

# Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI

Yang Liu, *Member, IEEE*, Weixing Chen, Yongjie Bai, Xiaodan Liang, *Senior Member, IEEE*, Guanbin Li, *Member, IEEE*, Wen Gao, *Fellow, IEEE*, Liang Lin, *Fellow, IEEE*

**Abstract**—Embodied Artificial Intelligence (Embodied AI) is crucial for achieving Artificial General Intelligence (AGI) and serves as a foundation for various applications (e.g., intelligent mechatronics systems, smart manufacturing) that bridge cyberspace and the physical world. Recently, the emergence of Multi-modal Large Models (MLMs) and World Models (WMs) have attracted significant attention due to their remarkable perception, interaction, and reasoning capabilities, making them a promising architecture for embodied agents. In this survey, we give a comprehensive exploration of the latest advancements in Embodied AI. Our analysis firstly navigates through the forefront of representative works of embodied robots and simulators, to fully understand the research focuses and their limitations. Then, we analyze four main research targets: 1) embodied perception, 2) embodied interaction, 3) embodied agent, and 4) sim-to-real adaptation, covering state-of-the-art methods, essential paradigms, and comprehensive datasets. Additionally, we explore the complexities of MLMs in virtual and real embodied agents, highlighting their significance in facilitating interactions in digital and physical environments. Finally, we summarize the challenges and limitations of embodied AI and discuss potential future directions. We hope this survey will serve as a foundational reference for the research community. The associated project can be found at [https://github.com/HCPLab-SYSU/Embodied\\_AI\\_Paper\\_List](https://github.com/HCPLab-SYSU/Embodied_AI_Paper_List).

**Index Terms**—Embodied AI, Cyber Space, Physical World, Multi-modal Large Models, Agents, Mechatronic Intelligence

## I. INTRODUCTION

EMBODIED AI was initially proposed from the Embodied Turing Test by Alan Turing in 1950 [1] and has wide

This work is supported in part by the National Key R&D Program of China under Grant 2021ZD0111601; in part by the open research fund of Pengcheng Laboratory under Grant 2025KF1B0050; in part by the National Natural Science Foundation of China under Grant 62436009, Grant 62322608, and Grant 62301532; and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515011874 and Grant 2023A1515011530. (*Corresponding author: Liang Lin*)

Yang Liu, Weixing Chen, Yongjie Bai are with the School of Computer Science and Engineering, Sun Yat-sen University, China, and Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, China. (E-mail: liuy856@mail.sysu.edu.cn, chen867820261@gmail.com, baiyu8581@gmail.com)

Xiaodan Liang is with the Shenzhen Campus of Sun Yat-sen University, China, Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, China, and Peng Cheng Laboratory, Shenzhen, China. (E-mail: xdliang328@gmail.com)

Guanbin Li and Liang Lin are with the School of Computer Science and Engineering, Sun Yat-sen University, China, Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, China, and Peng Cheng Laboratory, Shenzhen, China. (E-mail: liguanbin@mail.sysu.edu.cn, linliang@ieee.org)

Wen Gao is with the Peng Cheng Laboratory, Shenzhen, China, and also with the Institute of Digital Media, Peking University, Beijing, China. (E-mail: wgao@pku.edu.cn)

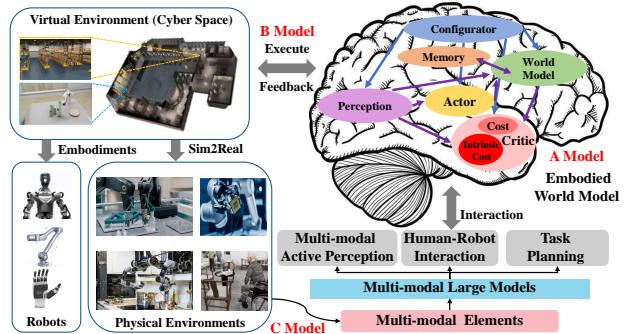


Fig. 1. The framework of the embodied agent based on MLMs and WMs, incorporates the ABC model, which stands for AI brain, Body, and Cross-modal sensors. The embodied agent is equipped with an embodied world model as the A model, enabling it to understand the virtual-physical environment. Through the C model, it actively perceives multi-modal elements, enhancing its situational awareness. Meanwhile, the B model endows the agent execute actions, and interact with humans while utilizing tools effectively.

applications including robotics, healthcare, and smart manufacturing. Embodied AI is designed to determine whether agents can display intelligence that is not just limited to solving abstract problems in a virtual environment (cyber space<sup>1</sup>), but that is also capable of navigating the complexity and unpredictability of the physical world. For example, embodied AI enhances mechatronic systems by integrating with physical components for adaptive, real-world interactions, which enables systems to learn, perceive, and execute autonomously, improving efficiency and functionality, as shown in Fig. 1. The agents in the cyber space are generally referred to as disembodied AI, while those in the physical space are embodied AI (Table I). Recent advances in Multi-modal Large Models (MLMs) have injected strong perception, interaction and planning capabilities to embodied models, to develop general-purpose embodied agents and robots that actively interact with virtual and physical environments. Therefore, the embodied agents are widely considered as the best carriers for MLMs. The recent representative embodied models are RT-2 [2] and RT-H [3]. Nevertheless, the capabilities of long-term memory, understanding complex intentions, and the decomposition of complex tasks are limited for current MLMs.

To achieve Artificial General Intelligence (AGI), the development of embodied AI stands as a fundamental avenue. Different from conversational agents like ChatGPT [4], embodied AI believes that the true AGI can be achieved by controlling physical embodiments and interacting with both

<sup>1</sup>The agents are the foundation of both disembodied and embodied AI. The agents can exist in both cyber and physical spaces, integrated with various entities. The entities include not only robots but also other devices.

TABLE I  
COMPARISON BETWEEN DISEMBODIED AI AND EMBODIED AI.

Type	Environment	Physical Entities	Description	Representative Agents
Disembodied AI	Cyber Space	No	Cognition and physical entities are disentangled	ChatGPT [4], RoboGPT [8]
Embodied AI	Physical Space	Robots, Cars, Other devices	Cognition is integrated into physical entities	RT-1 [9], RT-2 [10], RT-H [3]

simulated and physical environments [5], [6]. As we stand at the forefront of AGI-driven innovation, it is crucial to delve deeper into the realm of embodied AI, unraveling their complexities, evaluating their current developmental stage, and contemplating the potential trajectories they may follow in the future. Nowadays, embodied AI contains various key techniques across Computer Vision (CV), Natural Language Processing (NLP), and robotics, with the most representative being embodied perception, embodied interaction, embodied agents, and sim-to-real robotic control [7]. Therefore, it is imperative to capture the evolving landscape of embodied AI in the pursuit of AGI through a comprehensive survey.

Embodied agent is the most prominent basis of embodied AI. For an embodied task, the embodied agent must fully understand the human intention in language instructions, actively explore the surrounding environments, comprehensively perceive the multi-modal elements from both virtual and physical environments, and execute appropriate actions for complex tasks [11], [12], as shown in Fig. 1. The rapid progress in multi-modal models exhibits superior versatility, dexterity, and generalizability in complex environments compared to traditional deep reinforcement learning approaches. Pre-trained visual representations from state-of-the-art vision encoders [13], [14] provide precise estimations of object class, pose, and geometry, which makes the embodied models thoroughly perceive complex and dynamic environments. Powerful Large Language Models (LLMs) make robots better understand the linguistic instructions from humans. Promising MLMs give a feasible approach for aligning the visual and linguistic representations of embodied robots. The world models [15], [16] exhibit remarkable simulation capabilities and promising comprehension of physical laws, which makes embodied models comprehensively understand both the physical and real environments. These innovations empower embodied agents to comprehensively perceive complex environment, interact with humans naturally, and execute tasks reliably.

Despite the intensive interest in harvesting the powerful perception and reasoning ability from MLMs, the research community is short of a comprehensive survey that can help sort out existing embodied AI studies, the challenges faced, as well as future research directions. In the era of MLMs, we aim to fill up this gap by performing a systematic survey of embodied AI across cyber space to physical world. We conduct the survey from different perspectives including embodied robots, simulators, four representative embodied tasks (visual active perception, embodied interaction, multi-modal agents and sim-to-real adaptation), and future research directions. We believe that this survey will provide a clear big picture of what we have achieved, and we could further achieve along this emerging yet very prospective research direction.

**Differences from previous works:** Although there have been several survey papers [5], [6], [17], [18] for embodied AI, most of them are outdated as they were published before the

era of MLMs, which started around 2023. To the best of our knowledge, there are only two survey papers [6], [18] after 2023, which focused on vision-language-action models and embodied AI system for smart manufacturing, respectively. Embodied AI, with AI brain, Body and Cross-modal sensors, is first proposed in [18], which is also the first work to propose technical architecture of embodied AI system for future smart manufacturing in the era of foundation model. However, the MLMs, WMs and embodied agents are not fully considered in previous surveys. Additionally, recent developments in embodied robots and simulators are also overlooked. To address the scarcity of comprehensive survey papers in this rapidly developing field, we propose this comprehensive survey that covers representative embodied robots, simulators, and four main research tasks: embodied perception, embodied interaction, embodied agents, and sim-to-real adaptation. In summary, the main contributions of this work are threefold:

- To the best of our knowledge, this is the first comprehensive survey of embodied AI from the perspective of the alignment of cyber and physical spaces based on MLMs and WMs, offering novel insights about methodologies, benchmarks, challenges, and applications.
- We categorize and summarize embodied AI into several essential parts including robots, simulators, and four main research tasks: embodied perception, embodied interaction, embodied agents and sim-to-real adaptation, which serve as a detailed taxonomy of embodied AI.
- To facilitate the development of robust, general-purpose embodied agents, we propose a new dataset standard ARIES (All Robots In One) and a unified large-scale ARIES dataset, encompassing approximately 3 million episodes collected from 258 series and 321,064 tasks.

The rest of this survey is organized as follows. Section 2 introduces embodied robots. Section 3 describes general and real-scene embodied simulators. Section 4 introduces embodied perception, including active visual perception and visual language navigation. Section 5 introduces embodied interaction. Section 6 introduces embodied agents including the embodied multi-modal foundation model and embodied task planning. Section 7 introduces sim-to-real adaptation including embodied world model, data collection and training. In Section 8, we discuss promising research directions.

## II. EMBODIED ROBOTS

Embodied agents interact with the physical environment, including robots, smart appliances, and autonomous vehicles, etc. Fixed-base robots, shown in Fig. 2 (a), are used in laboratory automation and industry due to their precision, e.g., Franka Emika panda [19], [20], Kuka iiwa [21], [22], and Sawyer [23], [24]. Wheeled robots, depicted in Fig. 2 (b), are efficient in logistics and warehousing due to their simple structure and low cost, like Kiva and Jackal robots [25]. They face challenges on uneven terrain. Tracked robots, shown in

TABLE II

GENERAL SIMULATOR. **HFPS**: HIGH-FIDELITY PHYSICAL SIMULATION; **HQGR**: HIGH-QUALITY GRAPHICS RENDERING; **RRL**: RICH ROBOT LIBRARY; **DLS**: DEEP LEARNING SUPPORT; **LSPC**: LARGE-SCALE PARALLEL COMPUTING; **ROS**: TIGHT INTEGRATION WITH ROS; **MSS**: MULTIPLE SENSOR SIMULATION; **CP**: CROSS-PLATFORM NAV: ROBOT NAVIGATION **AD**: AUTO DRIVING; **RL**: REINFORCEMENT LEARNING **LSPS**: LARGE-SCALE PARALLEL SIM **MR**: MULTI-ROBOT SYSTEMS **RS**: ROBOT SIMULATION. ○ INDICATES THAT THE SIMULATOR EXCELS AT THIS ASPECT.

Simulator	Year	HFPS	HQGR	RRL	DLS	LSPC	ROS	MSS	CP	Physics Engine	Main Applications
Genesis [35]	2024	○	○	○	○	○	○	○	○	Custom	RL, LSPS, RS
Isaac Sim [36]	2023	○	○	○	○	○	○	○	○	PhysX	Nav, AD
Isaac Gym [37]	2019	○	○	○	○	○	○	○	○	PhysX	RL,LSPS
Gazebo [38]	2004	○	○	○	○	○	○	○	○	ODE, Bullet, Simbody, DART	Nav,MR
PyBullet [39]	2017	○	○	○	○	○	○	○	○	Bullet	RL,RS
Webots [40]	1996	○	○	○	○	○	○	○	○	ODE	RS
MuJoCo [41]	2012	○	○	○	○	○	○	○	○	Custom	RL, RS
Unity ML-Agents [42]	2017	○	○	○	○	○	○	○	○	Custom	RL, RS
AirSim [43]	2017	○	○	○	○	○	○	○	○	Custom	Drone sim, AD, RL
MORSE [44]	2015	○	○	○	○	○	○	○	○	Bullet	Nav, MR
V-REP (CoppeliaSim) [45]	2013	○	○	○	○	○	○	○	○	Bullet, ODE, Vortex, Newton	MR, RS

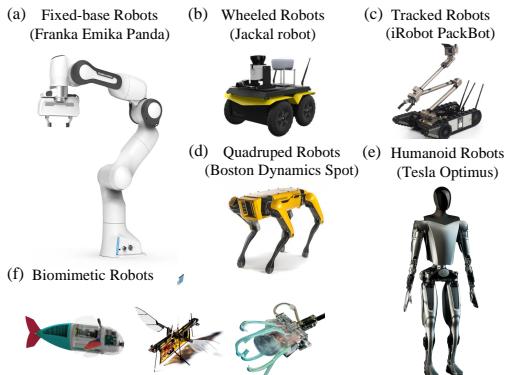


Fig. 2. The Embodied Robots include Fixed-base Robots, Quadruped Robots, Humanoid Robots, Wheeled Robots, Tracked Robots, and Biomimetic Robots.

Fig. 2 (c), are ideal for off-road tasks such as agriculture and disaster recovery. Their track systems provide stability on soft terrains [26] Quadruped robots, illustrated in Fig. 2 (d), excel in complex terrain exploration and rescue missions. Examples include Unitree Robotics' A1 and Go1, and Boston Dynamics Spot. Humanoid robots, shown in Fig. 2 (e), mimic human movements and behaviors to provide personalized services. Their dexterous hands enable them to perform intricate tasks [27], [28]. With LLM, these robots are anticipated to enhance efficiency and safety in manufacturing, healthcare, and services [29]. Biomimetic robots, depicted in Fig. 2 (f), replicate the movements and functions of natural organisms. This simulation aids in operating within complex environments and boosts energy efficiency by emulating biological mechanisms [30], [31]. Examples include fish-like [32], insect-like [33], and soft-bodied robots [34].

### III. EMBODIED SIMULATORS

Embodied simulators are crucial for embodied AI due to their cost-effectiveness, safety features, scalability, rapid prototyping, and accessibility for research. They allow for controlled experimentation, data generation for training and evaluation, and standardized benchmarks. To facilitate interaction with the environment, it's essential to build realistic simulations by considering physical characteristics, object properties, and their interactions. This section will introduce the commonly used simulation platforms in two parts: the general simulator based on underlying simulation and the simulator based on real scenes.



Fig. 3. Examples of General Simulators. The MuJoCo's figure is from [46].

#### A. General Simulator

The physical interactions and dynamic changes present in real environments are irreplaceable. However, deploying embodied models in the physical world often incurs high costs and faces numerous challenges. General-purpose simulators provide a virtual environment that closely mimics the physical world, allowing for algorithm development and model training, which offers significant cost, time, and safety advantages.

**Isaac Sim** [36] is an advanced simulation platform for robotics and AI research. It has high-fidelity physical simulation, real-time ray tracing, an extensive library of robotic models, and deep learning support. Its application scenarios include autonomous driving, industrial automation, and human-robot interaction. **Gazebo** [47] is an open-source simulator for robotics research. It has extensive robot libraries, and tight integration with ROS. It supports the simulation of various sensors and offers numerous pre-built robot models and environments. It is mainly used for robot navigation and control and multi-robot systems. **PyBullet** [39] is the python interface for the Bullet physics engine. It is easy to be used and has diverse sensor simulation and deep learning integration. PyBullet supports real-time physical simulation, including rigid body dynamics, collision detection, and constraint solving. Moreover, the newly launched **Genesis** [35] has differentiable physics engine and impressive generative capabilities. Table. II presents the key features and primary application scenarios of 11 general-purpose simulators. Fig. 3 shows the visualization effects of the general simulators.

#### B. Real-Scene Based Simulators

Achieving universal embodied agents in household activities is a primary focus in embodied AI. These embodied agents need to deeply understand human daily life and perform complex embodied tasks such as navigation and interaction in indoor environments. To meet the demands of these complex tasks, the simulated environments need to be close to the



Fig. 4. Examples of Real-Scene Based Simulators.

real world, which places high demands on the complexity and realism of the simulators. These simulators mostly collect data from the real world, create photorealistic 3D assets, and build scenes using 3D game engines like UE5 and Unity. The rich and realistic scenes make simulators based on real world environments the top choice in household activities.

**SAPIEN** [48] stands out for its design, specifically tailored to simulate interactions with joint objects like doors, cabinets, and drawers. **VirtualHome** [49] is notable for its unique environment graph, which facilitates high-level embodied planning based on natural language descriptions of environments. While **AI2-ThOR** [50] offers a wealth of interactive scenes, these interactions, similar to those in VirtualHome, are script-based and lack real physical interactions. This design suffices for embodied tasks not requiring fine-grained interactions. Both **iGibson** [51] and **TDW** [52] provide fine-grained embodied control and highly simulated physical interactions. iGibson excels in offering abundant and realistic large-scale scenes, making it suitable for complex and long-term mobile operations, whereas TDW allows greater user freedom in scene expansion and features unique audio and flexible fluid simulations, making it indispensable for related simulation scenarios. **Matterport3D** [53], a foundational 2D-3D visual dataset, is widely used and extended in embodied AI benchmarks. Although the embodied agent in Habitat lacks interaction capabilities, its extensive indoor scenes, user-friendly interfaces, and open framework make it highly regarded in embodied navigation. **InfiniteWorld** [54] focuses on unified and scalable simulation framework and implemented various improvements and the latest implicit asset reconstruction, as well as natural language-driven scene generation and editing. It provides strong support for complex robotic interactions through distributed collaboration, AI assistance, and Human-in-the-Loop.

Besides, automated simulation scene construction is greatly beneficial for obtaining high-quality embodied data. **RoboGen** [55] customizes tasks from randomly sampled 3D assets through LLMs, thereby creating scenes and automatically training agents; **HOLODECK** [56] can automatically customize corresponding high-quality simulation scenes in AI2-THOR based on human instructions; **PhyScene** [57] generates interactive and physically consistent high-quality 3D scenes based on conditional diffusion. The Allen Institute for Artificial Intelligence expanded AI2-THOR and proposed **ProcTHOR** [58], which can automatically generate simulated scenes with sufficient interactivity, diversity, and rationality.

#### IV. EMBODIED PERCEPTION

The “north stars” of the future of visual perception is embodied-centric visual reasoning and social intelligence [59].

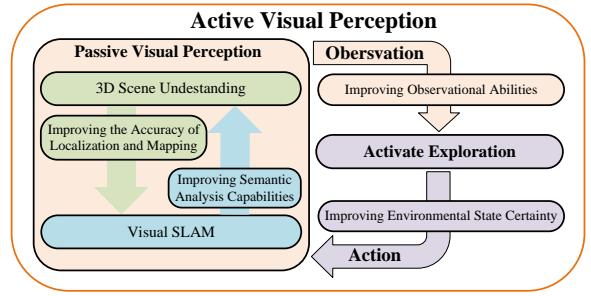


Fig. 5. The schematic diagram of active visual perception. Visual SLAM and 3D Scene Understanding provide the foundation for passive visual perception, while active exploration provides activeness to the passive perception system. These elements work collaboratively for the active visual perception system.

Unlike merely recognizing objects in images, agent with embodied perception must move in the physical world and interact with the environment. This requires a deeper understanding of 3D space and dynamic environments. Embodied perception requires visual perception and reasoning, understanding the 3D relations within a scene, and predicting and performing complex tasks based on visual information.

##### A. Active Visual Perception

Active visual perception systems require fundamental capabilities such as state estimation, scene perception, and environment exploration. As shown in Fig. 5, these capabilities have been extensively studied within the domains of Visual Simultaneous Localization and Mapping (vSLAM) [94], [95], 3D Scene Understanding [96], and Active Exploration [11]. These research areas contribute to developing robust active visual perception systems, facilitating improved environmental interaction and navigation in complex, dynamic settings. We briefly introduce these three components and summarize the methods mentioned in each part in Table III.

1) *Visual Simultaneous Localization and Mapping*: Simultaneous Localization and Mapping (SLAM) aims to determine a robot’s position within an unknown environment while simultaneously constructing a map of the environment [97]. Range-based SLAM [98], [99] relies on rangefinders, such as laser scanners, radar, and sonar, to generate point cloud representations. However, this approach is costly and provides limited environmental information. In contrast, Visual SLAM (vSLAM) [94], [95] employs on-board cameras to capture image frames and build environmental representations. Its advantages include low hardware costs, high accuracy in small-scale scenarios, and the ability to capture rich environmental details. Classical vSLAM can be broadly categorized into Traditional vSLAM and Semantic vSLAM [95]. Traditional vSLAM uses image data and multi-view geometry to estimate a robot’s pose and construct low-level maps (e.g., sparse, semi-dense, or dense point clouds) through methods like filter-based approaches (e.g., MonoSLAM [60]), keyframe-based methods (e.g., ORB-SLAM [61]), and direct tracking techniques (e.g., LSD-SLAM [62]). However, low-level maps do not directly correspond to objects, making them challenging for robots to interpret and utilize. Semantic vSLAM addresses this limitation by integrating semantic information, enhancing robots’ ability to perceive and navigate unexplored environments.

TABLE III  
THE COMPARISON OF THE ACTIVE VISUAL PERCEPTION METHODS.

Function	Type	Methods
vSLAM	Traditional vSLAM	MonoSLAM [60], ORB-SLAM [61], LSD-SLAM [62]
	Semantic vSLAM	SLAM++ [63], QuadricSLAM [64], So-SLAM [65], SG-SLAM [66], OVD-SLAM [67], GS-SLAM [68]
3D Scene Understanding	Projection-based	MV3D [69], PointPillars [70], MVCNN [71]
	Voxel-based	VoxNet [72], SSCNet [73], MinkowskiNet [74], SSCNs [75], Embodiedscan [76]
	Point-based	PointNet [77], PointNet++ [78], PointMLP [79], PointTransformer [80], Swin3d [81], PT2 [82], 3D-VisTA [83], LEO [84], PQ3D [85], PointMamba [86], Mamba3D [87]
Active Exploration	Interacting with the environment	Pinto et al. [88], Tatiya et al. [89]
	Changing the viewing direction	Jayaraman et al. [90], NeU-NBV [91], Hu et al. [92], Fan et al. [93]

2) *3D Scene Understanding*: 3D scene understanding [100] aims to distinguish objects' semantics, identify their locations, and infer the geometric attributes from 3D scene data [101], which is fundamental in autonomous driving [102], robot navigation [103], and human-computer interaction [104] etc. A scene may be recorded as 3D point clouds using 3D scanning tools like LiDAR or RGB-D sensors. Unlike images, point clouds are sparse, disordered, and irregular. Recent advances in deep learning for 3D scene understanding can be categorized into projection-based, voxel-based, and point-based methods. Concretely, projection-based methods (e.g., MV3D [69], PointPillars [70], MVCNN [71]) project 3D points onto various image planes and employ 2D CNN-based backbones for feature extraction. Voxel-based methods convert point clouds into regular voxel grids to facilitate 3D convolution operations (e.g., VoxNet [72], SSCNet [73]), and some works improve their efficiency through sparse convolution (e.g., MinkowskiNet [74], SSCNs [75], Embodiedscan [76]). In contrast, point-based methods process point clouds directly (e.g., PointNet [77], PointNet++ [78], PointMLP [79]). Recently, to achieve model scalability, Transformers-based (e.g., PointTransformer [80], Swin3d [81], PT2 [82], 3D-VisTA [83], LEO [84], PQ3D [85]) and Mamba-based (e.g., PointMamba [86], Mamba3D [87]) architectures have emerged. Notably, PQ3D [85] enhances scene understanding by seamlessly integrating features from point clouds, multi-view images, and voxels.

3) *Active Exploration*: The 3D scene understanding methods allow robots to passively perceive the environment, with static information acquisition and decision-making regardless of scene changes. Thus, while passive perception is essential, it must be complemented by active exploration, enabling robots to dynamically interact with and perceive their surroundings. The relationship between them is shown in Fig. 5. Current methods addressing active perception focus on interacting with the environment [88], [89] or by changing the viewing direction to obtain more visual information [90]–[93].

For example, Pinto et al. [88] proposed a curious robot that learns visual representations through physical interaction with the environment rather than relying solely on dataset category labels. To address the challenge of interactive object perception across robots with varying morphologies, Tatiya et al. [89] proposed a multi-stage projection framework that transfers implicit knowledge through learned exploratory interactions. Recognizing the challenge of autonomously capturing informative observations, Jayaraman et al. [90] proposed a reinforcement learning method where an agent learns to actively acquire informative visual observations by reducing its uncertainty about unobserved parts of its environment. NeU-NBV

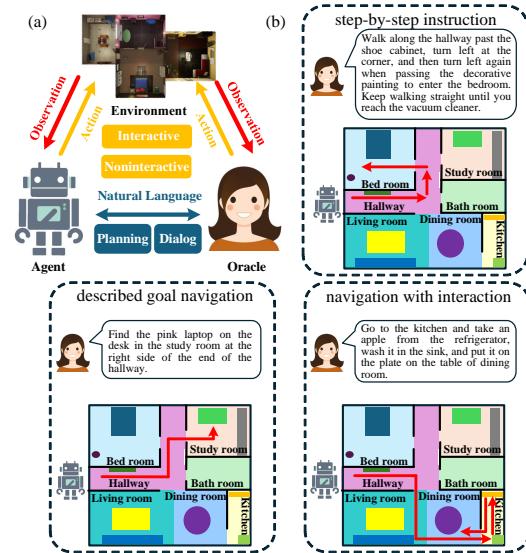


Fig. 6. (a) Overview of VLN. The embodied agent communicates with humans through natural language. Humans issue instructions to the embodied agent, who completes tasks such as planning and dialog. Subsequently, through collaborative cooperation or the embodied agent's independent actions, actions are made in interactive or non-interactive environments based on visual observations and instructions, (b) Different tasks of VLN.

[91] introduced a mapless planning framework that iteratively positions an RGB camera to capture the most informative images of an unknown scene. Hu et al. [92] developed a robot exploration algorithm that predicts the value of future states using a state value function. To address the issue of accidental input in open-world environments, Fan et al. [93] treated active recognition as a sequential evidence-gathering process, providing step-by-step uncertainty quantification and reliable prediction under evidence combination theory.

### B. Visual Language Navigation

Visual Language Navigation (VLN) is an essential task, aiming at navigating in unseen environments following linguistic instructions. VLN requires robots to understand complex and diverse visual observations and meanwhile interpret instructions at different granularities. The input typically consists of two parts: visual information and natural language instructions. The visual information can either be a video of past trajectories or a set of historical-current observation images. The natural language instructions include the target that the agent needs to reach or the task that the agent is expected to complete. The agent must use the above information to select one or a series of actions from a list of candidates to fulfill the requirements of the natural language instructions. This process could be represented as  $Action = \mathcal{M}(O, H, I)$ , where  $Action$  is the chosen action or a list of action candidates,  $O$  is the

TABLE IV

COMPARISON OF DIFFERENT VLN DATASETS. **M3D**: MATTERPORT3D, **AT**: AI2-THOR, **OG**: OMNIGIBSON, **I**: INDOOR, **D**: DISCRETE, **O**: OUTDOOR, **C**: CONTINUOUS, **SBS**: STEP-BY-STEP INSTRUCTIONS, **DGN**: DESCRIBED GOAL NAVIGATION, **DDN**: DEMAND-DRIVEN NAVIGATION, **NWI**: NAVIGATION WITH INTERACTION, **LSNWl**: LONG-SPAN NAVIGATION WITH INTERACTION, **D&O**: DIALOG AND ORACLE

Dataset	Year	Simulator	Environment	Feature	Size
R2R [105]	2018	M3D	I, D	Sbs	21,567
R4R [106]	2019	M3D	I, D	Sbs	200,000+
VLN-CE [107]	2020	Habitat	I, C	Sbs	-
TOUCHDOWN [108]	2019	-	O, D	Sbs	9,326
REVERIE [109]	2020	M3D	I, D	DGN	21,702
SOON [110]	2021	M3D	I, D	DGN	3,848
DDN [111]	2023	AT	I, C	DDN	30,000+
ALFRED [112]	2020	AT	I, C	NwI	25,743
OVMM [113]	2023	Habitat	I, C	NwI	7,892
BEHAVIOR-1K [114]	2023	OG	I, C	LSNWl	1,000
CVDN [115]	2020	M3D	I, D	D&O	2,050
DialFRED [116]	2022	AT	I, C	D&O	53,000

current observation,  $H$  is the historical information, and  $I$  is the natural language instruction.

1) *Datasets*: In VLN, natural language instructions can be a series of detailed action descriptions, a fully described goal, or just a roughly described task, even only the demands of human. The tasks that embodied agents need to complete maybe just a single navigation, or navigation with interaction, or multiple navigation tasks that need to be completed in sequence. These differences bring different challenges to VLN, and many different datasets have been built. Based on these differences, we introduce some important VLN datasets.

**Room to Room (R2R)** [105] is a VLN dataset based on Matterport3D. In R2R, embodied agents navigate according to step-by-step instructions, choosing the next adjacent navigation graph node to advance based on visual observations until they reach the target location. **Room-for-Room** [106] extends the paths in R2R to longer trajectories, which requires stronger long-distance instruction and history alignment capabilities of embodied agents. **VLN-CE** [107] extends R2R and R4R to continuous environments, embodied agents can move freely in the scene. Different from the above datasets based on indoor scenes, the **TOUCHDOWN** dataset [108] is created based on Google Street View. In TOUCHDOWN, embodied agents follow instructions to navigate in the street view rendering simulation of New York City to find the specified object. Similar to R2R, the **REVERIE** dataset [109] is also built based on the Matterport3D simulator. REVERIE requires embodied agents to accurately locate the distant invisible target object specified by concise, human-annotated high-level natural language instructions. In **SOON** [110], agents receive a long and complex instruction from coarse to fine to find the target object in the 3D environment. During navigation, agents first search a larger area, and then gradually narrow the search range according to the visual scene and instructions. **DDN** [111] moves a step further beyond these datasets, only providing human demands without specifying explicit objects. The agent needs to navigate through the scene to find objects.

**ALFRED** dataset [112] is based on the AI2-THOR simulator. In ALFRED, embodied agents need to understand environmental observations and complete household tasks in an interactive environment according to coarse-grained and fine-grained instructions. The task in **OVMM** [113] dataset is to pick any object in any unseen environment and place it

TABLE V  
COMPARISON OF VLN METHODS.

Method	Model	Year	Feature
<b>Memory-Understanding Based</b>	LVERG [117]	2020	Graph Learning
	CMG [118]	2020	Adversarial Learning
	RCM [119]	2021	Reinforcement learning
	FILM [120]	2022	Semantic Map
	LM-Nav [121]	2022	Graph Learning
	HOP [122]	2022	History Modeling
	NaviLLM [123]	2024	Large Model
	FSTT [124]	2024	Test-Time Augmentation
	DiscussNav [125]	2024	Large Model
	GOAT [126]	2024	Causal Learning
<b>Future-Prediction Based</b>	VER [127]	2024	Environment Encoder
	NaVid [128]	2024	Large Model
	LookBY [129]	2018	Reinforcement Learning
	NvEM [130]	2021	Environment Encoder
	BGBL [131]	2022	Graph Learning
<b>Others</b>	Mic [132]	2023	Large Model
	HNR [133]	2024	Environment Encoder
	ETPNav [134]	2024	Graph Learning
<b>Others</b>	MCR-Agent [135]	2023	Multi-Level Model
	OVLM [136]	2023	Large Model

in a specified location. OVMM provides a simulation based on Habitat and a framework for implementation in the real world. **Behavior-1K** dataset [114] is based on human needs, comprising 1,000 long-sequence, complex, skill-dependent daily tasks. Agents need to complete long-span navigation-interaction tasks which contain thousands of low-level action steps based on visual information and language instructions. These complex tasks require strong capabilities of understanding and memory. **CVDN** [115] requires embodied agents to navigate to the target based on dialogue history, and ask questions for help to decide the next action when uncertain. **DialFRED** [116], an extension of ALFRED, allows agents to ask questions during the navigation and interaction process to get help. These datasets introduce additional oracles, and embodied agents need to obtain more information beneficial to navigation by asking questions.

2) *Method*: VLN has made great strides recently with the astonishing performance of LLMs, the direction and focus of VLN have been profoundly influenced. Nevertheless, the VLN methods can be divided into two directions: **Memory-Understanding Based** and **Future-Prediction Based**.

Memory-Understanding based methods focus on the perception and understanding of the environment, as well as model design based on historical observations or trajectories, which is a method based on past learning. Future-Prediction based methods pay more attention to modeling, predicting, and understanding the future state, which is a method for future learning. Since VLN can be regarded as a partially observable Markov decision process, where future observations depend on the current environment and actions of the intelligent agent, historical information has important significance for navigation decisions, especially long-span navigation decisions, hence Memory-Understanding based methods have always been the mainstream of VLN. However, Future-Prediction based methods still have important significance. Its essential understanding of the environment has great value in VLN in continuous environments, especially with the rise of the concept of world model, Future-Prediction based methods are receiving more and more attention from researchers.

**Memory-Understanding based.** Graph-based learning is an essential part of the memory-understanding based method.

It usually represents the navigation process in the form of a graph, where the information obtained by the agent at each time step is encoded as nodes of the graph. The agent obtains global or partial navigation graph information as a representation of the historical trajectory. LVERG [117] encoded the language information and visual information of each node separately, design a new language and visual entity relationship graph to model the inter-modal relationship between text and vision, and the intra-modal relationship between visual entities. LM-Nav [121] used a goal-conditioned distance function to infer connections between original observation sets and construct a navigation graph, and extracted landmarks from the instructions through a LLM. Although HOP [122] is not based on graph learning, it requires to model time-ordered information at different granularities, thereby achieving a deep understanding of historical trajectories and memories.

The navigation graph discretizes the environment, but concurrently understanding and encoding the environment is also important. FILM [120] used RGB-D observations and semantic segmentation to gradually build a semantic map from 3D voxels during the navigation. VER [127] quantified the physical world into structured 3D units through 2D-3D sampling, providing fine-grained geometric details and semantics.

Different learning schemes explore how to utilize historical trajectories and memories better. Through adversarial learning, CMG [118] alternated between imitation learning and exploration encouragement schemes, effectively strengthening the understanding of instructions and historical trajectories, shortening the difference between training and inference. GOAT [126] directly trained unbiased models through Backdoor Adjustment Causal Learning (BACL) and Frontdoor Adjustment Causal Learning (FACL), conducted contrastive learning with vision, navigation history, and their combination to instructions, enabling the agent to make fuller use of information. The enhanced cross-modal matching method proposed by RCM [119] used goal-oriented external rewards and instruction-oriented internal rewards to perform cross-modal grounding globally and locally and learns from its own historical good decisions through self-supervised imitation learning. FSTT [124] introduced TTA into VLN and optimizes the model in terms of gradients and model parameters at two scales of time steps and tasks, effectively improving model performance.

The specific application of large models in Memory-Understanding based methods is to understand the representation of historical memory and to understand the environment and tasks based on its extensive world knowledge. NaviLLM [123] integrated the historical observation sequence into the embedding space through the visual encoder, inputs the multi-modal information of the fusion encoding into the LLM and fine-tunes it, reaching the state-of-the-art on multiple benchmarks. NaVid [128] made improvements in the encoding of historical information, achieves different degrees of information retention on historical observations and current observations through different degrees of pooling. LH-VLN [137] proposed the NavGen platform, the long-horizon navigation benchmark, and the Multi-Granularity Dynamic Memory (MGDM) module to enhance task evaluation and model adaptability in dynamic environments.

**Future-Prediction Based.** Graph-based learning is also widely used in Future-Prediction based methods. BGBL [131] and ETPNav [134] used a similar method to design a waypoint predictor that can predict movable path points in a continuous environment based on the observation of the current navigation graph node. They aim to migrate complex navigation in a continuous environment to node-to-node navigation in a discrete environment, thereby bridging the performance gap from discrete environments to continuous environments.

Improving the understanding and perception of the future environment through environmental encoding is also one of the research directions for predicting and exploring the future. NvEM [130] used a theme module and a reference module to perform fusion encoding of neighbor views from the global and local perspectives. This is actually an understanding and learning of future observations. HNR [133] used a large-scale pre-trained hierarchical neural radiation representation model to directly predict the visual representation of the future environment rather than pixel-level images using three-dimensional feature space encoding, and builds a navigable future path tree based on the representation of the future environment. They predict the future environment from different levels, providing effective references for navigation decisions.

Some reinforcement learning methods are also applied to predict and explore future states. LookBY [129] employed reinforcement prediction to enable the prediction module to imitate the world and forecast future states and rewards. This allows the agent to directly map “current observations” and “predictions of future observations” to actions, achieving state-of-the-art performance at the time. The rich world knowledge and zero-shot performance of large models provide many possibilities for Future-Prediction based methods. MiC [132] required the LLM to directly predict the target and its possible location from the instructions and provides navigation instructions through the description of scene perception. This method requires LLMs to fully exert its ‘imagination’ and build an imagined scene through prompts.

In addition, there are some methods that both learn from the past and for the future. MCR-Agent [135] designed a three-layer action strategy, which requires the model to predict the target from the instructions, predict the pixel-level mask for the target to be interact, and learn from the previous navigation decision; OVLM [136] required the LLMs to predict the corresponding operations and landmark sequences for the instructions. During the navigation process, the visual language map will be continuously updated and maintained, and the operations will be linked to the waypoints on the map.

## V. EMBODIED INTERACTION

Embodied interaction refer to scenarios where agents interact with humans and the environment in physical or simulated space. The typical embodied interaction tasks are Embodied Question Answering (EQA) and embodied grasping.

### A. Embodied Question Answering

For EQA task, the agent needs to explore the environment from a first-person perspective to gather information necessary

TABLE VI  
COMPARISON OF DIFFERENT EQA DATASETS.

Dataset	Year	Type	Data Sources	Simulator	Query Creation	Answer	Size
EQA v1 [138]	2018	Active EQA	SUNCG	House3D	Rule-Based	open-ended	5,000+
MT-EQA [139]	2019	Active EQA	SUNCG	House3D	Rule-Based	open-ended	19,000+
MP3D-EQA [140]	2019	Active EQA	MP3D	Simulator based on MINOS	Rule-Based	open-ended	1,136
IQUAD V1 [141]	2018	Interactive EQA	-	AI2THOR	Rule-Based	multi-choice	75,000+
VideoNavQA [142]	2019	Episodic Memory EQA	SUNCG	House3D	Rule-Based	open-ended	101,000
SQA3D [143]	2022	QA only	ScanNet	-	Manual	multi-choice	33,400
K-EQA [144]	2023	Active EQA	-	AI2THOR	Rule-Based	open-ended	60,000
OpenEQA [145]	2024	Active EQA, Episodic Memory EQA	ScanNet, HM3D	Habitat	Manual	open-ended	1,600+
HM-EQA [146]	2024	Active EQA	HM3D	Habitat	VLM	multi-choice	500
S-EQA [147]	2024	Active EQA	-	VirtualHome	LLM	binary	-
EXPRESS-Bench [148]	2025	Exploration-aware EQA	HM3D	Habitat	VLM	open-ended	2,044



Fig. 7. The gray box displays the scenes an agent observes during exploration. The other boxes show various types of question answering tasks. Except for the task of answering questions based on episodic memory, the agent ceases exploration once it has gathered sufficient information to answer the question.

to answer the given questions. An agent with autonomous exploration and decision-making capabilities must not only consider which actions to take to explore the environment but also determine when to stop exploring to answer questions. Existing works focus on different types of questions, some of which are shown in Fig. 7. In this section, we will introduce the existing datasets, discuss the related methods, describe the metrics used to evaluate model performance, and address the remaining limitations of this task.

1) *Datasets:* We briefly introduce several embodied question answering datasets, which are summarized in Table VI.

**EQA v1** [138] is the first dataset designed for EQA. Built on synthetic 3D indoor scenes from the SUNCG dataset [73] within the House3D [149] simulator, EQA v1 comprises four types of questions: location, color, color\_room, and preposition. Similar to EQA v1, **MT-EQA** [139] is built in House3D using SUNCG by executing functional programs consisting of some basic operations. However, it further extends the single-object question answering task to a multi-object setting. Six types of questions are designed, involving the comparison of color, distance, and size between multiple objects. **MP3D-EQA** [140] is built on a simulator developed based on MINOS [150] using the Matterport3D dataset [151], expanding the question-answering task to a realistic 3D environment. Referring to EQA v1, MP3D-EQA utilizes three types of templates: location, color, and color\_room, generating a total of 1,136

questions in 83 home environments. **IQUAD V1** [141] is built upon AI2-THOR and consists of three types of questions: existence, counting, and spatial relationships. Unlike other datasets, answering IQUAD V1 questions requires the agent to have a good understanding of affordances and interact with the dynamic environment. **VideoNavQA** [142] decouples the visual reasoning from the navigation aspect of the EQA problem. In this task, the agent accesses videos corresponding to exploration trajectories with sufficient information to answer questions. **SQA3D** [143] simplifies protocol (QA only) while still preserving the function of benchmarking embodied scene understanding, enabling more complex, knowledge-intensive questions and a much larger scale of data collection.

Unlike previous datasets that explicitly specify target objects in questions, **K-EQA** [144] features complex questions with logical clauses and knowledge-related phrases, requiring prior knowledge to answer. **OpenEQA** [145] is the first open-vocabulary dataset for EQA, supporting both episodic memory and active exploration cases. The episodic memory EQA (EM-EQA) tasks involve an agent developing an understanding of the environment from its episodic memory to answer questions. In active EQA (A-EQA) tasks, the agent answers questions by taking exploratory actions to gather necessary information. Utilizing GPT4-V, **HM-EQA** [146] is constructed in the Habitat simulator using HM3D. It includes 500 questions across 267 different scenes, which can be roughly categorized into identification, counting, existence, status, and location. **S-EQA** [147] leverages GPT-4 in VirtualHome for data generation and employs cosine similarity calculations to decide whether to retain the generated data, thereby enhancing dataset diversity. In S-EQA, answering questions requires the assessment of a collection of consensus objects and states to reach an existential “Yes/No” answer. **EXPRESS-Bench** [148] is the largest exploration-aware EQA dataset that consists of 777 exploration trajectories and 2,044 samples. It also introduces novel evaluation metrics to ensure faithful assessment.

2) *Methods:* The embodied question answering task mainly involves navigation and question-answering subtasks, with implementation methods broadly categorized into two types: neural network-based and LLMs/VLMs-based.

**Neural Network Methods.** In early work, researchers mainly addressed the embodied question answering task by building deep neural networks. They trained and fine-tuned these models using techniques such as imitation learning and reinforcement learning to improve performance.

The EQA task was first proposed by Das et al. [138]. In their work, the agent consists of four main modules:

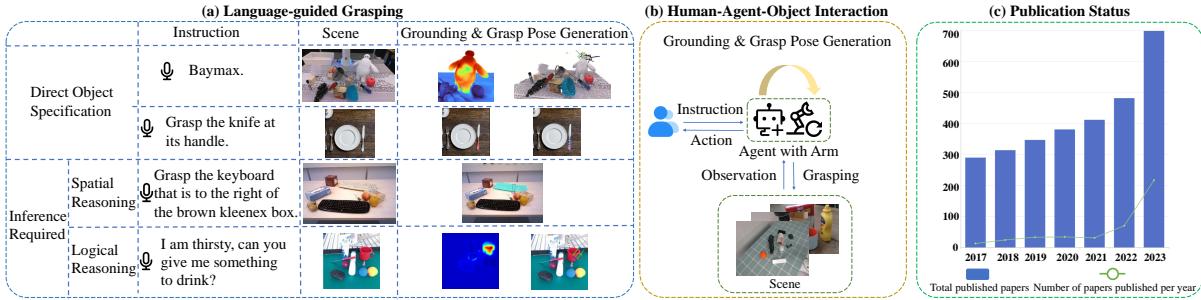


Fig. 8. The overview of the embodied grasping task. (a) demonstrates examples of language-guided grasping for different types of tasks, (b) provides an overview of human-agent-object interaction, (c) shows Google Scholar search results for topics of “Language-guided Grasping”.

vision, language, navigation, and answering. These modules are primarily constructed using traditional neural building blocks: Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Some subsequent works [152] retained modules like the question answering module proposed by Das et al. [138] and improved the model. Additionally, Wu et al. [152] proposed integrating the navigation and QA modules into a unified SGD training pipeline for joint training, thereby avoiding employing deep reinforcement learning to simultaneously train the separately trained navigation and question answering modules. From the perspective of task singularity, several works [153] expanded the task to include multiple objectives and multi-agent, respectively, making it necessary for the model to store and integrate the information obtained by the agent’s exploration through methods such as feature extraction and scene reconstruction. Considering the interaction between the agent and the dynamic environment, Gordon et al. [141] introduced the Hierarchical Interactive Memory Network. There is also a limitation in previous works where agents are unable to use external knowledge to answer complex questions and lack knowledge of the explored parts of the scene. To address this, Tan et al. [144] leveraged the neural program synthesis method and the table converted from the knowledge and 3D scene graphs, allowing the action planner to access object-related information. Additionally, an approach based on Monte Carlo Tree Search (MCTS) is used to determine the next location for the agent.

**LLMs/VLMs Methods.** Majumdar et al. [145] used LLMs and VLMs for episodic memory EQA (EM-EQA) and Active EQA (A-EQA) tasks. For EM-EQA task, they considered Blind LLMs, Socratic LLMs with language descriptions of the episodic memory, Socratic LLMs with descriptions of the constructed scene graph, and VLMs processing multiple scene frames. The A-EQA task extended EM-EQA methods with frontier-based exploration (FBE) [154] for problem-independent environment exploration. Some works [146], [155] employed frontier-based exploration method to identify areas for subsequent exploration and to build semantic maps. They ended the exploration early utilizing conformal prediction or image-text matching to avoid over-exploration. Patel et al. [156] emphasized the question answering aspect of the task. They leveraged multiple LLM-based agents to explore the environment and enable them to independently answer questions with “yes” or “no” answers. These individual responses are utilized to train a Central Answer Model, to aggregate responses and generate robust answers.

**3) Limitations:** a) **Dataset:** Constructing datasets requires substantial manpower and resources. Additionally, there are still few large-scale datasets, and the metrics for evaluating model performance vary across different datasets, complicating the testing and comparison of performance, b) **Model:** Despite the advancements brought by LLMs, the performance of these models still lags significantly behind human levels. Future work may focus more on effectively storing environmental information explored by agents and guiding them to plan actions based on environmental memory and questions, while also enhancing the interpretability of their actions.

### B. Embodied Grasping

Embodied interaction includes not only question-answering but also performing tasks like grasping and placing objects based on human instructions. Embodied grasping combines traditional kinematic methods [157], [158] with large models such as LLMs and vision-language models, enabling multi-sensory perception and reasoning for task execution. Figure 8 (b) shows an overview of human-agent-object interactions where embodied grasping is performed.

**1) Datasets:** Recently, a substantial number of grasping datasets [159]–[163] have been generated. These datasets typically contain annotated grasping data based on images (RGB, depth), point clouds, or 3D scenes. With the advent of MLMs and the application of foundational language models to robotic grasping, there is an urgent need for datasets that include linguistic text. Consequently, existing datasets have been extended or reconstructed to create semantic-grasping datasets [165]–[167]. These datasets are instrumental in studying grasping models grounded in language, enabling agents to develop a broad understanding of semantics.

Traditional grasping datasets encompass data for both single objects [159] and cluttered scenes [164], providing stable grasp annotations (4-DOF or 6-DOF) that conform to kinematics for each object. These data can be collected from real desktop environments [159], typically including RGB, depth, and point cloud data, or from virtual environments [162], which include image data, point clouds, or scene models. While these datasets are useful for grasping models, they lack semantic information. To bridge this gap, these datasets have been augmented or extended with semantic expressions [165], [168], thereby linking language, vision, and grasping. By incorporating semantic information, agents can better understand and execute grasping tasks. This enhancement allows for the

TABLE VII  
EMBODIED GRASPING DATASETS.

Dataset	Year	Type	Modality	Grasp Label	Gripper Finger	Objects	Grasps	Scenes	Language
Cornell [159]	2011	Real	RGB-D	Rect.	2	240	8K	Single	✗
Jacquard [160]	2018	Sim	RGB-D	Rect.	2	11K	1.1M	Single	✗
6-DOF GraspNet [161]	2019	Sim	3D	6D	2	206	7.07M	Single	✗
ACRONYM [162]	2021	Sim	3D	6D	2	8872	17.7M	Multi	✗
MultiGripperGrasp [163]	2024	Sim	3D	-	2-5	345	30.4M	Single	✗
OCID-Grasp [164]	2021	Real	RGB-D	Rect.	2	89	75K	Multi	✗
OCID-VLG [165]	2023	Real	RGB-D,3D	Rect.	2	89	75K	Multi	✓
ReasoningGrasp [166]	2024	Real	RGB-D	6D	2	64	99.3M	Multi	✓
CapGrasp [167]	2024	Sim	3D	-	5	1.8K	50K	Single	✓

development of more sophisticated and semantically aware grasping models, facilitating more intuitive and effective interaction with the environment. Table VII presents the datasets described above, including traditional grasping datasets and language-based grasping datasets.

2) *Language-guided grasping*: The concept of language-guided grasping [165], [166], [168], which has evolved from this integration, combines MLMs to provide agents with the capability of semantic scene reasoning. This allows the agent to execute grasping operations based on implicit or explicit human instructions. Figure 8 (c) illustrates the publication trends in recent years on the topic of language-guided grasping. With the advancement of LLMs, researchers have shown increasing interest in this topic. Currently, grasping research is increasingly focused on open-world scenarios, emphasizing the open-set generalization [169] methods. By leveraging the generalization capabilities of MLMs, robots can perform grasping tasks in open-world environments with greater intelligence and efficiency.

In language-guided grasping, semantics can originate from explicit instructions [169], [170] and implicit instructions [166], [167]. Explicit instructions clearly specify the category of the object to be grasped, such as a banana or an apple. Implicit instructions, however, require reasoning to identify the object or a part of the object to be grasped, involving spatial reasoning and logical reasoning.

Spatial reasoning [165] refers to instructions that may include the spatial relationship of the object or part to be grasped, necessitating the inference of grasping posture based on the spatial relationships of objects within the scene. For example, “Grasp the keyboard that is to the right of the brown kleenex box” involves understanding and inferring the spatial arrangement of objects. Logical reasoning [166], on the other hand, involves instructions that may contain logical relationships requiring inference to discern human intent and subsequently grasp the target. For instance, “I am thirsty, can you give me something to drink?” would prompt the agent to potentially hand over a glass of water or a bottle of a beverage. The agent must ensure that the liquid does not spill during the handover, thus generating a reasonable grasping posture.

In both cases, the integration of semantic understanding with spatial and logical reasoning enables the agent to perform complex grasping tasks effectively and accurately. Figure 8 (a) depicts various types of language-guided grasping tasks.

3) *End-to-End Approaches*: CLIPORT [168] is a language-conditioned imitation learning agent that combines the vision-language pre-trained model CLIP with the Transporter Net to create an end-to-end dual-stream architecture for semantic understanding and grasp generation. It is trained using a large

number of expert demonstration data collected from virtual environments, enabling the agent to perform semantically guided grasping. Based on the OCID dataset, CROG [165] proposes a vision-language-grasping dataset and introduces a competitive end-to-end baseline. It leverages CLIP’s visual foundation capabilities to learn grasp synthesis directly from image-text pairs. Reasoning Grasping [166] introduces the first reasoning grasping benchmark dataset based on the GraspNet-1 Billion dataset and proposes an end-to-end reasoning grasping model. The model integrates multimodal LLMs with vision-based robotic grasping frameworks to generate grasps based on semantics and vision. SemGrasp [167] is a method for semantic-based grasp generation that incorporates semantic information into grasp representations to generate dexterous hand grasp postures. It introduces a discrete representation aligning grasp space with semantic space, enabling the generation of grasp postures according to language instructions.

4) *Modular Approaches*: F3RM [169] seeks to elevate CLIP’s text-image priors into 3D space, using extracted features for language localization followed by grasp generation. It combines precise 3D geometry with rich semantics from 2D foundational models, utilizing features extracted from CLIP to specify objects for manipulation through free-text natural language. It demonstrates the ability to generalize to unseen expressions and new object categories. GaussianGrasper [170] utilizes a 3D Gaussian field to achieve language-guided grasping tasks. The proposed methodology begins with the construction of a 3D Gaussian field, followed by feature distillation. Subsequently, language-based localization is performed using the extracted features. Finally, grasp pose generation is carried out based on a SOTA pre-trained grasping network [171]. It integrates open-vocabulary semantics with precise geometry, enabling grasping based on language instructions.

These approaches advance language-guided grasping by using end-to-end and modular frameworks, enhancing robotic agents’ ability to perform complex grasping tasks from natural language instructions. Embodied grasping improves robots’ intelligence and utility in-home services and industrial manufacturing. However, current methods face limitations, including reliance on extensive data and poor generalization. Future research aims to enhance agent generality, enabling robots to understand complex semantics, grasp a variety of unseen objects, and tackle intricate tasks.

## VI. EMBODIED AGENT

An agent is defined as an autonomous entity capable of perceiving its environment and acting to achieve specific objectives. Recent advancements in MLMs have further expanded

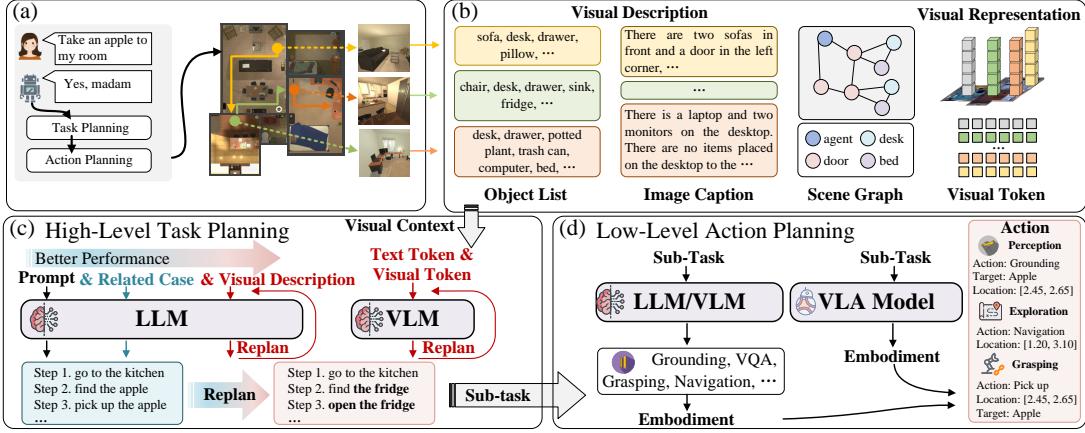


Fig. 9. The architecture of the embodied agent based on embodied multimodal foundation model, which consists of visual perception module, high-level task planning module, and low-level action planning module.

the application of agents to practical scenarios. When these MLM-based agents are embodied in physical entities, they can effectively transfer their capabilities from virtual space to physical world, thereby becoming embodied agents [172].

To enable embodied agents to operate in the information-rich and complex real world, the embodied agents have been developed to show strong multimodal perception, interaction and planning capabilities, as shown in Fig. 9. To complete a task, embodied agents typically involves the following process: 1) decomposing the abstract and complex task into specific subtasks, which is referred to as high-level Embodied Task Planning. 2) gradually implementing these subtasks by effectively utilizing Embodied Perception and Embodied Interaction models or leveraging the Foundation Model's policy function, named low-level Embodied Action Planning. It is worth noting that task planning involves thinking before acting, and is therefore typically considered in cyber space. In contrast, action planning must account for effective interaction with the environment and feedback on this information to the task planner to adjust task planning. Thus, it is crucial for embodied agents to align and generalize their abilities from the cyber space to the physical world.

#### A. Embodied Task Planning

Traditional embodied task planning methods are usually based on explicit rules and logical reasoning. For example, symbolic planning algorithms such as PDDL [173], and search algorithms like MCTS [174] and A\* [175], are used to generate plans. However, these methods often rely on predefined rules, constraints, and heuristics that are rigid and may not adapt well to dynamic or unforeseen changes. With the popularity of LLMs, many works have attempted to use LLMs for planning or to combine traditional methods with LLMs, leveraging the rich embedded world knowledge for reasoning and planning without the need for handcrafted definitions, greatly enhancing the model's generalization capabilities.

1) *Planning utilizing the Emergent Capabilities of LLMs:* Before the scale-up of natural language models, task planners were similarly implemented by training models like BERT on embodied instruction datasets such as Alfred [176] and Alfworld [177], as demonstrated by FILM [178]. However,

this approach was limited by the examples in the training set and could not effectively align with the physical world. Nowadays, thanks to the emergent capabilities of LLMs, they can decompose abstract tasks using their internal world knowledge and chain-of-thought reasoning, similar to how humans reason through task completion steps before acting. For example, Translated LM [179] and Inner Monologue [180] can break down complex tasks into manageable steps and devise solutions using their internal logic and knowledge systems without additional training. Similarly, the multi-agent collaboration framework ReAd [181] efficiently self-refined plans via different prompts. Additionally, some approaches abstract past successful examples into a series of skills stored in a memory bank to consider during inference and improve planning success rates [182]–[184]. Some works utilized code as the reasoning medium instead of natural language, where task planning is generated as code based on the available API library [185], [186]. Furthermore, multi-turn reasoning can effectively correct potential hallucinations in task planning. For instance, Socratic Models [187] and Socratic Planner [188] used Socratic questioning to derive reliable planning.

However, during task planning, potential failures may occur during execution, often resulting from the planner not fully accounting for the complexity of the real environment and the difficulty of task execution [180], [189]. Due to a lack of visual information, planned subtasks may deviate from the actual scenario, leading to task failure. Therefore, integrating visual information into planning or replanning during execution is necessary. This approach can significantly enhance the accuracy and feasibility of task planning, better addressing the challenges of real-world environments.

2) *Planning utilizing the visual information from embodied perception model:* Based on the above discussion, it is important to integrate visual information into task planning (or replanning). In this process, object labels, locations, or descriptions provided by visual input can offer critical references for task decomposition and execution by LLMs. Through visual information, LLMs can more accurately identify target objects and obstacles in the current environment, thereby optimizing task steps or modifying subtask objectives. Some works use an object detector to query the objects present in the

environment during task execution and feed this information back to the LLM, allowing it to modify unreasonable steps in the current plan [187], [189], [190]. RoboGPT considers the different names of similar objects within the same task, further improving the feasibility of replanning [8]. However, the information provided by labels is still too limited. Can further scene information be provided? SayPlan [191] proposes using hierarchical 3D scene graphs to represent the environment, effectively mitigating the challenges of task planning in large, multi-floor, and multi-room settings. Similarly, ConceptGraphs [192] also adopts 3D scene graphs to provide environmental information to LLMs. Compared to SayPlan, it offers more detailed open-world object detection and presents task planning in a code-based format, which is more efficient and better suited to the demands of complex tasks.

However, limited visual information can result in an agent's inadequate understanding of its environment. While LLMs are provided with visual cues, they often fail to capture the environment's complexity and dynamic changes, leading to misunderstandings and task failures. For example, if a towel is locked in a bathroom cabinet, the agent might repeatedly search the bathroom without considering this possibility [8]. To address this, more robust algorithms must be developed to integrate multiple sensory data, enhancing the agent's environmental understanding. Additionally, leveraging historical data and contextual reasoning, even when visual information is limited, can aid the agent in making reasonable judgments and decisions. This approach of multimodal integration and context-based reasoning not only increases task execution success rates but also provides new perspectives for the advancement of embodied artificial intelligence.

*3) Planning utilizing the VLMs:* Compared to converting environmental information into text using external visual models, VLM models can capture visual details in latent space, particularly contextual information that is difficult to represent with object labels. VLMs can discern rules underlying visual phenomena; for instance, even if a towel is not visible in the environment, it can be inferred that the towel might be stored in a cabinet. This process essentially demonstrates how abstract visual features and structured textual features can be more effectively aligned in latent space. In EmbodiedGPT [193], the Embodied-Former module aligns embodied, visual, and textual information, effectively considering the agent's state and environmental information during task planning. Unlike EmbodiedGPT, which directly uses third-person perspective images, LEO [194] encodes 2D egocentric images and 3D scenes into visual tokens. This method effectively perceives 3D world information and executes tasks accordingly. Similarly, the EIF-Unknown model utilizes Semantic Feature Maps extracted from Voxel Features as visual tokens, which are input along with text tokens into a trained LLaVA model for task planning [195]. Furthermore, embodied multimodal foundation models, or VLA models, have been extensively trained with large datasets in studies like the RT series [2], [9], PaLM-E [196], and Matcha [197] to achieve alignment of visual and textual features in embodied scenarios.

However, task planning is only the first step for an agent in completing an instruction task. Subsequent action planning

determines whether the task can be accomplished. In the experiments from RoboGPT [8], the accuracy of task planning reached 96%, but the overall task completion rate was only 60%, limited by the performance of the low-level planner. Therefore, whether an embodied agent can transition from the cyber space of "imagining how tasks are completed" to the physical world of "interacting with the environment and completing tasks" hinges on effective action planning.

### B. Embodied Action Planning

The distinction between task planning and action planning highlights that action planning must address real-world uncertainties due to the insufficient granularity of task planning subtasks [198]. Action planning can be achieved by: 1) using pre-trained embodied models to complete subtasks via APIs, or 2) leveraging the VLA model's capabilities. The results from action planning are fed back to refine task planning.

*1) Action utilizing APIs:* A common approach involves providing LLMs with definitions of well-trained policy models to understand and use them effectively for specific tasks [189], [199]. By generating code, LLMs can abstract tools into a function library, allowing better handling of sub-tasks [186]. Reflexion adjusts these tools during execution to improve generalization [200]. DEPS enables LLMs to learn and combine various skills through zero-shot learning [201]. The hierarchical planning paradigm simplifies development by separating high-level task planning from specific action execution through policy models. This modularity allows independent development, testing, and optimization, enhancing flexibility and maintainability. While it enables adaptability to various tasks and environments, reliance on external policy models can introduce latency and affect performance, making the quality of these models crucial for overall agent effectiveness.

*2) Action utilizing VLA model:* Different from previous approach that task planning and action execution are performed within the same system, this paradigm leverages the capabilities of embodied multimodal foundation models for planning and executing actions, reducing communication latency and improving system response speed and efficiency. In VLA models, the tight integration of perception, decision-making, and execution modules allows the system to handle complex tasks and adapt to changes in dynamic environments more efficiently. This integration also facilitates real-time feedback, enabling the agent to self-adjust strategies, thereby enhancing the robustness and adaptability of task execution [10], [193], [202]. However, this paradigm is undoubtedly more complex and costly, particularly when dealing with intricate or long-term tasks. Additionally, a key issue is that an action planner, without an embodied world model, cannot simulate physical laws using only the internal knowledge of an LLM. This limitation hinders the agent to accurately and effectively complete various tasks in the physical world, preventing the seamless transfer from cyber space to physical world.

*3) Scalability in Diverse Environments:* Scalability in embodied agents involves adapting to increased complexity in larger and more diverse environments through robust perception, efficient decision-making, and resource optimization.

Strategies include hierarchical SLAM for mapping, multi-modal perception, and energy-efficient edge computing. Collaborative scalability is enhanced via multi-agent systems and decentralized communication, while generalization relies on domain adaptation to operate in new environments.

## VII. SIM-TO-REAL ADAPTATION

Sim-to-Real adaptation in embodied AI refers to the process of transferring capabilities or behaviors learned in simulated environments (cyber space) to real-world scenarios (physical world). It involves validating and improving the effectiveness of algorithms, models, and control strategies developed in simulation to ensure they perform robustly and reliably in physical environments. To achieve sim-to-real adaptation, embodied world models, data collection and training methods, and embodied control algorithms are three essential components.

### A. Embodied World Model

Sim-to-Real involves creating simulation-based world models that closely resemble real-world environments. These models predict the next state to make decisions and are trained from scratch on physical world data, unlike VLA models which are pre-trained on large-scale datasets and fine-tuned with real-world data. World models are effective for structured tasks like autonomous driving and object sorting but are less suited for unstructured, complex tasks.

Learning world models is promising of the physical simulation field. Compared to traditional simulation methods, it offers significant advantages, such as the ability to reason about interactions with incomplete information, meet real-time computation requirements, and improve prediction accuracy over time. The predictive capability of such world models is crucial, enabling robots to develop the physical intuition necessary to operate in the human world. As shown in Fig. 10, according to the learning pipeline of the world environment, they can be divided into Generation-based methods, Prediction-based methods and Knowledge-driven methods.

1) *Generation-based Methods*: As the scale of models and data increases, generative models have demonstrated the ability to understand and generate images (e.g., World Models [203]), videos (e.g., Sora [16], Pandora [204]), point clouds (e.g., 3D-VLA [205]) or other formats of data (e.g., DWM [206]) that conform to physical laws. This capability suggests that generative models can internalize world knowledge. Specifically, after exposure to large datasets, these models not only capture statistical properties but also simulate physical and causal relationships through their intrinsic structures. Thus, generative models function as more than just pattern recognition tools; they exhibit characteristics of world models. The embedded world knowledge in these models can be harnessed to enhance the performance of other models. By leveraging this knowledge, we can improve model generalization, robustness, adaptability to new environments, and predictive accuracy on unseen data [204], [205]. However, generative models also have certain limitations and drawbacks. For instance, they may produce inaccurate or distorted outputs when faced with biased data distributions or insufficient training

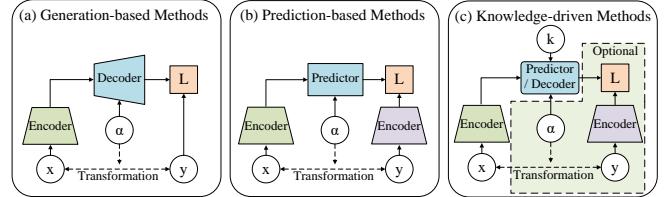


Fig. 10. Embodied world models can be roughly divided into three type. (a) **Generation-based Methods** learn the transformation relation between the input space and the output space using an autoencoder framework. (b) **Prediction-based Methods** are more general frameworks where a world model is trained in latent space. (c) **Knowledge-driven Methods** inject artificially constructed knowledge into the model, giving the model world knowledge to obtain output that meets the given knowledge constraints. Note that the components within the dashed line are optional.

data. Moreover, these models require substantial computational resources and time for training and often lack interpretability, which hinders practical application. While generative models have shown promise in generating content that adheres to physical laws, challenges such as improving efficiency, enhancing interpretability, and mitigating data bias must be addressed for broader application. Continued research is likely to unlock further value and potential in these models.

2) *Prediction-based Methods*: The prediction-based world model predicts and understands the environment by constructing and utilizing internal representations. By reconstructing corresponding features in the latent space based on provided conditions, it captures deeper semantics and associated world knowledge. This model maps input information to a latent space and operates within that space to extract and utilize high-level semantic information, thereby enabling the robots to perceive the essential representation of the world environment (e.g., I-JEPA [15], MC-JEPA [207], A-JEPA [208], Point-JEPA [209], IWM [210]) and more accurately perform embodied downstream tasks (e.g., iVideoGPT [211], IRASim [212], STP [213], MuDreamer [214]). Latent features, unlike pixel-level information, can abstract and decouple various forms of knowledge, enabling models to handle complex tasks and scenes more effectively while enhancing generalization [215]. In spatiotemporal modeling, for instance, a world model predicts an object's post-interaction state by integrating its current state, the nature of the interaction, and internal knowledge. Specifically, embodied world models generate dynamic environmental predictions by combining perceptual information with prior knowledge, relying on both sensory data and inherent world knowledge to accurately infer and predict environmental changes [211], [213], [214]. This process considers the current state of objects alongside their historical and contextual information.

Similarly, leveraging the world knowledge embedded in its representations can further enhance the model's perception and robustness [15], [207], [210], [216]. By operating in latent space, it is expected that robots can maintain high performance in different environments at a lower cost [214]. The key to this approach lies in abstract processing and knowledge decoupling, enabling efficient adaptation to complex situations. However, such models may exhibit limitations and instability when dealing with previously unseen environments and conditions. Additionally, the world knowledge decoupled

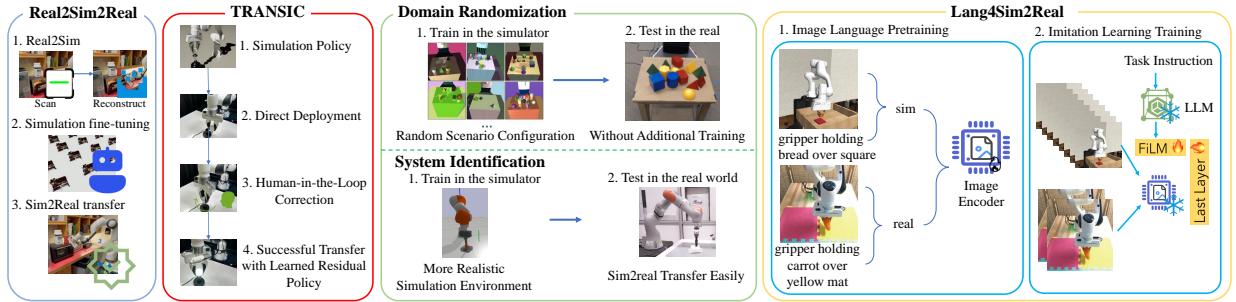


Fig. 11. Five pipelines to achieve sim-to-real gap. “**Real2Sim2Real**” reduces the gap by reconstructing real scenes. “**TRANSIC**” compensates for the sim-to-real transfer gap through human-corrected interventions. “**Domain Randomization**” enhances model transfer adaptability by simulating environmental diversity. “**System Identification**” improves sim-to-real environment similarity, thereby mitigating the sim-real gap. “**Lang4Sim2Real**” uses natural language to bridge two domains, learning invariant image representations and reducing visual gaps.

in the latent space may have interpretability issues.

**3) Knowledge-driven Methods:** Knowledge-driven world models inject artificially constructed knowledge into the models, endowing them with world knowledge. This method has shown broad application potential in the field of embodied AI. For example, in the real2sim2real approach [217], real-world knowledge is used to build physics-compliant simulators, which are then used to train robots, enhancing model robustness and generalization capabilities. Additionally, artificially constructing common sense or physics-compliant knowledge and applying them to generative models or simulators is a common strategy (e.g., ElastoGen [218], One-2-3-45 [219], PLoT [220]). This approach imposes more physically accurate constraints on the model, enhancing its reliability and interpretability in generation tasks. These constraints ensure the model’s knowledge is both accurate and consistent, reducing uncertainty during training and application. Some approaches combine artificially created physical rules with LLMs or MLMs. By leveraging the commonsense capabilities of LLMs and MLMs, these approaches (e.g., Holodeck [56], LEGENT [221], GRUtopia [222]) generate diverse and semantically rich scenes through automatic spatial layout optimization. This greatly advances the development of general-purpose embodied agents by training them in novel and diverse environments.

**4) Limitations:** Current limitations of world models include handling the complexity and variability of real-world environments, such as high-dimensional sensory inputs, dynamic and stochastic elements, and long-term dependencies. Many models struggle with generalization across unseen scenarios, often requiring extensive training data and computational resources. Additionally, sim-to-real transfer remains problematic, as simulated environments fail to fully capture real-world physics and noise. These limitations can be addressed by integrating more realistic simulators, incorporating multimodal sensory inputs, and using hierarchical or modular architectures. Improving data efficiency, enhancing transfer learning techniques, and incorporating real-world priors can also enable more accurate predictions and adaptive decision-making. Recently, the Cosmos Platform [223] was proposed by NVIDIA, which contains autoregressive and diffusion models for Text-to-World and Video-to-World generation, to accelerate the development of physical AI systems. It may give us some inspirations about how to build an embodied world model.

## B. Data Collection and Training

For sim-to-real adaptation, the high-quality data is important. Traditional data collection methods involve expensive equipment and precise operations, which are time-consuming and labor-intensive. Recently, some efficient and cost-effective methods have been proposed for high-quality demonstration data collection and training. We discuss data collection methods in both real-world and simulated environments.

**1) Real-World Data:** Training large, high-capacity models with rich datasets has shown great success. This approach is also promising for robotics, where large, diverse datasets can enhance generalization and adaptability. Open X-Embodiment [202] provides data from 22 robots with 527 skills and 160,266 tasks in domestic settings. UMI [224] offers a framework for collecting dynamic, bimanual data using a handheld gripper. Mobile ALOHA [225] enables data collection for full-body mobile manipulation tasks. Human-agent collaboration [226] improves data quality and efficiency by combining human input with agent refinement processes.

**2) Simulated Data:** Data collection in real-world settings is resource-intensive and time-consuming, making simulation-based data collection an attractive alternative. Simulated environments allow for automated, efficient data collection. For example, CLIPORT [168] and Transporter Networks [227] utilized Pybullet simulator data for training and successfully transferred models to real-world applications. GAPartNet [228] developed a large-scale dataset with detailed part-level annotations for better object interaction in both simulations and reality. SemGrasp [167] created the CapGrasp dataset for semantically rich hand grasping in virtual environments.

**3) Sim-to-Real Paradigms:** Recently, several sim-to-real paradigms have been introduced, to mitigate the need for extensive and costly real-world demonstration data by conducting extensive learning in simulation environments, followed by migration to real-world settings. This section outlines five paradigms for sim-to-real transfer, as shown in Fig. 11.

Real2Sim2real [229] improves imitation learning by using reinforcement learning in a “digital twin” simulation to develop strategies, which are then transferred to the real world. TRANSIC [230] reduces the sim-to-real gap through real-time human intervention and residual policy training based on corrected behaviors. Domain Randomization [231]–[233] increases model generalization by varying simulation parameters to cover real-world conditions. System Identifica-



Fig. 12. Exemplar tasks from ARIO, where the top row indicates the task category while the text at the bottom row provides task instructions.

tion [234], [235] creates accurate simulations of real-world scenes to ensure smooth transitions from simulation to reality. Lang4sim2real [236] leverages natural language descriptions to bridge the sim-to-real gap, improving model generalization with cross-domain image representations.

4) **ARIO (All Robots in One)**: Despite the seemingly unified structure of pre-training datasets like Open X-embodiment, several critical issues remain unresolved. These issues include the absence of comprehensive sensory modalities—no dataset currently integrates images, 3D vision, text, tactile, and auditory inputs simultaneously. Additionally, the lack of a unified format in multi-robot datasets complicates data processing and loading. Furthermore, there is an incompatibility in representing diverse control objects across different robotic platforms, insufficient data volume that hinders large-scale pretraining, and a scarcity of datasets that combine both simulated and real data, which is essential for addressing the sim-to-real gap.

To overcome these challenges, we propose ARIO (All Robots In One), which is a new dataset standard that optimizes existing datasets and facilitates the development of more versatile and general-purpose embodied agents. The ARIO standard [237] records control and motion data from robots with different morphologies in a unified format. ARIO's unified format accommodates variable data from diverse robot types, ensuring precise timestamps. This standard enables users to efficiently train high-performing, generalizable embodied AI models, positioning ARIO as the ideal format for embodied AI datasets. Building on the ARIO standard, a unified large-scale ARIO dataset is further developed, which comprises approximately 3 million episodes collected from 258 series and 321,064 tasks. Fig. 12 shows exemplar tasks from ARIO.

The ARIO dataset addresses the limitations of current datasets and facilitates the development of robust, general-purpose embodied agents. By providing a cohesive framework for data collection and representation, ARIO paves the way for the development of more powerful and versatile embodied agents, capable of navigating and interacting with the physical world in complex and diverse ways.

5) **Real-world Deployments of Embodied AI Systems**: Embodied AI systems have made significant strides across various scenes. In healthcare, robots like the Da Vinci Surgical System and Moxi automate tasks such as surgery precision and supply deliveries. In logistics, Amazon Robotics and Boston Dynamics' Stretch improve efficiency in warehousing and transportation. Manufacturing benefits from AI-driven robots like Fanuc and ABB, enhancing precision and collaboration.

Nevertheless, sim-to-real adaptation faces significant challenges, including the domain gap between simulation and real-world data distributions, the complexity of dynamic real-world interactions, and the limited diversity of training data. Models often overfit to simulations, struggle with real-world sensor noise, and fail to handle unexpected events.

### VIII. CHALLENGES AND FUTURE DIRECTIONS

Despite of the rapid progress of embodied AI, it faces several challenges and presents exciting future directions.

**High-quality Robotic Datasets:** Obtaining sufficient real world robotic data remains a significant challenge. Collecting this data is both time-consuming and resource-intensive. Relying solely on simulation data worsens the sim-to-real gap problem. Creating diverse real world robotic datasets necessitates close and extensive collaboration among various institutions. Additionally, the development of more realistic and efficient simulators is essential for improving the quality of simulated data. For building generalizable embodied models capable of cross-scenario and cross-task applications in robotics, it is essential to construct large-scale datasets, leveraging high-quality simulated environment data to assist real world data.

**Long-Horizon Task Execution:** Executing single instructions can often entail long-horizon tasks for robots, exemplified by commands like “clean the kitchen”, which involve activities such as rearranging objects, sweeping floors, wiping tables, and more. Accomplishing such tasks successfully necessitates the robot’s ability to plan and execute a sequence of low-level actions over extended time spans. While current high-level task planners have shown initial success, they often prove inadequate in diverse scenarios due to their lack of tuning for embodied tasks. Addressing this challenge requires the development of efficient planners equipped with robust perception capabilities and much commonsense knowledge. To balance the trade-off between planning complexity and real-time adaptability, we can combine lightweight monitor modules for high-frequency monitoring, and two adapters for subtask and path adaptation reasoning at a lower frequency.

**Causal Reasoning:** Existing data-driven embodied agents make decisions based on intrinsic data correlations. However, this approach does not allow the models to truly understand the causal relations between knowledge, behavior, and environment, resulting in biased strategies. This makes it difficult to ensure that they can operate in real-world environments in a robust and reliable manner. Therefore, it is important for embodied agents to be driven by world knowledge, capable of autonomous causal reasoning. By understanding the world through interaction and learning its workings via abductive reasoning, we can further enhance the adaptability, decision reliability, and generalization capabilities of embodied agents in complex real-world environments. For embodied tasks, it is necessary to establish spatial-temporal causal relations across modalities through interactive instructions and state predictions. Moreover, agents need to understand the affordances of objects to achieve adaptive task planning in dynamic scenes.

**Unified Evaluation Benchmark:** While numerous benchmarks exist for evaluating low-level control policies, they vary

significantly in the assessed skills. Furthermore, the objects and scenes in these benchmarks are typically limited by simulator constraints. To comprehensively evaluate embodied models, the benchmark should encompass a diverse range of skills using realistic simulators. Many benchmarks for high-level task planners focus on assessing planning capability through question-answering tasks. However, a more desirable approach involves evaluating both the high-level task planner and the low-level control policy together for executing long-horizon tasks and measuring success rates, rather than relying solely on isolated assessments of the planner.

**Security and Privacy:** Embodied agents face significant security challenges when deployed in sensitive or private spaces. These agents often rely on LLMs for decision-making, which introduces new vulnerabilities. For instance, LLMs are susceptible to backdoor attacks like word injection, scenario manipulation, and knowledge injection, which can lead to dangerous outcomes such as autonomous vehicles accelerating towards obstacles or robots performing hazardous actions. To mitigate these risks, we can evaluate potential attack vectors and develop more robust defenses. Additionally, the secure prompting, state management, and safety validation mechanisms can be used to enhance the security and robustness.

## IX. CONCLUSION

Embodied AI allows agents to sense, perceive, and interact with various objects from both cyber space and physical world, which exhibits its vital significance toward achieving AGI. This survey extensively reviews embodied robots, simulators, four representative embodied tasks: visual active perception, embodied interaction, embodied agents and sim-to-real adaptation, and future research directions. The comparative summary of the embodied robots, simulators, datasets, and approaches provides a clear picture of the recent development in embodied AI, which greatly benefits the future research along this emerging and promising research direction.

## REFERENCES

- [1] C. Machinery, “Computing machinery and intelligence-am turing,” *Mind*, vol. 59, no. 236, p. 433, 1950.
- [2] 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 *CoRL*, 2023, pp. 2165–2183.
- [3] S. Belkhale, T. Ding, T. Xiao, P. Serbanet, Q. Vuong, J. Tompson, Y. Chebotar, D. Dwibedi, and D. Sadigh, “Rt-h: Action hierarchies using language,” *arXiv preprint arXiv:2403.01823*, 2024.
- [4] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, and Y. Tang, “A brief overview of chatgpt: The history, status quo and potential future development,” *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, 2023.
- [5] 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*, vol. 6, no. 2, pp. 230–244, 2022.
- [6] 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.
- [7] H. Gao, Y. Liu, W. Sun, and X. Yu, “Adaptive wavelet tracking control of dual-linear-motor-driven gantry stage with suppression of crossbeam rotation,” *IEEE/ASME TMECH*, 2023.
- [8] Y. Chen, W. Cui, Y. Chen, M. Tan, X. Zhang, D. Zhao, and H. Wang, “Robogpt: an intelligent agent of making embodied long-term decisions for daily instruction tasks,” *arXiv preprint arXiv:2311.15649*, 2023.
- [9] 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.
- [10] 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 *CoRL*, 2023, pp. 2165–2183.
- [11] Y. Hu, Q. Xie, V. Jain, J. Francis, J. Patrikar, N. Keetha, S. Kim, Y. Xie, T. Zhang, Z. Zhao *et al.*, “Toward general-purpose robots via foundation models: A survey and meta-analysis,” *arXiv preprint arXiv:2312.08782*, 2023.
- [12] R. McCarthy, D. C. Tan, D. Schmidt, F. Acer, N. Herr, Y. Du, T. G. Thuruthel, and Z. Li, “Towards generalist robot learning from internet video: A survey,” *arXiv preprint arXiv:2404.19664*, 2024.
- [13] 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 *ICML*. PMLR, 2021, pp. 8748–8763.
- [14] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *ICML*. PMLR, 2023, pp. 19730–19742.
- [15] 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 *CVPR*, 2023, pp. 15619–15629.
- [16] 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.
- [17] R. Pfeifer and F. Iida, “Embodied artificial intelligence: Trends and challenges,” *Lecture notes in computer science*, pp. 1–26, 2004.
- [18] 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 TMECH*, pp. 1–11, 2024.
- [19] Z. Liao, G. Jiang, F. Zhao, Y. Wu, Y. Yue, and X. Mei, “Dynamic skill learning from human demonstration based on the human arm stiffness estimation model and riemannian dmp,” *IEEE/ASME TMECH*, vol. 28, no. 2, pp. 1149–1160, 2023.
- [20] X. Zhao, Y. Zhang, W. Ding, B. Tao, and H. Ding, “A dual-arm robot cooperation framework based on a nonlinear model predictive cooperative control,” *IEEE/ASME TMECH*, vol. 29, no. 5, pp. 3993–4005, 2024.
- [21] C. Li, F. Liu, Y. Wang, and M. Buss, “Data-informed residual reinforcement learning for high-dimensional robotic tracking control,” *IEEE/ASME TMECH*, pp. 1–11, 2024.
- [22] V. Ortenzi, N. Marturi, M. Mistry, J. Kuo, and R. Stolk, “Vision-based framework to estimate robot configuration and kinematic constraints,” *IEEE/ASME TMECH*, vol. 23, no. 5, pp. 2402–2412, 2018.
- [23] S. K. Kommuri, S. Han, and S. Lee, “External torque estimation using higher order sliding-mode observer for robot manipulators,” *IEEE/ASME TMECH*, vol. 27, no. 1, pp. 513–523, 2022.
- [24] E. Spyrapos-Papastavridis, P. R. N. Childs, and J. S. Dai, “Passivity preservation for variable impedance control of compliant robots,” *IEEE/ASME TMECH*, vol. 25, no. 5, pp. 2342–2353, 2020.
- [25] P. R. Wurman, R. D’Andrea, and M. Mountz, “Coordinating hundreds of cooperative, autonomous vehicles in warehouses,” *AI magazine*, vol. 29, no. 1, pp. 9–9, 2008.
- [26] Z. Shao and J. Zhang, “Vision-based adaptive trajectory tracking control of wheeled mobile robot with unknown translational external parameters,” *IEEE/ASME TMECH*, vol. 29, no. 1, pp. 358–365, 2024.
- [27] D. A. Bennett, S. A. Dalley, D. Truex, and M. Goldfarb, “A multi-grasp hand prosthesis for providing precision and conformal grasps,” *IEEE/ASME TMECH*, vol. 20, no. 4, pp. 1697–1704, 2015.
- [28] W. Chen, C. Xiong, and S. Yue, “Mechanical implementation of kinematic synergy for continual grasping generation of anthropomorphic hand,” *IEEE/ASME TMECH*, vol. 20, no. 3, pp. 1249–1263, 2015.
- [29] 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,” *NeurIPS*, vol. 36, 2024.
- [30] B. Siciliano, O. Khatib, and T. Kröger, *Springer handbook of robotics*. Springer, 2008, vol. 200.
- [31] Y. Yang, Z. He, P. Jiao, and H. Ren, “Bioinspired soft robotics: How do we learn from creatures?” *IEEE Reviews in Biomedical Engineering*, 2022.

- [32] R. K. Katzschmann, J. DelPreto, R. MacCurdy, and D. Rus, "Exploration of underwater life with an acoustically controlled soft robotic fish," *Science Robotics*, vol. 3, no. 16, p. eaar3449, 2018.
- [33] G. C. de Croon, J. Dupeyroux, S. B. Fuller, and J. A. Marshall, "Insect-inspired ai for autonomous robots," *Science robotics*, vol. 7, no. 67, p. eab16334, 2022.
- [34] N. R. Sinatra, C. B. Teeple, D. M. Vogt, K. K. Parker, D. F. Gruber, and R. J. Wood, "Ultralight manipulation of delicate structures using a soft robotic gripper," *Science Robotics*, vol. 4, no. 33, p. eaax5425, 2019.
- [35] G. Authors, "Genesis: A universal and generative physics engine for robotics and beyond," December 2024. [Online]. Available: <https://github.com/Genesis-Embodied-AI/Genesis>
- [36] NVIDIA, "Nvidia isaac sim: Robotics simulation and synthetic data," 2023. [Online]. Available: <https://developer.nvidia.com/isaac-sim>
- [37] V. Makovychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.
- [38] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *IROS*, vol. 3, 2004, pp. 2149–2154.
- [39] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," 2016.
- [40] Cyberbotics, "Webots: open-source robot simulator." [Online]. Available: <https://github.com/cyberbotics/webots>
- [41] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *IROS*, 2012, pp. 5026–5033.
- [42] A. Juliani, V.-P. Berges, E. Teng, A. Cohen, J. Harper, C. Elion, C. Goy, Y. Gao, H. Henry, M. Mattar, and D. Lange, "Unity: A general platform for intelligent agents," *arXiv preprint arXiv:1809.02627*, 2020.
- [43] 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.
- [44] ISAE-SUPAERO, "Morse: the modular open robots simulator engine."
- [45] E. Rohmer, S. P. Singh, and M. Freese, "V-rep: A versatile and scalable robot simulation framework," in *IROS*, 2013, pp. 1321–1326.
- [46] C. Wang, Q. Zhang, Q. Tian, S. Li, X. Wang, D. Lane, Y. Petillot, and S. Wang, "Learning mobile manipulation through deep reinforcement learning," *Sensors*, vol. 20, no. 3, p. 939, 2020.
- [47] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *IROS*, vol. 3, 2004, pp. 2149–2154.
- [48] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, L. Yi, A. X. Chang, L. J. Guibas, and H. Su, "Sapien: A simulated part-based interactive environment," in *CVPR*, Jun 2020.
- [49] X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, and A. Torralba, "Virtualhome: Simulating household activities via programs," in *CVPR*, Jun 2018.
- [50] E. Kolve, R. Mottaghi, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv: Computer Vision and Pattern Recognition, arXiv: Computer Vision and Pattern Recognition*, Dec 2017.
- [51] B. Shen, F. Xia, C. Li, R. Martín-Martín, L. Fan, G. Wang, C. Pérez-D'Arpino, S. Buch, S. Srivastava, L. Tchapmi, M. Tchapmi, K. Vainio, J. Wong, L. Fei-Fei, and S. Savarese, "igibson 1.0: A simulation environment for interactive tasks in large realistic scenes," in *IROS*, 2021, pp. 7520–7527.
- [52] C. Gan, J. Schwartz, S. Alter, M. Schrimpf, J. Traer, J. Freitas, J. Kubilius, A. Bhandwaldar, N. Haber, M. Sano, K. Kim, E. Wang, D. Mrowca, M. Lingelbach, A. Curtis, K. Feigelis, D. Bear, D. Gutfreund, D. Cox, J. DiCarlo, J. McDermott, J. Tenenbaum, and D. Yamins, "Threedeworld: A platform for interactive multi-modal physical simulation," *NeurIPS*, Dec 2021.
- [53] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," in *3DV*, Oct 2017.
- [54] P. Ren, M. Li, Z. Luo, X. Song, Z. Chen, W. Liu, Y. Yang, H. Zheng, R. Xu, Z. Huang *et al.*, "Infiniteworld: A unified scalable simulation framework for general visual-language robot interaction," *arXiv preprint arXiv:2412.05789*, 2024.
- [55] Y. Wang, Z. Xian, F. Chen, T.-H. Wang, Y. Wang, K. Fragkiadaki, Z. Erickson, D. Held, and C. Gan, "Robogen: Towards unleashing infinite data for automated robot learning via generative simulation," *arXiv:2311.01455*, Nov 2023.
- [56] Y. Yang, F.-Y. Sun, L. Weihs, E. VanderBilt, A. Herrasti, W. Han, J. Wu, N. Haber, R. Krishna, L. Liu *et al.*, "Holodeck: Language guided generation of 3d embodied ai environments," in *CVPR*, 2024, pp. 16 227–16 237.
- [57] Y. Yang, B. Jia, P. Zhi, and S. Huang, "Physcene: Physically interactable 3d scene synthesis for embodied ai," in *CVPR*, 2024.
- [58] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, J. Salvador, K. Ehsani, W. Han, E. Kolve, A. Farhadi, A. Kembhavi, and R. Mottaghi, "ProcTHOR: Large-Scale Embodied AI Using Procedural Generation," in *NeurIPS*, 2022, outstanding Paper Award.
- [59] L. Fei-Fei and R. Krishna, "Searching for computer vision north stars," *Daedalus*, vol. 151, no. 2, pp. 85–99, 2022.
- [60] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE TPAMI*, vol. 29, no. 6, pp. 1052–1067, 2007.
- [61] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE TRO*, vol. 31, no. 5, pp. 1147–1163, 2015.
- [62] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *ECCV*. Springer, 2014, pp. 834–849.
- [63] 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 *CVPR*, 2013, pp. 1352–1359.
- [64] L. Nicholson, M. Milford, and N. Sürdnerhauf, "Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam," *IEEE RAL*, vol. 4, no. 1, pp. 1–8, 2018.
- [65] Z. Liao, Y. Hu, J. Zhang, X. Qi, X. Zhang, and W. Wang, "So-slam: Semantic object slam with scale proportional and symmetrical texture constraints," *IEEE RAL*, vol. 7, no. 2, pp. 4008–4015, 2022.
- [66] S. Cheng, C. Sun, S. Zhang, and D. Zhang, "Sg-slam: A real-time rgbd visual slam toward dynamic scenes with semantic and geometric information," *IEEE TIM*, vol. 72, pp. 1–12, 2022.
- [67] J. He, M. Li, Y. Wang, and H. Wang, "Ovd-slam: An online visual slam for dynamic environments," *IEEE Sensors Journal*, 2023.
- [68] 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 *CVPR*, 2024, pp. 19 595–19 604.
- [69] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *CVPR*, 2017, pp. 1907–1915.
- [70] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *CVPR*, 2019, pp. 12 697–12 705.
- [71] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *ICCV*, 2015, pp. 945–953.
- [72] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *IROS*, 2015, pp. 922–928.
- [73] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *CVPR*, 2017, pp. 1746–1754.
- [74] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *CVPR*, 2019, pp. 3075–3084.
- [75] B. Graham, M. Engelcke, and L. Van Der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *CVPR*, 2018, pp. 9224–9232.
- [76] 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 *CVPR*, 2024, pp. 19 757–19 767.
- [77] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, 2017, pp. 652–660.
- [78] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *NeurIPS*, vol. 30, 2017.
- [79] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu, "Rethinking network design and local geometry in point cloud: A simple residual mlp framework," *arXiv preprint arXiv:2202.07123*, 2022.
- [80] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *ICCV*, 2021, pp. 16 259–16 268.
- [81] Y.-Q. Yang, Y.-X. Guo, J.-Y. Xiong, Y. Liu, H. Pan, P.-S. Wang, X. Tong, and B. Guo, "Swin3d: A pretrained transformer backbone for 3d indoor scene understanding," *arXiv preprint arXiv:2304.06906*, 2023.

- [82] X. Wu, Y. Lao, L. Jiang, X. Liu, and H. Zhao, “Point transformer v2: Grouped vector attention and partition-based pooling,” *NeurIPS*, vol. 35, pp. 33 330–33 342, 2022.
- [83] Z. Zhu, X. Ma, Y. Chen, Z. Deng, S. Huang, and Q. Li, “3d-vista: Pre-trained transformer for 3d vision and text alignment,” in *ICCV*, 2023, pp. 2911–2921.
- [84] J. Huang, S. Yong, X. Ma, X. Linghu, P. Li, Y. Wang, Q. Li, S.-C. Zhu, B. Jia, and S. Huang, “An embodied generalist agent in 3d world,” in *ICML*, 2024.
- [85] Z. Zhu, Z. Zhang, X. Ma, X. Niu, Y. Chen, B. Jia, Z. Deng, S. Huang, and Q. Li, “Unifying 3d vision-language understanding via promptable queries,” *arXiv preprint arXiv:2405.11442*, 2024.
- [86] D. Liang, X. Zhou, X. Wang, X. Zhu, W. Xu, Z. Zou, X. Ye, and X. Bai, “Pointmamba: A simple state space model for point cloud analysis,” *arXiv preprint arXiv:2402.10739*, 2024.
- [87] X. Han, Y. Tang, Z. Wang, and X. Li, “Mamba3d: Enhancing local features for 3d point cloud analysis via state space model,” *arXiv preprint arXiv:2404.14966*, 2024.
- [88] L. Pinto, D. Gandhi, Y. Han, Y.-L. Park, and A. Gupta, “The curious robot: Learning visual representations via physical interactions,” in *ECCV*, 2016, pp. 3–18.
- [89] G. Tatya, J. Francis, and J. Sinapov, “Transferring implicit knowledge of non-visual object properties across heterogeneous robot morphologies,” in *ICRA*, 2023, pp. 11 315–11 321.
- [90] D. Jayaraman and K. Grauman, “Learning to look around: Intelligently exploring unseen environments for unknown tasks,” in *CVPR*, 2018, pp. 1238–1247.
- [91] L. Jin, X. Chen, J. Rückin, and M. Popović, “Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering,” in *IROS*, 2023, pp. 11 305–11 312.
- [92] Y. Hu, J. Geng, C. Wang, J. Keller, and S. Scherer, “Off-policy evaluation with online adaptation for robot exploration in challenging environments,” *IEEE RAL*, 2023.
- [93] L. Fan, M. Liang, Y. Li, G. Hua, and Y. Wu, “Evidential active recognition: Intelligent and prudent open-world embodied perception,” in *CVPR*, 2024, pp. 16 351–16 361.
- [94] S. Mokssit, D. B. Licea, B. Guermah, and M. Ghogho, “Deep learning techniques for visual slam: A survey,” *IEEE Access*, vol. 11, pp. 20 026–20 050, 2023.
- [95] K. Chen, J. Zhang, J. Liu, Q. Tong, R. Liu, and S. Chen, “Semantic visual simultaneous localization and mapping: A survey,” *arXiv preprint arXiv:2209.06428*, 2022.
- [96] P. K. Vinodkumar, D. Karabulut, E. Avots, C. Ozcinar, and G. Anbarjafari, “A survey on deep learning based segmentation, detection and classification for 3d point clouds,” *Entropy*, vol. 25, no. 4, p. 635, 2023.
- [97] H. Durrant-Whyte and T. Bailey, “Simultaneous localization and mapping: part i,” *IEEE robotics & automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.
- [98] X. Zhang, J. Lai, D. Xu, H. Li, and M. Fu, “2d lidar-based slam and path planning for indoor rescue using mobile robots,” *Journal of Advanced Transportation*, vol. 2020, no. 1, p. 8867937, 2020.
- [99] J. Ruan, B. Li, Y. Wang, and Z. Fang, “Gp-slam+: real-time 3d lidar slam based on improved regionalized gaussian process map reconstruction,” in *IROS*, 2020, pp. 5171–5178.
- [100] J. Luo, Y. Liu, W. Chen, Z. Li, Y. Wang, G. Li, and L. Lin, “Dspnet: Dual-vision scene perception for robust 3d question answering,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [101] Z. Wei, J. Lin, Y. Liu, W. Chen, J. Luo, G. Li, and L. Lin, “3dafford-splat: Efficient affordance reasoning with 3d gaussians,” *arXiv preprint arXiv:2504.11218*, 2025.
- [102] V. Mittal, “Attnounder: Talking to cars with attention,” in *ECCV*. Springer, 2020, pp. 62–73.
- [103] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, and L. Zhang, “Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation,” in *CVPR*, 2019, pp. 6629–6638.
- [104] C. Bermejo, L. H. Lee, P. Chojecki, D. Przewozny, and P. Hui, “Exploring button designs for mid-air interaction in virtual reality: A hexa-metric evaluation of key representations and multi-modal cues,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. EICS, pp. 1–26, 2021.
- [105] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sunderhauf, I. Reid, S. Gould, and A. van den Hengel, “Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments,” in *CVPR*, Jun 2018.
- [106] V. Jain, G. Magalhaes, A. Ku, A. Vaswani, E. Ie, and J. Baldridge, “Stay on the path: Instruction fidelity in vision-and-language navigation,” in *ACL*, Jan 2019.
- [107] J. Krantz, E. Wijmans, A. Majumdar, D. Batra, and S. Lee, *Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments*, Jan 2020, p. 104–120.
- [108] H. Chen, A. Suhr, D. Misra, N. Snavely, and Y. Artzi, “Touchdown: Natural language navigation and spatial reasoning in visual street environments,” in *CVPR*, Jun 2019.
- [109] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. van den Hengel, “Reverie: Remote embodied visual referring expression in real indoor environments,” in *CVPR*, Jun 2020.
- [110] F. Zhu, X. Liang, Y. Zhu, Q. Yu, X. Chang, and X. Liang, “Soon: Scenario oriented object navigation with graph-based exploration,” in *CVPR*, Jun 2021.
- [111] H. Wang, A. G. H. Chen, X. Li, M. Wu, and H. Dong, “Find what you want: Learning demand-conditioned object attribute space for demand-driven navigation,” *NeurIPS*, 2023.
- [112] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, “Alfred: A benchmark for interpreting grounded instructions for everyday tasks,” in *CVPR*, Jun 2020.
- [113] S. Yenamandra, A. Ramachandran, K. Yadav, A. Wang, M. Khanna, T. Gervet, T.-Y. Yang, V. Jain, A. Clegg, J. Turner, Z. Kira, M. Savva, A. Chang, D. Chaplot, D. Batra, R. Mottaghi, Y. Bisk, and C. Paxton, “Homerobot: Open-vocabulary mobile manipulation,” in *NeurIPS*, Jun 2023.
- [114] 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,” in *CoRL*. PMLR, 2023, pp. 80–93.
- [115] J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer, “Vision-and-dialog navigation,” in *CoRL*. PMLR, 2020, pp. 394–406.
- [116] X. Gao, Q. Gao, R. Gong, K. Lin, G. Thattai, and G. Sukhatme, “Dialfred: Dialogue-enabled agents for embodied instruction following,” *arXiv:2202.13330*, 2022.
- [117] Y. Hong, C. Rodriguez, Y. Qi, Q. Wu, and S. Gould, “Language and visual entity relationship graph for agent navigation,” *NeurIPS*, vol. 33, pp. 7685–7696, 2020.
- [118] W. Zhang, C. Ma, Q. Wu, and X. Yang, “Language-guided navigation via cross-modal grounding and alternate adversarial learning,” *IEEE TCSVT*, vol. 31, pp. 3469–3481, 2020.
- [119] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, and L. Zhang, “Vision-language navigation policy learning and adaptation,” *IEEE TPAMI*, vol. 43, no. 12, pp. 4205–4216, 2021.
- [120] S. Y. Min, D. S. Chaplot, P. K. Ravikumar, Y. Bisk, and R. Salakhutdinov, “FILM: Following instructions in language with modular methods,” in *ICLR*, 2022.
- [121] D. Shah, B. Osinski, B. Ichter, and S. Levine, “Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action,” in *CoRL*, 2022.
- [122] Y. Qiao, Y. Qi, Y. Hong, Z. Yu, P. Wang, and Q. Wu, “Hop: History-and-order aware pretraining for vision-and-language navigation,” in *CVPR*, 2022, pp. 15 397–15 406.
- [123] D. Zheng, S. Huang, L. Zhao, Y. Zhong, and L. Wang, “Towards learning a generalist model for embodied navigation,” in *CVPR*, Jun 2024.
- [124] J. Gao, X. Yao, and C. Xu, “Fast-slow test-time adaptation for online vision-and-language navigation,” 2024.
- [125] Y. Long, X. Li, W. Cai, and H. Dong, “Discuss before moving: Visual language navigation via multi-expert discussions,” in *ICRA*, 2024.
- [126] R. D. M. S. C. L. Q. C. Liuyi Wang, Zongtao He, “Vision-and-language navigation via causal learning,” in *CVPR*, Jun 2024.
- [127] Y. Y. Rui Liu, Wenguang Wang, “Volumetric environment representation for vision-language navigation,” in *CVPR*, Jun 2024.
- [128] J. Zhang, K. Wang, R. Xu, G. Zhou, Y. Hong, X. Fang, Q. Wu, Z. Zhang, and W. He, “Navid: Video-based vlm plans the next step for vision-and-language navigation,” *ArXiv*, vol. abs/2402.15852, 2024.
- [129] X. E. Wang, W. Xiong, H. Wang, and W. Y. Wang, “Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation,” in *ECCV*, 2018.
- [130] D. An, Y. Qi, Y. Huang, Q. Wu, L. Wang, and T. Tan, “Neighbor-view enhanced model for vision and language navigation,” in *ACM MM*, 2021, pp. 5101–5109.
- [131] Y. Hong, Z. Wang, Q. Wu, and S. Gould, “Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation,” in *CVPR*, 2022, pp. 15 418–15 428.

- [132] Y. Qiao, Y. Qi, Z. Yu, J. Liu, and Q. Wu, "March in chat: Interactive prompting for remote embodied referring expression," in *ICCV*, October 2023, pp. 15 758–15 767.
- [133] Z. Wang, X. Li, J. Yang, Y. Liu, J. Hu, M. Jiang, and S. Jiang, "Look-ahead exploration with neural radiance representation for continuous vision-language navigation," in *CVPR*, 2024.
- [134] 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 TPAMI*, 2024.
- [135] S. Bhambri, B. Kim, and J. Choi, "Multi-level compositional reasoning for interactive instruction following," in *AAAI*, vol. 37, no. 1, 2023, pp. 223–231.
- [136] C. Xu, H. T. Nguyen, C. Amato, and L. L. Wong, "Vision and language navigation in the real world via online visual language mapping," *ArXiv*, vol. abs/2310.10822, 2023.
- [137] X. Song, W. Chen, Y. Liu, W. Chen, G. Li, and L. Lin, "Towards long-horizon vision-language navigation: Platform, benchmark and method," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [138] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," in *CVPR*, 2018, pp. 1–10.
- [139] L. Yu, X. Chen, G. Gkioxari, M. Bansal, T. L. Berg, and D. Batra, "Multi-target embodied question answering," in *CVPR*, 2019, pp. 6309–6318.
- [140] E. Wijmans, S. Datta, O. Maksymets, A. Das, G. Gkioxari, S. Lee, I. Essa, D. Parikh, and D. Batra, "Embodied question answering in photorealistic environments with point cloud perception," in *CVPR*, 2019, pp. 6659–6668.
- [141] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "Iqa: Visual question answering in interactive environments," in *CVPR*, 2018, pp. 4089–4098.
- [142] C. Cangea, E. Belilovsky, P. Liò, and A. Courville, "Videonavqa: Bridging the gap between visual and embodied question answering," *arXiv preprint arXiv:1908.04950*, 2019.
- [143] X. Ma, S. Yong, Z. Zheng, Q. Li, Y. Liang, S.-C. Zhu, and S. Huang, "Sqa3d: Situated question answering in 3d scenes," in *ICLR*, 2023.
- [144] S. Tan, M. Ge, D. Guo, H. Liu, and F. Sun, "Knowledge-based embodied question answering," *IEEE TPAMI*, 2023.
- [145] A. Majumdar, A. Ajay, X. Zhang, P. Putta, S. Yenamandra, M. Henaff, S. Silwal, P. Mcvay, O. Maksymets, S. Arnaud *et al.*, "Openeqa: Embodied question answering in the era of foundation models," in *CVPR*, 2024, pp. 16 488–16 498.
- [146] A. Z. Ren, J. Clark, A. Dixit, M. Itkina, A. Majumdar, and D. Sadigh, "Explore until confident: Efficient exploration for embodied question answering," *arXiv preprint arXiv:2403.15941*, 2024.
- [147] V. S. Dorbala, P. Goyal, R. Piramuthu, M. Johnston, D. Manocha, and R. Ghanadhan, "S-eqa: Tackling situational queries in embodied question answering," *arXiv preprint arXiv:2405.04732*, 2024.
- [148] K. Jiang, Y. Liu, W. Chen, J. Luo, Z. Chen, L. Pan, G. Li, and L. Lin, "Beyond the destination: A novel benchmark for exploration-aware embodied question answering," 2025.
- [149] Y. Wu, Y. Wu, G. Gkioxari, and Y. Tian, "Building generalizable agents with a realistic and rich 3d environment," *arXiv preprint arXiv:1801.02209*, 2018.
- [150] M. Savva, A. X. Chang, A. Dosovitskiy, T. Funkhouser, and V. Koltun, "Minos: Multimodal indoor simulator for navigation in complex environments," *arXiv preprint arXiv:1712.03931*, 2017.
- [151] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.
- [152] Y. Wu, L. Jiang, and Y. Yang, "Revisiting embodiedqa: A simple baseline and beyond," *IEEE TIP*, vol. 29, pp. 3984–3992, 2020.
- [153] S. Tan, W. Xiang, H. Liu, D. Guo, and F. Sun, "Multi-agent embodied question answering in interactive environments," in *ECCV*, 2020, pp. 663–678.
- [154] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proceedings IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*, 1997, pp. 146–151.
- [155] K. Sakamoto, D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe, "Map-based modular approach for zero-shot embodied question answering," *arXiv preprint arXiv:2405.16559*, 2024.
- [156] B. Patel, V. S. Dorbala, and A. S. Bedi, "Embodied question answering via multi-llm systems," *arXiv preprint arXiv:2406.10918*, 2024.
- [157] H. Liu, S. K. Sampath, N. Wang, and C. Yang, "Multifingered grasp planning based on gaussian process implicit surface and its partial differentials," *IEEE/ASME TMECH*, vol. 29, no. 5, pp. 3522–3533, 2024.
- [158] L. Chen, P. Huang, Y. Li, and Z. Meng, "Edge-dependent efficient grasp rectangle search in robotic grasp detection," *IEEE/ASME TMECH*, vol. 26, no. 6, pp. 2922–2931, 2021.
- [159] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *ICRA*, 2011, pp. 3304–3311.
- [160] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *IROS*, 2018, pp. 3511–3516.
- [161] A. Mousavian, C. Eppner, and D. Fox, "6-dof grapsnet: Variational grasp generation for object manipulation," in *ICCV*, 2019, pp. 2901–2910.
- [162] C. Eppner, A. Mousavian, and D. Fox, "Acronym: A large-scale grasp dataset based on simulation," in *ICRA*, 2021, pp. 6222–6227.
- [163] L. F. C. Murillo, N. Khargonkar, B. Prabhakaran, and Y. Xiang, "Multigrippergrasp: A dataset for robotic grasping from parallel jaw grippers to dexterous hands," *arXiv preprint arXiv:2403.09841*, 2024.
- [164] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgbd," in *ICRA*, 2021, pp. 13 452–13 458.
- [165] G. Tzafas, Y. Xu, A. Goel, M. Kasaei, Z. Li, and H. Kasaei, "Language-guided robot grasping: Clip-based referring grasp synthesis in clutter," *arXiv preprint arXiv:2311.05779*, 2023.
- [166] S. Jin, J. Xu, Y. Lei, and L. Zhang, "Reasoning grasping via multimodal large language model," *arXiv preprint arXiv:2402.06798*, 2024.
- [167] K. Li, J. Wang, L. Yang, C. Lu, and B. Dai, "Semgrasp: Semantic grasp generation via language aligned discretization," *arXiv preprint arXiv:2404.03590*, 2024.
- [168] M. Shridhar, L. Manuelli, and D. Fox, "Cliprot: What and where pathways for robotic manipulation," in *CoRL*. PMLR, 2022, pp. 894–906.
- [169] W. Shen, G. Yang, A. Yu, J. Wong, L. P. Kaelbling, and P. Isola, "Distilled feature fields enable few-shot language-guided manipulation," in *7th Annual CoRL*, 2023.
- [170] Y. Zheng, X. Chen, Y. Zheng, S. Gu, R. Yang, B. Jin, P. Li, C. Zhong, Z. Wang, L. Liu *et al.*, "Gaussiangrasper: 3d language gaussian splatting for open-vocabulary robotic grasping," *arXiv preprint arXiv:2403.09637*, 2024.
- [171] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE TRO*, 2023.
- [172] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou *et al.*, "The rise and potential of large language model based agents: A survey," *arXiv preprint arXiv:2309.07864*, 2023.
- [173] D. McDermott, M. Ghallab, A. E. Howe, C. A. Knoblock, A. Ram, M. M. Veloso, D. S. Weld, and D. E. Wilkins, "Pddl-the planning domain definition language," 1998.
- [174] N. C. Metropolis and S. M. Ulam, "The monte carlo method," *Journal of the American Statistical Association*, vol. 44 247, pp. 335–41, 1949.
- [175] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Trans. Syst. Sci. Cybern.*, vol. 4, pp. 100–107, 1968.
- [176] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "Alfred: A benchmark for interpreting grounded instructions for everyday tasks," in *CVPR*, 2020, pp. 10 740–10 749.
- [177] M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler, and M. Hausknecht, "Alfworld: Aligning text and embodied environments for interactive learning," *arXiv preprint arXiv:2010.03768*, 2020.
- [178] S. Y. Min, D. S. Chaplot, P. Ravikumar, Y. Bisk, and R. Salakhutdinov, "Film: Following instructions in language with modular methods," *arXiv preprint arXiv:2110.07342*, 2021.
- [179] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *ICML*. PMLR, 2022, pp. 9118–9147.
- [180] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar *et al.*, "Inner monologue: Embodied reasoning through planning with language models," in *CoRL*. PMLR, 2023, pp. 1769–1782.
- [181] Y. Zhang, S. Yang, C. Bai, F. Wu, X. Li, X. Li, and Z. Wang, "Towards efficient llm grounding for embodied multi-agent collaboration," *arXiv preprint arXiv:2405.14314*, 2024.
- [182] G. Sarch, Y. Wu, M. Tarr, and K. Fragkiadaki, "Open-ended instructable embodied agents with memory-augmented large language models," in *EMNLP*, 2023, pp. 3468–3500.

- [183] 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.
- [184] P. Sharma, A. Torralba, and J. Andreas, "Skill induction and planning with latent language," in *Annual Meeting of the Association for Computational Linguistics*, 2021.
- [185] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progpprompt: Generating situated robot task plans using large language models," in *ICRA*, 2023, pp. 11523–11530.
- [186] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. R. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," *ICRA*, pp. 9493–9500, 2022.
- [187] A. Zeng, M. Attarian, K. M. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. S. Ryoo, V. Sindhwani, J. Lee *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," in *ICLR*, 2023.
- [188] S. Shin, J. Kim, G.-C. Kang, B.-T. Zhang *et al.*, "Socratic planner: Inquiry-based zero-shot planning for embodied instruction following," *arXiv preprint arXiv:2404.15190*, 2024.
- [189] 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," *ICCV*, pp. 2986–2997, 2022.
- [190] Z. Wu, Z. Wang, X. Xu, J. Lu, and H. Yan, "Embodied task planning with large language models," *arXiv preprint arXiv:2307.01848*, 2023.
- [191] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. D. Reid, and N. Sünderhauf, "Sayplan: Grounding large language models using 3d scene graphs for scalable task planning," in *CoRL*, 2023.
- [192] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, C. Gan, C. M. de Melo, J. B. Tenenbaum, A. Torralba, F. Shkurti, and L. Paull, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," *ArXiv*, vol. abs/2309.16650, 2023.
- [193] 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," *NeurIPS*, vol. 36, 2024.
- [194] J. Huang, S. Yong, X. Ma, X. Linghu, P. Li, Y. Wang, Q. Li, S.-C. Zhu, B. Jia, and S. Huang, "An embodied generalist agent in 3d world," in *ICML*, 2024.
- [195] Z. Wu, Z. Wang, X. Xu, J. Lu, and H. Yan, "Embodied instruction following in unknown environments," 2024.
- [196] 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 *ICML*, 2023, pp. 8469–8488.
- [197] X. Zhao, M. Li, C. Weber, M. B. Hafez, and S. Wermter, "Chat with the environment: Interactive multimodal perception using large language models," in *IROS*, 2023, pp. 3590–3596.
- [198] Z. Zhao, S. Cheng, Y. Ding, Z. Zhou, S. Zhang, D. Xu, and Y. Zhao, "A survey of optimization-based task and motion planning: From classical to learning approaches," *IEEE/ASME TMECH*, pp. 1–27, 2024.
- [199] 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 *CoRL*, 2023, pp. 287–318.
- [200] N. Shinn, B. Labash, and A. Gopinath, "Reflexion: an autonomous agent with dynamic memory and self-reflection," *ArXiv*, vol. abs/2303.11366, 2023.
- [201] Z. Wang, S. Cai, A. Liu, X. Ma, and Y. Liang, "Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents," *ArXiv*, vol. abs/2302.01560, 2023.
- [202] Q. Vuong, S. Levine, H. R. Walke, K. Pertsch, A. Singh, R. Doshi, C. Xu, J. Luo, L. Tan, D. Shah *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," in *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023.
- [203] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.
- [204] J. Xiang, G. Liu, Y. Gu, Q. Gao, Y. Ning, Y. Zha, Z. Feng, T. Tao, S. Hao, Y. Shi *et al.*, "Pandora: Towards general world model with natural language actions and video states," *arXiv preprint arXiv:2406.09455*, 2024.
- [205] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, "3d-vla: A 3d vision-language-action generative world model," *arXiv preprint arXiv:2403.09631*, 2024.
- [206] Z. Ding, A. Zhang, Y. Tian, and Q. Zheng, "Diffusion world model," *arXiv preprint arXiv:2402.03570*, 2024.
- [207] A. Bardes, J. Ponce, and Y. LeCun, "Mc-jepa: A joint-embedding predictive architecture for self-supervised learning of motion and content features," *arXiv preprint arXiv:2307.12698*, 2023.
- [208] Z. Fei, M. Fan, and J. Huang, "A-jepa: Joint-embedding predictive architecture can listen," *arXiv preprint arXiv:2311.15830*, 2023.
- [209] A. Saito and J. Poovvancheri, "Point-jepa: A joint embedding predictive architecture for self-supervised learning on point cloud," *arXiv preprint arXiv:2404.16432*, 2024.
- [210] Q. Garrido, M. Assran, N. Ballas, A. Bardes, L. Najman, and Y. LeCun, "Learning and leveraging world models in visual representation learning," *arXiv preprint arXiv:2403.00504*, 2024.
- [211] J. Wu, S. Yin, N. Feng, X. He, D. Li, J. Hao, and M. Long, "Videogpt: Interactive videogpts are scalable world models," *arXiv preprint arXiv:2405.15223*, 2024.
- [212] F. Zhu, H. Wu, S. Guo, Y. Liu, C. Cheang, and T. Kong, "Irasim: Learning interactive real-robot action simulators," *arXiv preprint arXiv:2406.14540*, 2024.
- [213] J. Yang, B. Liu, J. Fu, B. Pan, G. Wu, and L. Wang, "Spatiotemporal predictive pre-training for robotic motor control," *arXiv preprint arXiv:2403.05304*, 2024.
- [214] M. Burchi and R. Timofte, "Mudreamer: Learning predictive world models without reconstruction," *arXiv preprint arXiv:2405.15083*, 2024.
- [215] 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.
- [216] A. Dawid and Y. LeCun, "Introduction to latent variable energy-based models: A path towards autonomous machine intelligence," *arXiv preprint arXiv:2306.02572*, 2023.
- [217] V. Lim, H. Huang, L. Y. Chen, J. Wang, J. Ichnowski, D. Seita, M. Laskey, and K. Goldberg, "Real2sim2real: Self-supervised learning of physical single-step dynamic actions for planar robot casting," in *ICRA*, 2022, pp. 8282–8289.
- [218] Y. Feng, Y. Shang, X. Feng, L. Lan, S. Zhe, T. Shao, H. Wu, K. Zhou, H. Su, C. Jiang *et al.*, "Elastogen: 4d generative elastodynamics," *arXiv preprint arXiv:2405.15056*, 2024.
- [219] M. Liu, C. Xu, H. Jin, L. Chen, M. Varma T, Z. Xu, and H. Su, "One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization," *NeurIPS*, vol. 36, 2024.
- [220] L. Wong, G. Grand, A. K. Lew, N. D. Goodman, V. K. Mansinghka, J. Andreas, and J. B. Tenenbaum, "From word models to world models: Translating from natural language to the probabilistic language of thought," *arXiv preprint arXiv:2306.12672*, 2023.
- [221] Z. Cheng, Z. Wang, J. Hu, S. Hu, A. Liu, Y. Tu, P. Li, L. Shi, Z. Liu, and M. Sun, "Legend: Open platform for embodied agents," *arXiv preprint arXiv:2404.18243*, 2024.
- [222] H. Wang, J. Chen, W. Huang, Q. Ben, T. Wang, B. Mi, T. Huang, S. Zhao, Y. Chen, S. Yang *et al.*, "Grutopia: Dream general robots in a city at scale," *arXiv preprint arXiv:2407.10943*, 2024.
- [223] N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, P. Chattopadhyay, Y. Chen, Y. Cui, Y. Ding *et al.*, "Cosmos world foundation model platform for physical ai," *arXiv preprint arXiv:2501.03575*, 2025.
- [224] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," *arXiv preprint arXiv:2402.10329*, 2024.
- [225] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.
- [226] S. Luo, Q. Peng, J. Lv, K. Hong, K. R. Driggs-Campbell, C. Lu, and Y.-L. Li, "Human-agent joint learning for efficient robot manipulation skill acquisition," *arXiv preprint arXiv:2407.00299*, 2024.
- [227] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani *et al.*, "Transporter networks: Rearranging the visual world for robotic manipulation," in *CoRL*. PMLR, 2021, pp. 726–747.
- [228] H. Geng, H. Xu, C. Zhao, C. Xu, L. Yi, S. Huang, and H. Wang, "Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts," in *CVPR*, 2023, pp. 7081–7091.
- [229] M. Torne, A. Simeonov, Z. Li, A. Chan, T. Chen, A. Gupta, and P. Agrawal, "Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation," *arXiv preprint arXiv:2403.03949*, 2024.
- [230] Y. Jiang, C. Wang, R. Zhang, J. Wu, and L. Fei-Fei, "Transic: Sim-to-real policy transfer by learning from online correction," *arXiv preprint arXiv:2405.10315*, 2024.

- [231] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, “Domain randomization for transferring deep neural networks from simulation to the real world,” in *IROS*, 2017, pp. 23–30.
- [232] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray *et al.*, “Learning dexterous in-hand manipulation,” *IJRR*, vol. 39, no. 1, pp. 3–20, 2020.
- [233] J. Matas, S. James, and A. J. Davison, “Sim-to-real reinforcement learning for deformable object manipulation,” in *CoRL*. PMLR, 2018, pp. 734–743.
- [234] M. Kaspar, J. D. M. Osorio, and J. Bock, “Sim2real transfer for reinforcement learning without dynamics randomization,” in *IROS*, 2020, pp. 4383–4388.
- [235] W. Yu, J. Tan, C. K. Liu, and G. Turk, “Preparing for the unknown: Learning a universal policy with online system identification,” *arXiv preprint arXiv:1702.02453*, 2017.
- [236] A. Yu, A. Foote, R. Mooney, and R. Martín-Martín, “Natural language can help bridge the sim2real gap,” *arXiv preprint arXiv:2405.10020*, 2024.
- [237] Z. Wang, H. Zheng, Y. Nie, W. Xu, Q. Wang, H. Ye, Z. Li, K. Zhang, X. Cheng, W. Dong *et al.*, “All robots in one: A new standard and unified dataset for versatile, general-purpose embodied agents,” *arXiv preprint arXiv:2408.10899*, 2024.



**Yang Liu** (M’21) is currently an Associate Professor working at the School of Computer Science and Engineering, Sun Yat-sen University. He received his Ph.D. degree from Xidian University in 2019. His current research interests include multi-modal reasoning, causality learning and embodied AI. He is the recipient of the First Prize of the Third Guangdong Province Young Computer Science Academic Show. He has authorized and co-authorized more than 40 papers in top-tier academic journals and conferences such as TPAMI, TIP, CVPR and ICCV.



**Weixing Chen** has received the B.S. degree from the college of Medicine and Biological Information Engineering, Northeastern University, in 2020 and M.S. degree from Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences in 2023. He is currently a Ph.D. student at the School of Computer Science and Engineering, Sun Yat-sen University. His main interests include multi-modal learning, causal relation discovery, and embodied ai. He has been serving as a reviewer for numerous academic journals and conferences such as TNNLS, NeurIPS, ICML, ICLR, MICCAI, and ACM MM.



**Yongjie Bai** received the B.S. degree from the School of Computer Science and Technology, Dalian University of Technology, in 2024. He is currently a Ph.D. student at the School of Computer Science and Engineering, Sun Yat-sen University. His main research interests include embodied AI, robot learning, and multi-modal learning.



**Xiaodan Liang** (Senior Member, IEEE) is currently a Professor at Sun Yat-sen University. She was a postdoc researcher in the machine learning department at Carnegie Mellon University, working with Prof. Eric Xing, from 2016 to 2018. She received her PhD degree from Sun Yat-sen University in 2016, advised by Liang Lin. She has published several cutting-edge projects on human-related analysis, including human parsing, pedestrian detection and instance segmentation, 2D/3D human pose estimation, and activity recognition.



**Guanbin Li** (M’15) is currently a Professor in School of Computer Science and Engineering, Sun Yat-Sen University. He received his PhD degree from the University of Hong Kong in 2016. His current research interests include computer vision, image processing, and deep learning. He is a recipient of ICCV 2019 Best Paper Nomination Award. He has authorized and co-authorized on more than 100 papers in top-tier academic journals and conferences. He serves as an area chair for the conference of VISAPP. He has been serving as a reviewer for numerous academic journals and conferences such as TPAMI, IJCV, TIP, TMM, TCyb, CVPR, ICCV, ECCV and NeurIPS.



**Wen Gao** (Fellow, IEEE) received the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991. He is currently a Professor of computer science with the School of Electronic Engineering and Computer Science, Institute of Digital Media, Peking University, Beijing, China. Before joining Peking University, he was a Professor of computer science with the Harbin Institute of Technology, Harbin, China, from 1991 to 1995, and a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has authored or coauthored extensively, including five books and more than 600 technical articles in refereed journals and conference proceedings in the areas of image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interfaces, and bioinformatics. He is a Member of the China Engineering Academy.



**Liang Lin** (Fellow, IEEE) is currently a Full Professor with Sun Yat-sen University, Guangzhou, China. From 2008 to 2010, he was a Postdoctoral Fellow with the University of California, Los Angeles, Los Angeles, CA, USA. From 2016 to 2018, he led the SenseTime R&D teams to develop cutting-edge and deliverable solutions for computer vision, data analysis and mining, and intelligent robotic systems. He has authored or coauthored more than 100 papers in top-tier academic journals and conferences, such as 15 papers in *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* and *International Journal of Computer Vision*, and more than 60 papers in CVPR, ICCV, NeurIPS, and IJCAI. He was an Associate Editor for *IEEE TRANSACTIONS ON MULTIMEDIA*, *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, and was an Area/Session Chair for numerous conferences, such as CVPR, ICCV, AAAI, ICME, and ICMR. He was the recipient of the Annual Best Paper Award by Pattern Recognition (Elsevier) in 2018, Best Paper Diamond Award at IEEE ICME 2017, Best Paper Runner-Up Award at ACM NPAR 2010, Google Faculty Award in 2012, Best Student Paper Award at IEEE ICME 2014, and Hong Kong Scholars Award in 2014. He is a Fellow of IAPR, AAIA, and IET.