Embodied AI: From LLMs to World Models

Tongtong Feng, Xin Wang*, *Member, IEEE,* Yu-Gang Jiang, *Fellow, IEEE,* Wenwu Zhu*, *Fellow, IEEE*

***Abstract*—Embodied Artificial Intelligence (AI) is an intelligent system paradigm for achieving Artificial General Intelligence (AGI), serving as the cornerstone for various applications and driving the evolution from cyberspace to physical systems. Recent breakthroughs in Large Language Models (LLMs) and World Models (WMs) have drawn significant attention for embodied AI. On the one hand, LLMs empower embodied AI via semantic reasoning and task decomposition, bringing high-level natural language instructions and low-level natural language actions into embodied cognition. On the other hand, WMs empower embodied AI by building internal representations and future predictions of the external world, facilitating physical law-compliant embodied interactions. As such, this paper comprehensively explores the literature in embodied AI from basics to advances, covering both LLM driven and WM driven works. In particular, we first present the history, key technologies, key components, and hardware systems of embodied AI, as well as discuss its development via looking from unimodal to multimodal angle. We then scrutinize the two burgeoning fields of embodied AI, i.e., embodied AI with LLMs/multimodal LLMs (MLLMs) and embodied AI with WMs, meticulously delineating their indispensable roles in end-to-end embodied cognition and physical laws-driven embodied interactions. Building upon the above advances, we further share our insights on the necessity of the joint MLLM-WM driven embodied AI architecture, shedding light on its profound significance in enabling complex tasks within physical worlds. In addition, we examine representative applications of embodied AI, demonstrating its wide applicability in real-world scenarios. Last but not least, we point out future research directions of embodied AI that deserve further investigation.**

***Index Terms*—Embodied AI, LLMs, World Models**

I. Introduction

Embodied Artificial Intelligence (AI) originated from the Embodied Turing Test by Alan Turing in 1950 [1], which is designed to explore whether agents can imitate human intelligence to achieve Artificial General Intelligence (AGI). Among them, agents that only solve abstract problems in digital world (cyberspace) are generally defined as disembodied AI, while those that also can interact with the physical world are regarded as embodied AI. Embodied AI builds on foundational insights from cognitive science and neuroscience [2], [3], which claims that intelligence emerges from the dynamic coupling of perception, cognition, and interaction. As shown in Fig. 1, embodied AI includes three key components in a closed-loop manner, i.e., 1) active perception (sensor-driven environmental observation), 2) embodied cognition (historical experience-driven cognition updating), and 3) dynamic interaction (actuator-mediated action control). Besides, hardware embodiment [4]–[6] is also critical due to escalating computational and energy demands, particularly under latency and power constraints of devices in real-world deployment scenarios.

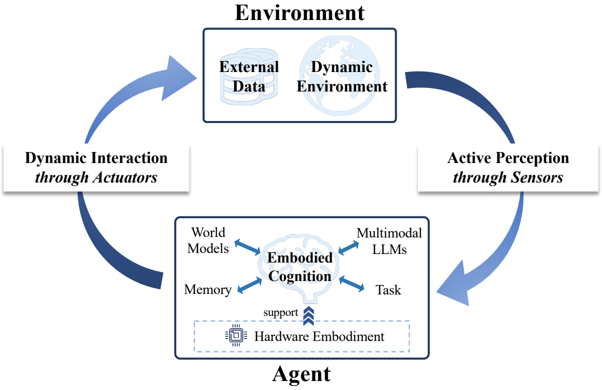

Fig. 1. The concept of embodied AI.

The development of embodied AI has evolved from unimodal to multimodal paradigm. In early stage, embodied AI is primarily studied through focusing on individual components with single modality such as vision, language, or action, where the perception, cognition, or interaction component is driven by one sensory input [7], [8], e.g., perception tends to be dominated by the visual modality [9], cognition tends to be dominated by the language modality [10], [11], and interaction tends to be dominated by the action modality [12], [13]. Although these methods perform well within individual components, they are limited by the narrow scope of information provided by each modality and the inherent gaps between modalities across components. The continued development of embodied AI witnesses the limitations of unimodal approaches, promoting a significant shift toward integration of multiple sensory modalities [14]–[16]. As such, multimodal embodied AI [17], [18] naturally arises to create more adaptive, flexible, and robust agents capable of performing complex tasks in dynamic environments.

Large Language Models (LLMs) empower embodied AI via semantic reasoning [19] and task decomposition [20], [21], bringing high-level natural language instructions and low-level natural language actions into embodied cognition. Representative LLM driven works include SayCan [22], which i) provides a real-world pretrained natural language action library to constrain LLMs from proposing infeasible and contextually inappropriate actions; ii) uses LLMs to convert natural language instructions into natural language action sequences; and iii) utilizes value functions to verify the feasibility of natural

The invited paper

*Corresponding author

Tongtong Feng, Xin Wang and Wenwu Zhu are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: fengtongtong, xin_wang, wwzhu@tsinghua.edu.cn). Xin Wang and Wenwu Zhu are also with Beijing National Research Center for Information Science and Technology. Yu-Gang Jiang is with the Institute of Trustworthy Embodied AI, Fudan University, Shanghai 200433, China. (e-mail: ygj@fudan.edu.cn).

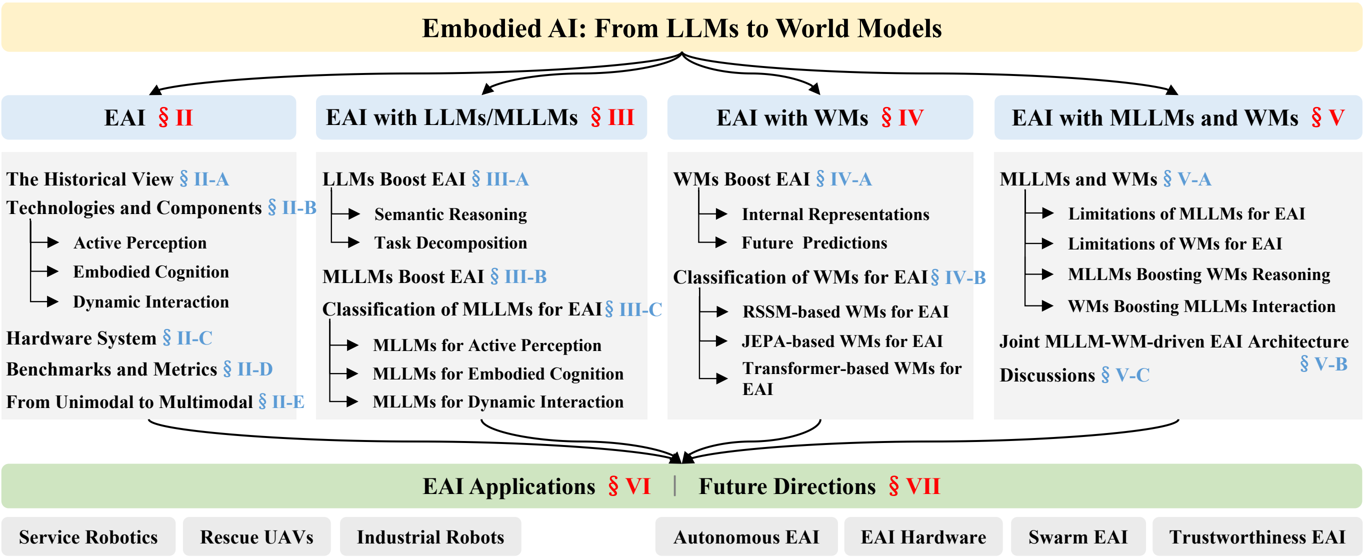
Embodied AI: From LLMs to World Models
EAI § II
EAI with LLMs/MLLMs § III
EAI with WMs § IV
EAI with MLLMs and WMs § V
The Historical View § II-A
Technologies and Components § II-B
Active Perception
Embodied Cognition
Dynamic Interaction
Hardware System § II-C
Benchmarks and Metrics § II-D
From Unimodal to Multimodal § II-E
LLMs Boost EAI § III-A
Semantic Reasoning
Task Decomposition
MLLMs Boost EAI § III-B
Classification of MLLMs for EAI § III-C
MLLMs for Active Perception
MLLMs for Embodied Cognition
MLLMs for Dynamic Interaction
WMs Boost EAI § IV-A
Internal Representations
Future Predictions
Classification of WMs for EAI § IV-B
RSSM-based WMs for EAI
JEPA-based WMs for EAI
Transformer-based WMs for EAI
MLLMs and WMs § V-A
Limitations of MLLMs for EAI
Limitations of WMs for EAI
MLLMs Boosting WMs Reasoning
WMs Boosting MLLMs Interaction
Joint MLLM-WM-driven EAI Architecture § V-B
Discussions § V-C
EAI Applications § VI
Future Directions § VII
Service Robotics
Rescue UAVs
Industrial Robots
Autonomous EAI
EAI Hardware
Swarm EAI
Trustworthiness EAI

Fig. 2. This paper comprehensively introduces the basics of Embodied AI (EAI) and the latest advancements of EAI with LLMs/MLLMs and WMs. MLLMs enable contextual task reasoning but overlook physical constraints, while WMs excel at physics-aware simulation but lack high-level semantics. Building upon the above advances, this paper proposes a joint MLLM-WM-driven EAI architecture. Finally, this paper discusss applications and future directions of EAI.

language action sequences in a particular physical environment. These works suggest that LLMs are extremely useful to robots which aim at acting upon high-level, temporally extended instructions expressed in natural language. However, LLMs are only a part of the entire embodied AI system (e.g., embodied cognition), which is limited by a fixed natural language action library and a specific physical environment, making it difficult for LLM driven embodied AI to achieve adaptive expansion for new robots and environments.

Recent breakthroughs in Multimodal LLMs (MLLMs) [23], [24] and World Models (WMs) [25]–[27] have opened up a new frontier in embodied AI research. MLLMs can act on the entire embodied AI system, bridging high-level multimodal inputting and low-level motor action sequences into end-to-end embodied applications. Semantic reasoning [28]–[30] leverages MLLMs' cross-modal comprehension to interpret semantics from visual, auditory, or tactile inputs, e.g., identifying objects, inferring spatial relationships, predicting environmental dynamics. Concurrently, task decomposition [31]–[33] employs MLLMs' sequential logic to break complex objectives into sub-tasks while dynamically adapting plans based on sensor feedback. However, MLLMs often fail to ground predictions in physics-compliant dynamics [34] and exhibit poor real-time adaptation [35] to environmental feedback.

On the other hand, WMs empower embodied AI by building internal representations [36]–[40] and making future predictions [41]–[44] of the external world. Such WM driven embodied AI is able to facilitate physical law-compliant embodied interactions in dynamic environments. Internal representations compress rich sensory inputs into structured latent spaces, capturing object dynamics, physics laws, and spatial structures, as well as allowing agents to reason about "what exists" and "how things behave" in their surroundings. Simultaneously, future predictions simulate potential rewards of sequence actions across multiple time horizons aligned with physical laws, thereby preempting risky or inefficient behaviors. However, WM driven approaches struggle with open-ended semantic reasoning [45] and lack the ability of generalizable task decomposition [26] without explicit priors.

Building upon the above advances, we further share our insights on the necessity of developing a joint MLLM-WM driven embodied AI architecture, shedding light on its profound significance in enabling complex tasks within physical worlds. MLLMs enable contextual task reasoning but overlook physical constraints, while WMs excel at physics-aware simulation but lack high-level semantics. The joint of MLLM and WM can bridge semantic intelligence with grounded physical interaction. For instance, EvoAgent [46] designs an autonomous-evolving agent with a joint MLLM-WM driven embodied AI architecture, which can autonomously complete various long-horizon tasks across environments through self-planning, self-reflection, and self-control, without human intervention. We believe that designing joint MLLM-WM driven embodied AI architectures will dominate next-generation embodied systems, bridging the gap between specialized AI agents and general physical intelligence.

We summarize the representative applications of embodied AI as service robotics, rescue UAVs, industrial Robots, and others etc., demonstrating its wide applicability in real-world scenarios. We also point out potential future directions of embodied AI, including but not limited to autonomous embodied AI, embodied AI hardware, and swarm embodied AI etc.

As shown in Fig. 2, the rest of this paper is organized as follows. Section II introduces the history, key technologies, key components, and hardware system of embodied AI, discussing the development of embodied AI from unimodal to multimodal angle. Section III presents embodied AI with LLMs/MLLMs, and Section IV presents embodied AI with WMs. Section V introduces our insights on designing a joint MLLM-WM driven embodied AI architecture. Section VI briefly examines applications of embodied AI. Potential future directions are discussed in Section VII.

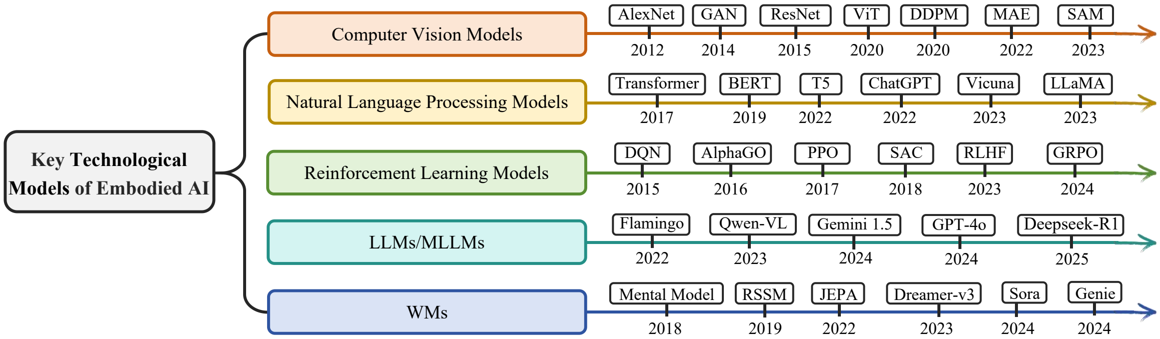

Fig. 3. Key technological models of embodied AI. Advancements in Computer Vision (CV) models, Natural Language Processing (NLP) models, Reinforcement Learning (RL) models, LLMs/MLLMs, and WMs have driven progress in embodied AI.

II. Embodied AI

This section provides a comprehensive overview of embodied AI. We first take a historical view to introduce the development of embodied AI in Subsection II-A. Based on technological advancements in five foundational areas related to embodied AI, Subsection II-B and Subsection II-C further review the developmental trajectories of core modules in software algorithms and hardware design, respectively. Finally, Subsection II-E discusses an overall analysis of the developmental trends from unimodal to multimodal.

A. The Historical View

The historical evolution of embodied AI reflects successive transitions from early philosophical foundations to technological breakthroughs in robotics and the rise of learning-driven paradigms, while recent progress in LLMs and WMs is driving an ongoing shift toward the next phase of development.

The theoretical roots of embodied AI trace to 1950, when Turing introduced the foundational idea that intelligence is inherently linked to physical experience [47]. In the 1980s, cognitive science further formalized this view. Lakoff and Johnson emphasized that human cognition arises from bodily experience rather than disembodied symbolic computation [48], while Harnad's symbol grounding problem highlighted the necessity of connecting symbolic representations to sensorymotor reality [49]. Technological advances in robotics during the late 1980s and 1990s brought these ideas into practice. Brooks proposed the subsumption architecture [50], [51], promoting behavior-based control through layered, reactive modules grounded in sensorimotor loops. The Cog project [52] advanced this line by constructing humanoid robots capable of developmental learning, imitation, and social interaction. Recently, the success of the learning-driven paradigm has driven the shift in embodied AI from motion control of robots to adaptive interaction [53]. In particular, the development of deep learning enables robots to learn complex nonlinear mappings from raw sensor data to action policy, significantly improving navigation and manipulation tasks [54], [55].

While embodied AI has made notable advances, achieving self-reflection intelligence in dynamic, uncertain environments remains a key challenge. Recent progress in LLMs/MLLMs [23], [24] and WMs [25]–[27] have progressively shown promise in overcoming these challenges.

B. The Key Technologies and Components

Before discussing the ongoing changes, we systematically review the development of key technologies and components.

1) Key Technologies of Embodied AI: The rapid development of embodied AI is closely tied to advances in foundational technological models such as Computer Vision (CV) models, Natural Language Processing (NLP) models, Reinforcement Learning (RL) models, LLMs/MLLMs, and WMs (as shown in Fig. 3), which can significantly enhance the capabilities of agents in perception, cognition and interaction.

Specifically, Classic models in computer vision, such as AlexNet [56], GAN [57], ResNet [58], ViT [59], DDPM [60], MAE [61], and SAM [62] provide the perceptual foundation for embodied agents to interpret high-dimensional sensory inputs in complex environments. In the field of NLP, the evolution from foundational architectures like Transformer [63], BERT [64], and T5 [65] to large-scale systems such as ChatGPT [66], Vicuna [67], and LLaMA [68], has equipped embodied agents with stronger capabilities in language understanding, task planning, and instruction following. RL offers the core algorithmic framework for agents to learn through interaction with their environments. Representative approaches include DQN [69], AlphaGo [70], PPO [71], SAC [72], RLHF [73], and GRPO [74].

Beyond these classical fields, one of the most promising directions in embodied AI lies in the integration of LLMs/MLLMs with WMs. LLMs and MLLMs (like Flamingo [20], Qwen-VL [75], Gemini-1.5 [76], GPT-4o [77], and Deepseek-R1 [78]) provide agents with the ability to understand instructions, reason over multimodal inputs, and generalize across tasks and environments. In contrast, WMs (like Mental Model [26], RSSM [79], JEPA [27], Dreamer-v3 [80], Sora [81], and Genie [36]) enable agents to model and predict environmental dynamics, supporting imagination-based planning and anticipatory decision-making in dynamic and uncertain environments.

2) Key Components of Embodied AI: Driven by advances in these key technologies, embodied AI has experienced rapid progress. In the following, we present a structured overview of developments in three key components.

a) Active Perception: Active perception refers to the agent selectively acquiring information from environmental observations [16], [82], [83]. Existing active perception meth-

TABLE I
COMPARISON OF THREE CATEGORIES OF ACTIVE PERCEPTION METHODS INCLUDING VISUAL SLAM, 3D SCENE UNDERSTANDING, AND ACTIVE ENVIRONMENT EXPLORATION.

Category	Method	Year	Sensor Type	Feature Type	Applicable Scenarios
Visual SLAM	CoSLAM [90]	2012	RGB-D	Geometric + Volumetric	Dynamic SLAM
	SLAM++ [93]	2013	RGB-D	Semantic	Object-level Mapping
	ORB-SLAM [8]	2015	RGB-D + Stereo	Geometric	Dynamic SLAM
	DS-SLAM [94]	2018	RGB-D	Geometric + Semantic	Dynamic SLAM
	TwistSLAM [95]	2022	RGB-D + Stereo	Geometric + Semantic	Dynamic SLAM
	GS-SLAM [96]	2024	RGB-D	Volumetric	Object-level Mapping
3D Scene Understanding	Gaudi [97]	2022	RGB	Volumetric	General Scene Understanding
	Clip2Scene [98]	2023	RGB + Point Cloud	Multimodal	Language-guided Scene Understanding
	OpenScene [99]	2023	RGB + Point Cloud	Multimodal	General Scene Understanding
	Lexicon3D [100]	2024	RGB-D	Semantic	Language-guided Scene Understanding
	GraphDreamer [101]	2024	RGB	Topological + Semantic	Structured Scene Reasoning
	HUGS [102]	2024	RGB-D	Multimodal	General Scene Understanding
	RegionPLC [103]	2024	RGB + Point Cloud	Multimodal	Language-guided Scene Understanding
Active Environment Exploration	MAX [104]	2019	RGB	Semantic	Semantic-guided Exploration
	Active Neural SLAM [105]	2020	RGB-D	Volumetric	Geometry-based Exploration
	APT [106]	2021	RGB	Semantic	Semantic-guided Exploration
	Conan [107]	2023	RGB	Topological	Geometry-based Exploration
	DBMF-BPI [108]	2023	RGB-D	Volumetric	Geometry-based Exploration
	ActiveRIR [109]	2024	RGB + Audio	Multimodal	Cross-modal Active Perception

ods can be roughly divided into three categories: visual SLAM, 3D scene understanding, and active environment exploration. To offer an effective perspective on active perception approaches, as summarized in Table I, we analyze representative methods along three practical dimensions: sensor type, feature type, and applicable scenarios.

Visual SLAM. Simultaneous Localization and Mapping (SLAM) is a pivotal technology enabling agents to both localize themselves and construct environmental maps in unknown environments [9], [84]. As a foundational technology of active perception, visual SLAM has been extensively studied [85], [86]. According to Wang et al. [87], existing methods fall into geometric-based and semantic-based categories. Geometric methods exploit spatial or temporal cues [8], such as dense scene flow [88], [89], triangulation consistency [90], and graph structure [91], [92], performing well in static settings but struggling with dynamic scenes. In contrast, semantic methods improve localization and mapping in dynamic environments by leveraging high-level information. Representative early methods include SLAM++ [93], integrating object-level semantics, and DS-SLAM [94], applying deep learning to dynamic scene understanding. Recent models such as TwistSLAM [95] and GS-SLAM [96] further enhance robustness by combining geometric optimization with semantic or generative modeling.

3D Scene Understanding. Scene understanding focuses on enabling agents to perceive, segment, and reason about complex environments in a structured and semantically meaningful way. Recent works have advanced this field by integrating vision-language models and generative priors. Early efforts like Gaudi [97] introduced generative models for 3D-aware scene synthesis. Clip2Scene [98] and OpenScene [99] leveraged vision-language embeddings to facilitate label-efficient and open-vocabulary 3D understanding. Structured scene understanding is further enhanced by Lexicon3D [100] and GraphDreamer [101], which model object-level relations in 3D space through structured representations such as scene graphs or semantic lexicons. Meanwhile, region-level multimodal grounding techniques, exemplified by HUGS [102] and RegionPLC [103], incorporate prompts and spatial grounding to achieve fine-grained, goal-conditioned 3D perception. These methods advance holistic, language-aligned 3D understanding.

Active Environment Exploration. Active exploration focuses on enabling agents to autonomously acquire informative observations through interaction with the environment. Early approaches relied on building explicit or implicit environmental models. Representative model-based methods include MAX [104] and Active Neural SLAM [105], which leverage predictive modeling and mapping to support efficient navigation in unseen spaces. In contrast, APT [106] and DBMF-BPI [108] focus on model-free exploration through direct environmental interaction to reduce reliance on explicit modeling. Recent efforts further enhance exploration capabilities by incorporating multimodal perception [109] and semantic reasoning [107].

b) Embodied Cognition: Embodied cognition refers to the emergence of internal representations and reasoning capabilities during the interaction, driven by the agent's self-reflection on its perception and accumulated experience [147]–[149]. This component forms the core of embodied AI, en-

TABLE II
COMPARISON OF THREE CATEGORIES OF EMBODIED COGNITION METHODS: TASK-DRIVEN SELF-PLANNING, MEMORY-DRIVEN SELF-REFLECTION, AND EMBODIED MULTIMODAL FOUNDATION MODELS. I, L AND P INDICATE THE IMAGE, LANGUAGE AND POINT CLOUD MODALITIES, RESPECTIVELY.

Category	Method	Year	Input Modalities	Cognition Type	Reasoning Mode	Output
Task-driven Self-planning	L3P [110]	2021	I + L	Planner	Neural + Symbolic	Action
	LLM-Planner [111]	2023	I + L	Planner	Neural + Symbolic	Action
	Egoplaner [112]	2023	I	Planner	Symbolic	Action
	AutoAct [113]	2024	L	Planner	Neural	Action
	RPG [114]	2024	I + L	Planner	Neural	Policy
	ETPNav [115]	2024	I + L	Planner	Neural + Symbolic	Policy
Memory-driven Self-reflection	Reflexion [116]	2023	L	Memory	Beam + Replay	Policy
	Reflect [117]	2023	I + L	Memory	Neural + Symbolic	Policy
	RILA [118]	2024	L	Memory	Neural	Policy
	Optimus-1 [119]	2024	I + L	Memory	Neural	Policy
	EvoAgent [46]	2025	I + L	Memory	Neural	Policy
	REMAC [120]	2025	L	Memory	Neural + Symbolic	Policy
Embodied Multimodal Foundation Models	SayCan [121]	2022	I + L	Planner + Aligner	Neural	Answer + Action
	GATO [122]	2022	I + L + P	Aligner	Neural	Action
	EmbodiedGPT [123]	2023	I + L	Aligner	Neural	Answer + Action
	Kosmos-2 [124]	2023	I + L	Aligner	Neural	Answer
	MultiPLY [125]	2024	I + L	Aligner	Neural	Answer
	ManipLLM [28]	2024	I + L	Aligner	Neural	Answer + Action

abling agents to perform task planning [150], causal inference [151], and long-horizon reasoning [152], [153]. Recent studies of embodied cognition primarily focus on three aspects: task-driven self-planning, memory-driven self-reflection, and embodied multimodal foundation models. Table II presents representative methods analyzed from four perspectives: input modalities, cognition type, reasoning mode, and output type. These dimensions reflect how embodied agents perceive information, form internal models and conduct reasoning.

Task-driven Self-Planning. In task-driven self-planning, agents autonomously generate structured plans based on task goals, environmental context, and internal knowledge, without explicit human instructions [154]–[156]. Structured learning is a classical solution that develops latent planning spaces or direct policy mappings, achieving high efficiency within training distributions but lacking robustness to out-of-distribution scenarios. Representative approaches include L3P [110], Egoplaner [112], and ETPNav [115]. Recent advances incorporate LLMs or generative models into self-planning. LLM-Planner [111] and AutoAct [113] integrate LLMs into planning by grounding language-guided reasoning into various tasks, while RPG [114] offers a generative perspective, aiming to unify planning and content creation through multimodal reasoning.

Memory-driven Self-Reflection. Memory-driven self-reflection enables agents to leverage past experiences for long-horizon reasoning, error correction, and self-improvement [46], [157]. Early studies focus on memory processing, including fixed-size replay buffers [158]–[160] and differentiable memory architectures [161], [162]. Recent advances introduce reflective mechanisms, where agents summarize or verbalize past experiences to guide future decisions. Reflexion [116] and Reflect [117] enable agents to iteratively self-correct by integrating verbalized feedback into action planning, while RILA [118] extends reflective reasoning to multimodal semantic navigation. Beyond individual reflection, Optimus-1 [119] and REMAC [120] integrate multimodal or multi-agent memory to support long-horizon collaboration. EvoAgent [46] further advances this direction by coupling continual world modeling with a memory-driven planner, enabling fully autonomous evolution across sequential tasks.

Embodied Multimodal Foundation Models. In the era of MLLMs, embodied multimodal foundation models [163]–[165] have emerged as one of the most promising solutions for unifying planning, reasoning, and other embodied cognitive capabilities. Recent progress is driven by both data construction and model development. Data efforts focus on constructing high-quality benchmarks to support scalable and cognitively meaningful evaluation, such as MuEP [166], ECBench [167], MFE-ETP [168], and EmbodiedBench [18]. On the model side, recent advances include affordance-grounded agents (e.g., SayCan [121] and GATO [122]) that align language understanding with embodied action spaces, vision-language pretraining approaches (like EmbodiedGPT [123] and Kosmos-2 [124]) that promote scalable embodied reasoning, and object-centric designs (such as MultiPLY [125] and ManipLLM [28]) that enhance manipulation and interaction capabilities. These models collectively aim to build transferable and generalizable embodied AI.

TABLE III
COMPARISON OF THREE CATEGORIES OF DYNAMIC INTERACTION METHODS INCLUDING ACTION CONTROL, BEHAVIORAL INTERACTION, AND COLLABORATIVE DECISION-MAKING, ACROSS INPUT MODALITIES, INTERACTION TYPE, MODELING PARADIGM, AND TASK TYPE. I, L, S, P, AND T DENOTE IMAGE, LANGUAGE, STATE, PROPRIOCEPTION, AND TRAJECTORY, RESPECTIVELY. IL DENOTES IMITATION LEARNING.

Category	Method	Year	Input Modalities	Interaction Type	Learning Paradigm	Task Type
Action Control	MineDojo [126]	2022	I + L	High-level Planning	LLM	Instruction Following
	PaLM-E [14]	2023	I + L + P	Low-level Control	MLLM	Embodied Manipulation
	RT-2 [24]	2023	I + L	Low-level Control	VLA	Embodied Manipulation
	OpenVLA [127]	2024	I + L	Low-level Control	VLA	Embodied Manipulation
	Cogagent [128]	2024	I + L	Low-level Control	MLLM	Instruction Following
	Octo [129]	2024	I + L + P	Low-level Control	VLA	Embodied Manipulation
	CrossFormer [130]	2024	I + L	Low-level Control	VLA	Embodied Manipulation
	HPT [131]	2024	I + L	Low-level Control	VLA	Embodied Manipulation
Behavioral Interaction	GAIL [132]	2016	T	Behavioral	IL	Trajectory Learning
	MGAIL [133]	2017	T	Behavioral	IL	Trajectory Learning
	TrafficSim [134]	2021	T	Behavioral	RL	Trajectory Learning
	TrajGen [135]	2022	I + T	Behavioral	RL	Trajectory Learning
	Behavior-1K [136]	2023	I	Trajectory	IL	Behavior Understanding
	AgentLens [137]	2024	I + S	Trajectory	IL	Behavior Understanding
	ECL [138]	2024	I + L	High-level Planning	IL	Embodied Manipulation
Collaborative Decision	QMIX [139]	2018	S	Behavioral	RL	Cooperative Decision
	Qtran [140]	2019	S	Behavioral	RL	Cooperative Decision
	QPLEX [141]	2019	S	Behavioral	RL	Cooperative Decision
	MAT [142]	2022	S	Behavioral	RL	Cooperative Decision
	CoELA [143]	2024	I + L	Low-level Control	LLM	Cooperative Manipulation
	AgentVerse [144]	2024	L	High-level Planning	LLM	Agent Society Simulation
	MetaGPT [145]	2024	L	High-level Planning	LLM	Agent Society Simulation
	Combo [146]	2024	L	High-level Planning	LLM	Cooperative Planning

c) Dynamic Interaction: Dynamic interaction refers to the process in which an agent influences the environment through actions or behaviors grounded in its perception and cognition [169], [170]. Existing research highlights the significance of this capability in enabling agents not only to respond but also to change their surroundings [171], [172]. Studies on dynamic interaction encompass action control, behavioral interaction, and collaborative decision-making. To better understand existing methods, we analyze representative approaches from four perspectives, including input modalities, interaction type, learning paradigm, and task type, as shown in Table III. These dimensions reflect how agents sense the environment, determine the level and structure of interaction, and generate appropriate behaviors in dynamic multi-agent or human-in-the-loop scenarios.

Action Control. Action control generates motor commands for embodied interaction. Early methods were based on control theory with dynamic system modeling [173], [174] or RL via trial and error [175], [176]. The former is effective for structured or repetitive tasks, while the latter is adaptable to high-dimensional, nonlinear problems. Recent advances mainly follow three directions. Vision-language-action (VLA) models, such as PaLM-E [14], RT-2 [24], OpenVLA [127], and CogAgent [128], integrate language-guided reasoning for flexible control and have been comprehensively reviewed by Ma et al. [177]. Open-ended frameworks like MineDojo [126] promote continual skill acquisition from open-world knowledge. In addition, Cross-embodiment learning, including CrossFormer [130], HPT [131], and Octo [129], aim to unify policy learning across diverse robots and modalities.

Behavioral Interaction. The behavior of an agent is composed of a sequence of actions. Compared to action control, it emphasizes high-level control through meaningful action patterns, enabling agents to interact in a flexible and goal-directed manner. Recent advances mainly fall into two directions. Imitation learning, including GAIL [132], MGAIL [133], TrafficSim [134], and TrajGen [135], enables efficient acquisition and simulation of complex behaviors. BEHAVIOR-1K [136] provides a large-scale benchmark for evaluating behavior generalization across 1,000 embodied tasks. Behavior-aware enhancement methods, such as AgentLens [137] and ECL [138], improve policy robustness and interpretability. Despite these advances, achieving reliable long-horizon behavioral interaction under sparse feedback remains challenging.

Collaborative Decision. Collaborative decision focuses on coordinating multiple agents to achieve shared goals, which is essential for multi-agent systems and human-robot collaboration [178]–[180]. Multi-agent RL is a classical solution, with methods like QTRAN [140], QPLEX [141], and Qatten [139] addressing cooperation via centralized training with decentralized execution. MAT [142] reframes MARL as a sequence modeling problem to mitigate scalability limitations in multi-agent RL. Recent advances integrate LLMs and WMs to enhance multi-agent collaboration. MetaGPT [145], CoELA [143], and AgentVerse [144] leverage LLMs for task reasoning and coordination, while COMBO [146] composes modular WMs to support scalable collaborative embodied decision.

C. Hardware

As embodied AI evolves, model complexity and size have grown, increasing computational and energy demands. Embodied systems, often operating in dynamic, real-world environments, face strict latency and power constraints—especially at the edge. Thus, developing hardware-friendly directions that maintain performance while optimizing efficiency is crucial for enabling responsive, energy-aware embodied agents. Hardware optimization in embodied AI typically includes four components: hardware-aware model compression, compiler-level optimization, domain-specific accelerators, and hardware-software co-design.

1) Hardware-aware Model Compression: Quantization and pruning [4] are key techniques for reducing model size and computational cost. In embodied agents, which frequently run on low-power embedded hardware, such techniques are vital for enabling fast and efficient inference. Quantization [181] maps weights and activations to lower bit-widths, while pruning [182] removes redundant parameters. To support real-world embodied tasks, such as robotic control or visual navigation, hardware efficiency metrics like power, performance, and area (PPA) can guide bit-width allocation or pruning ratios [183], enabling task-specific trade-offs between accuracy and deployability on physical platforms.

2) Compiler-level Optimization: Compilers bridge high-level embodied AI models and hardware execution. In real-time embodied systems, compiler toolchains are essential for efficient processing of sensor data and decision-making. TVM [5], built on LLVM [184] and CUDA, generates optimized code across platforms. These compilers transform computational graphs through operator fusion and redundant computation elimination [185], enabling responsive behavior. Mapping strategies like loop reordering and tiling enhance data locality, parallelism, and memory access [186], all of which are critical to maintaining low-latency inference in embodied agents.

3) Domain-specific Accelerators: With growing computational demands, domain-specific accelerators (DSAs) are a promising solution for embodied AI. These systems, from robots to AR/VR agents, benefit from fast, energy-efficient hardware tailored for frequent operations. Google's TPU [6], typically integrated with CPUs and GPUs via PCIe, accelerates key operations like matrix multiplication. FPGA-based accelerators [187] allow reconfigurability for adapting to new tasks or changing workloads; CGRA accelerators [188] improve structured, dataflow-heavy computations common in perception or control. Meanwhile, ASIC-based accelerators [189] offer high throughput and energy efficiency, ideal for deploying high-performance embodied models in real-world environments.

4) Hardware-software Co-design: Separating algorithm and hardware design can lower runtime efficiency. Hardware-software co-design addresses this through algorithm-system and algorithm-hardware co-optimization. Algorithm-system co-optimization focuses on how to take full advantage of GPU resources like tensor cores and CUDA cores to better support the algorithm [190]. Algorithm-hardware co-optimization aims to improve deployment efficiency by tuning both the model and the hardware architecture. For example, we can perform multi-objective optimization based on the types of operators in the network and the configuration parameters of the hardware [191]. We can also design different numerical quantization schemes along with matching hardware accelerators to better support embodied AI tasks [192].

D. Benchmarks and Evaluation Metrics

Standardized benchmarks and evaluation metrics are crucial for objectively assessing the performance of embodied AI systems. Widely adopted testbeds include Habitat [193], which provides photorealistic 3D indoor environments for navigation and interaction tasks, and ManiSkill [194], offering physics-based manipulation scenarios with diverse object sets. Simulation platforms like MuJoCo [195] enable precise control evaluation in continuous state-spaces, while EmbodiedBench [18] supports holistic evaluation of vision-driven agents across perception, cognition, and interaction. For UAV applications, AirSim [196], U2UData [197] and U2USim [198] provides high-fidelity aerial environments with dynamic obstacles. These testbeds vary in complexity: Habitat excels in visual realism, ManiSkill in object diversity, MuJoCo in physical accuracy, and EmbodiedBench in multimodal integration. Domain-specific benchmarks like BEHAVIOR-1K [136] further enable granular evaluation of 1,000 everyday activities under realistic constraints.

Key evaluation metrics span three critical dimensions: Task Success Rate measures completion accuracy of goal-oriented objectives (e.g., object manipulation or navigation) [24]; Real-time Responsiveness quantifies decision latency and adaptation speed to environmental changes [199]; and Energy Efficiency evaluates computational cost (FLOPS) and power consumption (Watts) during deployment [4]. Additional metrics include Path Length for navigation efficiency [105], Generalization Score for unseen scenarios [200], and Safety Violations for physical compliance [171]. For multi-agent systems, Coordination Efficiency [178] and Communication Overhead [201] provide critical insights. Standardized evaluation protocols like those in MFE-ETP [168] ensure fair cross-modal comparisons, though challenges remain in sim-to-real transfer validation [177].

E. From Unimodal to Multimodal

The development of embodied AI has evolved from unimodal to multimodal systems, as shown in Fig 4. Initially,

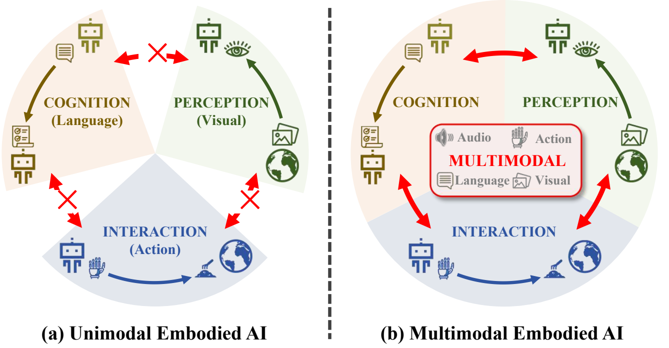

Fig. 4. Unimodal embodied AI and multimodal embodied AI. (a) Unimodal methods focus on specific modules of embodied AI. They are limited by the narrow scope of information provided by each modality and the inherent gaps between modalities across modules. (b) Multimodal embodied AI methods break these limitations and enable the mutual enhancement of the modules.

embodied AI was primarily concerned with individual modalities, such as vision, language, or action, where the perception, cognition, and interaction were driven by one sensory input [7], [8]. As the field matured, the limitations of unimodal embodied AI became apparent, and there has been a significant shift toward integrating multiple sensory modalities [14]–[16]. Multimodal embodied AI is now seen as crucial for creating more adaptive, flexible, and robust agents capable of performing complex tasks in dynamic environments [17], [18].

Unimodal embodied AI has benefited from rapid developments in fundamental areas such as computer vision, natural language processing, and reinforcement learning [12], [13]. These unimodal methods excel in dealing with a specific module in embodied AI. For example, computer vision techniques have driven advances in visual SLAM and 3D scene understanding in the active perception module [9]. Natural language processing techniques, especially LLM, have become popular solutions to address task planning and long-horizon reasoning in the embodied cognition module [10], [11]. Although unimodal embodied AI performs well in independent modules, it always faces two inherent limitations. On the one hand, the information contained in a single modality is limited, hindering the performance of perception, cognition, and interaction. For example, visual-only systems struggle to understand environments in dynamic or ambiguous settings, while auditory-based systems face challenges in real-world noise and signal processing [17], [202]. On the other hand, diverse and heterogeneous modalities hinder information transfer and sharing among modules. The agent's perception of the environment fails to facilitate the formation of its cognition, while the evolution of cognition fails to facilitate the interaction with the environment.

In contrast, multimodal embodied AI has emerged as a more promising paradigm [18]. By integrating data from multiple sensing modalities, such as visual, auditory, and olfactory feedback, these methods can provide a more holistic and precise understanding of the environment. More importantly, multimodal embodied AI can facilitate deeper integration among perception, cognition, and interaction. Recent advances in MLLMs and WMs enable agents to more effectively handle multiple modalities, promising to improve the capabilities of embodied AI [44], [81], [203]. The integration of these models is considered a key step toward enabling multimodal embodied AI in dynamic, uncertain environments.

III. Embodied AI with LLMs/MLLMs

This section provides a comprehensive overview of embodied AI with LLMs/MLLMs. We first elaborate in detail how LLMs boost embodied AI in Subsection III-A and how MLLMs boost embodied AI in Subsection III-B. Then we discuss the classification of MLLMs for embodied AI in Subsection III-C.

A. LLMs Boost Embodied AI

LLMs empower embodied AI via semantic reasoning and task decomposition, bringing high-level natural language instructions and low-level natural language actions into embodied cognition.

1) Semantic Reasoning: Semantic reasoning [19], [204], [205] leverages LLMs to interpret semantics from text instructions by analyzing linguistic patterns [206], contextual relationships [207], and implicit knowledge [208]. Through transformer architectures [63], LLMs map input tokens to latent representations, enabling hierarchical abstraction of meaning across syntactic and pragmatic levels. They employ attention mechanisms to weigh relevant semantic cues while suppressing noise, facilitating logical inference and analogical reasoning. By integrating world knowledge from pretraining corpora with task-specific prompts, LLMs dynamically construct conceptual graphs that align textual inputs with intended outcomes. This process supports multi-hop reasoning through probabilistic token prediction, resolving ambiguities by evaluating contextual coherence and semantic plausibility.

2) Task Decomposition: Task decomposition [20], [21] employs LLMs' sequential logic to break complex objectives into sub-tasks by hierarchically analyzing contextual dependencies and goal alignment. Leveraging chain-of-thought prompting, LLMs iteratively parse instructions into actionable steps, prioritizing interdependencies while resolving ambiguities through semantic coherence checks.

Representative works like SayCan [22] first provides a real-world pretrained natural language actions library, which is used to constrain LLMs to propose both feasible and contextually appropriate actions; then uses LLMs to convert natural language instructions into natural language action sequences; finally uses value functions to verify the feasibility of natural language action sequences in a particular physical environment. These works suggest that LLMs are extremely useful to robots aiming to act upon high-level, temporally extended instructions expressed in natural language. However, LLMs are only a part of the entire embodied AI system, which is limited by a fixed natural language actions library and a specific physical environment, and it is difficult to achieve adaptive expansion in new robots and environments.

B. MLLMs Boost Embodied AI

MLLMs can act on the entire embodied AI system and can solve LLMs' problems well by bridging high-level multimodal

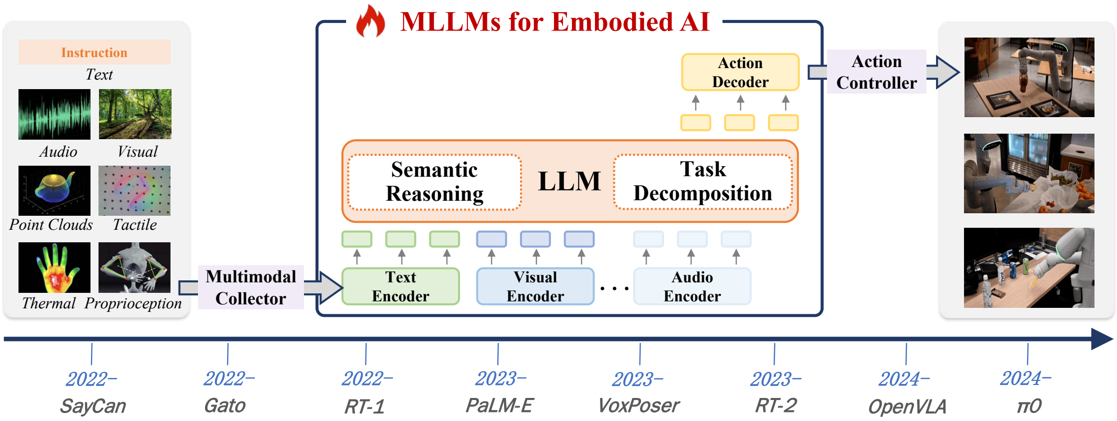

Fig. 5. The development roadmap of MLLMs for embodied AI. This roadmap highlights the key milestones in their conceptual and practical development.

inputting [209] and low-level motor action sequences [210] into end-to-end embodied applications (as shown in Fig. 5). Compared with LLMs, semantic reasoning [28]–[30] leverages MLLMs' cross-modal comprehension to interpret semantics from visual, auditory, or tactile inputs, e.g., identifying objects, inferring spatial relationships, or predicting environmental dynamics. Concurrently, task decomposition [31]–[33] employs MLLMs' sequential logic to break complex objectives into sub-tasks while dynamically adapting plans based on sensor feedback. MLLMs mainly include Vision-Language Models (VLMs) and Vision-Language-Action models (VLAs).

1) VLMs for Embodied AI: VLMs for embodied AI integrate visual and language instruction understanding to enable physical or virtual agents to perceive their environments in goal-driven tasks [211]–[213]. Representative works like PaLM-E [14] first train visual and language encodings end-to-end, in conjunction with a pre-trained large language model; then incorporate the results of real-world continuous sensor modalities encodings into VLMs and establish the link between words and percepts; finally, achieve multi-task completion through fixed action space mapping. For navigation, ShapeNet [214], which fine-tunes contrastive embeddings for 3D spatial reasoning, greatly reduces path planning errors. These works suggest that VLMs can combine perception and reasoning in embodied AI to solve a large number of tasks with fixed action spaces.

2) VLAs for Embodied AI: VLAs integrate multimodal inputs with low-level action control through differentiable pipelines. Representative works like RT-2 [24] first encode the robot's current image, language instructions, and robot actions at a specific timestep and convert them into text tokens; then use LLMs for semantic reasoning and task decomposition; finally, de-tokenizes generated tokens into the final action. Octo [129] pretrains on 100K robot demonstrations with language annotations, achieving cross-embodiment tool use. For dexterous manipulation, PerAct [215] utilizes 3D voxel representations to reach millimeter-level grasp accuracy. These works suggest that VLAs can act on the entire embodied AI system and achieve adaptive expansion in new robots and environments.

C. Classification of MLLMs for Embodied AI

MLLMs can empower active perception, embodied cognition, and dynamic interaction of embodied AI.

1) MLLMs for Active Perception: First, MLLMs can enhance 3D SLAM. By grounding visual observations into semantic representations, MLLMs augment traditional SLAM pipelines with high-level contextual information such as object categories, spatial relations, and scene semantics [216], [217]. Representative works like SEO-SLAM [218] utilize MLLMs to generate more specific and descriptive labels for objects, while dynamically updating a multiclass confusion matrix to mitigate biases in object detection. Second, MLLMs can enhance 3D scene understanding. Camera-based perception [30] remains the dominant setup in MLLM-driven embodied AI, as RGB inputs align naturally with the visual-language pretraining of many foundation models [219]–[221]. Representative works like EmbodiedGPT [123] leverage this synergy to map 2D visual inputs into semantically rich features aligned with language-based goals. Finally, MLLMs can enhance active environment exploration. MLLMs have also revolutionized how robots interact with their environments, particularly in feedback-driven closed-loop interactions. Representative works like LLM^3 [222] focus on structured motion-level feedback, which incorporates signals such as collision detections into the planning loop, allowing the model to iteratively revise symbolic action sequences. MART [223], on the other hand, leverages interaction feedback to improve retrieval quality.

2) MLLMs for Embodied Cognition: First, MLLMs can enhance task-driven self-planning [224]–[226]. Embodied agents with MLLMs can either directly map high-level goals to structured action sequences [31], or adopt an intermediate planning strategy that continually interacts with the environment to refine their plans [32]. Representative works like CoT-VLA [33] predict intermediate subgoal images that depict the desired outcomes of subtasks, helping the agent visualize and reason through each step of a complex task. Second, MLLMs can enhance memory-driven self-reflecting. MLLMs allow agents to learn from experience using this inherent memory module [129]. Representative works like Reflexion [116] enhance agent performance through self-generated linguistic feedback,

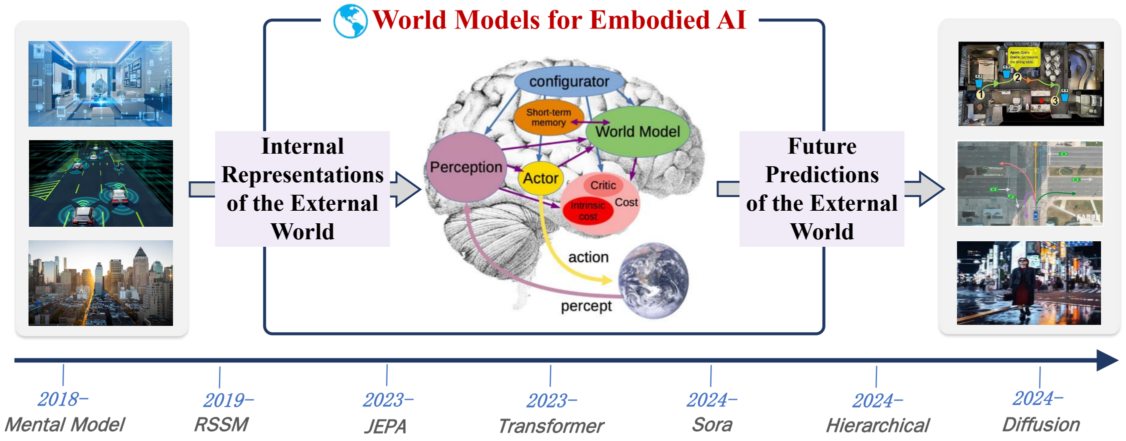

Fig. 6. The development roadmap of WMs for embodied AI. This roadmap highlights the key milestones in their conceptual and practical development.

which is stored in an episodic memory buffer and leveraged to guide future planning. Finally, MLLMs can enhance embodied multimodal foundation models. MLLMs can be adapted to the physical world through continued pretraining or fine-tuning in embodied settings. Representative works include Qwen-VL [75] and InternVL [227], along with models supporting broader modality alignment, such as Qwen2.5-Omni [228].

3) MLLMs for Dynamic Interaction: First, MLLMs can enhance action control. MLLMs have ability to decompose complex tasks into actionable subtasks [32]. To further produce continuous control signals for each subtask, MLLMs either generate actions autoregressively in a sequential manner [127], [229] or employ auxiliary policy heads to further process their internal representations [129]. Recent advances also explore generating executable code with MLLMs [230], enabling robots to follow interpretable and adaptable control policies. Second, MLLMs can enhance behavioral interaction. Through interaction with the environment, MLLMs are also capable of generating sequences of behavioral actions in a single step. Representative works like π-0 [31] combine a vision-language backbone with a flow-matching decoder to produce smooth, temporally extended behavioral trajectories. Finally, MLLMs can enhance collaborative decision-making. One line of research focuses on multi-agent systems that aim to achieve human-level coordination and adapt rapidly to unforeseen challenges [231]. For instance, Combo [146] introduces a novel framework that enhances cooperation among decentralized agents operating solely with egocentric visual observations. Other efforts investigate human-agent collaboration. VLAS [232] exemplifies this by aligning human verbal commands with visual context via a speech encoder and a LLaVA-style MLLM [233], enabling fluid and conversational human-agent interaction.

IV. Embodied AI with World Models

This section provides a comprehensive overview of embodied AI with WMs. We first elaborate in detail how WMs boost embodied AI in Subsection IV-A. Then we discuss the classification of WMs for embodied AI in Subsection IV-B.

A. World Models Boost Embodied AI

WMs empower embodied AI by building internal representations and future predictions of the external world (as shown in Fig. 6), facilitating physical law-compliant embodied interactions in dynamic environments.

1) Internal Representations of the External World: Internal representations compress rich sensory inputs into structured latent spaces, capturing object dynamics, physics laws, and spatial structures, allowing agents to reason about ”what exists” and ”how things behave” in their surroundings. These latent embeddings preserve hierarchical relationships [234] between entities and environments, mirroring the compositional nature of reality itself. The structured nature of these representations facilitates generalization across environments, as abstracted principles (like gravity or object permanence) transcend specific instances. Moreover, they support counterfactual reasoning [40] by maintaining disentangled variables for objects' intrinsic properties [38] and extrinsic relations [39], enabling flexible mental manipulation of individual components. This disentanglement also enhances sample efficiency in learning, as agents transfer knowledge between tasks, sharing latent factors. World models with rich internal representations, can introspect on their own uncertainty about environmental states and actively seek information to resolve ambiguities. By encoding temporal continuity and spatial topology [36], these models naturally enforce consistency constraints during planning, filtering physically implausible actions before execution. Ultimately, such structured latent spaces act as cognitive scaffolding for building causal understanding [37], mirroring how humans develop intuitive theories about their world through compressed sensory experiences.

2) Future Predictions of the External World: Future predictions simulate potential rewards of sequence actions across multiple time horizons aligned with physical laws, thereby preempting risky or inefficient behaviors [41], [42]. This predictive capacity bridges short-term actions with long-term goals [43], filtering out trajectories violating physical plausibility (e.g., walking through walls) or strategic coherence (e.g., depleting resources prematurely). Long-horizon prediction [44] allows adaptive balancing of exploration-exploitation trade-offs, simulating distant outcomes to avoid local optima while maintaining focus on actionable near-term steps. Crucially, these predictions incorporate uncertainty quantification [41], [235], distinguishing predictable regularities (daily patterns) from stochastic events (sudden changes) to optimize risk-aware planning. The simulation prediction improves sample efficiency [39], [236]–[238] by replacing costly trial-and-error with mental rehearsal, particularly valuable in safety-critical domains like autonomous driving or robotic surgery. Furthermore, continuous prediction-error minimization drives iterative model refinement [170], [239]–[241], creating self-correcting systems that align their internal physics simulators with observed reality. Such anticipatory capabilities ultimately grant artificial agents human-like foresight, transforming reactive responses into purposeful, future-optimized behaviors.

B. Classification of World Models for Embodied AI

Embodied AI with WMs can mainly be divided into three critical structures: the Recurrent State Space Model-based (RSSM-based) WMs for embodied AI, the Joint-Embedding Predictive Architecture-based (JEPA-based) WMs for embodied AI, and the Transformer-based WMs for embodied AI. Hierarchical-based WMs [242] and diffusion-based WMs [243] are similar to other structures and are shown in Fig. 6.

1) RSSM-based WMs for Embodied AI: RSSM constitutes the fundamental architecture underpinning the Dreamer algorithm family [41]–[44]. This framework enhances predictive capabilities in latent representations by acquiring temporal environment dynamics through visual inputs, subsequently enabling action selection via latent trajectory optimization. Through orthogonal decomposition of hidden states into probabilistic and deterministic components, the architecture explicitly accounts for both systematic patterns and environmental uncertainties. Its demonstrated effectiveness in robotic motion control applications has inspired numerous derivative studies building upon its theoretical framework.

2) JEPA-based WMs for Embodied AI: JEPA [27] provides a structure for developing autonomous machine intelligence systems. This architecture establishes mapping relationships between input data and anticipated outcomes through representation learning. Diverging from conventional generative approaches, JEPA operates in abstract latent spaces rather than producing pixel-wise reconstructions, thereby prioritizing semantic feature extraction over low-level signal synthesis. A key methodological foundation of JEPA [235] involves self-supervised training paradigms where neural networks learn to infer occluded or unobserved data segments. Such pretraining on extensive unlabeled datasets enables transfer learning across downstream applications, demonstrating enhanced generalization capabilities for both visual [244], [245] and non-visual domains [246].

3) Transformer-based WMs for Embodied AI: Originating in natural language processing research, the Transformer structure [63] fundamentally relies on attention mechanisms to process input sequences through parallelized context weighting. This design allows simultaneous computation of inter-element dependencies, overcoming the sequential processing constraints inherent in Recurrent Neural Networks (RNNs). Empirical evidence demonstrates superior performance in domains requiring persistent memory retention and explicit memory addressing for cognitive reasoning [247], which has propelled its adoption in reinforcement learning research since 2020. Existing advancements have successfully implemented WMs using Transformer variants [38], [40], [248], outperforming RSSM architectures in memory-intensive interactive scenarios [37]. Notably, Google's Genie framework [36] employs the Spatial-Temporal Transformer (ST-Transformer) [249] to create synthetic interactive environments through large-scale self-supervised video pretraining. This breakthrough establishes novel paradigms for actionable world modeling, revealing transformative potential for WMs development trajectories.

V. Embodied AI with MLLMs and WMs

This section provides a comprehensive overview of embodied AI with MLLMs and WMs. We first elaborate in detail on the limitations of MLLMs and WMs for embodied AI and explain how MLLMs boost WMs reasoning, and how WMs boost MLLMs interaction in Subsection V-A. Then we design a joint MLLM-WM-driven embodied AI architecture in Subsection V-B. Finally, we discuss the advantages and challenges of new architecture in Subsection V-C.

A. MLLMs and WMs

MLLMs enable contextual task reasoning but overlook physical constraints, while WMs excel at physics-aware simulation but lack high-level semantics. Their joint bridges semantic intelligence with grounded physical interaction.

1) The Limitations of MLLMs for Embodied AI (without WMs): MLLMs exhibit two critical limitations in embodied AI applications. First, they often fail to ground predictions [34] in physics-compliant dynamics, leading to impractical plans. For example, ignoring friction or material properties when manipulating objects may cause slippage or task failure. Second, their poor real-time adaptation to environmental feedback limits responsiveness [35]. While MLLMs excel at semantic task decomposition, they struggle to adaptively adjust actions when the environment changes dramatically. These limitations stem from their reliance on static, pre-trained knowledge rather than continuous physical interaction.

2) The Limitations of WMs for Embodied AI (without LLMs/MLLMs): WMs face limitations in abstract reasoning and generalization. They struggle with open-ended semantic tasks [45] due to their focus on physical simulation rather than contextual understanding. Additionally, WMs lack generalizable task decomposition [26] without explicit priors. For example, a WM model trained on rigid-object manipulation may fail to adapt to deformable materials without extensive retraining. Their predictive accuracy heavily depends on domain-specific interaction records, hindering scalability across diverse environments.

3) MLLMs Boosting WMs Reasoning: By leveraging cross-modal alignment and semantic grounding, MLLMs enable WMs to process complex environments dynamically, improving semantic reasoning, task decomposition, and human-robot interaction. 1) MLLMs can enrich WMs by fusing visual, auditory, and textual data into unified semantic representations.

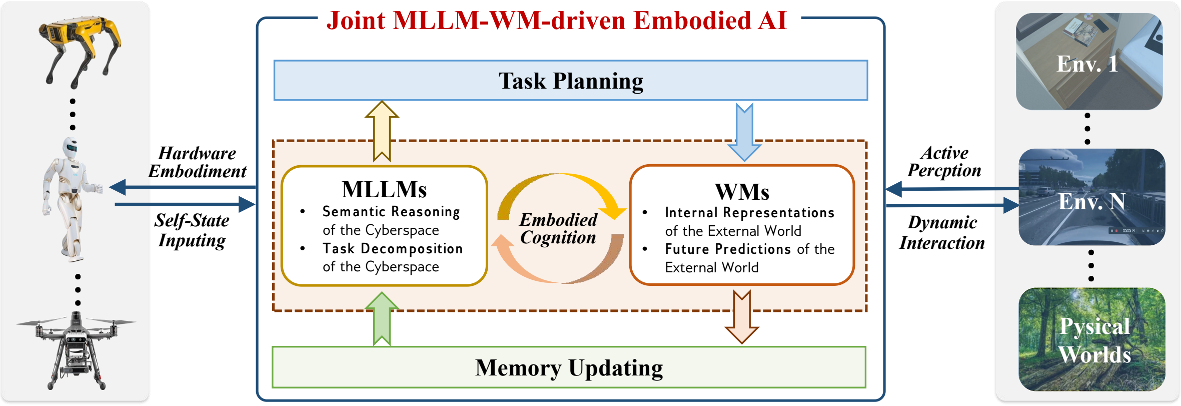
Joint MLLM-WM-driven Embodied AI
Task Planning
MLLMs
Semantic Reasoning of the Cyberspace
Task Decomposition of the Cyberspace
Embodied Cognition
WMs
Internal Representations of the External World
Future Predictions of the External World
Memory Updating
Hardware Embodiment
Self-State Inputing
Active Percption
Dynamic Interaction
Env. 1
Env. N
Pysical Worlds

Fig. 7. Embodied AI with MLLMs and WMs. MLLMs can enhance WMs by injecting semantic knowledge for task decomposition and long-horizon reasoning, while WMs can assist MLLMs by building the physical world's internal representations and future predictions, making joint MLLM-WM a promising architecture for embodied systems.

For instance, CLIP-based architectures [250] enable agents to align visual scenes with linguistic cues, reducing ambiguity in object recognition [251]. 2) MLLMs can augment WM's task decomposition capacity by decomposing high-level goals into executable sub-tasks. Models like GPT-4V [252] generate step-by-step plans using environmental context stored in WM. For robotic manipulation, Code-as-Policies [253] translates natural language instructions into code snippets, leveraging WM to track intermediate states. 3) MLLMs enable WMs to refine internal representations through human feedback. Techniques like Reinforcement Learning with Human Feedback (RLHF) [73] allow agents to update WM priors based on corrective inputs [116]. Those works in this Subsubsection are all possible ways for MLLMs to boost WMs reasoning, which is not achieved in existing works.

4) WMs boosting MLLMs Interaction: WMs can play a pivotal role in refining MLLMs by providing physical laws, spatio-temporal relationships, and closed-loop interaction experiences. WMs can mitigate MLLMs' inherent limitations in temporal coherence and environmental grounding, enabling more robust decision-making in dynamic embodied tasks. 1) WMs can provide MLLMs with explicit representations of physical laws (e.g., gravity, friction) and commonsense rules to constrain action proposals. For example, Physion++ [254] integrating WM-stored biomechanical models can be used to filter MLLM-generated robotic motions violating torque limits; RoboGuide [255] injects spatial occupancy maps into MLLM planners, preventing collisions during navigation. 2) WMs can stabilize MLLMs reasoning by maintaining spatio-temporal context during multimodal processing. For instance, MemPrompt [256] can use WM buffers to align visual object trajectories with linguistic descriptions, resolving ambiguities in cluttered environments; RoboMem [257] can leverage WM-prioritized attention to filter irrelevant sensory noise, improving MLLM-based scene understanding. 3) WMs can enable iterative refinement of MLLM outputs through closed-loop interaction. Reflexion [116] can store task-execution histories in WM, allowing MLLMs to correct kinematic errors using failure patterns [253]. Those works in this Subsubsection are all possible ways for WMs to boost MLLMs' decisions, which has not been achieved in existing works.

B. Joint MLLM-WM-driven Embodied AI Architecture

We propose a joint MLLM-WM-driven embodied AI architecture (as shown in Fig. 7), shedding light on their profound significance in enabling complex tasks within physical worlds. The specific workflow is as follows, with arrows highlighting the data exchange process.

1) Robots → Self-State Inputing → MLLMs/WMs → Hardware Embodiment → Robots: The process initiates with self-state inputting tracking proprioceptive metrics, such as degrees of freedom, number of sensors, etc. These metrics feed into both WMs and MLLMs: WMs use them to build internal representations of the agent's physical state, while MLLMs contextualize these states for task alignment. Hardware embodiment is focused on implementing WMs and MLLMs into physical devices to solve sim-to-real problems. This bidirectional flow ensures actions respect both mechanical limits and high-level goals.

2) MLLMs → Task Planning → WMs → Memory Updating → MLLMs: MLLMs decompose abstract instructions into subtasks. A forward arrow delivers this plan to WMs, which predict outcomes based on existing environmental modeling. During execution, WMs log outcomes into memory. A vertical arrow transmits these logs to memory updating modules, which structure memory into experiences, represent the forgetting of past task memories, the renewal of current task memories, and the prediction of future task memories. These are then fed back to MLLMs via an arrow, enriching their knowledge base. This enables lifelong learning, where past failures directly inform future planning.

3) Environments → Active Perception → MLLMs/WMs → Dynamic Interaction → Environments: WMs first drive active perception by predicting key environmental changes. Multimodal inputs are then used to construct an internal representation of the external world through WMs and semantic reasoning through MLLMs. Then, the task decomposition of MLLMs and future prediction of WMs enable action selection and environmental interaction. Adaptive perception and interaction of dynamic environments are achieved through continuous iteration.

TABLE IV
QUALITATIVE COMPARISON OF MLLM-ONLY, WM-ONLY, AND JOINT MLLM-WM ARCHITECTURES IN EMBODIED AI. LOW , MEDIUM , HIGH .

Performance	LLM/MLLM-only	WM-only	Joint MLLM-WM
Semantic Understanding	Advantages in contextual task reasoning and natural language understanding	Limited in open-ended semantic understanding	Combines high-level semantic abstraction with grounded contextual alignment
Task Decomposition	Sequential logic enables sub-task planning via language prompts	Lacks generalizable task decomposition mechanisms	Semantic plans refined through physical feasibility via joint planning-execution loop
Physics Compliance	Ignores physical constraints and dynamics in real-world interaction	Physics-aware simulation with temporal consistency	Enforces semantic-physical alignment for safe and executable plans
Future Prediction	Lacks imagination-based reasoning	Long-horizon multi-step prediction with uncertainty modeling	Combines symbolic foresight and physically grounded imagination
Real-time Interaction	Poor responsiveness to environmental feedback and significant reasoning latency	Supports real-time predictive control via future state simulation	Enables online adaptation through iterative plan refinement and memory updating
Memory Structure	Sparse and unstructured memory	Structured latent space encodes object dynamics and causal relationships	Integrates semantic memory and world modeling for lifelong learning and reflection
Scalability	Limited to pre-trained task space	Poor transfer to unseen tasks without retraining	Cross-task, cross-domain generalization through symbolic and sensorimotor synergy

C. Discussions

Joint MLLM-WM offer a promising architecture for embodied AI. As shown in TABLE IV, MLLMs excel in semantic reasoning, enabling high-level task decomposition, contextual understanding, and adaptive planning by leveraging multimodal inputs. Meanwhile, WMs provide grounded, physics-based simulations of environments, ensuring actions align with real-world constraints. This synergy allows agents to balance abstract reasoning with real-time physical interactions, enhancing decision-making in dynamic settings. For instance, MLLMs can generate task plans while WMs validate feasibility, enabling iterative refinement. Additionally, joint architectures support cross-modal generalization, improving robustness in partially observable or novel scenarios by bridging symbolic knowledge and sensorimotor experiences.

The challenges of joint MLLM-WM-driven embodied AI architecture include 1) real-time synchronization between MLLMs' high-latency semantic processing and WMs' physics-based representation, often leading to delayed responses in dynamic environments; 2) semantic-physical misalignment, where MLLM-generated plans violate unmodeled physical constraints; and 3) scalable memory management, as continuous updates to WM's internal states risk overwhelming MLLMs with irrelevant context. Additionally, training such systems requires vast multimodal datasets covering rare edge cases, while ensuring robustness against sensor noise and partial observability remains unsolved. These challenges need lightweight MLLMs inference, tighter feedback loops, and dynamic context-filtering mechanisms to minimize latency.

VI. EMBODIED AI APPLICATIONS

This section overviews the application of embodied AI in service robots, rescue robots, and other domains, highlighting trends in joint MLLMs and WMs to advance active perception, embodied cognition and dynamic interaction.

A. Service Robotics

Embodied AI is becoming an important technology in the service field. It helps service robots go beyond fixed rules and perform tasks in a flexible way using different types of information. Recent research [258], [259] highlights its flexible applications across various fields. In domestic settings, systems such as RT-2 [229] and SayCan [121] combine language instructions with robot control, allowing robots to do tasks such as stacking dishes or cooking. Few-shot learning methods like AED [260] acquire new skills from limited demonstrations. In healthcare, robots with multiple types of input can help with reminders, rehabilitation, and companionship. [261], [262]. In public environments, platforms like Habitat [193] and RT-X [263] support navigation and item delivery, even in changing environments, without needing special training for each task. This makes the system more general and useful in real life.

However, current approaches remain limited in handling long-horizon tasks. As illustrated in Fig. 7, the joint of WMs and MLLMs is emerging as a key strategy for enhancing the autonomy and long-term reasoning capabilities of service robots. The WM maintains an evolving environment model for planning and simulation, while the MLLM grounds commands like "clean up the living room" into adaptive subtasks. This collaboration supports flexible reasoning, goal adaptation, and robust real-world execution.

B. Rescue UAVs

Embodied AI technology technology is changing the way drones are used in disaster situations. Traditional drones are either manually controlled or rely on pre-built maps when in use, which leads to their inability to adapt to the environment independently. However, embodied drones [264], [265] can sense the environment in real time and respond to sudden changes. This ability makes them very useful in dangerous places like earthquake zones, forest fires, or floods. Recent studies show that embodied drones can perform many complex tasks. For instance, with the help of language models, they can understand and follow human voice instructions, helping drones quickly change their actions and enhancing their responses in emergency situations, such as "search near the collapsed bridge" [115], [266]–[269]. Secondly, some work use world models to simulate dangerous environments, which helps them avoid danger and plan a safer path [270]–[272]. Other studies explore how multiple drones can work together to find survivors and map damaged areas [201], [273], [274].

However, despite these advancements, current approaches remain limited in handling long-horizon reasoning and autonomous decision-making under uncertainty. As illustrated in Fig. 7, jointing WMs and MLLMs has emerged as a key strategy for further enhancing UAV autonomy. The WM maintain a continuously evolving spatiotemporal representation of the environment, supporting planning and risk prediction even in GPS-denied conditions. The MLLM grounds commands into structured subtasks based on the UAV's belief state. This coordination improves generalization, long-horizon reasoning, and high-level autonomy in mission-critical conditions.

C. Industrial Robots

Embodied AI is changing the way robots work in factories. With embodied AI, industrial robots [275] can make smarter decisions based on their surroundings. Traditional industrial robots are usually fixed in one place. They use special sensors and tools and are required to complete tasks with very high accuracy. Because of this, they are better at doing jobs that need the same movements again and again.

However, with embodied AI, these robots can do more than repeat actions. By combining MLLMs and WMs, industrial robots can adjust how hard they hold fragile objects, or find a new path when they meet an obstacle. This has already been used in real life. For example, robots in Tesla's factory can find and fix parts that are not lined up, without help from people.At JD, robots [276], [277] use different sensors to sort packages by size and address. In Tmall's warehouse [278], robots use thermal cameras, LiDAR, and RGB sensors to check for problems in the inventory and send alerts when something is wrong. These examples show that embodied AI is helping robots become more flexible, reliable, and smart in factories.

D. Other Applications

In addition to its use in homes, healthcare, and rescue missions, embodied AI is also being applied in educational, virtual, and space environments [279]. In smart manufacturing, it supports robots that can work together with humans, perform accurate assembly tasks, and adapt their actions based on changes in the workspace or human behavior [280]–[282]. With the help of visual and touch feedback, these robots can safely handle fragile items [283], [284]. In education, embodied AI is used in social robots that adjust their speech, gaze, and gestures according to the student's focus and emotions [285]–[287]. This helps create a more personalized learning experience and builds long-term trust between students and robots [288], [289]. In virtual environments, embodied agents learn to move, interact with objects, and complete tasks that require several steps. They also develop memory over time to improve their performance [290]. In space exploration, where conditions are unknown and communication with Earth is delayed, embodied AI allows robots to make decisions on their own and adapt to new surroundings [291]. These examples show that embodied AI is becoming more flexible and useful across many fields, helping machines see, act, and learn in both real and virtual worlds.

VII. Future Directions

As embodied AI moves from simulation to real-world deployment, future research must prioritize the development of unified and reliable systems across several core domains. Key directions include autonomous embodied AI, embodied AI hardware, swarm embodied AI, and evaluation benchmark.

A. Autonomous Embodied AI

The purpose of autonomous embodied AI is to enable agents to operate independently for a long time in a dynamic and open environment. Future research is expected to develop along several key directions. First, adaptive perception can give the system the ability to autonomously select input data, which can be achieved by dynamically choosing and integrating information from different sensory modalities. Second, Building on this foundation, building environmental awareness is essential. Environmental awareness helps agents quickly adapt to changes, predict the consequences of their actions, and transfer their behavior to new environments. It requires memory architectures that can capture spatiotemporal patterns and model causal relationships. Third, future systems should combine MLLMs with real-time physical interaction, which allows agents to bridge high-level language instructions with low-level control, and accurately model the real physical world.

B. Embodied AI Hardware

Future research in embodied AI hardware is expected to advance in the following four directions. First, hardware-aware model compression will continue to integrate techniques such as quantization and pruning with hardware performance metrics, enabling precise control over the trade-off between model accuracy and deployment efficiency. Second, graph-level compilation optimization will play a key role in bridging the gap between high-level embodied models and low-level hardware execution, which will focus on more effective operator fusion, scheduling strategies, and memory access efficiency to reduce execution overhead. Third, domain-specific accelerators will be increasingly tailored to the computational characteristics of embodied tasks. Reconfigurable architectures such as FPGA and CGRA offer flexibility and adaptability, while ASIC-based designs provide high efficiency and performance. Fourth, hardware-software co-design will become essential for eliminating mismatches between algorithm behavior and hardware architecture. Joint optimization of model structures and hardware architecture will be critical to achieving real-time, energy-efficient execution in embodied systems.

C. Swarm Embodied AI

Swarm embodied AI refers to the collaborative perception and decision-making of multiple agents. refers to the collaborative perception and decision-making of multiple agents. Because multiple agents can exhibit stronger capabilities when cooperating than a single agent, this kind of ”collective intelligence” has aroused the interest of many researchers and is also regarded as an important step for agents to approach humans. First of all, to enable multiple agents to cooperate smoothly, it is necessary to develop collaborative WMs. This model can establish a shared and dynamic environmental representation based on the observations of each agent, forming the basis of collective understanding. Secondly, multi-agent representation learning is very important. It can help the agent understand its own state and also comprehend the situations of other agents. This is the basis for communication and cooperation among agents. In addition, modeling social behavior among agents is also crucial. Role allocation and group decision-making can be better achieved through behavioral modeling. Finally, to seamlessly integrate into real-world applications, it is also important to design natural human-swarm interaction interfaces. It may include multimodal language foundations and get-based control methods, making it easier for humans to direct and guide the entire agent group.

D. Explainability and Trustworthiness Embodied AI

Explainability and trustworthiness represent a critical frontier for Embodied AI, essential for its safe, ethical, and widespread real-world deployment as agents increasingly interact physically with humans and dynamic environments. Future research must address several key challenges: Firstly, designing benchmarks that provide real-time, human-understandable justifications for agent actions, particularly during unexpected situations or failures, is crucial for user trust and debugging. Secondly, establishing robust mechanisms to ensure agents adhere to ethical principles and human values during autonomous decision-making, especially in morally ambiguous scenarios common in rescue or healthcare applications, requires significant advancement. Thirdly, creating verifiable safety guarantees and certification standards for agents operating in unstructured physical settings, mitigating risks associated with unpredictable interactions, remains an open problem. Finally, enhancing robustness against adversarial attacks, sensor noise, and distribution shifts, ensuring reliable performance despite uncertainties inherent in the real world, is fundamental for trustworthy operation. Addressing these multifaceted research problems in explainability and trustworthiness is paramount, as progress in this direction will unlock the full potential of Embodied AI by fostering user confidence, enabling responsible innovation, and facilitating regulatory acceptance.

E. Other Directions

Several new directions may influence the future development of embodied AI. One important direction is lifelong learning. Agents need to continuously learn new skills without forgetting what they have already learned. Only in this way can they adapt to the dynamic environment and maintain the accuracy of the previously completed tasks. Another key direction is human-in-the-loop learning. Human feedback is very important supervisory information. A small amount of feedback can significantly improve the performance of an agent and make it more human-like. To achieve this goal, we need better methods to enable agents to understand human goals and preferences. Finally, as agents become more autonomous, moral decision-making becomes increasingly important. Future systems should learn to carefully identify moral hazard and follow human values. This will help ensure that the embedded artificial intelligence is both safe and reliable.

ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China No.62222209, Beijing National Research Center for Information Science and Technology under Grant No.BNR2023TD03006, China Postdoctoral Science Foundation under Grant No.2024M751688, Postdoctoral Fellowship Program of CPSF under Grant No.GZC20240827.

We thank Dr. Ren Wang for his contribution to Section II. We thank Dr. Yu-Wei Zhan for her contributions to Section VI and Section VII.

REFERENCES

[1] A. M. Turing, “Computing machinery and intelligence,” *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[2] R. Pfeifer and J. Bongard, *How the body shapes the way we think: a new view of intelligence*. MIT press, 2006.

[3] A. Clark, *Being there: Putting brain, body, and world together again*. MIT press, 1998.

[4] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding,” in *International Conference on Learning Representations*, 2015.

[5] T. Chen, T. Moreau, Z. Jiang, L. Zheng, E. Yan, H. Shen, M. Cowan, L. Wang, Y. Hu, L. Ceze *et al.*, “Tvm: An automated end-to-end optimizing compiler for deep learning,” in *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, 2018, pp. 578–594.

[6] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers *et al.*, “In-datacenter performance analysis of a tensor processing unit,” in *Proceedings of the 44th annual international symposium on computer architecture*, 2017, pp. 1–12.

[7] J. Engel, T. Schöps, and D. Cremers, “Lsd-slam: Large-scale direct monocular slam,” in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[8] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, “Orb-slam: A versatile and accurate monocular slam system,” *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[9] A. Macario Barros, M. Michel, Y. Moline, G. Corre, and F. Carrel, "A comprehensive survey of visual slam algorithms," *Robotics*, vol. 11, no. 1, p. 24, 2022.

[10] Z. Wu, Z. Wang, X. Xu, J. Lu, and H. Yan, "Embodied task planning with large language models," *arXiv preprint arXiv:2307.01848*, 2023.

[11] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, and J. M. Alvarez, "Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning," *arXiv preprint arXiv:2405.01533*, 2024.

[12] D. Jayaraman and K. Grauman, "Learning to look around: Intelligently exploring unseen environments for unknown tasks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1238–1247.

[13] B. Y. Lin, C. Huang, Q. Liu, W. Gu, S. Sommerer, and X. Ren, "On grounded planning for embodied tasks with language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 192–13 200.

[14] 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," in *International Conference on Machine Learning*. PMLR, 2023, pp. 8469–8488.

[15] W. Jiang, B. Lei, K. Ashton, and K. Daniilidis, "Multimodal llm guided exploration and active mapping using fisher information," *arXiv preprint arXiv:2410.17422*, 2024.

[16] Z. Wang, C. Chen, F. Luo, Y. Dong, Y. Zhang, Y. Xu, X. Wang, P. Li, and Y. Liu, "Actiview: Evaluating active perception ability for multimodal large language models," *arXiv preprint arXiv:2410.04659*, 2024.

[17] Y. Qin, E. Zhou, Q. Liu, Z. Yin, L. Sheng, R. Zhang, Y. Qiao, and J. Shao, "Mp5: A multi-modal open-ended embodied system in minecraft via active perception," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 16 307–16 316.

[18] R. Yang, H. Chen, J. Zhang, M. Zhao, C. Qian, K. Wang, Q. Wang, T. V. Koripella, M. Movahedi, M. Li *et al.*, "Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents," *arXiv preprint arXiv:2502.09560*, 2025.

[19] 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.

[20] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[21] L. Floridi and M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds and Machines*, 2020.

[22] 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.

[23] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: program generation for situated robot task planning using large language models," *Autonomous Robots*, vol. 47, no. 8, pp. 999–1012, 2023.

[24] 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.

[25] J. Ding, Y. Zhang, Y. Shang, Y. Zhang, Z. Zong, J. Feng, Y. Yuan, H. Su, N. Li, N. Sukiennik *et al.*, "Understanding world or predicting future? a comprehensive survey of world models," *arXiv preprint arXiv:2411.14499*, 2024.

[26] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.

[27] Y. LeCun, "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27," *Open Review*, vol. 62, no. 1, pp. 1–62, 2022.

[28] X. Li, M. Zhang, Y. Geng, H. Geng, Y. Long, Y. Shen, R. Zhang, J. Liu, and H. Dong, "Manipllm: Embodied multimodal large language model for object-centric robotic manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[29] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A survey of embodied ai: From simulators to research tasks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2022.

[30] T. Brophy, D. Mullins, A. Parsi, J. Horgan, E. Ward, P. Denny, C. Eising, B. Deegan, M. Glavin, and E. Jones, "A review of the impact of rain on camera-based perception in automated driving systems," *IEEE Access*, 2023.

[31] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π0: A vision-language-action flow model for general robot control," *URL https://arxiv.org/abs/2410.24164*, 2024.

[32] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, "Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation," *arXiv preprint arXiv:2409.01652*, 2024.

[33] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn *et al.*, "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models," *arXiv preprint arXiv:2503.22020*, 2025.

[34] Y. Liu, J. Smith, and Q. Chen, "Task alignment in embodied ai: Bridging semantics and physical constraints," *Transactions on Robotics*, vol. 40, pp. 1–15, 2024.

[35] D. Driess, Y. Huang, and A. e. a. Zeng, "Vision-language-action models: Unified frameworks for embodied execution," in *Conference on Robot Learning (CoRL)*, 2023, pp. 1234–1245.

[36] J. Bruce, M. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps *et al.*, "Genie: Generative interactive environments," 2024.

[37] C. Chen, Y.-F. Wu, J. Yoon, and S. Ahn, "Trans-dreamer: Reinforcement learning with transformer world models," 2022.

[38] J. Robine, M. Hoffmann, T. Uelwer, and S. Harmeling, "Transformer-based world models are happy with 100k interactions," in *ICLR*, 2023.

[39] X. Wang, Z. Zhu, G. Huang, B. Wang, X. Chen, and J. Lu, "World-dreamer: Towards general world models for video generation via predicting masked tokens," 2024.

[40] W. Wu, Z. Li, Y. He, M. Z. Shou, C. Shen, L. Cheng, Y. Li, T. Gao, D. Zhang, and Z. Wang, "Paragraph-to-image generation with information-enriched diffusion model," *arXiv preprint arXiv:2311.14284*, 2023.

[41] D. Hafner, T. P. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," in *ICLR*, 2020.

[42] D. Hafner, T. P. Lillicrap, M. Norouzi, and J. Ba, "Mastering atari with discrete world models," in *ICLR*, 2021.

[43] M. Okada and T. Taniguchi, "Dreamingv2: Reinforcement learning with discrete world models without reconstruction," in *IROS*, 2022.

[44] P. Wu, A. Escontrela, D. Hafner, P. Abbeel, and K. Goldberg, "Day-dreamer: World models for physical robot learning," in *Conference on robot learning*. PMLR, 2023, pp. 2226–2240.

[45] A. Clark, *Whatever Next? Predictive Brains and Embodied Cognition*. MIT Press, 2013.

[46] T. Feng, X. Wang, Z. Zhou, R. Wang, Y. Zhan, G. Li, Q. Li, and W. Zhu, "Evoagent: Agent autonomous evolution with continual world model for long-horizon tasks," *arXiv preprint arXiv:2502.05907*, 2025.

[47] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[48] G. Lakoff and M. Johnson, *Metaphors we live by*. University of Chicago press, 1980.

[49] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.

[50] R. Brooks, "A robust layered control system for a mobile robot," *IEEE journal on robotics and automation*, vol. 2, no. 1, pp. 14–23, 1986.

[51] R. A. Brooks, "Intelligence without representation," *Artificial intelligence*, vol. 47, no. 1-3, pp. 139–159, 1991.

[52] R. A. Brooks, C. Breazeal, M. Marjanović, B. Scassellati, and M. M. Williamson, "The cog project: Building a humanoid robot," in *International workshop on computation for metaphors, analogy, and agents*. Springer, 1998, pp. 52–87.

[53] R. Chrisley, "Embodied artificial intelligence," *Artificial intelligence*, vol. 149, no. 1, pp. 131–150, 2003.

[54] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, no. 39, pp. 1–40, 2016.

[55] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[57] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[58] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[59] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2020.

[60] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[61] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[62] 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,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[63] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017.

[64] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[65] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[66] 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.

[67] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, “Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality,” *See https://vicuna. lmsys. org (accessed 14 April 2023)*, vol. 2, no. 3, p. 6, 2023.

[68] 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.

[69] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning,” *arXiv preprint arXiv:1312.5602*, 2013.

[70] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, “Mastering the game of go with deep neural networks and tree search,” *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

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

[72] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *International conference on machine learning*. Pmlr, 2018, pp. 1861–1870.

[73] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, “Training language models to follow instructions with human feedback,” *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[74] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, “Deepseekmath: Pushing the limits of mathematical reasoning in open language models,” *arXiv preprint arXiv:2402.03300*, 2024.

[75] 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.

[76] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang *et al.*, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” *arXiv preprint arXiv:2403.05530*, 2024.

[77] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, “Gpt-4o system card,” *arXiv preprint arXiv:2410.21276*, 2024.

[78] 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.

[79] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, “Learning latent dynamics for planning from pixels,” in *International conference on machine learning*. PMLR, 2019, pp. 2555–2565.

[80] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, “Mastering diverse domains through world models,” *arXiv preprint arXiv:2301.04104*, 2023.

[81] Z. Zhu, X. Wang, W. Zhao, C. Min, N. Deng, M. Dou, Y. Wang, B. Shi, K. Wang, C. Zhang *et al.*, “Is sora a world simulator? a comprehensive survey on general world models and beyond,” *arXiv preprint arXiv:2405.03520*, 2024.

[82] D. S. Chaplot, M. Dalal, S. Gupta, J. Malik, and R. R. Salakhutdinov, “Seal: Self-supervised embodied active learning using exploration and 3d consistency,” *Advances in neural information processing systems*, vol. 34, pp. 13 086–13 098, 2021.

[83] M. F. Ahmed, K. Masood, V. Fremont, and I. Fantoni, “Active slam: A review on last decade,” *Sensors*, vol. 23, no. 19, p. 8097, 2023.

[84] M. R. U. Saputra, A. Markham, and N. Trigoni, “Visual slam and structure from motion in dynamic environments: A survey,” *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, pp. 1–36, 2018.

[85] H. Durrant-Whyte and T. Bailey, “Simultaneous localization and mapping: part i,” *IEEE robotics & automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[86] T. Bailey and H. Durrant-Whyte, “Simultaneous localization and mapping (slam): Part ii,” *IEEE robotics & automation magazine*, vol. 13, no. 3, pp. 108–117, 2006.

[87] Y. Wang, Y. Tian, J. Chen, K. Xu, and X. Ding, “A survey of visual slam in dynamic environment: The evolution from geometric to semantic approaches,” *IEEE Transactions on Instrumentation and Measurement*, 2024.

[88] P. F. Alcantarilla, J. J. Yebes, J. Almazán, and L. M. Bergasa, “On combining visual slam and dense scene flow to increase the robustness of localization and mapping in dynamic environments,” in *2012 IEEE international conference on robotics and automation*. IEEE, 2012, pp. 1290–1297.

[89] D.-H. Kim and J.-H. Kim, “Effective background model-based rgb-d dense visual odometry in a dynamic environment,” *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1565–1573, 2016.

[90] D. Zou and P. Tan, “Coslam: Collaborative visual slam in dynamic environments,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 2, pp. 354–366, 2012.

[91] W. Dai, Y. Zhang, P. Li, Z. Fang, and S. Scherer, “Rgb-d slam in dynamic environments using point correlations,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 373–389, 2020.

[92] Z.-J. Du, S.-S. Huang, T.-J. Mu, Q. Zhao, R. R. Martin, and K. Xu, “Accurate dynamic slam using crf-based long-term consistency,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 4, pp. 1745–1757, 2020.

[93] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, “Slam++: Simultaneous localisation and mapping at the level of objects,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1352–1359.

[94] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, “Ds-slam: A semantic visual slam towards dynamic environments,” in *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2018, pp. 1168–1174.

[95] M. Gonzalez, E. Marchand, A. Kacete, and J. Royan, “Twistslam: Constrained slam in dynamic environment,” *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6846–6853, 2022.

[96] C. Yan, D. Qu, D. Xu, B. Zhao, Z. Wang, D. Wang, and X. Li, “Gs-slam: Dense visual slam with 3d gaussian splatting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 595–19 604.

[97] M. A. Bautista, P. Guo, S. Abnar, W. Talbott, A. Toshev, Z. Chen, L. Dinh, S. Zhai, H. Goh, D. Ulbricht *et al.*, “Gaudi: A neural architect for immersive 3d scene generation,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 102–25 116, 2022.

[98] R. Chen, Y. Liu, L. Kong, N. Chen, X. Zhu, Y. Ma, T. Liu, and W. Wang, “Towards label-free scene understanding by vision foundation models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 75 896–75 910, 2023.

[99] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser *et al.*, “Openscene: 3d scene understanding with open vocabularies,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 815–824.

[100] Y. Man, S. Zheng, Z. Bao, M. Hebert, L. Gui, and Y.-X. Wang, “Lexicon3d: Probing visual foundation models for complex 3d scene understanding,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 76 819–76 847, 2024.

[101] G. Gao, W. Liu, A. Chen, A. Geiger, and B. Schölkopf, “Graphdreamer: Compositional 3d scene synthesis from scene graphs,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 295–21 304.

[102] H. Zhou, J. Shao, L. Xu, D. Bai, W. Qiu, B. Liu, Y. Wang, A. Geiger, and Y. Liao, “Hugs: Holistic urban 3d scene understanding via gaussian splatting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 336–21 345.

[103] J. Yang, R. Ding, W. Deng, Z. Wang, and X. Qi, “Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 823–19 832.

[104] P. Shyam, W. Jaśkowski, and F. Gomez, “Model-based active exploration,” in *International conference on machine learning*. PMLR, 2019, pp. 5779–5788.

[105] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, “Learning to explore using active neural slam,” in *8th International Conference on Learning Representations, ICLR 2020*, 2020.

[106] H. Liu and P. Abbeel, “Behavior from the void: Unsupervised active pre-training,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 459–18 473, 2021.

[107] M. Xu, G. Jiang, W. Liang, C. Zhang, and Y. Zhu, “Active reasoning in an open-world environment,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 11 716–11 736, 2023.

[108] A. Russo and A. Proutiere, “Model-free active exploration in reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 54 740–54 753, 2023.

[109] A. Somayazulu, S. Majumder, C. Chen, and K. Grauman, “Activerir: Active audio-visual exploration for acoustic environment modeling,” in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 13 830–13 836.

[110] L. Zhang, G. Yang, and B. C. Stadie, “World model as a graph: Learning latent landmarks for planning,” in *International conference on machine learning*. PMLR, 2021, pp. 12 611–12 620.

[111] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, “Llm-planner: Few-shot grounded planning for embodied agents with large language models,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 2998–3009.

[112] X. Zhou, Z. Wang, H. Ye, C. Xu, and F. Gao, “Ego-planner: An esdf-free gradient-based local planner for quadrotors,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 478–485, 2020.

[113] S. Qiao, N. Zhang, R. Fang, Y. Luo, W. Zhou, Y. E. Jiang, C. Lv, and H. Chen, “Autoact: Automatic agent learning from scratch for qa via self-planning,” *arXiv preprint arXiv:2401.05268*, 2024.

[114] L. Yang, Z. Yu, C. Meng, M. Xu, S. Ermon, and B. Cui, “Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms,” in *Forty-first International Conference on Machine Learning*, 2024.

[115] D. An, H. Wang, W. Wang, Z. Wang, Y. Huang, K. He, and L. Wang, “Etpnav: Evolving topological planning for vision-language navigation in continuous environments,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[116] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, “Reflexion: Language agents with verbal reinforcement learning,” *Advances in Neural Information Processing Systems*, 2023.

[117] Z. Liu, A. Bahety, and S. Song, “Reflect: Summarizing robot experiences for failure explanation and correction,” *arXiv preprint arXiv:2306.15724*, 2023.

[118] Z. Yang, J. Liu, P. Chen, A. Cherian, T. K. Marks, J. Le Roux, and C. Gan, “Rila: Reflective and imaginative language agent for zero-shot semantic audio-visual navigation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 251–16 261.

[119] Z. Li, Y. Xie, R. Shao, G. Chen, D. Jiang, and L. Nie, “Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks,” *arXiv preprint arXiv:2408.03615*, 2024.

[120] P. Yuan, A. Ma, Y. Yao, H. Yao, M. Tomizuka, and M. Ding, “Remac: Self-reflective and self-evolving multi-agent collaboration for long-horizon robot manipulation,” *arXiv preprint arXiv:2503.22122*, 2025.

[121] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian *et al.*, “Do as i can, not as i say: Grounding language in robotic affordances,” in *Conference on robot learning*. PMLR, 2023, pp. 287–318.

[122] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg *et al.*, “A generalist agent,” *arXiv preprint arXiv:2205.06175*, 2022.

[123] Y. Mu, Q. Zhang, M. Hu, W. Wang, M. Ding, J. Jin, B. Wang, J. Dai, Y. Qiao, and P. Luo, “Embodiedgpt: Vision-language pre-training via embodied chain of thought,” *Advances in Neural Information Processing Systems*, 2023.

[124] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei, “Kosmos-2: Grounding multimodal large language models to the world,” *arXiv preprint arXiv:2306.14824*, 2023.

[125] 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,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[126] L. Fan, G. Wang, Y. Jiang, A. Mandlekar, Y. Yang, H. Zhu, A. Tang, D.-A. Huang, Y. Zhu, and A. Anandkumar, “Minedojo: Building open-ended embodied agents with internet-scale knowledge,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 343–18 362, 2022.

[127] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, “Openvla: An open-source vision-language-action model,” *arXiv preprint arXiv:2406.09246*, 2024.

[128] W. Hong, W. Wang, Q. Lv, J. Xu, W. Yu, J. Ji, Y. Wang, Z. Wang, Y. Dong, M. Ding *et al.*, “Cogagent: A visual language model for gui agents,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 281–14 290.

[129] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu *et al.*, “Octo: An open-source generalist robot policy,” *arXiv preprint arXiv:2405.12213*, 2024.

[130] R. Doshi, H. Walke, O. Mees, S. Dasari, and S. Levine, “Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation,” *arXiv preprint arXiv:2408.11812*, 2024.

[131] L. Wang, X. Chen, J. Zhao, and K. He, “Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 124 420–124 450, 2024.

[132] J. Ho and S. Ermon, “Generative adversarial imitation learning,” *Advances in neural information processing systems*, vol. 29, 2016.

[133] N. Baram, O. Anschel, I. Caspi, and S. Mannor, “End-to-end differentiable adversarial imitation learning,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 390–399.

[134] S. Suo, S. Regalado, S. Casas, and R. Urtasun, “Trafficsim: Learning to simulate realistic multi-agent behaviors,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 400–10 409.

[135] Q. Zhang, Y. Gao, Y. Zhang, Y. Guo, D. Ding, Y. Wang, P. Sun, and D. Zhao, “Trajgen: Generating realistic and diverse trajectories with reactive and feasible agent behaviors for autonomous driving,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 24 474–24 487, 2022.

[136] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun *et al.*, “Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation,” *Conference on Robot Learning*, pp. 80–93, 2023.

[137] J. Lu, B. Pan, J. Chen, Y. Feng, J. Hu, Y. Peng, and W. Chen, “Agentlens: Visual analysis for agent behaviors in llm-based autonomous systems,” *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[138] B. Chen, J. Kang, P. Zhong, Y. Liang, Y. Sheng, and J. Wang, “Embodied contrastive learning with geometric consistency and behavioral awareness for object navigation,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 4776–4785.

[139] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, “Monotonic value function factorisation for deep multi-agent reinforcement learning,” *Journal of Machine Learning Research*, vol. 21, no. 178, pp. 1–51, 2020.

[140] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, and Y. Yi, “Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning,” in *International conference on machine learning*. PMLR, 2019, pp. 5887–5896.

[141] J. Wang, Z. Ren, T. Liu, Y. Yu, and C. Zhang, “Qplex: Duplex dueling multi-agent q-learning,” *arXiv preprint arXiv:2008.01062*, 2020.

[142] M. Wen, J. Kuba, R. Lin, W. Zhang, Y. Wen, J. Wang, and Y. Yang, "Multi-agent reinforcement learning is a sequence modeling problem," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 509–16 521, 2022.

[143] 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.

[144] W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C. Qian, C.-M. Chan, Y. Qin, Y. Lu, R. Xie *et al.*, "Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents," *arXiv preprint arXiv:2308.10848*, vol. 2, no. 4, p. 6, 2023.

[145] S. Hong, X. Zheng, J. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou *et al.*, "Metagpt: Meta programming for multi-agent collaborative framework," *arXiv preprint arXiv:2308.00352*, vol. 3, no. 4, p. 6, 2023.

[146] H. Zhang, Z. Wang, Q. Lyu, Z. Zhang, S. Chen, T. Shu, B. Dariush, K. Lee, Y. Du, and C. Gan, "Combo: Compositional world models for embodied multi-agent cooperation," *arXiv preprint arXiv:2404.10775*, 2024.

[147] A. Clark, "An embodied cognitive science?" *Trends in cognitive sciences*, vol. 3, no. 9, pp. 345–351, 1999.

[148] R. W. Gibbs Jr, *Embodiment and cognitive science*. Cambridge University Press, 2005.

[149] T. Ziemke, "What's that thing called embodiment?" in *Proceedings of the 25th Annual Cognitive Science Society*. Psychology Press, 2013, pp. 1305–1310.

[150] X. Liu, H. Palacios, and C. Muise, "Egocentric planning for scalable embodied task achievement," *Advances in Neural Information Processing Systems*, vol. 36, pp. 54 586–54 613, 2023.

[151] K. C. Stocking, A. Gopnik, and C. Tomlin, "From robot learning to robot understanding: Leveraging causal graphical models for robotics," in *Conference on Robot Learning*. PMLR, 2022, pp. 1776–1781.

[152] P. Sermanet, T. Ding, J. Zhao, F. Xia, D. Dwibedi, K. Gopalakrishnan, C. Chan, G. Dulac-Arnold, S. Maddineni, N. J. Joshi *et al.*, "Robovqa: Multimodal long-horizon reasoning for robotics," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 645–652.

[153] J. Zhang, L. Tang, Y. Song, Q. Meng, H. Qian, J. Shao, W. Song, S. Zhu, and J. Gu, "Fltrnn: Faithful long-horizon task planning for robotics with large language models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6680–6686.

[154] S. Nayak, A. Morrison Orozco, M. Have, J. Zhang, V. Thirumalai, D. Chen, A. Kapoor, E. Robinson, K. Gopalakrishnan, J. Harrison *et al.*, "Long-horizon planning for multi-agent robots in partially observable environments," *Advances in Neural Information Processing Systems*, vol. 37, pp. 67 929–67 967, 2024.

[155] X. Zhang, H. Qin, F. Wang, Y. Dong, and J. Li, "Lamma-p: Generalizable multi-agent long-horizon task allocation and planning with lm-driven pddl planner," *arXiv preprint arXiv:2409.20560*, 2024.

[156] A. Padmakumar, J. Thomason, A. Shrivastava, P. Lange, A. Narayan-Chen, S. Gella, R. Piramuthu, G. Tur, and D. Hakkani-Tur, "Teach: Task-driven embodied agents that chat," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2017–2025.

[157] Z. Zhang, X. Bo, C. Ma, R. Li, X. Chen, Q. Dai, J. Zhu, Z. Dong, and J.-R. Wen, "A survey on the memory mechanism of large language model based agents," *arXiv preprint arXiv:2404.13501*, 2024.

[158] B. Bakker, "Reinforcement learning with long short-term memory," *Advances in neural information processing systems*, vol. 14, 2001.

[159] D. Ramani, "A short survey on memory based reinforcement learning," *arXiv preprint arXiv:1904.06736*, 2019.

[160] G. Zhu, Z. Lin, G. Yang, and C. Zhang, "Episodic reinforcement learning with associative memory," in *International Conference on Learning Representations*, 2020.

[161] A. Khan, C. Zhang, N. Atanasov, K. Karydis, V. Kumar, and D. D. Lee, "Memory augmented control networks," *arXiv preprint arXiv:1709.05706*, 2017.

[162] A. Pritzel, B. Uria, S. Srinivasan, A. P. Badia, O. Vinyals, D. Hassabis, D. Wierstra, and C. Blundell, "Neural episodic control," in *International conference on machine learning*. PMLR, 2017, pp. 2827–2836.

[163] Z. Xu, K. Wu, J. Wen, J. Li, N. Liu, Z. Che, and J. Tang, "A survey on robotics with foundation models: toward embodied ai," *arXiv preprint arXiv:2402.02385*, 2024.

[164] L. Ren, J. Dong, S. Liu, L. Zhang, and L. Wang, "Embodied intelligence toward future smart manufacturing in the era of ai foundation model," *IEEE/ASME Transactions on Mechatronics*, 2024.

[165] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. Hausman *et al.*, "Foundation models in robotics: Applications, challenges, and the future," *The International Journal of Robotics Research*, p. 02783649241281508, 2023.

[166] K. Li, B. Yu, Q. Zheng, Y. Zhan, Y. Zhang, T. Zhang, Y. Yang, Y. Chen, L. Sun, Q. Cao *et al.*, "Muep: A multimodal benchmark for embodied planning with foundation models [c]," in *Intemational Joint Conferences on Artificial Intelligence. IJCAI*, 2024, pp. 129–138.

[167] R. Dang, Y. Yuan, W. Zhang, Y. Xin, B. Zhang, L. Li, L. Wang, Q. Zeng, X. Li, and L. Bing, "Ecbench: Can multi-modal foundation models understand the egocentric world? a holistic embodied cognition benchmark," *arXiv preprint arXiv:2501.05031*, 2025.

[168] M. Zhang, X. Fu, J. Hao, P. Han, H. Zhang, L. Shi, H. Tang, and Y. Zheng, "Mfe-etp: A comprehensive evaluation benchmark for multi-modal foundation models on embodied task planning," *arXiv preprint arXiv:2407.05047*, 2024.

[169] F. Sun, R. Chen, T. Ji, Y. Luo, H. Zhou, and H. Liu, "A comprehensive survey on embodied intelligence: Advancements, challenges, and future perspectives," *CAAI Artificial Intelligence Research*, vol. 3, 2024.

[170] L. Jin and L. Jia, "Embodied world models emerge from navigational task in open-ended environments," *arXiv preprint arXiv:2504.11419*, 2025.

[171] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019.

[172] A. Ajoudani, A. M. Zanchettin, S. Ivaldi, A. Albu-Schäffer, K. Kosuge, and O. Khatib, "Progress and prospects of the human–robot collaboration," *Autonomous robots*, vol. 42, pp. 957–975, 2018.

[173] R. Kelly, "A tuning procedure for stable pid control of robot manipulators," *Robotica*, vol. 13, no. 2, pp. 141–148, 1995.

[174] P. Rocco, "Stability of pid control for industrial robot arms," *IEEE transactions on robotics and automation*, vol. 12, no. 4, pp. 606–614, 2002.

[175] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 6, pp. 2042–2062, 2017.

[176] B. Singh, R. Kumar, and V. P. Singh, "Reinforcement learning in robotic applications: a comprehensive survey," *Artificial Intelligence Review*, vol. 55, no. 2, pp. 945–990, 2022.

[177] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, "A survey on vision-language-action models for embodied ai," *arXiv preprint arXiv:2405.14093*, 2024.

[178] B. Jiang, Y. Xie, X. Wang, W. J. Su, C. J. Taylor, and T. Mallick, "Multi-modal and multi-agent systems meet rationality: A survey," *ICML Workshop on LLMs and Cognition*, 2024.

[179] Z. Zhu, M. Liu, L. Mao, B. Kang, M. Xu, Y. Yu, S. Ermon, and W. Zhang, "Madiff: Offline multi-agent learning with diffusion models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 4177–4206, 2024.

[180] M. Ma and L. Cheng, "A human-robot collaboration controller utilizing confidence for disagreement adjustment," *IEEE Transactions on Robotics*, 2024.

[181] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, "Smoothquant: Accurate and efficient post-training quantization for large language models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 38 087–38 099.

[182] H. Wang, Z. Zhang, and S. Han, "Spatten: Efficient sparse attention architecture with cascade token and head pruning," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021, pp. 97–110.

[183] K. Wang, Z. Liu, Y. Lin, J. Lin, and S. Han, "Haq: Hardware-aware automated quantization with mixed precision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8612–8620.

[184] C. Lattner and V. Adve, "Llvm: A compilation framework for lifelong program analysis & transformation," in *International symposium on code generation and optimization, 2004. CGO 2004.* IEEE, 2004, pp. 75–86.

[185] Q. Sun, X. Zhang, H. Geng, Y. Zhao, Y. Bai, H. Zheng, and B. Yu, "Gtuner: Tuning dnn computations on gpu via graph attention network," in *Proceedings of the 59th ACM/IEEE Design Automation Conference*, 2022, pp. 1045–1050.

[186] J. Cai, Z. Wu, S. Peng, Y. Wei, Z. Tan, G. Shi, M. Gao, and K. Ma, "Gemini: Mapping and architecture co-exploration for large-scale dnn chiplet accelerators," in *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2024, pp. 156–171.

[187] J. Liu, S. Zeng, L. Ding, W. Soedarmadji, H. Zhou, Z. Wang, J. Li, J. Li, Y. Dai, K. Wen *et al.*, "Flightvgm: Efficient video generation model inference with online sparsification and hybrid precision on fpgas," in *Proceedings of the 2025 ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, 2025, pp. 2–13.

[188] J. Qin, T. Xia, C. Tan, J. Zhang, and S. Q. Zhang, "Picachu: Plug-in cgra handling upcoming nonlinear operations in llms," in *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 2025, pp. 845–861.

[189] H. Genc, S. Kim, A. Amid, A. Haj-Ali, V. Iyer, P. Prakash, J. Zhao, D. Grubb, H. Liew, H. Mao *et al.*, "Gemmini: Enabling systematic deep-learning architecture evaluation via full-stack integration," in *2021 58th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2021, pp. 769–774.

[190] Y. Lin, H. Tang, S. Yang, Z. Zhang, G. Xiao, C. Gan, and S. Han, "Qserve: W4a8kv4 quantization and system co-design for efficient llm serving," *arXiv preprint arXiv:2405.04532*, 2024.

[191] M. Wang, Y. Meng, C. Tang, W. Zhang, Y. Qin, Y. Yao, Y. Li, T. Feng, X. Wang, X. Guan *et al.*, "Jaq: Joint efficient architecture design and low-bit quantization with hardware-software co-exploration," *arXiv preprint arXiv:2501.05339*, 2025.

[192] L. Zou, W. Zhao, S. Yin, C. Bai, Q. Sun, and B. Yu, "Bie: bi-exponent block floating-point for large language models quantization," in *Forty-first International Conference on Machine Learning*, 2024.

[193] 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," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9339–9347.

[194] T. Mu, Z. Ling, F. Xiang, D. Yang, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su, "Maniskill: Learning-from-demonstrations benchmark for generalizable manipulation skills," *International Conference on Learning Representations*, 2021.

[195] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012.

[196] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics*, 2017.

[197] T. Feng, X. Wang, F. Han, L. Zhang, and W. Zhu, "U2udata: A large-scale cooperative perception dataset for swarm uavs autonomous flight," in *ACM Multimedia 2024*, 2024.

[198] F. Han, L. Zhang, X. Wang, K.-A. Zhao, Y. Zhong, Z. Su, T. Feng, and W. Zhu, "U2usim - a uav telepresence simulation platform with multi-agent sensing and dynamic environment," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, p. 11258–11260.

[199] D. Driess, Y. Huang, A. Zeng, P. Florence, F. Tombari, A. Wahid, Q. Vuong, K. Hausman, M. Heo, U. Lee *et al.*, "Vision-language-action models: Unified frameworks for embodied execution," in *Conference on Robot Learning*, 2023, pp. 1234–1245.

[200] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[201] Y. Zhang, S. Yang, C. Bai, F. Wu, X. Li, Z. Wang, and X. Li, "Towards efficient llm grounding for embodied multi-agent collaboration," *arXiv preprint arXiv:2405.14314*, 2024.

[202] T. Wang, X. Mao, C. Zhu, R. Xu, R. Lyu, P. Li, X. Chen, W. Zhang, K. Chen, T. Xue *et al.*, "Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[203] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-drive world models for autonomous driving," in *European Conference on Computer Vision*. Springer, 2024, pp. 55–72.

[204] B. Zhao, Z. Wang, J. Fang, C. Gao, F. Man, J. Cui, X. Wang, X. Chen, Y. Li, and W. Zhu, "Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning," *arXiv preprint arXiv:2504.12680*, 2025.

[205] Z. Song, X. Wang, Z. Qian, H. Chen, L. Huang, H. Xue, and W. Zhu, "Modularized self-reflected video reasoner for multimodal llm with application to video question answering," in *Forty-second International Conference on Machine Learning*.

[206] C. Ge, X. Wang, Z. Zhang, H. Chen, J. Fan, L. Huang, H. Xue, and W. Zhu, "Dynamic mixture of curriculum lora experts for continual multimodal instruction tuning," *arXiv preprint arXiv:2506.11672*, 2025.

[207] B. Huang, F. He, Q. Wang, H. Chen, G. Li, Z. Feng, X. Wang, and W. Zhu, "Neighbor does matter: Curriculum global positive-negative sampling for vision-language pre-training," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 8005–8014.

[208] Y. Zhang, X. Wang, H. Chen, J. Fan, W. Wen, H. Xue, H. Mei, and W. Zhu, "Large language model with curriculum reasoning for visual concept recognition," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6269–6280.

[209] Z. Pan, X. Wang, Y. Zhang, H. Chen, K. M. Cheng, Y. Wu, and W. Zhu, "Modular-cam: Modular dynamic camera-view video generation with llm," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025, pp. 6363–6371.

[210] B. Huang, X. Wang, H. Chen, Z. Song, and W. Zhu, "Vtimellm: Empower llm to grasp video moments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 271–14 280.

[211] X. Wang, Z. Pan, Y. Zhou, H. Chen, C. Ge, and W. Zhu, "Curriculum co-disentangled representation learning across multiple environments for social recommendation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 36 174–36 192.

[212] X. Wang, H. Chen, Z. Pan, Y. Zhou, C. Guan, L. Sun, and W. Zhu, "Automated disentangled sequential recommendation with large language models," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–29, 2025.

[213] X. Wang, H. Chen, Y. Zhou, J. Ma, and W. Zhu, "Disentangled representation learning for recommendation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 408–424, 2022.

[214] J. Thomason, M. Shridhar, Y. Bisk, C. Paxton, and L. Zettlemoyer, "Language grounding with 3d objects," in *Conference on robot learning*. PMLR, 2022, pp. 1691–1701.

[215] S. Wang, M. Han, Z. Jiao, Z. Zhang, Y. N. Wu, S.-C. Zhu, and H. Liu, "Llmˆ 3: Large language model-based task and motion planning with motion failure reasoning," pp. 12 086–12 092, 2024.

[216] H. Chen, X. Wang, X. Lan, H. Chen, X. Duan, J. Jia, and W. Zhu, "Curriculum-listener: Consistency-and complementarity-aware audio-enhanced temporal sentence grounding," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3117–3128.

[217] X. Wang, Z. Pan, H. Chen, and W. Zhu, "Divico: Disentangled visual token compression for efficient large vision-language model," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[218] J. Hong, R. Choi, and J. J. Leonard, "Learning from feedback: Semantic enhancement for object slam using foundation models," *arXiv preprint arXiv:2411.06752*, 2024.

[219] J. Yang, R. Tan, Q. Wu, R. Zheng, B. Peng, Y. Liang, Y. Gu, M. Cai, S. Ye, J. Jang *et al.*, "Magma: A foundation model for multimodal ai agents," *arXiv preprint arXiv:2502.13130*, 2025.

[220] H. Chen, X. Wang, Y. Zhang, Y. Zhou, Z. Zhang, S. Tang, and W. Zhu, "Disenstudio: Customized multi-subject text-to-video generation with disentangled spatial control," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 3637–3646.

[221] H. Chen, Y. Zhang, S. Wu, X. Wang, X. Duan, Y. Zhou, and W. Zhu, "Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation," *arXiv preprint arXiv:2305.03374*, 2023.

[222] S. Wang, M. Han, Z. Jiao, Z. Zhang, Y. N. Wu, S.-C. Zhu, and H. Liu, "Llmˆ 3: Large language model-based task and motion planning with motion failure reasoning," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[223] J. Yue, X. Xu, B. F. Karlsson, and Z. Lu, "Mllm as retriever: Interactively learning multimodal retrieval for embodied agents," *arXiv preprint arXiv:2410.03450*, 2024.

[224] Y. Yao, X. Wang, Y. Qin, Z. Zhang, W. Zhu, and H. Mei, "Data-augmented curriculum graph neural architecture search under distribution shifts," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 15, 2024, pp. 16 433–16 441.

[225] Y. Qin, X. Wang, Z. Zhang, H. Chen, and W. Zhu, "Multi-task graph neural architecture search with task-aware collaboration and curriculum," *Advances in neural information processing systems*, vol. 36, pp. 24 879–24 891, 2023.

[226] X. Lan, Y. Yuan, H. Chen, X. Wang, Z. Jie, L. Ma, Z. Wang, and W. Zhu, "Curriculum multi-negative augmentation for debiased video grounding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1213–1221.

[227] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, 2024.

[228] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang *et al.*, "Qwen2. 5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[229] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[230] X. Wang, Y. Chen, L. Yuan, Y. Zhang, Y. Li, H. Peng, and H. Ji, "Executable code actions elicit better llm agents," in *Forty-first International Conference on Machine Learning*, 2024.

[231] D. Wu, X. Wei, G. Chen, H. Shen, X. Wang, W. Li, and B. Jin, "Generative multi-agent collaboration in embodied ai: A systematic review," *arXiv preprint arXiv:2502.11518*, 2025.

[232] W. Zhao, P. Ding, M. Zhang, Z. Gong, S. Bai, H. Zhao, and D. Wang, "Vlas: Vision-language-action model with speech instructions for customized robot manipulation," *arXiv preprint arXiv:2502.13508*, 2025.

[233] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, 2023.

[234] J. W. Forrester, "Counterintuitive behavior of social systems," *Theory and decision*, vol. 2, no. 2, pp. 109–140, 1971.

[235] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 619–15 629.

[236] R. Yang, P. Srivastava, and S. Mandt, "Diffusion probabilistic modeling for video generation," *Entropy*, vol. 25, no. 10, p. 1469, 2023.

[237] J. Cho, F. D. Puspitasari, S. Zheng, J. Zheng, L.-H. Lee, T.-H. Kim, C. S. Hong, and C. Zhang, "Sora as an agi world model? a complete survey on text-to-video generation," *arXiv preprint arXiv:2403.05131*, 2024.

[238] D. Li, Y. Fang, Y. Chen, S. Yang, S. Cao, J. Wong, M. Luo, X. Wang, H. Yin, J. E. Gonzalez *et al.*, "Worldmodelbench: Judging video generation models as world models," *arXiv preprint arXiv:2502.20694*, 2025.

[239] J. Xiang, T. Tao, Y. Gu, T. Shu, Z. Wang, Z. Yang, and Z. Hu, "Language models meet world models: Embodied experiences enhance language models," *Advances in neural information processing systems*, vol. 36, pp. 75 392–75 412, 2023.

[240] P. Mazzaglia, T. Verbelen, B. Dhoedt, A. Courville, and S. Rajeswar, "Genrl: Multimodal-foundation world models for generalization in embodied agents," *Advances in Neural Information Processing Systems*, vol. 37, pp. 27 529–27 555, 2024.

[241] T. Gupta, W. Gong, C. Ma, N. Pawlowski, A. Hilmkil, M. Scetbon, M. Rigter, A. Famoti, A. J. Llorens, J. Gao *et al.*, "The essential role of causality in foundation world models for embodied ai," *arXiv preprint arXiv:2402.06665*, 2024.

[242] N. Hansen, J. SV, V. Sobal, Y. LeCun, X. Wang, and H. Su, "Hierarchical world models as visual whole-body humanoid controllers," *arXiv preprint arXiv:2405.18418*, 2024.

[243] E. Alonso, A. Jelley, V. Micheli, A. Kanervisto, A. J. Storkey, T. Pearce, and F. Fleuret, "Diffusion for world modeling: Visual details matter in atari," *Advances in Neural Information Processing Systems*, vol. 37, pp. 58 757–58 791, 2024.

[244] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas, "Revisiting feature prediction for learning visual representations from video," 2024.

[245] A. Bardes, J. Ponce, and Y. LeCun, "Mc-jepa: A joint-embedding predictive architecture for self-supervised learning of motion and content features," 2023.

[246] Z. Fei, M. Fan, and J. Huang, "A-jepa: Joint-embedding predictive architecture can listen," 2024.

[247] A. Banino, A. P. Badia, R. Koster, M. J. Chadwick, V. F. Zambaldi, D. Hassabis, C. Barry, M. M. Botvinick, D. Kumaran, and C. Blundell, "Memo: A deep network for flexible combination of episodic models," 2020.

[248] V. Micheli, E. Alonso, and F. Fleuret, "Transformers are sample-efficient world models," in *ICLR*, 2023.

[249] M. Xu, W. Dai, C. Liu, X. Gao, W. Lin, G.-J. Qi, and H. Xiong, "Spatial-temporal transformer networks for traffic flow forecasting," 2021.

[250] 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," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[251] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra *et al.*, "Language is not all you need: Aligning perception with language models," *Advances in Neural Information Processing Systems*, 2023.

[252] OpenAI, "Gpt-4 technical report," *arXiv:2303.08774*, 2023.

[253] 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," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9493–9500.

[254] H.-Y. Tung, M. Ding, Z. Chen, D. Bear, C. Gan, J. Tenenbaum, D. Yamins, J. Fan, and K. Smith, "Physion++: Evaluating physical scene understanding that requires online inference of different physical properties," *Advances in Neural Information Processing Systems*, vol. 36, pp. 67 048–67 068, 2023.

[255] X. Shi, D. Li, P. Zhao, Q. Tian, Y. Tian, Q. Long, C. Zhu, J. Song, F. Qiao, L. Song *et al.*, "Are we ready for service robots? the openloris-scene datasets for lifelong slam," pp. 3139–3145, 2020.

[256] R. Wadhawan, H. Bansal, K.-W. Chang, and N. Peng, "Contextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models," *arXiv preprint arXiv:2401.13311*, 2024.

[257] H. Sun, Y. Zhuang, L. Kong, B. Dai, and C. Zhang, "Adaplanner: Adaptive planning from feedback with language models," *Advances in neural information processing systems*, vol. 36, pp. 58 202–58 245, 2023.

[258] W. Li, X. Meng, Z. Zhao, Z. Liu, C. Chen, and H. Wang, "Lot: A transformer-based approach based on channel state information for indoor localization," *IEEE Sensors Journal*, vol. 23, no. 22, pp. 28 205–28 219, 2023.

[259] Y. Shen, H. Liu, P. Liu, R. Xia, T. Yao, Y. Sun, and T. Feng, "Detach: Cross-domain learning for long-horizon tasks via mixture of disentangled experts," *arXiv preprint arXiv:2508.07842*, 2025.

[260] J.-F. Yeh, K.-H. Hung, P.-C. Lo, C. M. Chung, T.-H. Wu, H.-T. Su, Y.-T. Chen, and W. Hsu, "Aed: Adaptable error detection for few-shot imitation policy," *Advances in Neural Information Processing Systems*, vol. 37, pp. 136 805–136 836, 2024.

[261] G. Canal, C. Torras, and G. Alenyà, "Are preferences useful for better assistance? a physically assistive robotics user study," *ACM Transactions on Human-Robot Interaction*, vol. 10, no. 4, pp. 1–19, 2021.

[262] R. Kachouie, S. Sedighadeli, R. Khosla, and M.-T. Chu, "Socially assistive robots in elderly care: a mixed-method systematic literature review," *International Journal of Human-Computer Interaction*, vol. 30, no. 5, pp. 369–393, 2014.

[263] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning*, 2022, pp. 991–1002.

[264] T. Feng, Q. Li, X. Wang, M. Wang, G. Li, and W. Zhu, "Multi-weather cross-view geo-localization using denoising diffusion models," in *Proceedings of the 2nd Workshop on UAVs in Multimedia: Capturing the World from a New Perspective*, 2024, pp. 35–39.

[265] T. Feng, X. Wang, F. Han, L. Zhang, and W. Zhu, "U2udata-2: A scalable swarm uavs autonomous flight dataset for long-horizon tasks," in *ArXiv*, 2025.

[266] Y. Zhang, Z. Ma, J. Li, Y. Qiao, Z. Wang, J. Chai, Q. Wu, M. Bansal, and P. Kordjamshidi, "Vision-and-language navigation today and tomorrow: A survey in the era of foundation models," *arXiv preprint arXiv:2407.07035*, 2024.

[267] F. Flammini, C. Alcaraz, E. Bellini, S. Marrone, J. Lopez, and A. Bondavalli, "Towards trustworthy autonomous systems: Taxonomies and future perspectives," *IEEE Transactions on Emerging Topics in Computing*, 2022.

[268] L. Chen, H. Liang, Y. Pan, and T. Li, "Human-in-the-loop consensus tracking control for uav systems via an improved prescribed performance approach," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 6, pp. 8380–8391, 2023.

[269] G. Lin, H. Li, C. K. Ahn, and D. Yao, "Event-based finite-time neural control for human-in-the-loop uav attitude systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 10 387–10 397, 2022.

[270] J. Pesonen, T. Hakala, V. Karjalainen, N. Koivumäki, L. Markelin, A.-M. Raita-Hakola, J. Suomalainen, I. Pölönen, and E. Honkavaara, "Detecting wildfires on uavs with real-time segmentation trained by larger teacher models," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 5166–5176.

[271] S. Javaid, H. Fahim, B. He, and N. Saeed, "Large language models for uavs: Current state and pathways to the future," *IEEE Open Journal of Vehicular Technology*, 2024.

[272] Q. Chen, T. Wang, Z. Yang, H. Li, R. Lu, Y. Sun, B. Zheng, and C. Yan, "Sdpl: Shifting-dense partition learning for uav-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 11 810–11 824, 2024.

[273] X. Zhou, X. Wen, Z. Wang, Y. Gao, H. Li, Q. Wang, T. Yang, H. Lu, Y. Cao, C. Xu *et al.*, "Swarm of micro flying robots in the wild," *Science Robotics*, vol. 7, no. 66, p. eabm5954, 2022.

[274] Q. Chen, T. Wang, R. Lu, Y. Liu, B. Zheng, and Z. Zheng, "Scale-adaptive uav geo-localization via height-aware partition learning," *arXiv preprint arXiv:2412.11535*, 2024.

[275] W. Zheng, J. Yang, J. Chen, J. He, P. Li, D. Sun, C. Chen, and X. Meng, "Cross-temporal knowledge injection with color distribution normalization for remote sensing change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[276] H. Qin, J. Xiao, D. Ge, L. Xin, J. Gao, S. He, H. Hu, and J. G. Carlsson, "Jd. com: Operations research algorithms drive intelligent warehouse robots to work," *INFORMS Journal on Applied Analytics*, vol. 52, no. 1, pp. 42–55, 2022.

[277] R. Bogue, "The role of robots in logistics," *Industrial Robot: the international journal of robotics research and application*, vol. 51, no. 3, pp. 381–386, 2024.

[278] I. F. A. Prawira, A. H. Habbe, I. Muda, R. M. Hasibuan, and A. Umbrajkaar, "Robot as staff: Robot for alibaba e-commerce warehouse process," in *2023 International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2023, pp. 1619–1623.

[279] H. Liu, X. Huang, J. Gu, J. Shi, N. He, and T. Feng, "Tcdformer-based momentum transfer model for long-term sports prediction," *Expert Systems with Applications*, p. 128310, 2025.

[280] T. Feng, Q. Qi, J. Wang, J. Liao, and J. Liu, "Timely and accurate bitrate switching in http adaptive streaming with date-driven i-frame prediction," *IEEE Transactions on Multimedia*, vol. 25, pp. 3753–3762, 2022.

[281] T. Feng, H. Sun, Q. Qi, J. Wang, and J. Liao, "Vabis: Video adaptation bitrate system for time-critical live streaming," *IEEE Transactions on Multimedia*, vol. 22, no. 11, pp. 2963–2976, 2019.

[282] T. Feng, Q. Qi, L. Guo, and J. Wang, "Meta-uad: A meta-learning scheme for user-level network traffic anomaly detection," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[283] T. Jin and X. Han, "Robotic arms in precision agriculture: A comprehensive review of the technologies, applications, challenges, and future prospects," *Computers and Electronics in Agriculture*, vol. 221, p. 108938, 2024.

[284] H. Xue, J. Ren, W. Chen, G. Zhang, Y. Fang, G. Gu, H. Xu, and C. Lu, "Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation," *arXiv preprint arXiv:2503.02881*, 2025.

[285] J. Ye, X. Wen, Y. Wei, Y. Xu, K. Liu, and H. Shan, "Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*. IEEE, 2023, pp. 1–5.

[286] J. Ye, B. Cao, and H. Shan, "Emotional face-to-speech," in *International Conference on Machine Learning*, 2025.

[287] J. Ye, Y. Wei, X. Wen, C. Ma, Z. Huang, K. Liu, and H. Shan, "Emo-dna: Emotion decoupling and alignment learning for cross-corpus speech emotion recognition," in *Proceedings of the 31st ACM International Conference on Multimedia, MM*, 2023, pp. 5956–5965.

[288] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science robotics*, vol. 3, no. 21, p. eaat5954, 2018.

[289] G. Lampropoulos, "Social robots in education: Current trends and future perspectives," *Information*, vol. 16, no. 1, p. 29, 2025.

[290] 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.

[291] L. Qunzhi, M. Chao, P. Jing, W. Jinqiao, W. Zhiliang, Z. Guibo, L. Yongjian, D. Boyuan, and W. Jie, "Research on the application of embodied intelligence technology in space exploration," in *2024 IEEE International Conference on Control Science and Systems Engineering (ICCSSE)*. IEEE, 2024, pp. 149–155.

Tongtong Feng is currently a postdoctoral researcher at the Department of Computer Science and Technology, Tsinghua University. He got his Ph.D. degree in Computer Science and Technology from Beijing University of Posts and Telecommunications. His research interests include Embodied AI, World Model, and Multimedia Intelligence. He has published over 15 high-quality research papers in top journals and conferences, including IEEE TMM, ESWA, ACM Multimedia and AAAI, etc. He got the Best Paper Nomination of ACM Multimedia 2024.

Xin Wang is currently an Associate Professor at the Department of Computer Science and Technology, Tsinghua University. He got both his Ph.D. and B.E degrees in Computer Science and Technology from Zhejiang University, China. He also holds a Ph.D. degree in Computing Science from Simon Fraser University, Canada. His research interests include multimedia intelligence, machine learning and its applications in multimedia big data analysis. He has published over 150 high-quality research papers in top journals and conferences including IEEE TPAMI, IEEE TKDE, ACM TOIS, ICML, NeurIPS, ACM KDD, ACM Web Conference, ACM SIGIR and ACM Multimedia etc., winning three best paper awards. He is the recipient of 2020 ACM China Rising Star Award, 2022 IEEE TCMC Rising Star Award and 2023 DAMO Academy Young Fellow.

Yu-Gang Jiang (Fellow, IEEE) received the PhD degree in Computer Science from City University of Hong Kong in 2009 and worked as a Postdoctoral Research Scientist at Columbia University, New York, during 2009-2011. He is currently a Distinguished Professor of Computer Science at Fudan University, Shanghai, China. His research lies in the areas of multimedia, computer vision, embodied AI and trustworthy AI. His research has led to the development of innovative AI tools that have been used in many practical applications like defect detection for high-speed railway infrastructures. His open-source video analysis toolkits and datasets such as CU-VIREO374, CCV, THUMOS, FCVID and WildDeepfake have been widely used in both academia and industry. He currently serves as Chair of ACM Shanghai Chapter and Associate Editor of several international journals. For contributions to large-scale and trustworthy video analysis, he was elected to Fellow of IEEE, IAPR and CCF.

Wenwu Zhu is currently a Professor in the Department of Computer Science and Technology at Tsinghua University, the Vice Dean of Beijing National Research Center for Information Science and Technology. Prior to his current post, he was a Senior Researcher and Research Manager at Microsoft Research Asia. He was the Chief Scientist and Director at Intel Research China from 2004 to 2008. He worked at Bell Labs New Jersey as Member of Technical Staff during 1996-1999. He received his Ph.D. degree from New York University in 1996.

His research interests include graph machine learning, curriculum learning, data-driven multimedia, big data. He has published over 400 referred papers, and is inventor of over 100 patents. He received ten Best Paper Awards, including ACM Multimedia 2012 and IEEE Transactions on Circuits and Systems for Video Technology in 2001 and 2019.

He serves as the EiC for IEEE Transactions on Circuits and Systems for Video Technology (2024-2025), the EiC for IEEE Transactions on Multimedia (2017-2019) and the Chair of the steering committee for IEEE Transactions on Multimedia (2020-2022). He serves as General Co-Chair for ACM Multimedia 2018 and ACM CIKM 2019. He is an AAAS Fellow, IEEE Fellow, ACM Fellow, SPIE Fellow, and a member of Academia Europaea.