulation using sophisticated material modeling techniques.

Environment reconstruction-centric simulators: Environment reconstruction-centric simulators are designed to replicate real-world environments with high fidelity, enabling tasks such as navigation, exploration, and spatial reasoning. These simulators combine multi-modal data, including video, audio, and LiDAR, to create a comprehensive representation of the environment. Notable simulators include AI2-THOR [38], which offers interactive indoor scenes with physical properties, Matterport3D [39], featuring a large-scale 2D-3D dataset for embodied navigation and Habitat [40], which supports high-performance, parallel 3D simulations. Furthermore, simulators like ThreeDWorld (TDW) [41] and iGibson [42] advance this category by integrating high-quality visual and physical effects, making them ideal for multi-modal learning and perception applications.

3.3. Real-world data

Real-world datasets are critical for enabling embodied agents to perceive and operate effectively within physical environments. Derived directly from real-world scenarios, these datasets are indispensable for capturing the intricate complexities and uncertainties inherent in real-world interactions. Unlike their simulated counterparts, real-world datasets provide fine-grained physical properties, dynamic environmental conditions, and nuanced human-object interactions, serving as a foundational resource for grounding embodied intelligence in the physical world.

Object datasets: Object datasets are foundational to tasks such as robotic perception, environment modeling, and manipulation. These datasets often include synthesized or real-world scanned objects with labels, meshes, point clouds, and other annotations. For example, OmniObject3D [43] contains 6,000 scanned objects spanning 190 categories, providing high-fidelity shapes and multi-view imagery. Such datasets are invaluable for enabling downstream tasks like object reconstruction, novel view synthesis, and robotic interaction.

Human datasets: Human datasets focus on activities captured from first- or third-person perspectives in diverse environments, such as kitchens or outdoor settings. These datasets include real-world information about human-object interactions and physical laws embedded in human dynamics. For example, Ego-Exo4D [44] comprises 1,422 hours of multi-view videos from over 800 participants in 13 cities. These datasets enable agents to learn about human-centric interactions, which are critical for real-world applications like service robots and assistive technologies.

Robotic datasets: Robotic datasets document the task performance of robots in either real-world or simulated environments. The collection of these datasets, however, remains a resource-intensive and time-consuming endeavor, often necessitating teleoperation systems and skilled operators. For exmample, RT-X [45] project collected data from 22 robots across 21 institutions, showcasing 527 distinct skills executed over 160,266 tasks. Such large-scale datasets have significantly contributed to advancements in zero-shot generalization and cross-morphology learning.

3.4. Challenges and future directions

Despite their importance, real-world datasets face inherent limitations due to the time-intensive and resource-demanding nature of their acquisition processes, which restrict both the quantity and quality of data available. Conversely, simulated or internet datasets, while cost-effective and scalable, often lack the nuanced complexity and fidelity of real-world environments. This trade-off presents a significant challenge in embodied AI research, where bridging the gap between simulated and real-world data is critical for developing robust and adaptable systems.

Modern approaches to addressing these challenges increasingly adopt hybrid methodologies that combine real-world and simulation-based datasets during training. By utilizing the detailed richness of real-world datasets while offsetting their limitations with simulation-based data, researchers are exploring advanced techniques such as transfer learning to effectively harmonize these two domains. For example, transformer-based models pre-trained on simulated data can be fine-tuned with real-world datasets, thereby improving generalization across diverse environments.

4. EMBODIED ROBOTIC CONTROL AND SKILL ACQUISITION USING REINFORCEMENT LEARNING

Building on the integration of real-world and simulated datasets, reinforcement learning (RL) has emerged as a powerful framework for enabling robots to acquire complex control and skill-learning capabilities. By leveraging interactions with their environment and learning from trial-and-error feedback, RL allows embodied AI systems to develop adaptive policies that generalize across diverse and dynamic settings. This RL approach in Fig. 5 is par-

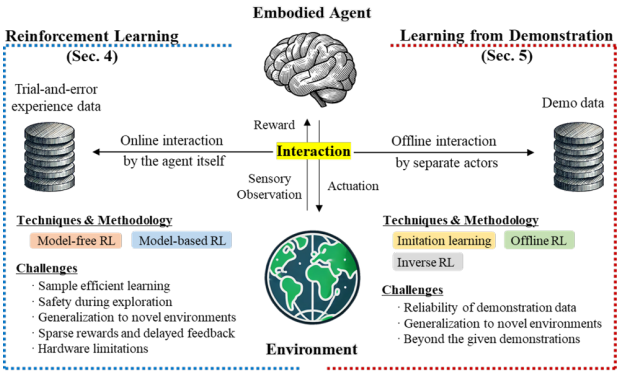


Fig. 5. Approaches for learning control and skill acquisition in embodied agents.

ticularly advantageous in scenarios where pre-collected datasets may be insufficient or impractical, as RL facilitates on-the-fly learning in real-time conditions. RL-driven policies have demonstrated remarkable success, notably in addressing tasks that demand agility, precision, and adaptability, including locomotion, manipulation, and autonomous navigation.

4.1. Techniques and methodologies

RL aims to optimize policy models using experience data obtained through trial and error. RL algorithms have evolved into two main approaches: model-free RL and model-based RL. Popular model-free RL methods, such as proximal policy optimization (PPO) [46] and soft actor-critic (SAC) [47], optimize policies by estimating state-action cumulative rewards from experience data. The recently proposed BRO algorithm [48] has improved sample efficiency and policy performance in model-free RL by incorporating structural enhancements and regularization techniques in critic networks. Model-based RL learns state-transition models from experience data, enabling optimal control policies based on state prediction. Since trained state-transition models allow for imagination rollouts, model-based RL generally requires fewer trial-and-error interactions than model-free RL, improving sample efficiency. Model-based RL approaches commonly employ techniques such as model predictive control (MPC) [49] and multi-step loss frameworks [50] for policy optimization. Notably, temporal difference model predictive control (TD-MPC) [51,52] combines temporal difference learning with MPC, demonstrating high sample efficiency and the ability to learn complex tasks in robotic systems.

Beyond these primary RL methodologies, several advancements have been made to facilitate RL applications in physical robotic systems and complex tasks. Sim-to-real transfer techniques allow policies trained in simulated environments to be effectively deployed in real-world systems. Domain randomization [53] and zero-shot transfer methods [54] enable successful policy application in real-world settings despite discrepancies between simulation and reality. Hierarchical RL [55-57] decomposes complex tasks into subtasks, facilitating high-dimensional robotic learning, and multi-modal RL [58] integrates visual, auditory, and tactile data, allowing robots and physical systems to make more intelligent and refined decisions.

4.2. Applications and case studies

RL has demonstrated remarkable achievements across various robotic applications. For example, Boston Dynamics' Spot and ANYmal robots leverage RL-based algorithms to learn adaptive and energy-efficient locomotion patterns [59]. These robots successfully traverse uneven terrains and dynamically adjust to environmental changes. Additionally, hierarchical RL frameworks have been employed to train robots for environment-aware locomotion skills [60]. Recently, RL has also been successfully applied to train robots for advanced movements, such as parkour maneuvers [61].

In robotic manipulation, RL has played a crucial role in achieving precision control. OpenAI's Dactyl project [62] demonstrated the ability to manipulate a Rubik's Cube using domain randomization, highlighting RL's potential in high-precision manipulation tasks. Affordance learning techniques have been employed to generalize manipulation policies across objects with diverse shapes and material properties [63,64]. Furthermore, multi-modal RL has enhanced object recognition and handling, making robotic grasping and sorting tasks more efficient [58].

RL is also actively studied in autonomous driving applications [65], where it is used for lane merging, obstacle avoidance, and urban navigation. In the medical field, RL is being applied to surgical robotics, enabling precise and adaptive minimally invasive procedures. RL-based optimization techniques have been also used to improve robotic systems in clinical settings [66].

4.3. Challenges and future directions

Despite its successes, RL still faces several challenges. Sample efficiency remains a significant limitation, as RL requires extensive interaction with the environment, which can be time-consuming and potentially harmful to physical systems. Research efforts are focused on improving sample efficiency in both model-free and model-based RL approaches, with an emphasis on hybrid methods that integrate the strengths of both paradigms. Additionally, Sim-to-Real techniques are essential for the practical application of RL in physical environments, and ongoing research aims to bridge the gap between simulation and reality for more reliable deployment.

Ensuring safety during exploration is another critical challenge, particularly in human-robot collaborative settings. Safe RL frameworks and constrained exploration techniques are being developed to mitigate risks during policy learning [67-69]. To enhance RL generalization

to new environments, researchers are investigating meta-learning and curriculum learning approaches [70,71]. Furthermore, sparse rewards and delayed feedback create difficulties in policy optimization, necessitating innovations in intrinsic motivation and reward-shaping methodologies [72]. Hardware limitations, including actuation delays and limited computational resources, pose additional challenges in the deployment of RL systems in physical robots [54]. Addressing these constraints through resource-aware RL frameworks and energy-efficient algorithms is critical for future embodied AI applications.

Future research directions will focus on developing lifelong learning frameworks. To build human-level intelligent embodied agents, RL must incorporate continuous learning capabilities. Research on lifelong RL or continual RL aims to develop algorithms that allow robots to sequentially acquire new skills while retaining previously learned ones, much like humans do [73,74]. Multi-agent RL and human-in-the-loop RL methods are gaining traction, offering promising avenues for improving learning efficiency through collaboration and human feedback [75-77]. Finally, advancements in dynamic reward shaping and contextual learning using LLMs hold potential for enhancing RL efficiency and intuitive learning [78,79]. Moving beyond traditional handcrafted reward functions, applying LLMs to automated reward engineering can make RL-based robotic learning more intuitive and effective.

5. LEARNING ROBOTIC CONTROL AND SKILLS FROM DEMONSTRATIONS

In addition to RL, learning robotic control and skills from demonstrations in Fig. 5 is another promising approach for high-level task control in embodied agents. This approach enables policy acquisition by leveraging expert data obtained from human demonstrations or other sources. By imitating expert behavior or learning from stored interaction datasets, robots can bypass or mitigate challenges associated with RL, such as reward engineering and extensive exploration requirements. This method has been successfully applied across various industries to achieve human-level, high-level robotic control. In particular, hardware and software developments in mobile robot-based data collection platforms have enhanced the ease of acquiring expert demonstration data. These advancements have significantly improved the availability and quality of expert demonstration datasets, leading to remarkable improvements in policy learning for complex manipulation tasks.

5.1. Techniques and methodologies

Three major methodological approaches have been studied for learning from demonstrations: imitation learning, offline RL, and inverse RL.

Imitation learning plays a crucial role in learning high-level robotic task policies using high-quality data from expert demonstrations. One of the simplest approaches, behavioral cloning [80], maps observations to actions using supervised learning. Generative adversarial imitation learning (GAIL) [81] extends this concept by incorporating adversarial training to improve policy robustness. These methods have been effectively applied in robotic manipulation, locomotion, and navigation tasks. Recently, vision-language models (VLMs) and transformer-based models have been utilized to handle complex observation-to-action mappings, demonstrating the potential to learn high-level tasks that were previously unachievable by robots [82-84].

Offline RL bridges imitation learning and conventional RL by training policies on pre-existing datasets of interactions. Unlike imitation learning, offline RL accommodates both expert demonstrations and control failure data, such as trial-and-error interactions. A key challenge in offline RL is handling dataset biases and distributional shifts. Algorithms such as conservative Q-learning (CQL) [85] and behavior-regularized actor-critic (BRAC) [86,87] address these challenges, enhancing the performance and generalization of learned policies. Advanced methods, such as normalized actor-critic [88] and multi-modal offline RL [89,90], have further expanded the potential of learning from given datasets. Decision transformer [91], a model architecture that utilizes transformers for learning complex action sequences from observations, has demonstrated the potential of improving offline RL performance by increasing model complexity.

Inverse RL takes a different approach by inferring underlying reward functions from expert behaviors rather than directly mimicking actions. Unlike standard RL, which relies on predefined reward functions, inverse RL extracts meaningful rewards from expert demonstrations, enabling better generalization across different environments. Techniques such as maximum entropy inverse RL [92] and adversarial inverse RL [93] have demonstrated success in deriving robust reward functions across diverse tasks and environments.

5.2. Applications and case studies

Practical applications of learning from demonstrations span various fields. RoboTurk [94], which leverages crowd-sourced human demonstrations, has trained robots for complex manipulation tasks such as object sorting and assembly. In the medical field, imitation learning frameworks have been employed to teach surgical robots precise and adaptive control techniques for minimally invasive procedures. A notable example is the da Vinci Surgical Research Kit system, which demonstrates how learning from expert demonstrations enhances accuracy and efficiency in robotic-assisted surgery [95].

Autonomous driving systems frequently employ offline

RL and imitation learning to handle diverse scenarios, including pedestrian interactions and navigation in complex urban environments [96,97]. In addition, integration of these techniques has been suggested to improve the robustness of self-driving vehicles in real-world scenarios [98]. In applications requiring reward function inference, IRL techniques, such as maximum entropy IRL, have been instrumental in designing robust control systems for dynamic tasks, such as multi-agent collaboration and human-robot interaction, ensuring more adaptive and responsive robotic behavior.

5.3. Challenges and future directions

Despite its advantages, learning from demonstrations presents several challenges that must be addressed for broader practical applications in embodied agent control and skill acquisition. Firstly, improving the accessibility and efficiency of expert data collection is crucial. Developing advanced data collection platforms, such as ALOHA [99], ALOHA 2 [100], and Mobile ALOHA [101], facilitates large-scale dataset acquisition for bimanual teleoperation tasks in an economical and ergonomic manner.

Furthermore, ensuring the quality and diversity of demonstration data is essential for robust policy learning. Techniques such as data augmentation [102,103] and reward relabeling [104] can help mitigate some of these limitations. Additionally, enabling robots to generalize beyond the constraints of demonstration datasets remains an open problem. Advanced IRL techniques focusing on scalable reward inference and adversarial robustness are necessary to expand the applicability of these methods.

Future advancements in hybrid methodologies integrating RL, IRL, and imitation learning are expected to play a pivotal role in overcoming these challenges. For example, integrating transformer-based architectures like decision transformer [91] into policy learning could enhance temporal dependencies and context understanding in robotic control. Additionally, developing self-supervised and meta-learning approaches could improve adaptability to novel tasks with minimal human supervision. Overcoming these barriers will enable robots to seamlessly integrate human-like adaptability and expertise into their control policies, leading to broader deployment in real-world applications.

6. EFFICIENCY IMPROVEMENTS IN NEURAL ARCHINTECTURES

Both RT-1 and RT-2 are Transformer-based models that utilize key-value caches for all tokens in an input sequence, allowing embodied AI to capture long-term dependencies crucial for sequential decision-making. However, their quadratic computational complexity poses a major challenge for real-time robotic control, where efficiency is critical. This limitation often necessitates restricting sequence length or reducing input resolution, potentially hindering generalization in dynamic environments.

6.1. Deep state space models for embodied AI

To address the high computational cost of Transformers, alternative models can be categorized into two main approaches: 1) methods that retain self-attention while achieving subquadratic complexity through approximation or by limiting the number of considered tokens, and 2) methods that incorporate memory states to sequentially scan the input sequence with subquadratic complexity. The first category includes sparse attention [105,106], which limits the tokens considered within a sequence based on specific patterns, and linear attention [107,108], which approximates the original softmax attention computation to achieve linear complexity. The second category encompasses long short-term memory (LSTM)-based methods [109] focused on gating mechanisms and state space models (SSMs) [110-114] that emphasize the use of systems. Due to their highly reduced computational and memory requirements, the second approach can be particularly well-suited for embodied AI applications that demand real-time or internet-free predictions in low-cost computing environments.

6.2. Techniques and methodologies

Specifically, SSM layers parameterize linear state-space systems, which have been widely used in robotics and control theory [115], demonstrating matching or exceeding performance to Transformer in addressing long-range dependency problems [116]. The linearity of the systems enables fast training through parallel scan [112] or frequency-domain inference [117] to compute the output sequence, as illustrated in Fig. 6. For the inference on long sequences, the computational stability of SSMs can be guaranteed by utilizing diagonal systems and directly constraining poles of the systems during training

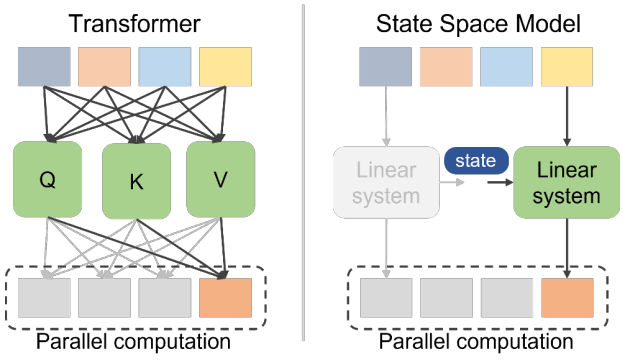


Fig. 6. Schematic comparison of inference mechanisms in training Transformers and SSMs.

[111,112,118]. The early models with input-independent systems [110-112,117] have demonstrated effectiveness in modeling sequences with strong continuous properties, such as audio data. In contrast, models like Mamba [113] and Mamba-2 [114] leverage input-dependent systems to compress input sequences into states more effectively, enabling their potential application to domains with strong discrete characteristics, such as language.

In embodied AI, SSM-based approaches have been explored as a low-cost alternative for various tasks, including manipulation [119], navigation [120,121], and imitation learning [122,123]. In RL, the Decision Transformer [124], which generates actions via reward-conditioned sequence modeling, has inspired research into Decision Mamba [125-127], a series of methods that replace the Transformer mechanism with Mamba while retaining the same overall approach.

6.3. Challenges and future directions

While state-based modeling provides substantial efficiency gains, SSMs face limitations in tasks that rely on context-dependent information retrieval and in-context learning [116]. To overcome these challenges, hybrid SSM-Attention architectures have been proposed by arranging SSM and attention heads serially or in parallel [116,128-130]. These recent studies on hybrid models indicate that SSM heads and attention heads can complement weaknesses by balancing information usage and efficiency.

Additionally, various sensors generate multi-modal data in applications like navigation within embodied AI environments [131]. Therefore, models must efficiently process this data to perform tasks effectively. To address this, extensions of Mamba have been proposed for fusing different modalities [131-133]. Similar to previous approaches where multi-modal tokens are defined as a single sequence, integrating SSMs allows for implicit fusion within the model, significantly enhancing inference speed compared to attention-based models [134]. Focusing on the application of SSMs in various modalities and evaluating performance on longer sequences is essential, along with efforts to apply these models in real-world environments.

7. DISCUSSIONS

Despite significant advancements, several challenges remain in embodied AI, particularly in reinforcement learning (RL), imitation learning, and real-world deployment. RL continues to suffer from sample inefficiency, necessitating extensive interactions that are costly and impractical for physical systems. Hybrid approaches combining model-free and model-based RL and Sim-to-Real transfer techniques are critical for bridging the gap between simulation and reality. Similarly, learning from demonstrations faces limitations in data accessibility, quality, and generalization. Enhancing scalable data collection, reward inference, and augmentation strategies is essential for improving policy robustness.

Moreover, while Transformers have been widely adopted for sequence modeling, their quadratic computational complexity constrains real-time deployment. State-space models (SSMs) offer a more efficient alternative, incorporating insights from robotics and control theory. Research into SSM-Transformer hybrids and multi-modal architectures aims to balance computational efficiency with contextual reasoning, yet ensuring their robustness in dynamic environments remains an open problem. Future work integrating RL, inverse RL, imitation learning, adaptive architectures, and advances in self-supervised learning and decision transformers will be crucial for improving generalization and efficiency. Addressing these challenges will drive embodied AI's progress across robotics, healthcare, and autonomous systems, advancing the pursuit of more generalizable intelligence.

DECLARATIONS

Conflict of Interest

The authors declare that there is no competing financial interest or personal relationship that could have appeared to influence the work reported in this paper. Sangmoon Lee is a Senior Editor of International Journal of Control, Automation, and Systems. Senior Editor status has no bearing on editorial consideration.

Author Contributions

Wookyong Kwon: Investigation, Visualization, Writing-original draft, Funding acquisition. Seungmin Baek: Conceptualization, Investigation, Writing-original draft. Jongchan Baek: Conceptualization, Investigation, Writing-original draft. WooSang Shin: Conceptualization, Investigation, Writing-original draft. Minseon Gwak: Investigation, Writing-original draft. PooGyeon Park: Supervision, Conceptualization, Writing-review & editing. Sangmoon Lee: Project administration, Supervision, Visualization, Writing-original draft, Writing-review & editing, Funding acquisition.

Funding

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) though the Korean Government (MSIT) (No. RS-2025-02218631, Development of Reinforced Embodied Intelligence through Active Interaction in Real-World) and Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [25ZD1130, Regional Industry ICT Convergence Technology Advancement and Support Project in Daegu-GyeongBuk (Robot)].

REFERENCES

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint* arXiv:2108.07258, 2021.

[2] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint* arXiv:1810.04805, 2018.

[3] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1-113, 2023.

[4] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint* arXiv:2303.08774, 2023.

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint* arXiv:2302.13971, 2023.

[6] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, *et al.*, "Qwen technical report," *arXiv preprint* arXiv:2309.16609, 2023.

[7] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, *et al.*, "Deepseek-v3 technical report," *arXiv preprint* arXiv:2412.19437, 2024.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," *Proc. of International Conference on Machine Learning*, PMLR, pp. 8748-8763, 2021.

[9] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, *et al.*, "Gemini: A family of highly capable multimodal models," *arXiv preprint* arXiv:2312.11805, 2023.

[10] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp. 4015-4026, 2023.

[11] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *Proc. of International Conference on Machine Learning*, PMLR, pp. 8821-8831, 2021.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684-10695, 2022.

[13] "Learning to reason with LLMs," OpenAI, [Online]. Available: https://openai.com/index/learning-to-reason-with-llms. [Accessed: March 23, 2025].

[14] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint* arXiv:2501.12948, 2025.

[15] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," *Proc. of 2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 10608-10615, 2023.

[16] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar, "Voyager: An open-ended embodied agent with large language models," *arXiv preprint* arXiv:2305.16291, 2023.

[17] G. Tziafas and H. Kasaei, "Lifelong robot library learning: Bootstrapping composable and generalizable skills for embodied control with language models," *Proc. of 2024 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 515-522, 2024.

[18] T. Yoneda, J. Fang, P. Li, H. Zhang, T. Jiang, S. Lin, B. Picker, D. Yunis, H. Mei, and M. R. Walter, "Statler: State-maintaining language models for embodied reasoning," *Proc. of 2024 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 15083-15091, 2024.

[19] H. Zhang, W. Du, J. Shan, Q. Zhou, Y. Du, J. B. Tenenbaum, T. Shu, and C. Gan, "Building cooperative embodied agents modularly with large language models," *arXiv preprint* arXiv:2307.02485, 2023.

[20] Y. Zhou, L. Huang, Q. Bu, J. Zeng, T. Li, H. Qiu, H. Zhu, M. Guo, Y. Qiao, and H. Li, "Embodied understanding of driving scenarios," *Proc. of European Conference on Computer Vision*, Springer, pp. 129-148, 2024.

[21] Y. Yang, F.-Y. Sun, L. Weihs, E. VanderBilt, A. Herrasti, W. Han, J. Wu, N. Haber, R. Krishna, L. Liu, *et al.*, "Holodeck: Language guided generation of 3D embodied AI environments," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16227-16237, 2024.

[22] J. Huang, S. Yong, X. Ma, X. Linghu, P. Li, Y. Wang, Q. Li, S.-C. Zhu, B. Jia, and S. Huang, "An embodied generalist agent in 3D world," *arXiv preprint* arXiv:2311.12871, 2023.

[23] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, *et al.*, "Do as I can, not as I say: Grounding language in robotic affordances," *arXiv preprint* arXiv:2204.01691, 2022.

[24] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," *Proc. of 2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 9493-9500, 2023.

[25] J. Yang, Y. Dong, S. Liu, B. Li, Z. Wang, H. Tan, C. Jiang, J. Kang, Y. Zhang, K. Zhou, *et al.*, "Octopus: Embodied vision-language programmer from environmental feedback," *Proc. of European Conference on Computer Vision*, Springer, pp. 20-38, 2024.

[26] T. Xie, S. Zhao, C. H. Wu, Y. Liu, Q. Luo, V. Zhong, Y. Yang, and T. Yu, "Text2reward: Reward shaping with language models for reinforcement learning," *Proc. of the 12th International Conference on Learning Representations*, 2024.

[27] Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, "Eureka: Human-level reward design via coding large language models," *arXiv preprint* arXiv:2310.12931, 2023.

[28] M. Klissarov, P. D'Oro, S. Sodhani, R. Raileanu, P.-L. Bacon, P. Vincent, A. Zhang, and M. Henaff, "Motif: Intrinsic motivation from artificial intelligence feedback," *arXiv preprint* arXiv:2310.00166, 2023.

[29] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint* arXiv:2212.06817, 2022.

[30] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint* arXiv:2307.15818, 2023.

[31] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint* arXiv:2303.03378, 2023.

[32] P. Mazzaglia, T. Verbelen, B. Dhoedt, A. Courville, and S. Rajeswar, "Genrl: Multimodal-foundation world models for generalization in embodied agents," *arXiv preprint* arXiv:2406.18043, 2024.

[33] Y. Hong, Z. Zheng, P. Chen, Y. Wang, J. Li, and C. Gan, "Multiply: A multisensory object-centric embodied large language model in 3d world," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26406-26416, 2024.

[34] N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, P. Chattopadhyay, Y. Chen, Y. Cui, Y. Ding, *et al.*, "Cosmos world foundation model platform for physical AI," *arXiv preprint* arXiv:2501.03575, 2025.

[35] E. Coumans and Y. Bai, "Pybullet, a Python module for physics simulation for games, robotics and machine learning," 2016.

[36] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, *et al.*, "Sapien: A simulated part-based interactive environment," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11097-11107, 2020.

[37] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao, *et al.*, "Maniskill2: A unified benchmark for generalizable manipulation skills," *arXiv preprint* arXiv:2302.04659, 2023.

[38] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, M. Deitke, K. Ehsani, D. Gordon, Y. Zhu, *et al.*, "Ai2-thor: An interactive 3d environment for visual AI," *arXiv preprint* arXiv:1712.05474, 2017.

[39] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint* arXiv:1709.06158, 2017.

[40] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, *et al.*, "Habitat: A platform for embodied AI research," *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp. 9339-9347, 2019.

[41] C. Gan, J. Schwartz, S. Alter, D. Mrowca, M. Schrimpf, J. Traer, J. De Freitas, J. Kubilius, A. Bhandwaldar, N. Haber, *et al.*, "Threedworld: A platform for interactive multi-modal physical simulation," *arXiv preprint* arXiv:2007.04954, 2020.

[42] F. Xia, W. B. Shen, C. Li, P. Kasimbeg, M. E. Tchapmi, A. Toshev, R. Martín-Martín, and S. Savarese, "Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 713-720, 2020.

[43] T. Wu, J. Zhang, X. Fu, Y. Wang, J. Ren, L. Pan, W. Wu, L. Yang, J. Wang, C. Qian, *et al.*, "Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 803-814, 2023.

[44] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, *et al.*, "Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383-19400, 2024.

[45] A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan, *et al.*, "Open X-embodiment: Robotic learning datasets and RT-X models," *arXiv preprint* arXiv:2310.08864, 2023.

[46] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint* arXiv:1707.06347, 2017.

[47] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *Proc. of International Conference on Machine Learning*, PMLR, pp. 1861-1870, 2018.

[48] M. Nauman, M. Ostaszewski, K. Jankowski, P. Miłoś, and M. Cygan, "Bigger, regularized, optimistic: Scaling for compute and sample-efficient continuous control," *arXiv preprint* arXiv:2405.16158, 2024.

[49] C. E. Garcia, D. M. Prett, and M. Morari, "Model predictive control: Theory and practice – A survey," *Automatica*, vol. 25, no. 3, pp. 335-348, 1989.

[50] Y. Yang, K. Caluwaerts, A. Iscen, T. Zhang, J. Tan, and V. Sindhwani, "Data Efficient Reinforcement Learning for Legged Robots," *arXiv preprint* arXiv:1907.03613, 2019.

[51] N. Hansen, X. Wang, and H. Su, "Temporal difference learning for model predictive control," *arXiv preprint* arXiv:2203.04955, 2022.

[52] N. Hansen, H. Su, and X. Wang, "TD-MPC2: Scalable, robust world models for continuous control," *arXiv preprint* arXiv:2310.16828, 2023.

[53] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," *Proc. of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23-30, 2017.

[54] X. Gu, Y.-J. Wang, X. Zhu, C. Shi, Y. Guo, Y. Liu, and J. Chen, "Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning," *arXiv preprint* arXiv:2408.14472, 2024.

[55] O. Nachum, S. S. Gu, H. Lee, and S. Levine, "Data-efficient hierarchical reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[56] D. Jain, A. Iscen, and K. Caluwaerts, "Hierarchical reinforcement learning for Quadruped Locomotion," *arXiv preprint* arXiv:1905.08926, 2019.

[57] Z. Hou, J. Fei, Y. Deng, and J. Xu, "Data-efficient hierarchical reinforcement learning for robotic assembly control applications," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 11, pp. 11565-11575, 2020.

[58] J. Zhang, N. Gireesh, J. Wang, X. Fang, C. Xu, W. Chen, L. Dai, and H. Wang, "GAMMA: Graspability-aware mobile manipulation policy learning based on online grasping pose fusion," *arXiv preprint* arXiv:2309.15459, 2024.

[59] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, "Learning agile and dynamic motor skills for legged robots," *Science Robotics*, vol. 4, no. 26, 2019.

[60] X. B. Peng, G. Berseth, K. Yin, and M. Van De Panne, "Deeploco: Dynamic locomotion skills using hierarchical deep reinforcement learning," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1-13, 2017.

[61] E. Chane-Sane, J. Amigo, T. Flayols, L. Righetti, and N. Mansard, "Soloparkour: Constrained reinforcement learning for visual locomotion from privileged experience," *Proc. of Conference on Robot Learning*, 2024.

[62] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, *et al.*, "Learning dexterous in-hand manipulation," *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3-20, 2020.

[63] J. Borja-Diaz, O. Mees, G. Kalweit, L. Hermann, J. Boedecker, and W. Burgard, "Affordance learning from play for sample-efficient policy learning," *Proc. of 2022 International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 6372-6378, 2022.

[64] Y. Geng, B. An, H. Geng, Y. Chen, Y. Yang, and H. Dong, "Rlafford: End-to-end affordance learning for robotic manipulation," *Proc. of 2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 5880-5886, 2023.

[65] L. Wang, J. Liu, H. Shao, W. Wang, R. Chen, Y. Liu, and S. L. Waslander, "Efficient reinforcement learning for autonomous driving with parameterized skills and priors," *arXiv preprint* arXiv:2305.04412, 2023.

[66] Z. Zhou, A. Garg, D. Fox, C. Garrett, and A. Mandlekar, "SPIRE: Synergistic planning, imitation, and reinforcement learning for long-horizon manipulation," *arXiv preprint* arXiv:2410.18065, 2024.

[67] B. Thananjeyan, A. Balakrishna, S. Nair, M. Luo, K. Srinivasan, M. Hwang, J. E. Gonzalez, J. Ibarz, C. Finn, and K. Goldberg, "Recovery RL: Safe reinforcement learning with learned recovery zones," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4915-4922, 2021.

[68] A. Wachi, W. Hashimoto, X. Shen, and K. Hashimoto, "Safe exploration in reinforcement learning: A generalized formulation and algorithms," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[69] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," *Proc. of International Conference on Machine Learning*, PMLR, pp. 22-31, 2017.

[70] J. Beck, R. Vuorio, E. Z. Liu, Z. Xiong, L. Zintgraf, C. Finn, and S. Whiteson, "A survey of meta-reinforcement learning," *arXiv preprint* arXiv:2301.08028, 2023.

[71] S. Narvekar, B. Peng, M. Leonetti, J. Sinapov, M. E. Taylor, and P. Stone, "Curriculum learning for reinforcement learning domains: A framework and survey," *Journal of Machine Learning Research*, vol. 21, no. 181, pp. 1-50, 2020.

[72] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," *Proc. of International Conference on Machine Learning*, PMLR, pp. 2778-2787, 2017.

[73] D. Abel, A. Barreto, B. Van Roy, D. Precup, H. P. van Hasselt, and S. Singh, "A definition of continual reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 50377-50407, 2023.

[74] J. Baek, S. Baek, and S. Han, "Efficient multitask reinforcement learning without performance loss," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 10, pp. 14739-14753, 2023.

[75] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[76] X. Gao, J. Si, Y. Wen, M. Li, and H. Huang, "Reinforcement learning control of robotic knee with human-in-the-loop by flexible policy iteration," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5873-5887, 2021.

[77] C. O. Retzlaff, S. Das, C. Wayllace, P. Mousavi, M. Afshari, T. Yang, A. Saranti, A. Angerschmid, M. E. Taylor, and A. Holzinger, "Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities," *Journal of Artificial Intelligence Research*, vol. 79, pp. 359-415, 2024.

[78] Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, "Eureka: Human-level reward design via coding large language models," *arXiv preprint* arXiv:2310.12931, 2023.

[79] T. Xie, S. Zhao, C. H. Wu, Y. Liu, Q. Luo, V. Zhong, Y. Yang, and T. Yu, "Text2Reward: Reward shaping with language models for reinforcement learning," *Proc. of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[80] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Advances in Neural Information Processing Systems*, vol. 1, 1988.

[81] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[82] X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Liu, *et al.*, "Vision-language foundation models as effective robot imitators," *arXiv preprint* arXiv:2311.01378, 2023.

[83] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "RT-1: Robotics transformer for real-world control at scale," *arXiv preprint* arXiv:2212.06817, 2022.

[84] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint* arXiv:2307.15818, 2023.

[85] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative Q-learning for offline reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1179-1191, 2020.

[86] Y. Wu, G. Tucker, and O. Nachum, "Behavior regularized offline reinforcement learning," *arXiv preprint* arXiv:1911.11361, 2019.

[87] C. Zhang, S. Kuppannagari, and P. Viktor, "BRAC+: Improved behavior regularized actor critic for offline reinforcement learning," *Proc. of Asian Conference on Machine Learning*, PMLR, pp. 204-219, 2021.

[88] Y. Gao, H. Xu, J. Lin, F. Yu, S. Levine, and T. Darrell, "Reinforcement learning from imperfect demonstrations," *arXiv preprint* arXiv:1802.05313, 2018.

[89] M. Wang, Y. Jin, and G. Montana, "Learning on one mode: Addressing multi-modality in offline reinforcement learning," *arXiv preprint* arXiv:2412.03258, 2024.

[90] T. Zheng, G. Zhang, X. Qu, M. Kuang, S. W. Huang, and Z. He, "MORE-3S: Multimodal-based offline reinforcement learning with shared semantic spaces," *arXiv preprint* arXiv:2402.12845, 2024.

[91] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15084-15097, 2021.

[92] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey, *et al.*, "Maximum entropy inverse reinforcement learning.," *Proc. of AAAI Conference on Artificial Intelligence*, vol. 8, pp. 1433-1438, 2008.

[93] J. Fu, K. Luo, and S. Levine, "Learning robust rewards with adversarial inverse reinforcement learning," *arXiv preprint* arXiv:1710.11248, 2017.

[94] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, *et al.*, "RoboTurk: A crowdsourcing platform for robotic skill learning through imitation," *Proc. of Conference on Robot Learning*, PMLR, pp. 879-893, 2018.

[95] J. W. Kim, T. Z. Zhao, S. Schmidgall, A. Deguet, M. Kobilarov, C. Finn, and A. Krieger, "Surgical robot transformer (SRT): Imitation learning for surgical tasks," *arXiv preprint* arXiv:2407.12998, 2024.

[96] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," *Proc. of 2018 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 4693-4700, 2018.

[97] A. Kendall, J. Hawke, D. Janz, P. Mazur, D. Reda, J.-M. Allen, V.-D. Lam, A. Bewley, and A. Shah, "Learning to drive in a day," *Proc. of 2019 International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 8248-8254, 2019.

[98] Y. Lu, J. Fu, G. Tucker, X. Pan, E. Bronstein, R. Roelofs, B. Sapp, B. White, A. Faust, S. Whiteson, *et al.*, "Imitation is not enough: Robustifying imitation with reinforcement learning for challenging driving scenarios," *Proc. of 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, pp. 7553-7560, 2023.

[99] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, K. Ghasemipour, C. Finn, and A. Wahid, "ALOHA unleashed: A simple recipe for robot dexterity," *arXiv preprint* arXiv:2410.13126, 2024.

[100] J. Aldaco, T. Armstrong, R. Baruch, J. Bingham, S. Chan, K. Draper, D. Dwibedi, C. Finn, P. Florence, S. Goodrich, *et al.*, "ALOHA2: An enhanced low-cost hardware for bimanual teleoperation," *arXiv preprint* arXiv:2405.02292, 2024.

[101] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile ALOHA: Learning bimanual mobile manipulation using low-cost whole-body teleoperation," *Proc. of 8th Annual Conference on Robot Learning*, 2024.

[102] J. Jang, J. Han, and J. Kim, "K-mixup: Data augmentation for offline reinforcement learning using mixup in a koopman invariant subspace," *Expert Systems with Applications*, vol. 225, 120136, 2023.

[103] N. E. Corrado, Y. Qu, J. U. Balis, A. Labiosa, and J. P. Hanna, "Guided data augmentation for offline reinforcement learning and imitation learning," *arXiv preprint* arXiv:2310.18247, 2023.

[104] T. Yu, A. Kumar, Y. Chebotar, K. Hausman, C. Finn, and S. Levine, "How to leverage unlabeled data in offline reinforcement learning," *Proc. of International Conference on Machine Learning*, PMLR, pp. 25611-25635, 2022.

[105] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint* arXiv:2004.05150, 2020.

[106] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, *et al.*, "Big bird: Transformers for longer sequences," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17283-17297, 2020.

[107] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transformers with linear attention," *Proc. of International Conference on Machine Learning*, PMLR pp. 5156-5165, 2020.

[108] S. Yang, B. Wang, Y. Shen, R. Panda, and Y. Kim, "Gated linear attention transformers with hardware-efficient training," *Proc. of Forty-first International Conference on Machine Learning*, 2024.

[109] M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter, "XLSTM: Extended long short-term memory," *arXiv preprint* arXiv:2405.04517, 2024.

[110] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *Proc. of International Conference on Learning Representations*, 2022.

[111] A. Gu, A. Gupta, K. Goel, and C. Ré, "On the parameterization and initialization of diagonal state space models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 35971-35983, 2022.

[112] J. T. Smith *et al.*, "Simplified state space layers for sequence modeling," *Proc. of International Conference on Learning Representations*, 2023.

[113] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint* arXiv:2312.00752, 2023.

[114] T. Dao and A. Gu, "Transformers are ssms: Generalized models and efficient algorithms through structured state space duality," *Proc. of Forty-first International Conference on Machine Learning*, 2024.

[115] M. W. Spong, S. Hutchinson, and M. Vidyasagar, *Robot Modeling and Control*, John Wiley & Sons, 2020.

[116] R. Waleffe, W. Byeon, D. Riach, B. Norick, V. Korthikanti, T. Dao, A. Gu, A. Hatamizadeh, S. Singh, D. Narayanan, *et al.*, "An empirical study of mamba-based language models," *arXiv preprint* arXiv:2406.07887, 2024.

[117] R. N. Parnichkun, S. Massaroli, A. Moro, J. T. Smith, R. Hasani, M. Lechner, Q. An, C. Ré, H. Asama, S. Ermon, *et al.*, "State-free inference of state-space models: The transfer function approach," *arXiv preprint* arXiv:2405.06147, 2024.

[118] M. Gwak, S. Moon, J. Ko, and P. Park, "Layer-adaptive state pruning for deep state space models," *Proc. of the Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[119] J. Liu, M. Liu, Z. Wang, L. Lee, K. Zhou, P. An, S. Yang, R. Zhang, Y. Guo, and S. Zhang, "Robomamba: Multimodal state space model for efficient robot reasoning and manipulation," *arXiv preprint* arXiv:2406.04339, 2024.

[120] J. Wang, X. Guan, Z. Sun, T. Shen, D. Huang, F. Liu, and H. Cui, "Omega: Efficient occlusion-aware navigation for air-ground robots in dynamic environments via state space model," *IEEE Robotics and Automation Letters*, 2024.

[121] S. M. Mustafa, Z. A. Usmani, O. Rizvi, A. B. Memon, and M. M. Movania, "Context aware mamba-based reinforcement learning for social robot navigation," *Proc. of 2024 12th International Conference on Control, Mechatronics and Automation (ICCMA)*, IEEE, pp. 154-159, 2024.

[122] T. Tsuji, "Mamba as a motion encoder for robotic imitation learning," *arXiv preprint* arXiv:2409.02636, 2024.

[123] X. Jia, Q. Wang, A. Donat, B. Xing, G. Li, H. Zhou, O. Celik, D. Blessing, R. Lioutikov, and G. Neumann, "Mail: Improving imitation learning with mamba," *arXiv preprint* arXiv:2406.08234, 2024.

[124] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," *Advances in neural information processing systems*, vol. 34, pp. 15084-15097, 2021.

[125] T. Ota, "Decision mamba: Reinforcement learning via sequence modeling with selective state spaces," *arXiv preprint* arXiv:2403.19925, 2024.

[126] S. Huang, J. Hu, Z. Yang, L. Yang, T. Luo, H. Chen, L. Sun, and B. Yang, "Decision mamba: Reinforcement learning via hybrid selective sequence modeling," *arXiv preprint* arXiv:2406.00079, 2024.

[127] Q. Lv, X. Deng, G. Chen, M. Y. Wang, and L. Nie, "Decision mamba: A multi-grained state space model with self-evolution regularization for offline RL," *arXiv preprint* arXiv:2406.05427, 2024.

[128] O. Lieber, B. Lenz, H. Bata, G. Cohen, J. Osin, I. Dalmedigos, E. Safahi, S. Meirom, Y. Belinkov, S. Shalev-Shwartz, *et al.*, "Jamba: A hybrid transformer-mamba language model," *arXiv preprint* arXiv:2403.19887, 2024.

[129] L. Ren, Y. Liu, Y. Lu, Y. Shen, C. Liang, and W. Chen, "Samba: Simple hybrid state space models for efficient unlimited context language modeling," *arXiv preprint* arXiv:2406.07522, 2024.

[130] X. Dong, Y. Fu, S. Diao, W. Byeon, Z. Chen, A. S. Mahabaleshwarkar, S.-Y. Liu, M. Van Keirsbilck, M.-H. Chen, Y. Suhara, *et al.*, "Hymba: A hybrid-head architecture for small language models," *arXiv preprint* arXiv:2411.13676, 2024.

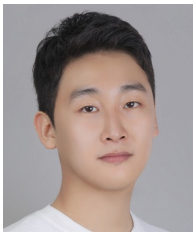
[131] Z. Wan, P. Zhang, Y. Wang, S. Yong, S. Stepputtis, K. Sycara, and Y. Xie, "Sigma: Siamese mamba network for multi-modal semantic segmentation," *arXiv preprint* arXiv:2404.04256, 2024.

[132] W. Li, H. Zhou, J. Yu, Z. Song, and W. Yang, "Coupled mamba: Enhanced multi-modal fusion with coupled state space model," *arXiv preprint* arXiv:2405.18014, 2024.

[133] Y. Wang, L. Cao, and H. Deng, "Mfmamba: A mamba-based multi-modal fusion network for semantic segmentation of remote sensing images," *Sensors*, vol. 24, no. 22, 7266, 2024.

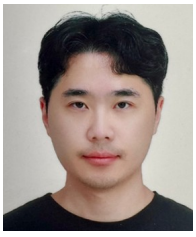
[134] H. Zhao, M. Zhang, W. Zhao, P. Ding, S. Huang, and D. Wang, "Cobra: Extending mamba to multi-modal large language model for efficient inference," *arXiv preprint* arXiv:2403.14520, 2024.

Wookyong Kwon received his B.S. degree in electronic and electrical engineering from Pohang University of Science and Technology (POSTECH), Pohang, Korea, in 2011. He received his M.S. and Ph.D. degrees from Graduate Institute of Ferrous Technology from POSTECH, in 2012 and 2017, respectively, where he was a Postdoctoral Researcher with the Department of Creative IT Engineering. He is currently a Senior Researcher with Electronics and Telecommunications Research Institute (ETRI).



Seungmin Baek received his B.S. degree in electrical engineering and a Ph.D. degree in convergence IT engineering from the Pohang University of Science and Technology (POSTECH), Pohang, Korea, in 2017 and 2022, respectively. His current research interests include controller design for robot manipulator, humanoid, and AI based control algorithm.

Jongchan Baek received his B.S. degree in mathematics from Sungkyunkwan University, Suwon, Korea, in 2014, and a Ph.D. degree in convergence IT engineering from Pohang University of Science and Technology (POSTECH), Pohang, Korea, in 2023. He is currently a researcher at Electronics and Telecommunications Research Institute (ETRI). His research interests include intelligent robot systems, machine learning, and reinforcement learning.



WooSang Shin received his B.S. degree in electronic engineering from Kyungpook National University, Korea, in 2009. He obtained both his M.S. and Ph.D. degrees from the same institution in 2020 and 2024, respectively. From 2018 to 2024, he worked as a student researcher at the Korea Institute of Industrial Technology (KITECH), focusing on AI-driven industrial inspection solutions, including smart inspection systems and anomaly detection. Following the completion of his Ph.D., he continued his research at KITECH as a postdoctoral researcher. Since 2025, he has been a Senior Researcher at Polaris3D, where his work centers on embodied AI. His research interests include artificial intelligence, computer vision, machine learning, generative models, and physical intelligence.

Minseon Gwak received her B.S. and M.S. degrees in electrical engineering from Pohang University of Science and Technology, Pohang, Korea, in 2019 and 2021, respectively. She is currently pursuing a Ph.D. degree in electrical engineering at Pohang University of Science and Technology. Her research interests include machine learning, signals and systems, deep state space models, and sequence modeling.

PooGyeon Park received his B.S. and M.S. degrees in control and instrumentation engineering from Seoul National University, Seoul, Korea, in 1988 and 1990, respectively, and a Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1995. Since 1996, he has been with the Division of Electrical Engineering, Pohang University of Science and Technology, Pohang, Korea, where he is currently a Professor. His research interests include the robust, LPV, and network-related control theories, delayed systems, fuzzy systems, and signal processing.

Sangmoon Lee received his B.S. degree in electronic engineering from Kyungpook National University, Daegu, Korea, in 1999, his M.S. and Ph.D. degrees in electronic engineering from Pohang University of Science and Technology (POSTECH), Pohang, Korea, in 2001 and 2006, respectively. Currently, he is a Professor at the School of Electronic and Electrical Engineering, Kyungpook National University. His main research interests include cyber-physical systems, networked control systems, fuzzy systems, reinforcement learning, model predictive control, and industrial applications.