A SURVEY OF EMBODIED ARTIFICIAL INTELLIGENCE DATA ENGINEERING

Xuan Xia[1], **Haoran Tong**[1], **Xing He**[1], **Bo Yu**[1], **Ning Ding**[1], **Xue Liu**[2], **Shaoshan Liu**[2*]
1.Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China
2.Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
{xiaxuan, tonghaoran, hexing, boyu, dingning, shaoshanliu}cuhk.edu.cn
Steve.liu@mbzuai.ac.ae

ABSTRACT

Embodied Artificial Intelligence (EAI) Data Engineering represents a transformative shift in the field of AI, focusing on developing systematic, standardized, scalable and goal-driven technical frameworks to meet the data requirements of EAI systems. This comprehensive overview explores the concept of EAI data, its production systems, standardization, production technologies, and optimization directions in data engineering for EAI. It highlights the importance of addressing data bottlenecks such as cost inefficiency, data silos, and evaluation void. The key components of EAI data engineering are outlined, including the design of data production systems, establishment of data standards, real-world data collection technologies, and simulation data generation technologies. The deployment and application of EAI data engineering in various fields such as manufacturing, mining, and the service industry are also explored. By providing an in-depth analysis of the current state of EAI data engineering and offering insights into its future optimization directions, this survey aims to serve as a valuable resource for researchers and practitioners in the field.

***K*eywords** Embodied Artificial Intelligence · Data Engineering · Data Collection · Data Generation · Teleopration · Simulation

1 Introduction

Embodied Artificial Intelligence (EAI) represents a transformative shift in AI, where intelligence is not just computed but enacted—emerging through perception, interaction, and continuous adaptation in the physical world [1]. A key trait of EAI systems is that they must operate in dynamic, uncertain, and multi-modal environments. This fundamental difference places unprecedented demands on data: it must be temporally coherent, sensorily rich, causally structured, and behaviorally relevant. The success of embodied agents hinges not merely on model architectures, but on the depth, diversity, and structure of the data they are trained on [2]. Meanwhile, with a total addressable market size over $10 tillion, data engineering has become a critical enabler of both scientific progress and economic impact [3, 2].

Scaling laws [4, 5] offer a guiding principle for the development of EAI: intelligence emerges from data. However, unlike the vast amounts of data already accumulated in fields such as Natural Language Processing (NLP) and autonomous driving, the data required when robots enter homes, warehouses, and factories is fundamentally different—it is data of physical interaction. The acquisition of such data, including motion trajectories, collision feedback, haptic sensations, lighting conditions, and friction, faces exponentially increasing difficulty and cost. Even tens of thousands of hours of real-world robotic interaction data fall far short of the scale seen in Large Language Models (LLMs). While LLMs consume trillions of tokens, the interaction data currently available for robots amounts to only a tiny fraction—equivalent to just one in a hundred thousand of what LLMs process.

Therefore, the demand for EAI data has driven the rapid development of technology in this field in recent years. As shown in Fig. 1, The current EAI data production exists in various modes, each with its own advantages and disadvantages in terms of equipment costs, labor costs, scene limitations, and computational consumption. More importantly, current methods are fragmented, unsustainable, and inconsistencies in data quality and universality, led

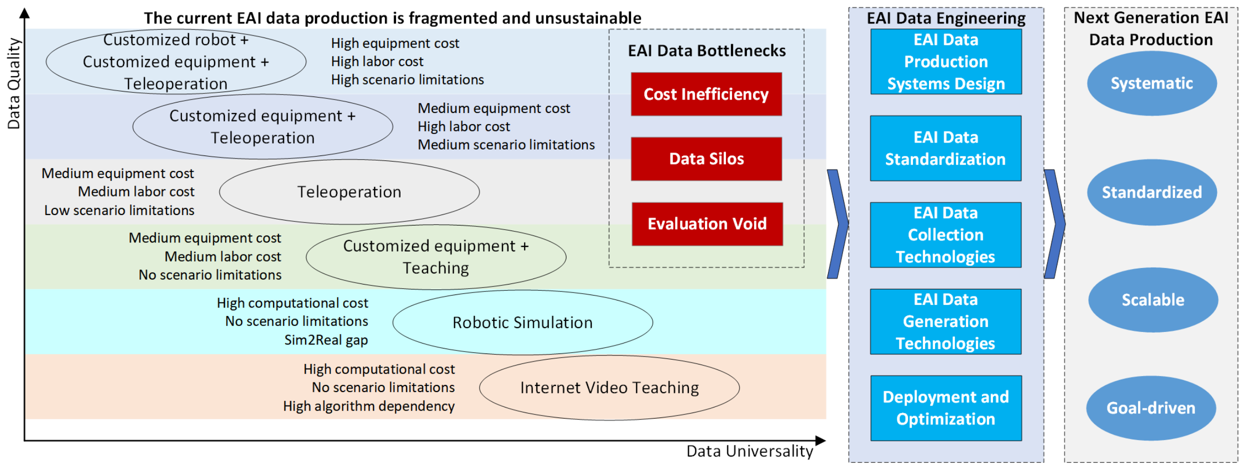

Figure 1: Current EAI data production methods are fragmented, unsustainable, and inconsistencies in data quality and universality, led to the current EAI data bottlenecks. EAI data engineering represents a significant shift from opportunistic EAI data production to next generation EAI data production to solve EAI data bottlenecks.

to the current EAI data bottlenecks (which will be detailed in Section 2.3). Solving these bottlenecks requires the use of systematic engineering methods to design new EAI data production pipelines. Therefore, we argue that data engineering is no longer a support task, but the foundation upon which scalable and generalizable EAI will be built. By mapping the current landscape and identifying methodological and infrastructural gaps, this survey aims to establish EAI data engineering as a first-class research frontier. We advocate for a shift from opportunistic EAI data production to systematic, standardized, scalable and goal-driven EAI data production, so as to unlock new opportunities for reproducible research, robust generalization, and inclusive innovation in EAI.

1.1 Concept of Embodied Artificial Intelligence Data

The concept of EAI originated in Alan Turing's seminal 1950 paper, "Computing Machinery and Intelligence."[6] In this paper, Turing envisioned two potential paths for the development of artificial intelligence: one focused on abstract computational intelligence (e.g., playing chess), and the other involving equipping machines with sensors to enable interaction with the physical world, humans, and their environment through a physical presence, hence achiving scalability [7, 8]. The latter approach constitutes what we now refer to as EAI.

EAI data refers to the multimodal sensory inputs and behavioral outputs that enable intelligent agents to perceive and interact with their environments. This data encompasses both physical-world observations collected by robotic sensors (e.g., LiDAR, cameras, force-torque sensors) and synthetic data generated through simulation platforms. The uniqueness of EAI data lies in its embodiment characteristics - it must capture spatiotemporal relationships between agents' actions and environmental changes. Physical agents produce real-world operational data through task execution, while digital agents generate simulated interaction data with programmed environments. Both data types share common structuring requirements for temporal alignment, action-effect pairing, and contextual annotation, but differ in fidelity and collection scalability. EAI data serves as the foundational resource for developing embodied cognition models, bridging the gap between abstract intelligence and physical/digital embodiment.

1.2 Related Surveys in the Field of Embodied Artificial Intelligence Data

As shown in Table 1, recent years have seen a surge in the publication of surveys related to EAI data. These surveys cover a wide range of topics, from teleoperation techniques [10, 12] to the Simulators [9, 11] and datasets [16, 17]. Notable contributions include comprehensive reviews on the use of internet video data for robot learning [14], task planning and code generation [15], and the integration of generative artificial intelligence [20]. These reviews highlight the rapid advancements and increasing complexity of data in the EAI field.

Despite the wealth of existing surveys, the current landscape of data-related technologies in EAI is fragmented and lacks a systematic approach. Existing surveys often focus on specific aspects or applications, but they do not provide a comprehensive and unified framework for understanding and guiding the production of EAI data. This gap necessitates the introduction of the concept of EAI Data Engineering. This new concept aims to offer a systematic and theoretical

Table 1: Related Surveys in the Field of EAI Data

Title	Year	Publication	Data Engineering Related Content
Toward next-generation learned robot manipulation [9]	2021	SCIENCE ROBOTICS	Data and simulation of manipulation
Teleoperation methods and enhancement techniques for mobile robots: A comprehensive survey [10]	2021	Robotics and Autonomous Systems	Teleoperation enhancement techniques
A Survey of Embodied AI: From Simulators to Research Tasks [11]	2022	IEEE TETCI	Simulation platform and embodied question answering data
Teleoperation of Humanoid Robots: A Survey [12]	2023	IEEE Transactions on Robotics	Teleoperation systems for humanoid robots
Multiple Mobile Robot Task and Motion Planning: A Survey [13]	2023	ACM Computing Surveys	Task and Motion Planning
Towards Generalist Robot Learning from Internet Video: A Survey [14]	2024	ArXiv	Learning from internet video data
Real-world robot applications of foundation models: a review [15]	2024	ADVANCED ROBOTICS	Task planning and code generation
A Survey of Imitation Learning: Algorithms, Recent Developments, and Challenges [16]	2024	IEEE TRANSACTIONS ON CYBERNETICS	Datasets of imitation learning
Robot learning in the era of foundation models: a survey [17]	2025	Neurocomputing	Datasets of manipulation, navigation, planning, and reasoning
A Survey of Robotic Navigation and Manipulation with Physics Simulators in the Era of Embodied AI [18]	2025	ArXiv	Simulators and benchmark datasets of navigation and manipulation
A Survey of Interactive Generative Video [19]	2025	ArXiv	Task planning and policy learning via generative simulation
Generative Artificial Intelligence in Robotic Manipulation: A Survey [20]	2025	ArXiv	Data, image, code, policy generation for manipulation

foundation for the production, management, and utilization of data in EAI. By proposing EAI Data Engineering, we can address the unique challenges and opportunities in this interdisciplinary field more effectively. A dedicated survey would not only synthesize the latest advancements and trends but also identify gaps and future directions, facilitating more efficient and effective research and development efforts.

1.3 Embodied Artificial Intelligence Data Engineering

As shown in Fig. 2, EAI data originates from various sources, ranging from the broadest category of internet data, to the intermediate layer of simulation data (including synthetic data), and finally to the rarest real-world data, forming the EAI data pyramid. Different EAI technological approaches have varying requirements for these types of data. EAI data engineering refers to a systematic technical framework designed to address the data requirements of EAI, encompassing the entire lifecycle of data production from design and development to management. Its core objective is to establish high-quality, multimodal datasets through standardized data collection and generation. Specifically, this engineering discipline covers the following key components:

Design of EAI Data Production Systems: The design of data production systems for EAI involves planning and constructing a framework capable of efficiently and accurately acquiring multimodal data tailored to the needs of robots. This design must comprehensively consider factors such as sensor configurations, data types, data collection frequency and precision, as well as data storage and preprocessing methods [21, 22].

Establishment of EAI Data Standards: EAI data standards refer to a set of norms and guidelines formulated to ensure the quality, consistency, and interoperability of data within EAI. These standards cover aspects such as data formats, annotation methods, quality control, privacy protection, and the integration of multimodal data, aiming to provide a unified framework for data collection, generation, storage, and sharing [23, 24]. By establishing clear data standards, the usability and reliability of data can be improved, fostering data sharing and collaboration across different systems and platforms.

Development of Real-World EAI Data Collection Technologies: They involve methods for directly acquiring multimodal data from physical environments using sensors, cameras, microphones, and other devices. These technologies capture information such as the robot's visual, auditory, tactile, and motion, as well as environmental objects, scenes, and human behaviors, providing realistic data support for EAI models [25, 26, 27].

Development of Simulation EAI Data Generation Technologies: They involve creating high-fidelity, diverse virtual environments and task scenarios through virtual simulation platforms to generate multimodal data. Leveraging advanced 3D modeling, physics engines, and generative artificial intelligence, these technologies can rapidly produce large volumes of high-quality training data, simulating various interactions and dynamic changes in the real world [28, 29, 30].

Application and Optimization: They involve designing and implementing data production solutions tailored to the specific needs of industries or domains such as healthcare, industrial manufacturing, and education. By continuously

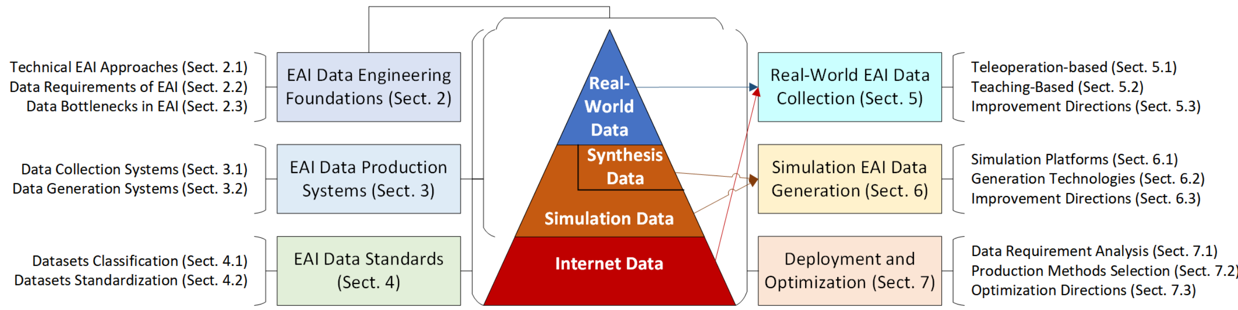

Figure 2: The components of EAI data engineering and the outline of this survey

optimizing data collection processes and systems, this approach aims to improve data quality, reduce costs, and enhance data availability and real-time performance [31, 32, 33].

This survey will introduce the content of EAI data engineering in the above order. Specifically, Section 2 begins with an overview of the foundations of EAI data engineering, Section 3 discusses EAI data production systems, Section 4 explores EAI data standards, Section 5 examines real-world EAI data collection technologies, Section 6 delves into simulation EAI data generation technologies, Section 7 focus on the application and optimization of EAI data engineering. The main contributions of this survey are summarized as follows:

- This survey introduces and formalizes the concept of **EAI Data Engineering**, framing it as a foundational discipline for enabling scalable, generalizable, and robust EAI.
- It provides a **theoretical explanation of the data bottleneck** in EAI, analyzing why current data practices constrain learning efficiency and task transfer.
- A **lifecycle architecture** for EAI data production is proposed, along with standardization principles that enable systematic data collection, generation, dataset construction, and quality assessment.
- Through a detailed survey, this article **maps real-world data collection technologies and simulation data generation technologies**, identifies trade-offs, and highlights recent advances that improve efficiency and diversity of EAI datasets.
- The work advocates for a shift toward **systematic, standardized, scalable and goal-driven EAI data production**, emphasizing their impact on data requirement analysis, data production methods selection, and optimization directions in industry and service industry.

2 Foundations of EAI Data Engineering

The primary challenge in EAI data engineering lies in overcoming bottlenecks encountered during the data collection process. The importance of data collection stems from a widespread consensus that, similar to the field of natural language processing, scaling laws are equally applicable in the domain of EAI. Guided by scaling laws, the development of EAI cannot proceed without support from extensive robotic data.

Previous research has confirmed that in imitation learning, the model's generalization ability over objects/ scenarios, success rate in a single scenario, and spatial generalization ability all indeed follow scaling laws [34, 35, 36]. However, no studies have yet revealed how scaling laws create bottlenecks in the production of EAI data or how they can guide researchers to improve data production efficiency. The difficulty in conducting such research lies in the fact that current EAI technological approaches are not unified. The diverse robot embodiments, model architectures, and data modalities make it challenging to quantify the impact of data. Therefore, this section attempts to conduct a qualitative analysis of data requirements based on the current EAI technological approaches and "Fast and Slow System" theory [37], in order to more intuitively identify the data bottlenecks in EAI.

2.1 Technical Approaches for EAI

As illustrated in Fig. 3, there are two main approaches to achieving EAI: hierarchical EAI and end-to-end EAI. Currently, there are three types of hierarchical approaches, leading to a total of four technical routes:

- *Hierarchical EAI (Type I):* A general-purpose or specialized "System II" handles high-level reasoning, planning, and decision-making, invoking the robot's function APIs to execute specific tasks such as localization and

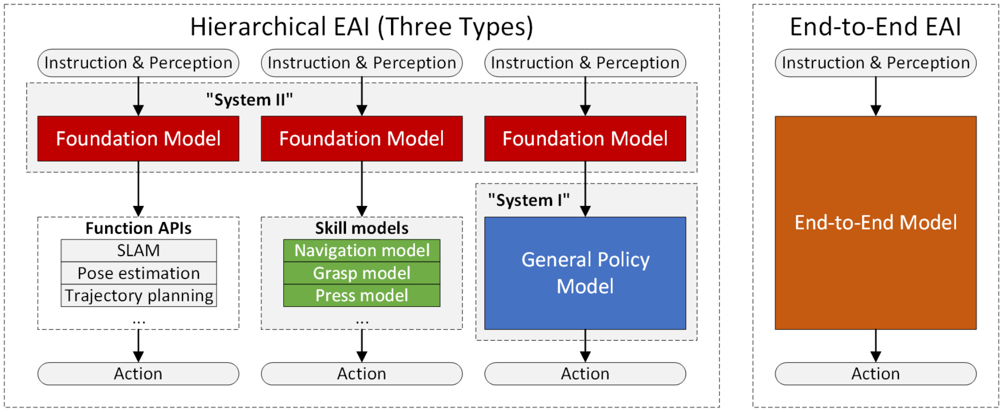

Figure 3: Technological approaches of EAI and their corresponding data demand models

navigation, pose estimation, trajectory control, etc. In this structure, all functionalities of the robot must be implemented by various functions that can be called upon by the "System II". It primarily relies on foundation models such as large language models (LLMs) and multimodal LLMs.

- *Hierarchical EAI (Type II):* A "System II" invoking specific skill models to perform policies like navigation, grasping, pressing, etc. Under this setup, the robot's skills are encapsulated into callable modules, and the "System II" does not need to provide exact control parameters to invoke these skills.
- *Hierarchical EAI (Type III):* While a "System II" is running, a general-purpose or specialized "System I" (a general policy model) takes care of low-level motion planning and control. In this architecture, the functions performed by the "System I" can be considered a collection of all skills.
- *End-to-End EAI:* End-to-end EAI is typically realized through a single model that directly learns from input to output without distinguishing between "System I" and "System II". Instead of intermediate API calls, the model outputs execution commands during inference.

Through the analysis of the technological approaches of EAI as discussed above, it can be identified that there are four models of data demand within EAI: foundation models, skill models, general policy models, and end-to-end models. The following section will analyze the data requirements of EAI from these four models.

2.2 Data Requirements of EAI Models

In the previous section, four data demanders for EAI were identified. As shown in Table 2, they can be analyzed in terms of training methods, data types, typical datasets, and typical models.

The training methods of foundation models predominantly involve pre-training and fine-tuning with specialized datasets. Examples include Generative Pre-training (GP) [38], Supervised Fine-Tuning (SFT) [39], Reinforcement Learning from Human Feedback (RLHF) [40], and Direct Preference Optimization (DPO) [41]. The data types used are mostly internet data and instruction-tuning datasets. Representative datasets include LLaVA-v1.5 [42] and RoboVQA [43]. Typical foundation models in EAI include VoxPoser [44] and ManipLLM [45].

The skill models are typically trained using Reinforcement Learning (RL) [46] and Imitation Learning (IL) [47]. These learning approaches require robotic operation data and perception data. Representative datasets include BC-Z [48] and ARIO [24]. Typical skill models include AnyGrasp [49] and Diffusion Policy Model (DPM) [50].

The training methods of general policy models may include RL and IL, as well as end-to-end vision-language-action (VLA) model learning. Consequently, the data types involved include perception data, operation data, and instruction data. Representative datasets include BridgeData V2 [51], and Open X-Embodiment [23]. Typical general policy models include InstructNav [52] and RDT [53].

The end-to-end training methods for EAI may involve learning based on VLA models. This implies that the data types used could encompass all of the above-mentioned categories. Consequently, the datasets utilized could also include all the aforementioned datasets, covering as many scenarios, tasks, and robot bodies as possible. Representative end-to-end training models include RT-2 [54] and OpenVLA [55].

Based on the above analysis, the data requirements of the four data demanders can be summarized as follows.

Table 2: Overview of Training Methods, Data Types, Typical Datasets, and Typical Models of Data Demanders

Data demanders	Foundation Models	Skill Models	General Policy Models	End-to-End Models
Training Methods	GP, SFT, RLHF, DPO, .etc.	RL, IL, .etc.	RL, IL, GP, SFT, .etc.	GP, SFT, RLHF, DPO, RL, IL, .etc.
Data Types	Internet data, Instruction data, .etc.	Operational data, Perceptual data, .etc.	Operational data, Perceptual data, Instruction fine-tuning data, .etc.	
Typical Datasets	LLaVA-v1.5, RoboVQA, .etc.	BC-Z, RoboTurk, BridgeData V2, Open X-Embodiment, ARIO .etc.		
Typical Models	VoxPoser, ManipLLM, .etc.	AnyGrasp, DPM, .etc.	InstructNav, RDT, .etc.	RT-2, GR-2, .etc.

- *"System II" Training: Common Sense of the Physical World + Robotics Domain Knowledge*. The former involves basic rules and common knowledge about the physical world that robots need to understand, such as gravity, friction, causal relationship. The latter includes specific commands and instructions that robots must comprehend in order to perform tasks within the robotics domain.
- *Skills Training: (Human Demonstrations + Robot Perception) × Multiple Scenarios*. Skill training first requires human demonstrations, followed by the integration of robot perception data to translate human demonstrations into robot-centered learning objectives. Furthermore, these data must cover a variety of scenarios to enable robots to learn generalizable skills across different environments.
- *"System I" Training: (Skill Training Data + Human Semantic Annotations) × Multiple Tasks*. Task execution relies on the combination of various skills, such as dishwashing, which integrates grasping, wiping, and squeezing. Therefore, this training requires skill training data on one hand, and human semantic annotations on the other, to understand which skills are needed for specific tasks. And data collection and annotation must be performed across multiple tasks to enhance the generalization ability.
- *End-to-End Training: ("System II" Training Data + "System I" Training Data) × Multiple Robot Bodies*. End-to-end training requires combining the training data from above. Moreover, this data must be applicable to a variety of robot models to achieve the generalization capability of end-to-end models across different scenarios, tasks, and robot models.

2.3 Data Bottlenecks in EAI

In the previous section, the specific data requirements of the four models were clarified. It can be observed that the data requirements for end-to-end training are the highest. Therefore, the data requirements for end-to-end training can be considered as the upper limit of the total data demand expectation for EAI. Let the total data demand expectation for EAI be denoted as D. This can be qualitative expressed as:

$$\begin{aligned} D &= (B + C) \times m \\ &= [B + (S + l) \times t] \times m \\ &= \{B + [(d + p) \times s + l] \times t\} \times m \end{aligned} \tag{1}$$

where D is the data demand for EAI, B is the data demand for the "System II", C is the data demand for the "System I", and m is the number of robot categories, S is the data demand for skills, l is the demand for human semantic annotation, d is the demand for human demonstration, p is the demand for robot perception, s is the number of scene categories, t is the number of task categories.

If we assume that the scaling laws still hold in the field of EAI, maximizing D is crucial to meet the data demand expectations of EAI models. The most effective ways to increase D are:

- Increase d and p: Enhance the volume of high-quality human demonstration data and robot perception data. Since these factors are multiplied with s, t, and m, even a slight increase can significantly boost D.
- Increase s, t, and m: Enrich the diversity of training scenarios, tasks, and robot categories. Increasing these amplification coefficients can multiply D.

In summary, increasing the availability of high-quality human demonstration and robot perception data, as well as enhancing the richness of training scenarios, tasks, and robot categories, are the two most effective approaches to meeting the data demand expectations of EAI models. However, the two theoretically most effective approaches encounter significant challenges in practice, which can be categorized as follows:

- *Cost Inefficiency*. The model's performance enhancement demands data in an exponential manner, whereas the real-world data that can be collected only grows linearly. This creates a huge cost pressure when it comes to obtaining high-quality human demonstrations and robot perception data. The costs involved are not limited to the design, manufacturing, and purchasing of data collection devices. They also cover robot adaptation, site maintenance, and long-term human resource investment. Although some video demonstration data reduces the cost of real-world data collection, and simulation and synthetic data provide significant supplementation, the cost of collecting high-quality teleoperation data is still prohibitive. Reducing this cost requires further technological innovation and optimization of data collection processes.

- *Data Silos*. The use of various data collection devices and technologies makes it difficult to gather data in a unified format across diverse scenarios, tasks, and robot bodies. As a result, EAI datasets are isolated from each other. This makes it difficult to share and integrate data across different systems. The absence of EAI models that can generalize across different robot bodies means that datasets will continue to exist in isolated states. Building more universal and compatible EAI models and data standards is necessary to break down these data silos and enhance data sharing and utilization efficiency.
- *Evaluation Void*. There is a lack of standards and theoretical guidance in the data collection process. It is hard to assess whether the collected data effectively enhances the value of the dataset. This leads to blind data collection, redundant construction, and waste of resources. Developing more scientific and reasonable evaluation metrics and standards is essential to improve data quality and promote the healthy development of EAI data engineering.

The cost inefficiency, data silos, and evaluation void are the three data bottlenecks in EAI. The purpose of EAI data engineering is to collect high-quality human demonstration and robot perception at a low cost across as many scenarios, tasks, and robot bodies as possible, in order to construct high-quality EAI datasets. **EAI data engineering is designed to address these three bottlenecks**.

3 Data Production Systems Design for EAI

The first step in conducting EAI data engineering is to design an EAI data production system. EAI data production consists of two aspects: real-world data collection and simulation data generation. Real-world data collection involves robots interacting directly with the external environment through sensors in actual settings to gather operational data and environmental feedback. This method can provide authentic and direct data. Simulation data generation refers to creating data through computer simulations or generative models. The primary advantage of this method is the ability to rapidly produce large amounts of data, thereby reducing costs.

The design of data production systems is crucial for addressing the EAI data bottleneck of cost efficiency. Effective EAI data engineering must strike a balance between high-fidelity real-world data collection, which provides invaluable insights but can be resource-intensive, and scalable, diverse simulation generation, which offers flexibility and scalability at a potentially lower cost. By integrating these two approaches, data production systems can optimize the trade-offs between data quality, cost, and scalability, thereby enhancing the overall efficiency and effectiveness of EAI data engineering.

3.1 Real-World Data Collection Systems

Real-world data collection systems can be categorized into teleoperation-based data collection systems (tele-DCS) and teaching-based data collection systems (teach-DCS), depending on the different methods of data collection. A more detailed classification and introduction of real-world data collection technologies will be provided in Section 5. Here, a brief introduction to their system architecture design will be given to help readers understand the basic principles of system operation.

3.1.1 Teleoperation-Based Data Collection Systems

As shown in Fig. 4, the basic hardware architecture of tele-DCS mainly consists of five major components. Teleoperation devices are used to output control parameters and receive feedback data, including control devices (such as joysticks for robot movement and direction control), display devices (such as monitors or virtual reality headsets), and a computing unit for processing operator inputs and feedback data. Communication devices are responsible for transmitting control parameters, feedback data, and collected data between the teleoperation devices, robot execution devices, data collection devices, and storage devices, ensuring low-latency and high-bandwidth communication. Execution devices, namely robots, are responsible for executing the operator's commands. Data collection devices obtain multimodal data of the robot and its environment in real-time, divided into sensors installed inside and outside the robot (the latter not always necessary), including visual, force, pose, and environmental sensors. Storage devices save the collected data to support subsequent analysis and playback, such as local storage devices and cloud storage.

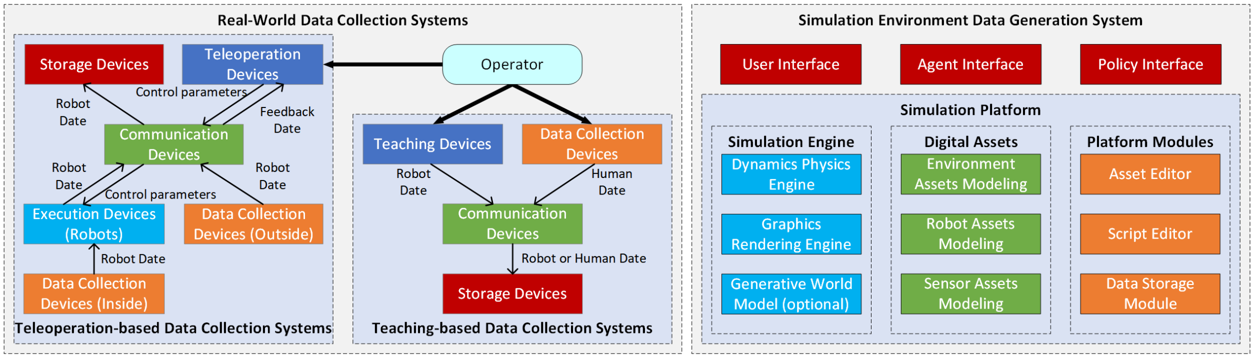

Figure 4: The basic architecture of real-world data collection systems and simulation data generation systems

3.1.2 Teaching-Based Data Collection Systems

Teach-DCS are used to record the teaching actions of human. The core purpose of teaching is to record the movements, postures, and environmental interaction information from robot (directly) or human (indirectly). Fig. 4 illustrates the hardware structure of a teach-DCS. It is more simplified compared to tele-DCS. Teaching devices can be categorized into three types, with the specific choice depending on factors such as task requirements, operational precision, environmental adaptability, and cost: (1) The robot itself as the teaching device, where operators directly manipulate the robot to complete the teaching tasks; (2) A part of the robot as the teaching device, such as the end effector of a robotic arm or a specific sensor module; and (3) Teaching devices or teaching data separate from the robot, where operators collect human teaching data using external devices, such as cameras or motion capture systems.

3.2 Simulation Data Generation Systems

Simulation data generation systems (SDGS) are tools used to simulate robot behavior in virtual environments and generate multimodal data. Compared to real-world data collection systems, they are pure software systems that omit the hardware part in development, offering advantages such as low cost and ease of use.

Generally speaking, SDGS do not exist in isolation but are part of a robot simulation system. In addition to data generation functions, a robot simulation system may also include functions for training, testing, and deploying models of robot perception, decision-making, control, and more. This section will not introduce the entire robot simulation system but will focus solely on the data generation aspect. As shown in the Fig. 4, the system is composed of multiple hierarchical key components.

3.2.1 Simulation Engine

The simulation engine is the core of the entire system, responsible for simulating the behavior of robots in a virtual environment. It includes a dynamics physics engine that simulates the physical behavior of robots interacting with the environment, including various forces such as gravity, friction, elasticity, and inertia, and their effects on the robot's motion state, ensuring that the physical phenomena in the simulation conform to real physical laws. Additionally, it features a graphics rendering engine that converts three-dimensional models or scenes into realistic two-dimensional images based on computer graphics and visual perception theories. This engine uses geometric data, texture data, lighting data, and other inputs to generate images that conform to real visual perception. Optionally, the system may also include a generative world model that generates descriptions and predictions of various scenarios, objects, and behaviors in the real or virtual world. This model simulates the physical properties and dynamic changes of the environment, providing decision-making support and behavioral planning capabilities for EAI agents.

3.2.2 Digital Assets

Digital assets are the basic elements of the simulation environment and include environment assets, robot assets, and sensor assets. Environment assets modeling involves the digital construction of terrains, buildings, indoor layouts, outdoor scenes, and various objects that robots may interact with and operate on. This requires not only accurate geometric shapes and dimensions but also the simulation of physical properties such as materials, textures, lighting, and shadow effects to ensure visual and physical realism. Robot assets modeling is divided into geometric modeling, which focuses on the robot's shape, structure, and position by determining its coordinate system, link lengths, and joint angles, and dynamics modeling, which analyzes the robot's kinematic and dynamic characteristics under physical conditions such as forces, motion, and acceleration. Sensor assets modeling aims to generate a mathematical model that accurately reflects the relationship between sensor inputs and outputs, including the functional relationships between mechanical behavior, displacement, strain, stress, or vibration characteristics and the measured quantities. These models can simulate the working principles of devices such as cameras, radar, and force sensors, as well as their interactions with robots or other objects.

3.2.3 Platform Modules

The construction of a simulation platform requires the addition of various platform modules. Here, only three core modules are introduced. Other non-core modules, such as the graphical user interface and communication modules, are not discussed here. The asset editor allows users to create, edit, and manage digital assets in an intuitive manner. The script editor allows users to write and edit scripts that control the behavior of the simulation. These scripts can define the actions of robots, dynamic changes in the environment, responses of sensors, etc. The data storage module saves various data generated during the simulation process.

3.2.4 System Interfaces

The simulation platform only provides a general digital modeling platform, and a SDGS can only be constructed by designing corresponding interfaces on its basis. These interfaces serve as the bridge for interaction between the system and external models, environments, or users. The User Interface allows the simulation platform to exchange data with external systems or users, defining the rules and protocols for data transmission to ensure that different systems or applications can be interconnected and exchange data, achieving data sharing and information interoperability. The Agent Interface enables the simulation platform to integrate various types of agents, such as robots controlled by LLMs, thereby achieving automated and intelligent processing of complex tasks, including path planning, high-level semantic understanding, long-range reasoning, and more. The Policy Interface can be connected to various robot policy models and algorithms, allowing users to control the behavior of robots or agents based on specific models, rules, or conditions, such as path planning under a specified policy or bimanual coordination under a specified trajectory generation policy.

4 Standardization for EAI Data

The standardization of EAI data is crucial for addressing the EAI data bottlenecks of data silos and evaluation void. In the intricate tapestry of EAI ecosystems, where diverse data sources and formats often lead to fragmented and incompatible datasets, standardization acts as the unifying thread. It harmonizes data structures, facilitates seamless interoperability, and ensures that datasets from various origins can be integrated and utilized cohesively [56]. Moreover, standardization provides a common framework for evaluating data quality and utility, thereby filling the evaluation void and enabling more reliable and consistent assessments of EAI models. The standardization of data in EAI can be divided into multiple aspects. This section will first introduce the classification of EAI datasets. Subsequently, it will propose standardization directions for EAI datasets.

4.1 Classification of EAI Datasets

EAI datasets can be classified as shown in Fig. 5. Among these, demonstration datasets and embodied question-answering (EQA) datasets can be used for training EAI models or agents. The former is primarily utilized for training the "System I", while the latter is used for training the "System II". Both types of datasets can also be combined for end-to-end model training. On the other hand, benchmark datasets are generally not involved in the training of EAI models but are instead used more for evaluating the performance of agents.

4.1.1 Demonstration Datasets

Demonstration datasets typically consist of a series of operational or movement examples that robots can learn from to acquire the skills needed to complete tasks. These can be further divided into manipulation demonstration datasets and locomotion demonstration datasets. The former focuses on robots learning how to perform tasks by observing human or robot manipulation behaviors, while the latter is centered on robots learning how to move and perform actions in space. Table 3 and Table 4 present statistical information on common demonstration datasets currently in use, respectively.

- *Manipulation Demonstration Datasets (MDD)*. Manipulation refers to a series of actions performed by humans or robots on objects, such as grasping, moving, rotating, placing, or adjusting the posture and position of objects to complete specific tasks. MDD usually contain a series of manipulation videos or action sequences

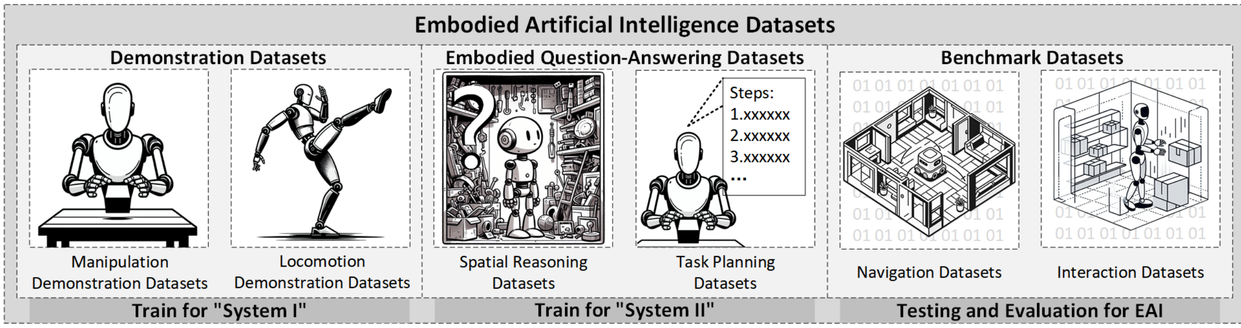

Figure 5: EAI datasets classification

Table 3: Common Manipulation Demonstration Datasets

Dataset	Data Form	Data Scale	Data Modality	Year
GraspNet-1Billion[57]	Real	97,280 images and 1.2B grasping	RGB-D	2020
RoboNet[58]	Real	162k trajectories and 15 million frames	Color Images	2020
ACRONYM[59]	Simulated	17.7M parallel grasping pairs	Point Cloud	2021
BridgeData[60]	Real	7,200 demonstrations	RGB-D	2021
AKB-48[61]	Real	100K generated RGB-D images	RGB-D	2022
BC-Z[48]	Real	25k demonstrations and 18k human video	RGB	2022
RT-1[62]	Real	130k robot demonstrations	RGB	2022
Grasp-Anything[63]	Simulated	1M samples and 600M grasping	Text / Image	2023
GAPartNet[64]	Simulated	8,489 instances	Point Cloud / RGB-D	2023
ManiSkill2[65]	Simulated	4M demonstration frames	Point Cloud / RGB-D	2023
ARNOLD[66]	Simulated	10,080 demonstrations	Text / RGB-D	2023
DexArt[67]	Simulated	6K point clouds for each object	Point Cloud	2023
BridgeData V2[51]	Real	60,096 trajectories, 50,365 teleoperation demonstrations, 9,731 deployments	RGB-D, Audio, Text and Haptic	2023
Open X-Embodiment[23]	Real	22 types of robots, over 1 million trajectories, 527 skills	Force Sensing Information / Point Cloud / RGB-D	2023
RH20T[68]	Real	147 tasks, 42 skills, 10,000 robot operation sequences and 110,000 corresponding human demonstration videos	RGB, Depth, Binocular Infrared, Haptic, Audio	2024
DROID[69]	Real	76k trajectories and 350 hours of interaction	RGB-D	2024
ARIO[24]	Real & Simulated	258 series and 321,064 tasks	RGB-D, Audio, Text and Haptic	2024
RoboMIND[70]	Real & Simulated	55,000 robot trajectories, 279 tasks, 61 types of objects	Text / RGB-D	2024
AgiBot World[71]	Real	Over 1 million trajectories of over 100 robots, over 100 scenes in five domains	RGB-D, Haptic	2025

carried out by humans or robots. These actions are meticulously recorded and annotated so that robots can analyze and learn how to execute these actions through machine learning algorithms. Since most manipulations are based on grasping, some MDD may exclusively contain grasping data.

- *Locomotion Demonstration Datasets (LDD)*. LDD focus on recording and providing full-body motion control data of robots or organisms when performing movement tasks, such as walking, running, jumping, crawling, and their variants under different environments and conditions. By capturing and recording key frames, joint angles, velocities, accelerations, and other information during the movement process, LDD provide the foundation for robots to learn how to move in three-dimensional space and maintain balance. Most current LDD are used for humanoid robots to meet specific task requirements.

The construction of MDD is a systematic process. It begins with defining clear manipulation tasks, then designing corresponding experimental scenarios in the real world or simulation environments. Data on robot-environment interactions are collected using teleoperation technologies, among others. These data are subsequently labeled and analyzed to extract key features and interaction patterns, ultimately being organized into a comprehensive dataset that includes information on environmental states, robot actions, object properties, and task outcomes. The sources of motion data in LDD mainly come in three forms: motion capture data, video-based human motion estimation, and synthetic data.

Table 4: Common Locomotion Demonstration Datasets

Dataset	Year	Data Source	Data Scale	Modalities
Human3.6M[72]	2014	Motion capture	3.6 million frames of 3D human pose data	2D and 3D skeletal joint positions, depth images, and video sequences
KIT Motion-Language Dataset[73]	2016	Motion capture	3,911 actions with 6,278 natural language annotations	3D skeletal joint positions, text
AMASS[74]	2019	Motion capture	Over 300 subjects and more than 11,000 movements	3D skeletal joint positions
HumanAct12[75]	2020	Synthetic	1,191 3D motion clips, totaling 90,099 poses	3D skeletal joint positions
HumanML3D[76]	2022	Motion capture & Synthetic	14,616 actions and 44,970 descriptions	3D skeletal joint positions, text
Humanoid-X[77]	2024	Pose estimation	163,800 action samples	Videos, text descriptions, 3D human poses, humanoid robot key points, and robot action sequences

Table 5: Common Embodied Question-Answering Datasets

Dataset	Year	Q&A Type	Q&A Mode	Data Form	Answer Type	Scale
EQA v1[78]	2018	SRD	Active EQA	Simulation	Open-ended	5,000+
VideoNavQA[79]	2019	TPD, SRD	Episodic Memory EQA	Simulation	Open-ended	101,000
SQA3D[80]	2022	TPD, SRD	QA only	Real	Multi-choice	33,400
K-EQA[81]	2023	SRD	Active EQA	Simulation	Open-ended	60,000
EgoPlan-Bench[82]	2023	TPD, SRD	Interactive EQA, Active EQA	Real	Open-ended	4,900+
OpenEQA[83]	2024	TPD, SRD	Active EQA, Episodic Memory EQA	Simulation	Open-ended	1,600+
HM-EQA[84]	2024	SRD	Active EQA	Simulation	Multi-choice	500
S-EQA[85]	2024	SRD	Active EQA	Simulation	Binary	-
MARPLE[86]	2024	TPD, SRD	Episodic Memory EQA	Simulation	Multi-choice	-
MFE-ETP[87]	2024	TPD, SRD	Interactive EQA	Manual	1,000+	
RoboVQA[43]	2024	TPD, SRD	QA only	Real	Open-ended	829,502
EmbSpatial-Bench[88]	2024	SRD	Interactive EQA, Active EQA	Real	Open-ended	3,640
EmbodiedCity[89]	2024	TPD, SRD	QA only	Simulation	Open-ended	50,400
V-IRL[90]	2024	SRD	QA only	Real	Multi-choice	-
VSI-Bench[91]	2024	SRD	QA only	Real	Single-choice	5,000+

4.1.2 Embodied Question-Answering (EQA) Datasets

EQA datasets are designed to train and evaluate a robot's ability to understand and answer questions related to the environment or tasks. These datasets are crucial for enhancing the interactivity and intelligence of robots. They can be further categorized into spatial reasoning datasets and task planning datasets. The former focuses on spatial cognition and reasoning, including the understanding and inference of object positions, orientations, and spatial relationships. The latter contains questions and answers that helping robots learn how to plan action steps based on given goals and constraints.

Since spatial reasoning is the foundation of task planning, task planning datasets generally include spatial reasoning datasets, but spatial reasoning datasets do not necessarily include task planning datasets. Table 5 presents statistical information on common EQA datasets currently in use.

- *Spatial Reasoning Datasets (SRD)*. SRD focus on enhancing agents' abilities to understand and manipulate objects in three-dimensional space. These datasets consist of a series of queries about object positions, orientations, and spatial relationships. Agents need to verify these spatial relationships through perception or exploration of the environment. The purpose of SRD is to train agents for precise spatial localization and path planning, which is essential for robots navigating and operating in complex environments. The functions of these datasets include providing rich spatial relationship information, simulating various spatial layouts, and evaluating agents' accuracy and efficiency in processing spatial information.
- *Task Planning Datasets (TPD)*. TPD provide a structured environment for agents to learn how to decompose complex tasks into a series of executable steps. These datasets typically include task descriptions, goals, constraints, and possible action plans. Agents learn how to effectively plan and execute tasks through interaction with the environment. The purpose of TPD is to train agents in decision-making and resource allocation to achieve specific goals. The functions of TPD include providing diverse task scenarios, simulating different environmental conditions, and evaluating agents' adaptability and efficiency.

The key steps in constructing an EQA dataset include selecting appropriate environment data (synthetic or real) and simulating the environment using a simulator (or directly operating based on the environment data); designing diverse question templates that cover various scenarios and object attributes; generating specific questions through programming or manual means, which can be assisted by rule-based methods or LLMs; determining the correct answers

Table 6: Common Benchmark Datasets

Dataset	Year	Type	Data Form	Agent	Sensors	Supported Tasks
nuScenes[92]	2020	ND	Real	Vehicle	RGB/Radar/Lidar	Autonomous Driving
VLN-CE[93]	2020	ND	Simulation	Robots	RGB/RGBD	Language Instruction, Navigation
Vis.Room Rearr.[94]	2021	ID	Simulation	Robots	RGB	Manipulation
ManipulaTHOR[95]	2021	ID	Simulation	Robots	RGBD	Manipulation
AVDN[96]	2022	ND	Real	Drone	RGB	Navigation
MetaDrive[97]	2022	ND	Simulation	Robots	RGBD/Lidar	Navigation
ProcTHOR-10k[98]	2022	ND, ID	Simulation	Robots	RGB/RGBD	Navigation, Manipulation
HomeRobot[99]	2023	ND, ID	Real	Robots	RGB/RGBD	Navigation, Manipulation
Arnold[66]	2023	ND, ID	Simulation	Robots	RGB/RGBD	Language Instruction, Manipulation
Behavior-1K[100]	2023	ND, ID	Simulation	Robots	RGB/RGBD	Navigation, Manipulation
AerialVLN[101]	2023	ND	Simulation	Drone	RGBD	VLN
MetaUrban[102]	2024	ND	Simulation	Vehicle	RGBD/Lidar/Pose	Autonomous Driving
GRUtopia[103]	2024	ND	Simulation	Robots	RGBD	Autonomous Driving
CityNav[104]	2024	ND	Real	Drone	RGBD	VLN
V-IRL[90]	2024	ND	Real	-	RGB	Navigation/QA/Planning
EmbodiedCity[89]	2024	ND	Simulation	ALL	RGBD/Lidar/Pose	Scene Understanding/QA/ Dialogue/Navigation/Planning
ET-Plan-Bench[105]	2024	ND, ID	Simulation	Robots	RGBD	Navigation/QA/Planning
EmboDiedBench[106]	2025	ND, ID	Simulation	Robots	RGBD	Scene Understanding/ Navigation/QA/Planning

for each question by analyzing the simulated environment or simulating the exploration behavior of an agent; manually annotating and verifying the generated questions and answers to ensure accuracy and consistency; and finally optimizing it based on agent's feedback, such as adjusting the difficulty of the questions, increasing diversity, or improving the simulated environment.

4.1.3 Benchmark Datasets

Benchmark datasets are used to evaluate the performance of robots in specific tasks or environments, providing researchers with a standardized testing platform. Benchmark datasets can be divided into navigation datasets and interaction datasets. The former is used to assess the navigation capabilities of agents in different environments, including indoor, outdoor, and complex terrain scenarios. The latter focuses on evaluating the interaction capabilities of agents with operable objects or other agents, including manipulation and transportation, tool use, and multi-agent collaboration. Table 7 shows commonly used benchmark datasets.

- *Navigation Datasets (ND).* ND are specifically designed to evaluate and enhance the autonomous navigation capabilities of agents in diverse environments. These datasets meticulously record the path planning, obstacle avoidance, and interaction behaviors of agents while performing navigation tasks. The purpose of ND is to simulate real-world navigation challenges, such as indoor, outdoor, and complex terrain scenarios, as well as dynamically changing environmental conditions. These datasets provide a standardized testing platform for evaluating the performance of different navigation strategies, thereby promoting the development and innovation of navigation technologies.
- *Interaction Datasets (ID).* Interaction datasets focus on evaluating the interaction capabilities of agents with operable objects and other agents. By providing a wealth of interaction scenarios and tasks, these datasets enable agents to practice and refine these fundamental skills in simulated or real environments. Through these datasets, researchers can develop and optimize interaction algorithms for agents, enabling them to interact more naturally and effectively with operable objects and other agents. These datasets also provide important benchmarks for assessing the efficiency and accuracy of agent.

It is worth noting that the above dataset classifications are not mutually exclusive. For example, benchmark datasets may include demonstration datasets or EQA datasets, and manipulation demonstration datasets can be combined with locomotion demonstration datasets.

4.2 Standardization of EAI Datasets

Standardization can enhance the universality and interoperability of datasets, enabling data produced by different companies or research institutions to be shared and open-sourced. A unified basic architecture for datasets helps to objectively and comprehensively assess data quality, thereby enabling standardized data management and promoting the construction of large-scale datasets, and improving the practicality and effectiveness of datasets.

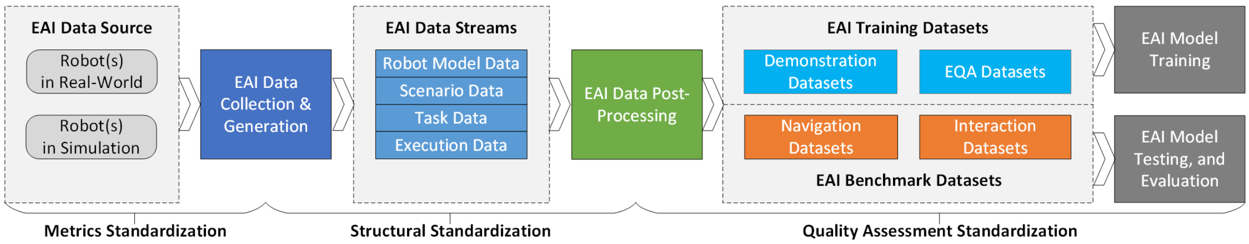

Figure 6: The entire lifecycle of EAI data and its corresponding three standardization phases

As shown in Fig. 6, the data in EAI training datasets originates from the execution processes of single or multiple robots performing different tasks in real or simulated environments. The construction of EAI datasets encompasses acquiring EAI data streams via data collection and generation technologies, followed by a series of post-processing steps, including classification, alignment, annotation, cleaning, and structuring. Through distinct construction processes, EAI data is shaped into both training datasets and benchmark datasets. Subsequently, after training the EAI model with the training data, it is imperative to utilize the EAI benchmark dataset for rigorous testing and evaluation. This sequence of activities constitutes the entire lifecycle of the EAI data experience. The standardization of this process involves three phases: metrics, structural, and quality assessment.

4.2.1 Metrics Standardization for EAI Datasets

The metrics standards are used to ensure the reliability and usability by setting minimum quality requirements for the production process [51, 70]. These standards encompass several key aspects: spatial metrics standards, which include the spatial motion (angular) resolution and accuracy of the robot itself and its joints, as well as the spatial perception resolution and accuracy of sensors; temporal metrics standards, which cover the duration of samples, the temporal resolution of the data, and temporal accuracy (system time synchronization error, data acquisition delay, temporal offset and drift error between different modalities of data); and other metrics standards, which involve the spatial positioning accuracy of tactile sensors, the sampling rate of acoustic sensors, the physical simulation accuracy of digital assets, and more.

4.2.2 Structural Standardization of EAI Datasets

EAI datasets should encompass a wide range of useful data modalities and be structured to maximize compatibility with subsequent model training, testing, and evaluation requirements. At a minimum, these datasets should include four types of data streams [23, 24]: Robot Model Data, which covers hardware and software versions of robots, sensors, simulators, and more; Scenario Data, which includes the type of scenario, maps, sensor calibrations, textual descriptions, and digital assets of simulated scenes; Task Data, which involves task descriptions, skill categorizations, initial states of the robot and its components, and attributes of objects to be manipulated; and Execution Data, which consists of motion data (e.g., position, velocity, angles), perception data (e.g., RGB, depth, point cloud), external perception data (e.g., visual motion capture), decision-making data, action annotations, and simulation execution parameters.

4.2.3 Quality Assessment Standardization for EAI Datasets

This part of standardization primarily focuses on two key areas: quantitative metrics and empirical metrics. Quantitative metrics provide objective, measurable criteria to evaluate datasets, including aspects such as completeness, consistency, accuracy, diversity, and balance [31, 33]. Empirical metrics, on the other hand, are often based on the performance of models trained on the dataset and offer insights into how well the data supports the intended applications, encompassing model performance, generalization ability, robustness, transferability, and user feedback [32]. Additionally, the construction of benchmark datasets is crucial for providing a standardized testing platform to assess the performance of EAI systems and ensure the practical applicability and effectiveness of the datasets [105, 106].

5 Real-World Data Collection Technologies for EAI

The improvements of real-world data collection technologies (RWDCT) is crucial for addressing the EAI data bottleneck of cost efficiency. Various RWDCT share a common goal: to collect data of the highest possible quality and universality in the most convenient manner. The more convenient the data collection process, the lower the cost; and the higher the quality and universality of the data, the more likely it is to eliminate data silos. However, data

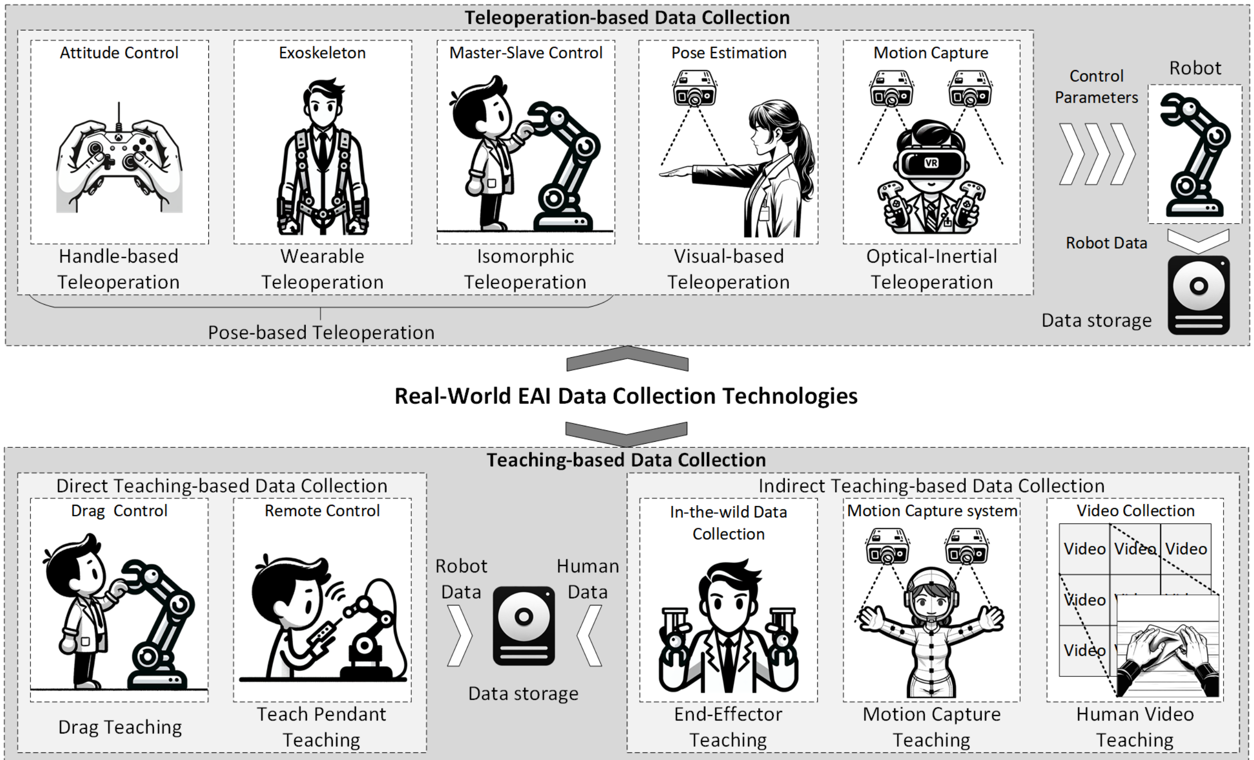

Figure 7: The classification of real-world data collection technologies for EAI

silos cannot be eliminated solely through improvements in RWDCT. Synchronized development of general models and robot bodies is also required to achieve this goal.

RWDCT for EAI can be categorized into teleoperation-based and teaching-based data collection. As shown in Fig. 7, these categories can be further divided into various specific methods. All technologies discussed in this section essentially involve the construction of the demonstration datasets introduced in Section 4.1.1.

5.1 Teleoperation-based Data Collection Technologies

Teleoperation, or telerobotics, refers to a method where a human operator controls a robot or mechanical system from a distance. The prefix "tele-" implies remote operation, allowing the operator to manipulate the robot's actions from a distant location. As shown in Fig. 7, teleoperation can be categorized into three types.

5.1.1 Pose-based Teleoperation Technologies

Pose-based teleoperation refers to the method where a human operator remotely controls a robot using devices that directly record pose data. These devices convert pose signals into control signals for the robot's movements. Among remote operation devices, pose-based systems are the most diverse. They can range from simple handheld controllers to wearable devices such as gloves, motion capture suits, or exoskeletons, and isomorphic teleoperation robots that form a master-slave structure with the controlled robot. Therefore, as shown in Fig. 7, these technologies can be further subdivided into three categories.

- *Handle-based Teleoperation*: Typically, such devices feature a simple structure and transmit the pose parameters of the end effector to the robot solely through a joystick-like device, such as HATO [107].
- *Wearable Teleoperation*: Such devices are generally presented in the form of exoskeletons and offer greater intuitiveness and naturalness, as it allows operators to directly control the robot through their own body movements, such as AirExo [108] and ACE[109].
- *Isomorphic Teleoperation*: Isomorphic teleoperation refers to the real-time replication of movements between two identical robots, such as Mobile ALOHA[25], GELLO[110], and HOMIE[111]. This involves setting one

robot as the master (operator) device and the other as the subordinate device. Since the dynamic structures of the two robots are identical, the complexity of control and motion replication is significantly reduced.

5.1.2 Visual-based Teleoperation Technologies

Visual-based teleoperation refers to the process of capturing an operator's movements using visual sensing technologies (such as RGB-D cameras) and then converting these movements into control commands to manipulate a robot. This method directly maps human actions to robot actions, allowing operators to easily and intuitively control robotic systems. It is suitable for cost-saving scenarios with lower precision requirements, such as DexPilot[112], AnyTeleop [113], HumanPlus[114], and DIME [**?**].

5.1.3 Optical-Inertial Teleoperation Technologies

Optical-inertial teleoperation is a sophisticated approach that integrates optical motion capture systems with inertial measurement units (IMUs) to remotely control robots. This method leverages the respective advantages of wearable teleoperation and vision-based teleoperation technologies to achieve more accurate, reliable, and continuous tracking of the operator's movements. Typical optical-inertial teleoperation systems include motion capture systems, virtual reality (VR)-based teleoperation platforms, and other integrated forms, such as Bunny-VisionPro [115], OmniH2O[114], and Mobile-TeleVision[116].

5.2 Teaching-Based Data Collection Technologies

Teaching-based data collection refers to the process where a human operator performs a task or a series of tasks, and the teaching data is then used to guide the robot in performing similar tasks. As shown in Fig. 7, teaching-based data collection methods can be divided into two categories.

5.2.1 Direct Teaching Technologies

Direct teaching, also known as hand-guided teaching, is characterized by its intuitive operation, making it suitable for simple teaching tasks. It does not require additional hardware, resulting in lower costs. However, its drawbacks include low teaching efficiency and limited applicable scenarios. Specific implementations of direct teaching include the following:

- *Drag Teaching*: Physically manipulates the robot's joints or end-effector to desired positions through manual guidance. It has been widely applied in various industrial robotic arms and assistive robotic arms.
- *Teach Pendant Teaching*: Allowing operators to directly control or program the robot via a handheld teach pendant, such as buttons, knobs, and touchscreens.

In terms of usage, teach pendant teaching is somewhat similar to handheld-based teleoperation. However, there are key differences. Teach pendant teaching is suited for programming setups and scenarios requiring precise control, with operators interacting directly and in close proximity to the robot. In contrast, handheld-based teleoperation emphasizes flexibility and safety, enabling remote, real-time control of the robot by the operator. A more intuitive distinction is that teach pendants are typically specialized devices manufactured according to robot vendor specifications, while handheld-based teleoperation devices are usually third-party, general-purpose tools, adhering to different interface standards and communication protocols.

5.2.2 Indirect Teaching Technologies

In indirect teaching, operators no longer directly manipulate the entire robot. In the field of data collection for EAI, three common indirect teaching methods exist:

- *End-Effector Teaching*: It involves operators completing teaching tasks by controlling the robot's end-effector, such as UMI [26] and Fast-UMI [117]. They transform the end-effector into a universal manipulation interface, enabling humans to hold it independently for data collection. Compared to using the entire robot for data collection, using only the end-effector allows for convenient data acquisition in various open environments. As a result, this technique is also referred to as "in-the-wild" data collection.
- *Motion Capture Teaching*: It refers to the process where an operator wears motion capture devices (such as data gloves [27] and motion capture suits [118]), and the system records the operator's movements to serve as teaching data for robots.

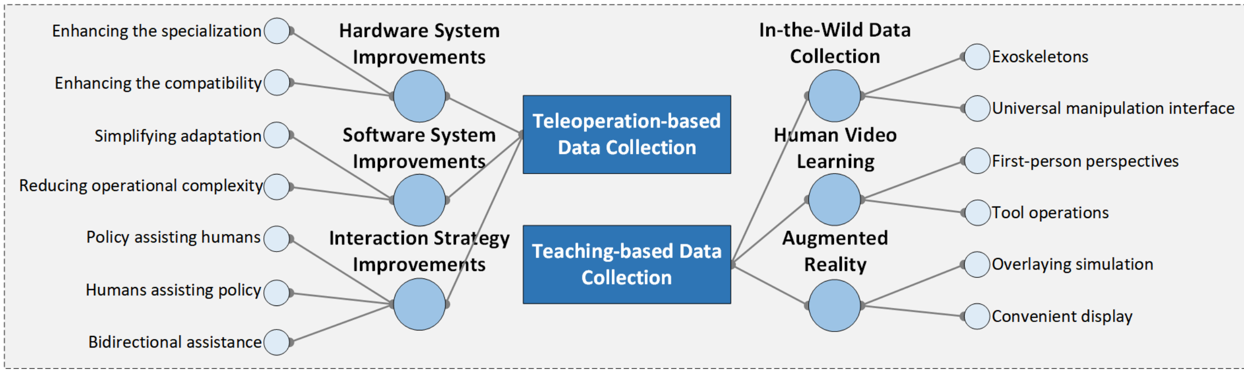

Figure 8: The improvement directions in real-world data collection technologies

- *Human Video Teaching*: It is an emerging robotic learning method aimed at completing complex tasks by imitating human behavior, without the need for manual programming or extensive robot data collection. The core of this approach lies in using human video demonstrations as a source of knowledge, enabling robots to understand and execute tasks while demonstrating strong generalization capabilities [119, 120]. This technology may be more cost-effective than expert demonstrations performed by robots [121, 122].

There is an essential difference between the data collected through indirect teaching and teleoperation: Indirect teaching collects human data, while teleoperation collects robot data. Indirect teaching data may not precisely correspond to robot movements and may not guarantee the usability of the collected data. For example, the range of motion in indirect teaching data may exceed the robot's workspace, or lacks tactile data. In contrast, teleoperation directly maps human actions to the robot and it can obtain all robot data.

It should be noted that the above technical classifications are not absolute. A data collection system can integrate multiple collection technologies (for example, HOMIE is an integration of wearable teleoperation and isomorphic teleoperation). In practice, people need to make comprehensive choices for the optimal data collection methods based on various aspects such as the adopted technical approach, collection efficiency, and cost.

5.3 Improvement Directions in Real-World Data Collection Technologies

5.3.1 Improvement Directions in Teleoperation-based Data Collection Technologies

Teleoperation is the most widely used method for data collection, directly producing robot data with significant research focus on its improvement. As shown in Fig. 8, improvements can be categorized into three main directions:

Hardware System Improvements aim to enhance the specialization and compatibility of teleoperation hardware systems. Specialization focuses on better adaptation for specific robot types, such as bimanual robots [123], quadruped robots [124], grippers [125], dexterous hands [126, 127], and humanoid robots [128, 129]. Compatibility improvements aim to work with various robot types [130, 131]. **Software System Improvements** are dedicated to simplifying teleoperation software adaptation and reducing user operational complexity. Examples include integrating multiple human-machine interaction interfaces [132, 133], improving device compatibility [134], and incorporating built-in motion mapping strategies to avoid singularities [135]. **Interaction Strategy Improvements** aim to address the inherent unreliability of human movements, which can introduce delays, jitter, and errors in teleoperation. To collect high-quality data, various strategies have been proposed: Policy assisting humans involves using existing policy models to autonomously perform repetitive actions during data collection or correct unreliable human teleoperation behaviors online, requesting human input only when uncertain [136, 137, 138]. Humans assisting policy leverages policy models to perform repetitive actions, with humans correcting and updating the model when it produces unreliable behaviors [139, 140, 141, 142, 143]. Bidirectional assistance combines these two modes, and may even incorporate adversarial strategies [144, 145, 146, 147].

5.3.2 Improvement Directions in teaching-based Data Collection Technologies

Teaching, especially indirect teaching, offers greater flexibility as it is not constrained by the robot's physical form. The focus of its improvement lies in leveraging this advantage while ensuring data quality. As shown in Fig. 8, improvements can be categorized into three main directions:

In-the-Wild Data Collection aims to develop affordable, lightweight, and user-friendly hardware devices for efficient data collection in open environments, such as exoskeletons and universal manipulation interface. The former focus on lightweight design [148] and wider variety compatibility of dexterous hands [149, 150]. The later focus on developing more versatile hardware and software systems, such as lighter structures [151], tactile supporting [152, 153, 154, 155], dexterous hands supporting [156, 157], join policy assistance [158], and adaptability to other robot forms [159, 160]. **Human Video Learning** focuses on more accurately and comprehensively learning human manipulation fundamentals from videos, such as precise first-person perspectives [161], fine tool operations [162], and generalization capabilities [163]. And **Augmented Reality** enhances human demonstrations' compatibility with robot dynamics, such as overlaying simulated robots onto the operation view using VR headsets [164, 165] or pads [166].

6 Simulation Data Generation Technologies for EAI

Simulation data generation refers to the process of generate data related to robot interactions within a simulated environment. This data serves as an important supplement to real-world data collection. **The improvements of simulation data generation technologies (SDGT) is crucial for addressing the EAI data bottleneck of cost efficiency and data silos**. The emergence of data silos in real-world data is significantly attributable to the absence of a unified data collection technology, which inherently exacerbates the discrepancies among diverse datasets. In contrast, SDGT can effectively mitigate these disparities to a considerable extent, thereby expanding the coverage across various scenarios, tasks, and robot bodies.

6.1 Introduction to Robotic Simulation Platforms

Over the past decade, the evolution of modern robotic simulation platforms can be categorized into two directions:

- *Realistic Visual Rendering*: This direction has been primarily driven by the demands of the film and gaming industries. In this area, simulation platforms focus on integrating advanced graphics tools to achieve photorealistic rendering of simulated scenes, such as Unity [167], Unreal Engine [168], and CryEngine [169]. These simulation platforms can improve the visual generalization ability of trained models. However, they may lack sufficiently realistic dynamics simulation, limiting their application in robotic simulations.
- *Realistic Dynamics Simulation*: This direction has been primarily driven by the research needs of academic institutions. In this area, simulation platforms emphasize integrating advanced dynamics solvers to achieve realistic physics simulations in virtual environments, such as MuJoCo [170], NVIDIA's PhysX [171] framework, Gazebo [172], and PyBullet [173]. However, early versions of these platforms often neglected visual realism, making them less suitable for the emerging field of EAI.

With the recent rise of EAI, many simulation platforms have begun to integrate realistic visual rendering with accurate dynamics simulation. For example, NVIDIA's Isaac series [174] combines its expertise in game rendering and robotic dynamics simulation, providing an excellent simulation environment for robots, so as SAPIEN [175] and Genesis [176].

6.2 Simulation Data Generation Classification for EAI

On the one hand, simulation data generation for EAI requires the use of computer simulation technologies to create virtual environments and scenarios that mimic the physical processes and interactions in the real world. On the other hand, it involves importing real-world data into the simulation platform and using algorithms, statistical models, or real-world data to synthesize new data that is statistically similar to real data. As shown in Fig. 8, simulation EAI data generation technologies can be divided into four types.

6.2.1 Trajectory Synthesis

It is used to generate trajectory data for the robot's body or end-effector in a simulation environment. The main process of trajectory synthesis includes path planning and motion control, which involves generating smooth, continuous, obstacle-avoiding paths from an initial position to a target position while ensuring that the body or end-effector moves accurately along the planned trajectory, satisfying constraints such as velocity, acceleration, and jitter. In practice, there are two main approaches:

- *Virtual Teleoperation-Based*: Virtual teleoperation refers to generating robotic behavioral control data by sending remote control commands to the simulation platform via teleoperation devices. Virtual teleoperation

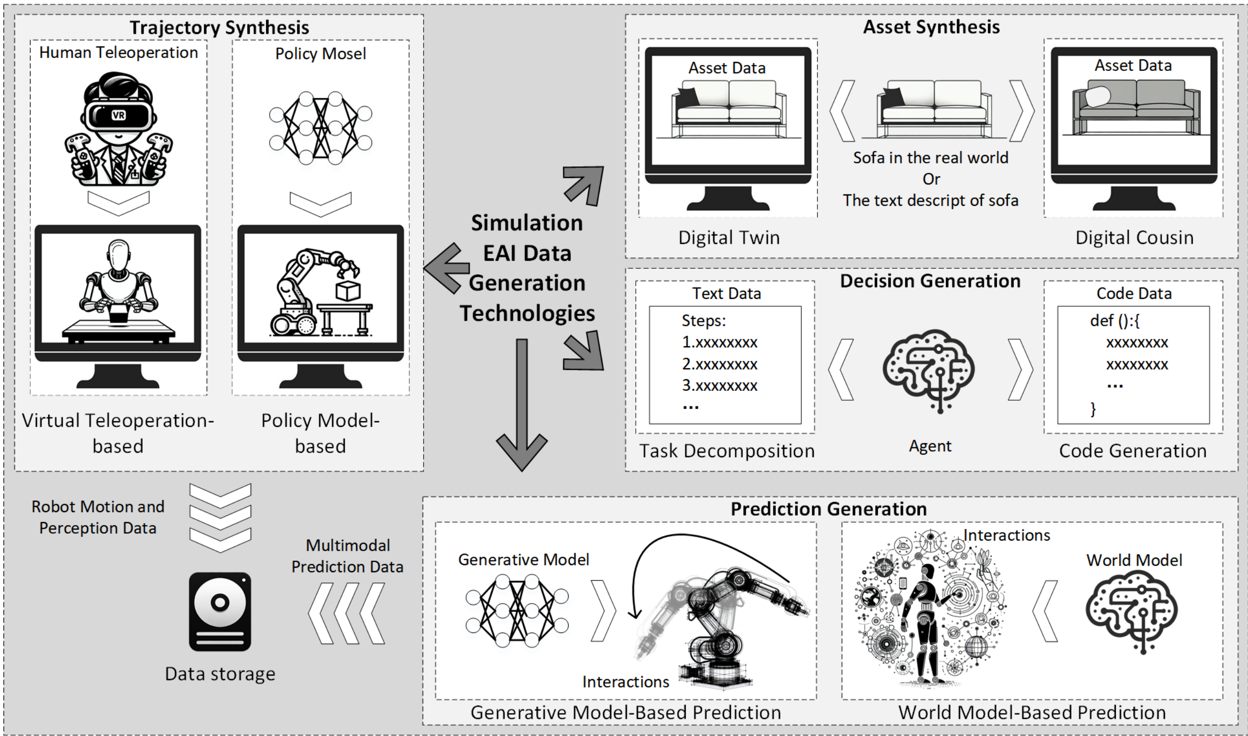

Figure 9: The classification of simulation EAI data generation technologies

allows operators to directly generate demonstration data within simulation environments, which can serve as seeds for subsequent large-scale data generation, such as MimicGen [28].

- *Policy Model-Based*: Compared to manually synthesizing trajectory data through virtual teleoperation, using an existing policy model to automatically synthesize large amounts of data in a simulator offers significant efficiency improvements and can construct a powerful data flywheel, such as DexMimicGen[177].

6.2.2 Asset Synthesis

It refers to the creation of virtual scenes and objects, particularly interactive objects in simulation environments, using generative AI and related technologies to support the training, simulation, and evaluation of robots. Asset synthesis is typically based on real-world scenes or objects to avoid generating arbitrary assets that deviate from reality. It often involve the processing of 3D reconstruction or 3D generation technologies, such as Neural Radiance Fields (NeRF) [178] and Gaussian Splatting [179] technologies. Asset synthesis methods can be categorized into two types:

- *Digital Twin-Based*: A digital twin is a virtual model created through digital means to precisely map and simulate physical entities or systems in the real world. In the field of EAI, digital twins are used to synthesize interchangeable objects that are as consistent as possible with their real-world counterparts, such as RoboGSim[180] and RoboTwin[181].
- *Digital Cousin-Based*: While digital twins minimize the discrepancy between simulated and real objects, they are costly to produce and cannot generalize across domains as virtual replicas of real scenes. To address these limitations, ACDC[29] introduced the concept of digital cousins. Unlike digital twins, they do not explicitly mimic real-world counterparts but still exhibit similar geometric and semantic functionalities, then reduce the cost of creating high-precision virtual environments.

6.2.3 Decision Generation

In the field of EAI, decision generation typically refers to the process of converting natural language instructions into executable action commands by fusing multimodal information (such as visual, auditory, and textual data) and leveraging the powerful language understanding and generation capabilities of LLMs. It is an essential component of hierarchical EAI. Generally, there are two levels of decision generation.

- *Task Decomposition*: It is a prerequisite for executing complex tasks, which involves breaking down task goals into a series of actionable sub-goals. This usually requires the use of LLM to perform reasoning and analysis based on input information and to formulate specific action plans in conjunction with a task planning module, such as COWP [182] and EAIB [183].
- *Code Generation*: It involves transforming natural language instructions into executable program code to achieve precise control over robot behavior. This typically requires the use of LLMs to generate control code, such as GenSim [184] and RoboCodeX [185]. Alternatively, the LLM can generate executable control code after completing task decomposition, based on the results of that decomposition.

Decision generation technology is widely applied in navigation tasks, complex task execution, and human-robot collaboration. For example, VLN models generate motion directions and target position information based on language descriptions and visual observations to guide robots in completing navigation tasks. The decision generation process is typically implemented through an agent, which integrates the LLMs into the simulation data generation system via the agent interface described in Section 3.2.4 to produce corresponding decision data.

6.2.4 Prediction Generation

The model's understanding of the physical world is often difficult to measure, so it is typically translated into the model's ability to predict the development of events or the changes and outcomes resulting from its interactions with the environment. To train this ability, on the one hand, a large amount of real-world data capturing physical change processes is required. On the other hand, specialized generation tools are needed to produce synthetic data that is difficult to collect in the real world. The latter approach is known as prediction generation technologies. This section further divides them into two categories:

- *Generative Model-Based Approaches*: They are increasingly being used in the field of EAI to generate richer and more realistic interaction scenarios. Specifically, these models are applied in human motion generation to produce human motion videos for robot policy learning, enabling robot manipulation strategies to generalize to new tasks [186]. Additionally, generative models are used for robot motion generation, creating video predictions of different robot embodiments in various scenes and tasks [187].
- *World Models-Based Approaches*: In this section, the term "world model" specifically refers to generative world models. These are technologies that utilize generative models to create virtual worlds, simulating the physical laws, dynamic changes, and interactive behaviors of the real world. The core of generative world models lies in their ability to generate rich, interactive virtual environments from minimal inputs, such as Genie [188] and Cosmos [189].

6.3 Improvement Directions in Simulation Data Generation Technology

As shown in Fig. 10, generated data serves as an effective supplement to real data, with rapid progress in several key areas. For **Enhanced Data Generation**, more simulation and synthetic data are generated based on existing real or simulation data. This includes Real2Sim, which efficiently generates simulation data from real-world teleoperation demonstrations to better learn skills [190, 191] and achieve scene generalization [192]; Sim2Syn, which generates synthetic data from human demonstrations in simulation environments to conform to physical laws [193], learn skills [194, 195], and generalize scenes [196]; Asset Generation, which produces higher precision assets [197, 198], richer morphological and visual diversity [199, 200, 201, 30], and more controllable generation [202]; and Decision Generation, which enhances reasoning accuracy through stricter physical constraints [203] and chain-of-thought techniques [204, 205].

For **Human Demonstration Data Conversion**, human operation demonstration data is directly converted into robot operation data in simulation environments. This includes Real2Sim conversion of real human operation demonstrations (e.g., bimanual dexterous operations from various perspectives) into simulation data to avoid cumbersome teleoperation [206, 207, 208, 209]; and Sim2Syn synthesis of robot operation data in simulation environments based on human demonstrations, further eliminating the need for real data collection [210, 211].

Finally, **World Simulators** focus on end-to-end simulation by constructing better world models, emphasizing richer generation [212], more realistic reconstruction [213], more realistic interaction experiences [214], and perception that conforms to physical laws [215].

While simulation data offers substantial reductions in equipment and labor costs, it necessitates confronting the escalating computational expenses and the persistent sim2real gap. The sim2real gap arises fundamentally from the inherent limitation that simulations can only asymptotically approach, but never fully replicate, the complexities of the

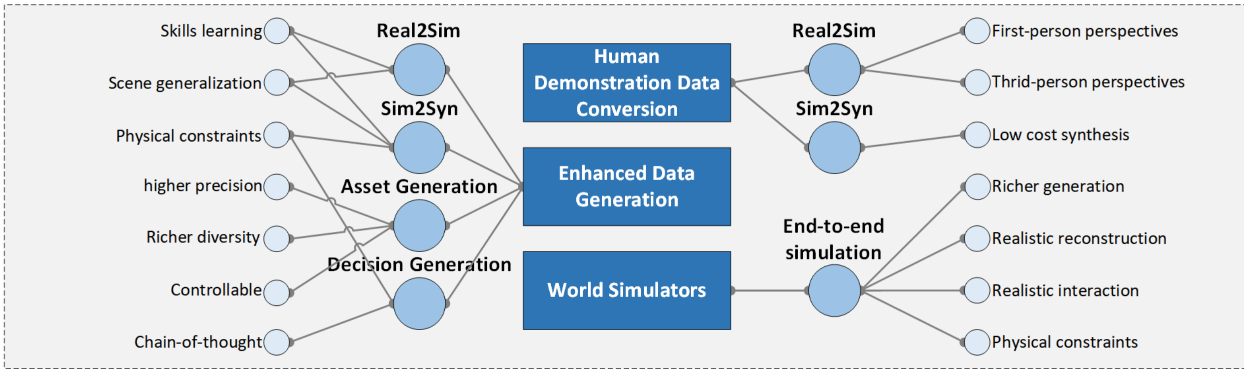

Figure 10: The improvement directions in simulation data generation technologies

real world. This asymptotic approximation may also be subject to scaling laws, implying that the cost of achieving higher fidelity in simulations could increase disproportionately. Consequently, merely enhancing simulation precision may not suffice to bridge the gap. A more effective strategy might involve constructing a world model that incorporates the sim2real gap as a learnable component, enabling the use of simulation outcomes to predict real-world behaviors more accurately.

The decision of whether to employ simulation data, and if so, determining the appropriate scale and proportion for its integration, remains an open and complex question within the field. These questions necessitate a multifaceted approach, involving the coordinated development and maturation of various technologies over an extended period. Only through such efforts can we hope to arrive at well-informed and effective solutions that balance the benefits and limitations of simulation data in relation to real-world applications.

7 The Application and Optimization of EAI Data Engineering

The application of EAI Data Engineering can be delineated into three distinct phases: the analysis of data requirements within specific application domains, the selection of appropriate data production methods, and the optimization of concrete deployments. Accordingly, this section will unfold discussions from these three perspectives, offering guidance to practitioners in the field.

7.1 Data Requirement Analysis of Industry and Service Industry

As shown in Table 7, applications in different fields have varying requirements for the core capabilities of EAI, and thus their most prioritized data needs also differ. This section defines the industrial field as covering four sectors: manufacturing, mining, utilities (including electricity, heat, gas, and water production and supply), and construction. Robots in these four fields are referred to as industrial robots. Generally, manufacturing robots are distinguished from robots in the other three subfields, which are collectively referred to as special robots.

Manufacturing robots have specific capability requirements to meet the demands of modern production processes. They need to possess autonomous learning and adaptability, enabling them to automatically adjust their operational processes in response to changes in tasks. In precision manufacturing sectors, such as electronics and semiconductor production, high-precision motion control capabilities are essential. Additionally, manufacturing robots must be able to quickly switch between production tasks to minimize changeover time.

Special robots, on the other hand, face unique challenges that require different capabilities. They need to have high environmental adaptability to operate in extreme conditions, such as high temperatures, high pressures, and toxic or hazardous environments. These robots must also be capable of executing complex tasks with higher uncertainty, such as disaster rescue, power inspection, and space exploration. Furthermore, special robots must prioritize safety and reliability to ensure stable operation in hazardous environments [216].

As listed in the Characteristics in Table 7, the industrial field is characterized by the fact that centuries of development have standardized the scenarios, tools, objects, and operations as much as possible. The remaining challenge is how to use EAI to generalize to the parts that are not standardized. Therefore, the application of EAI data engineering should focus on **goal-driven producing the corresponding prioritized data according to the**

Table 7: Goal-Driven Data Requirement Analysis of Industry and Service Industry

Field	Subfield		Characteristics	Core Capability Requirements	Most Needed Data
Industry	Manufacturing		Standardized scenarios, tools, objects, and operation	Production line adaptability, high-precision motion control	Domain knowledge data, manipulation data, and asset data
	Special field	Mining	Non-standardized scenarios, tools, objects, and operation	High environmental adaptability, safety and reliability	Domain knowledge data, manipulation and locomotion data
		Utilities	Non-standardized scenarios, but standardized tools, objects, and operation	High safety and reliability, customization, autonomous decision-making	Domain knowledge data, manipulation data, locomotion data, and asset data
		Construction	Standardized scenarios and tools, but non-standardized objects and operation	High environmental adaptability, safety and reliability	Domain knowledge data, manipulation data, locomotion data, and asset data
Service Industry	-		Highly diverse and dynamic, with different sub-sectors having varying requirements for robot capabilities	Strong perception,precise motion control, autonomous decision-making, emotion recognition, continuous learning and adaptation	Common sense data, manipulation and locomotion data, decision-making data, human-robot interaction and empirical data

characteristics and core capability requirements of different industrial fields, and improve the corresponding production efficiency.

The application fields of service industry include wholesale and retail trade, transportation, storage and postal services, accommodation and catering services, education, health and social work, culture, sports and entertainment, and healthcare. In the service industry, which is highly diverse and dynamic, with different sub-sectors having varying requirements for robot capabilities, the demands for the dynamics performance of robots and the intelligence of EAI models are both very high. As a result, the need for all types of data is almost equally important. To this end, **data production in the service industry needs to be closely synchronized with the development of robot bodies and EAI models**, and progress in tandem within the loop of "production-training-testing-improvement-reproduction".

7.2 Selection of EAI Data Production Methods

The choice of EAI data production methods is pivotal for the efficiency and effectiveness of data acquisition, as illustrated by the comparative analysis in Table 8. Each method presents unique advantages and challenges across various parameters such as equipment cost, labor cost, computational cost, application scope, productivity, data availability, and diversity. Understanding these attributes is fundamental for goal-driven selecting the most appropriate method for specific EAI applications.

Teleoperation-based data collection methods, offer medium to high data availability and diversity, making them suitable for applications requiring a broad spectrum of data. However, these methods also come with high labor costs and varying productivity levels, which might not be feasible for all projects. Especially, the application of teleoperation may be limited by the scenario, for example, the master-slave structure of isomorphic teleoperation may prevent robots from entering narrow spaces.

Indirect teaching-based methods offer a balance between high productivity and medium to high application scope. These methods are particularly useful for tasks that benefit from direct human demonstration. The medium to high data availability and diversity ensure that these methods can support a wide array of learning algorithms and models. However, this demonstration data cannot fully guarantee data availability, as the motion trajectories of the demonstration data may exceed the robot's workspace, causing data availability deterioration.

Simulation data generation stands out with its high application scope and productivity, thanks to its low computational and labor costs. This method is particularly advantageous for scenarios where real-world data collection is impractical or too costly. However, as discussed in Section 6, the sim2real gap remains a formidable barrier that is currently difficult to overcome, which diminishes data availability.

In conclusion, the choice of EAI data production methods should be guided by a thorough evaluation of the specific needs of the application, including the required data quality, diversity, and the available resources. **The core concept in selecting data production methods is to achieve the highest productivity while covering the broadest range of target scenarios**.

7.3 Optimization Directions of EAI Data Engineering

The field of EAI is still in rapid development. The current issues in EAI data engineering, as outlined in Section 2.3, revolve around the bottlenecks in EAI development. Any other technical issues can be traced back to these three

Table 8: Comparison of characteristics of different EAI data production methods when producing the same amount of data

Technical type	**Real-World Data Collection**									**Simulation Data Generation**
	Teleoperation-based					**Teaching-Based**				
	Pose-based			Visual-based	Optical-Inertial	Direct Teaching	Indirect Teaching			
	Handle-based	Wearable	Isomorphic				End-Effector	Motion Capture	Human Video	
Equipment cost	Medium	Medium	High	Medium	High	-	Medium	Medium	-	Low
Labor cost	High	High	High	High	High	High	Medium	Medium	-	Low
Computational cost	-	-	-	Low	Low	-	Low	Low	High	High
Application scope	Medium	Medium	Low	Medium	Medium	Low	High	High	High	High
Productivity	Low	Medium	Medium	Medium	Medium	Low	High	High	-	High
Data Availability	High	High	High	High	High	High	Medium	Medium	Low	Medium
Data diversity	Medium	Medium	Medium	Medium	Medium	Low	High	High	High	High

bottlenecks. However, breaking through these three bottlenecks requires systematic efforts, including the optimization of systems, standards, hardware, software, and applications.

- *Systematic EAI Data Production System*: This system should have high compatibility in both software and hardware to accommodate various data collection devices, robots, and simulation platforms. It needs to offer efficient data compression and transmission capabilities, along with a rich set of tools for automated dataset construction, storage, labeling, and management. Such a platform would streamline data production processes, improve data quality and usability, and support diverse EAI applications and research.
- *Scalable EAI Dataset Standards*: This involves defining unified data formats, annotation methods, and quality evaluation criteria. Standardized datasets enable better data sharing and exchange across different research institutions and enterprises, promote collaboration and integration in the EAI field, and facilitate comparative analysis and evaluation of different algorithms and models.
- *Integration of Data Production and Model Training and Testing*: It focus on one-to-many teleoperation data collection, enabling simultaneous data collection from multiple robots or environments through a single operation. Integrating online learning into automated data production allows the system to continuously learn and adapt. Additionally, optimizing large-scale data parallel transmission to improve bandwidth utilization can enhance the efficiency and scalability of real-world data collection systems.
- *EAI Data Production with Real-Sim Collaborative*: It focuses on narrowing the sim2real gap and employs bidirectional data enhancement to improve model generalization. Additionally, it supports interactive learning for seamless knowledge acquisition across both simulated and real-world scenarios. The approach also capitalizes on a data flywheel effect, where continuous cycles of data collection and model refinement boost EAI performance, ensuring robustness and adaptability in diverse environments.
- *Goal-Driven Specialized and Socialized EAI Data Production*: This concept involves the concurrent advancement of data production methods tailored for both specific, technical applications and broader, social interaction scenarios. This dual-track approach acknowledges the diverse requirements of EAI, where specialized data production caters to the unique challenges of technical tasks that demand high precision and reliability. Conversely, socialized data production addresses the complexities of human-robot interaction. By fostering the parallel growth of these two domains, the EAI field can more effectively develop datasets that are not only technically robust but also socially adept.
- *Open EAI Data Trading Platform*: It is envisioned as an inclusive and transparent ecosystem that facilitates the exchange of EAI data. This platform would encompass a comprehensive set of features including, but not limited to, a secure marketplace for data providers and consumers, standardized data formats for ease of integration, advanced data evaluation metrics to ensure quality, and mechanisms for data provenance and licensing to promote trust and compliance. Additionally, it would offer tools for data anonymization to protect privacy, algorithms for data matching to enhance discoverability, and protocols for secure transactions to safeguard intellectual property. The ultimate goal of such a platform would be to democratize access to high-quality EAI data, fostering innovation and collaboration across the field while adhering to ethical standards and legal regulations.

8 Conclusion

EAI Data Engineering plays a central role in advancing intelligent robotic systems by connecting theory with real-world deployment. This survey has underscored the importance of high-quality EAI data in shaping embodied agents' capabilities, covering system design, data standardization, collection, generation, and application. The challenges inherent in EAI Data Engineering, while formidable, present opportunities for innovation and improvement. The development of systematic EAI data production platforms, scalable standards, integration of data and model, real-sim collaborative, goal-driven production, and open data trading platforms are identified as key directions for future progress. These advancements are poised to enhance the efficiency, quality, and applicability of EAI data, thereby propelling the field forward.

To realize the full potential of EAI, the industry must shift from opportunistic data use to systematic, standardized, scalable and goal-driven EAI data engineering. As we look to the future, the continuous evolution of EAI Data Engineering will be instrumental in unlocking the full potential of embodied intelligence. By fostering collaboration across academia, industry, and government, and by leveraging cutting-edge technologies, we can surmount existing barriers and create more intelligent, adaptive, and human-centric robotic systems. The ultimate goal is to develop EAI systems that can seamlessly integrate into our daily lives, enhancing productivity, safety, and quality of life. This review serves as a testament to the dynamic and promising nature of EAI Data Engineering, inviting researchers and practitioners to contribute to this transformative journey.

9 Acknowledgments

This research was funded by Guangdong Basic and Applied Basic Research Foundation (2024A1515012026, 2021A1515110700, 2023A1515012570), National Natural Science Foundation of China (62106155), Longang District Shenzhen's "Ten Action Plan" (LGKCSDPT 2024002, 2024003, 2024004), and Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS).

References

[1] Wang Fan and Shaoshan Liu. Putting the smarts into robot bodies. *Communications of the ACM*, 68(3):6–8, 2025.

[2] Shaoshan Liu. The value of data in embodied artificial intelligence. *Communications of the ACM*, 2024.

[3] Shaoshan Liu. Shaping the outlook for the autonomy economy. *Communications of the ACM*, 67(6):10–12, 2024.

[4] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[5] Tim Pearce, Tabish Rashid, Dave Bignell, Raluca Georgescu, Sam Devlin, and Katja Hofmann. Scaling laws for pre-training agents and world models. *arXiv preprint arXiv:2411.04434*, 2024.

[6] Alan M Turing. *Computing machinery and intelligence*. Springer, 2009.

[7] Shuang Wu, Bo Yu, Shaoshan Liu, and Yuhao Zhu. Autonomy 2.0: The quest for economies of scale. *Communications of the ACM*, 68(4):28–32, 2025.

[8] Shaoshan Liu and Shuang Wu. A brief history of embodied artificial intelligence, and its future outlook. *Communications of the ACM*, 2024.

[9] Jinda Cui and Jeff Trinkle. Toward next-generation learned robot manipulation. *Science robotics*, 6(54):eabd9461, 2021.

[10] MD Moniruzzaman, Alexander Rassau, Douglas Chai, and Syed Mohammed Shamsul Islam. Teleoperation methods and enhancement techniques for mobile robots: A comprehensive survey. *Robotics and Autonomous Systems*, 150:103973, 2022.

[11] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

[12] Kourosh Darvish, Luigi Penco, Joao Ramos, Rafael Cisneros, Jerry Pratt, Eiichi Yoshida, Serena Ivaldi, and Daniele Pucci. Teleoperation of humanoid robots: A survey. *IEEE Transactions on Robotics*, 39(3):1706–1727, 2023.

[13] Luke Antonyshyn, Jefferson Silveira, Sidney Givigi, and Joshua Marshall. Multiple mobile robot task and motion planning: A survey. *ACM Computing Surveys*, 55(10):1–35, 2023.

[14] Robert McCarthy, Daniel CH Tan, Dominik Schmidt, Fernando Acero, Nathan Herr, Yilun Du, Thomas G Thuruthel, and Zhibin Li. Towards generalist robot learning from internet video: A survey. *arXiv preprint arXiv:2404.19664*, 2024.

[15] Kento Kawaharazuka, Tatsuya Matsushima, Andrew Gambardella, Jiaxian Guo, Chris Paxton, and Andy Zeng. Real-world robot applications of foundation models: A review. *Advanced Robotics*, 38(18):1232–1254, 2024.

[16] Maryam Zare, Parham M Kebria, Abbas Khosravi, and Saeid Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, 2024.

[17] Xuan Xiao, Jiahang Liu, Zhipeng Wang, Yanmin Zhou, Yong Qi, Shuo Jiang, Bin He, and Qian Cheng. Robot learning in the era of foundation models: A survey. *Neurocomputing*, page 129963, 2025.

[18] Lik Hang Kenny Wong, Xueyang Kang, Kaixin Bai, and Jianwei Zhang. A survey of robotic navigation and manipulation with physics simulators in the era of embodied ai. *arXiv preprint arXiv:2505.01458*, 2025.

[19] Jiwen Yu, Yiran Qin, Haoxuan Che, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Hao Chen, and Xihui Liu. A survey of interactive generative video. *arXiv preprint arXiv:2504.21853*, 2025.

[20] Kun Zhang, Peng Yun, Jun Cen, Junhao Cai, Didi Zhu, Hangjie Yuan, Chao Zhao, Tao Feng, Michael Yu Wang, Qifeng Chen, et al. Generative artificial intelligence in robotic manipulation: A survey. *arXiv preprint arXiv:2503.03464*, 2025.

[21] Kaiyuan Chen, Letian Fu, David Huang, Yanxiang Zhang, Lawrence Yunliang Chen, Huang Huang, Kush Hari, Ashwin Balakrishna, Ted Xiao, Pannag R Sanketi, et al. Robo-dm: Data management for large robot datasets. *arXiv preprint arXiv:2505.15558*, 2025.

[22] Xuan Xia, Bo Yu, Jialin Jiao, Xinmin Ding, Xing He, Haoran Tong, Yufei Lin, Tongyi Shen, Ning Ding, and Shaoshan Liu. Airspeed: An open-source universal data production platform for embodied artificial intelligence.

[23] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, and Acorn Pooley et al. Open x-embodiment: Robotic learning datasets and rt-x models. *ArXiv*, abs/2310.08864, 2023.

[24] Zhiqiang Wang, Hao Zheng, Yunshuang Nie, Wenjun Xu, Qingwei Wang, Hua Ye, Zhe Li, Kaidong Zhang, Xuewen Cheng, Wanxi Dong, Chang Cai, Liang Lin, Feng Zheng, and Xiaodan Liang. All robots in one: A new standard and unified dataset for versatile, general-purpose embodied agents. *ArXiv*, abs/2408.10899, 2024.

[25] Zipeng Fu, Tony Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *ArXiv*, abs/2401.02117, 2024.

[26] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *ArXiv*, abs/2402.10329, 2024.

[27] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Fei-Fei Li, and Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *ArXiv*, abs/2403.07788, 2024.

[28] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj S. Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. In *Conference on Robot Learning*, 2023.

[29] Tianyuan Dai, Josiah Wong, Yunfan Jiang, Chen Wang, Cem Gokmen, Ruohan Zhang, Jiajun Wu, and Fei-Fei Li. Automated creation of digital cousins for robust policy learning. *ArXiv*, abs/2410.07408, 2024.

[30] Leonardo Barcellona, Andrii Zadaianchuk, Davide Allegro, Samuele Papa, Stefano Ghidoni, and Efstratios Gavves. Dream to manipulate: Compositional world models empowering robot imitation learning with imagination. *arXiv preprint arXiv:2412.14957*, 2024.

[31] Joey Hejna, Suvir Mirchandani, Ashwin Balakrishna, Annie Xie, Ayzaan Wahid, Jonathan Tompson, Pannag Sanketi, Dhruv Shah, Coline Devin, and Dorsa Sadigh. Robot data curation with mutual information estimators. *arXiv preprint arXiv:2502.08623*, 2025.

[32] Haozhuo Li, Yuchen Cui, and Dorsa Sadigh. How to train your robots? the impact of demonstration modality on imitation learning. *arXiv preprint arXiv:2503.07017*, 2025.

[33] Shivin Dass, Alaa Khaddaj, Logan Engstrom, Aleksander Madry, Andrew Ilyas, and Roberto Martín-Martín. Datamil: Selecting data for robot imitation learning with datamodels. *arXiv preprint arXiv:2505.09603*, 2025.

[34] Hengkai Tan, Xuezhou Xu, Chengyang Ying, Xinyi Mao, Songming Liu, Xingxing Zhang, Hang Su, and Jun Zhu. Manibox: Enhancing spatial grasping generalization via scalable simulation data generation. *arXiv preprint arXiv:2411.01850*, 2024.

[35] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. *arXiv preprint arXiv:2410.18647*, 2024.

[36] Sebastian Sartor and Neil Thompson. Neural scaling laws for embodied ai. *arXiv preprint arXiv:2405.14005*, 2024.

[37] Daniel Kahneman. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.

[38] Alec Radford. Improving language understanding by generative pre-training. 2018.

[39] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[40] Daniel M. Ziegler, Nisan Stiennon, Jeff Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593, 2019.

[41] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290, 2023.

[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023.

[43] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, and Chan et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 645–652, 2024.

[44] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *ArXiv*, abs/2307.05973, 2023.

[45] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18061–18070, 2024.

[46] Damien Ernst and Arthur Louette. Introduction to reinforcement learning. 2024.

[47] Maryam Zare, Parham Mohsenzadeh Kebria, Abbas Khosravi, and Saeid Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, 54:7173–7186, 2023.

[48] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. *ArXiv*, abs/2202.02005, 2022.

[49] Haoshu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39:3929–3945, 2022.

[50] Bingyi Kang, Xiao Ma, Chao Du, Tianyu Pang, and Shuicheng Yan. Efficient diffusion policies for offline reinforcement learning. *ArXiv*, abs/2305.20081, 2023.

[51] Homer Rich Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Maximilian Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Ho Vuong, Andre Wang He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, 2023.

[52] Yuxing Long, Wenzhe Cai, Hongchen Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. In *Conference on Robot Learning*, 2024.

[53] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *ArXiv*, abs/2410.07864, 2024.

[54] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Krzysztof Choromanski, and Tianli Ding et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *ArXiv*, abs/2307.15818, 2023.

[55] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[56] Shaoshan Liu. Establishing standards for embodied ai. *Communications of the ACM*, 2024.

[57] Fang Hao-Shu, Wang Chenxi, Gou Minghao, and Lu Cewu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020.

[58] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *ArXiv*, abs/1910.11215, 2019.

[59] Clemens Eppner, Arsalan Mousavian, and Dieter Fox. Acronym: A large-scale grasp dataset based on simulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6222–6227. IEEE, 2021.

[60] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *ArXiv*, abs/2109.13396, 2021.

[61] L. Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Yong-Jin Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14789–14798, 2022.

[62] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, and Chelsea Finn et al. Rt-1: Robotics transformer for real-world control at scale. *ArXiv*, abs/2212.06817, 2022.

[63] An Dinh Vuong, Minh Nhat Vu, Hieu Le, Baoru Huang, Huynh Thi Thanh Binh, Thieu Vo, Andreas Kugi, and Anh Nguyen. Grasp-anything: Large-scale grasp dataset from foundation models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14030–14037. IEEE, 2024.

[64] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023.

[65] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Z. Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yuan Yao, Xiao Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. Maniskill2: A unified benchmark for generalizable manipulation skills. *ArXiv*, abs/2302.04659, 2023.

[66] Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai, Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20426–20438, 2023.

[67] Chen Bao, Helin Xu, Yuzhe Qin, and Xiaolong Wang. Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21190–21200, 2023.

[68] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 653–660. IEEE, 2024.

[69] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, and Lawrence Yunliang Chen et al. Droid: A large-scale in-the-wild robot manipulation dataset. *ArXiv*, abs/2403.12945, 2024.

[70] Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, and Yinuo Zhao et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *ArXiv*, abs/2412.13877, 2024.

[71] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xu Huang, Shu Jiang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[72] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[73] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4 4:236–252, 2016.

[74] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2019.

[75] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[76] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5142–5151, 2022.

[77] Jiageng Mao, Siheng Zhao, Siqi Song, Tianheng Shi, Junjie Ye, Mingtong Zhang, Haoran Geng, Jitendra Malik, Vitor Campanholo Guizilini, and Yue Wang. Learning from massive human videos for universal humanoid pose control. *ArXiv*, abs/2412.14172, 2024.

[78] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2018.

[79] Cătălina Cangea, Eugene Belilovsky, Pietro Lio', and Aaron C. Courville. Videonavqa: Bridging the gap between visual and embodied question answering. *ArXiv*, abs/1908.04950, 2019.

[80] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *ArXiv*, abs/2210.07474, 2022.

[81] Sinan Tan, Mengmeng Ge, Di Guo, Huaping Liu, and Fuchun Sun. Knowledge-based embodied question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11948–11960, 2023.

[82] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Rui-Lan Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking multimodal large language models for human-level planning. 2023.

[83] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, and Silwal et al. Openeqa: Embodied question answering in the era of foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16488–16498, 2024.

[84] Allen Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. Explore until confident: Efficient exploration for embodied question answering. *ArXiv*, abs/2403.15941, 2024.

[85] Vishnu Sashank Dorbala, Prasoon Goyal, Robinson Piramuthu, Michael Johnston, Reza Ghanadhan, and Dinesh Manocha. Is the house ready for sleeptime? generating and evaluating situational queries for embodied question answering. 2024.

[86] Emily Jin, Zhuoyi Huang, Jan-Philipp Fränken, Weiyu Liu, Hannah Cha, Erik Brockbank, Sarah A. Wu, Ruohan Zhang, Jiajun Wu, and Tobias Gerstenberg. Marple: A benchmark for long-horizon inference. *ArXiv*, abs/2410.01926, 2024.

[87] Min Zhang, Jianye Hao, Xian Fu, Peilong Han, Hao Zhang, Lei Shi, Hongyao Tang, and Yan Zheng. Mfe-etp: A comprehensive evaluation benchmark for multi-modal foundation models on embodied task planning. *ArXiv*, abs/2407.05047, 2024.

[88] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

[89] Chen Gao, Baining Zhao, Weichen Zhang, Jinzhu Mao, Jun Zhang, Zhiheng Zheng, Fanhang Man, Jianjie Fang, Zile Zhou, Jinqiang Cui, Xinlei Chen, and Yong Li. Embodiedcity: A benchmark platform for embodied agent in real-world city environment. *ArXiv*, abs/2410.09604, 2024.

[90] Jihan Yang, Runyu Ding, Ellis L Brown, Xiaojuan Qi, and Saining Xie. V-irl: Grounding virtual intelligence in real life. In *European Conference on Computer Vision*, 2024.

[91] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *ArXiv*, abs/2412.14171, 2024.

[92] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2020.

[93] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, 2020.

[94] Luca Weihs, Matt Deitke, Aniruddha Kembhavi, and Roozbeh Mottaghi. Visual room rearrangement. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5918–5927, 2021.

[95] Kiana Ehsani, Winson Han, Alvaro Herrasti, Eli VanderBilt, Luca Weihs, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Manipulathor: A framework for visual object manipulation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4495–4504, 2021.

[96] Yue Fan, Winson X. Chen, Tongzhou Jiang, Chun ni Zhou, Yi Zhang, and Xin Eric Wang. Aerial vision-and-dialog navigation. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

[97] Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3461–3475, 2023.

[98] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *ArXiv*, abs/2206.06994, 2022.

[99] Sriram Yenamandra, A. Ramachandran, Karmesh Yadav, Austin S. Wang, Mukul Khanna, Théophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander Clegg, John Turner, Zsolt Kira, Manolis Savva, Angel X. Chang, Devendra Singh Chaplot, Dhruv Batra, Roozbeh Mottaghi, Yonatan Bisk, and Chris Paxton. Homerobot: Open-vocabulary mobile manipulation. In *Conference on Robot Learning*, 2023.

[100] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, and Sanjana Srivastava et al. Behavior-1k: A human-centered, embodied ai benchmark with 1, 000 everyday activities and realistic simulation. *ArXiv*, abs/2403.09227, 2024.

[101] Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peifeng Wang, Yaning Zhang, and Qi Wu. Aerialvln: Vision-and-language navigation for uavs. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15338–15348, 2023.

[102] Wayne Wu, Honglin He, Jack He, Yiran Wang, Chenda Duan, Zhizheng Liu, Quanyi Li, and Bolei Zhou. Metaurban: An embodied ai simulation platform for urban micromobility. In *International Conference on Learning Representations*, 2024.

[103] Hanqing Wang, Jiahe Chen, Wensi Huang, Qingwei Ben, Tai Wang, Boyu Mi, Tao Huang, Siheng Zhao, Yilun Chen, Sizhe Yang, Peizhou Cao, Wenye Yu, Zichao Ye, Jialun Li, Junfeng Long, Zirui Wang, Huiling Wang, Ying Zhao, Zhongying Tu, Yu Qiao, Dahua Lin, and Jiangmiao Pang. Grutopia: Dream general robots in a city at scale. *ArXiv*, abs/2407.10943, 2024.

[104] Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daich Azuma, Yutaka Matsuo, and Nakamasa Inoue. Citynav: Language-goal aerial navigation dataset with geographic information. *ArXiv*, abs/2406.14240, 2024.

[105] Lingfeng Zhang, Yuening Wang, Hongjian Gu, Atia Hamidizadeh, Zhanguang Zhang, Yuecheng Liu, Yutong Wang, David Gamaliel Arcos Bravo, Junyi Dong, Shunbo Zhou, Tongtong Cao, Yuzheng Zhuang, Yingxue Zhang, and Jianye Hao. Et-plan-bench: Embodied task-level planning benchmark towards spatial-temporal cognition with foundation models. *ArXiv*, abs/2410.14682, 2024.

[106] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *ArXiv*, abs/2502.09560, 2025.

[107] Toru Lin, Yu Zhang, Qiyang Li, Haozhi Qi, Brent Yi, Sergey Levine, and Jitendra Malik. Learning visuotactile skills with two multifingered hands. *ArXiv*, abs/2404.16823, 2024.

[108] Hongjie Fang, Haoshu Fang, Yiming Wang, Jieji Ren, Jing Chen, Ruo Zhang, Weiming Wang, and Cewu Lu. Airexo: Low-cost exoskeletons for learning whole-arm manipulation in the wild. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15031–15038, 2023.

[109] Shiqi Yang, Minghuan Liu, Yuzhe Qin, Runyu Ding, Jialong Li, Xuxin Cheng, Ruihan Yang, Sha Yi, and Xiaolong Wang. Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation. *ArXiv*, abs/2408.11805, 2024.

[110] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and P. Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12156–12163, 2023.

[111] Qingwei Ben, Feiyu Jia, Jia Zeng, Junting Dong, Dahua Lin, and Jiangmiao Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *ArXiv*, abs/2502.13013, 2025.

[112] Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan D. Ratliff, and Dieter Fox. Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9164–9170, 2019.

[113] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. *ArXiv*, abs/2307.04577, 2023.

[114] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Conference on Robot Learning*, 2024.

[115] Runyu Ding, Yuzhe Qin, Jiyue Zhu, Chengzhe Jia, Shiqi Yang, Ruihan Yang, Xiaojuan Qi, and Xiaolong Wang. Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning. *ArXiv*, abs/2407.03162, 2024.

[116] Chenhao Lu, Xuxin Cheng, Jialong Li, Shiqi Yang, Mazeyu Ji, Chengjing Yuan, Ge Yang, Sha Yi, and Xiaolong Wang. Mobile-television: Predictive motion priors for humanoid whole-body control. *ArXiv*, abs/2412.07773, 2024.

[117] Zhaxizhuoma, Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Pengan Chen, Pingrui Zhang, Haoming Song, Delin Qu, Dong Wang, Zhigang Wang, Nieqing Cao, Yan Ding, Bin Zhao, and Xuelong Li. Fastumi: A scalable and hardware-independent universal manipulation interface with dataset. 2024.

[118] Wu Zhen and Lian Luan. Physical world to virtual reality–motion capture technology in dance creation. In *Journal of Physics: Conference Series*, volume 1828, page 012097, 2021.

[119] Moo Jin Kim, Jiajun Wu, and Chelsea Finn. Giving robots a hand: Learning generalizable manipulation with eye-in-hand human video demonstrations. *ArXiv*, abs/2307.05959, 2023.

[120] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *ArXiv*, abs/2410.06158, 2024.

[121] Chen Wang, Linxi (Jim) Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. In *Conference on Robot Learning*, 2023.

[122] Irmak Güzey, Yinlong Dai, Georgy Savva, Raunaq M. Bhirangi, and Lerrel Pinto. Bridging the human to robot dexterity gap through object-oriented rewards. *ArXiv*, abs/2410.23289, 2024.

[123] Haiqin Cui, Yifu Yuan, Yan Zheng, and Jianye Hao. Aharobot: A low-cost open-source bimanual mobile manipulator for embodied ai. *arXiv preprint arXiv:2503.10070*, 2025.

[124] Ri-Zhao Qiu, Yuchen Song, Xuanbin Peng, Sai Aneesh Suryadevara, Ge Yang, Minghuan Liu, Mazeyu Ji, Chengzhe Jia, Ruihan Yang, Xueyan Zou, et al. Wildlma: Long horizon loco-manipulation in the wild. *arXiv preprint arXiv:2411.15131*, 2024.

[125] Junda Huang, Kai Chen, Jianshu Zhou, Xingyu Lin, Pieter Abbeel, Qi Dou, and Yunhui Liu. Dih-tele: Dexterous in-hand teleoperation framework for learning multiobjects manipulation with tactile sensing. *IEEE/ASME Transactions on Mechatronics*, 2025.

[126] Kenneth Shaw, Yulong Li, Jiahui Yang, Mohan Kumar Srirama, Ray Liu, Haoyu Xiong, Russell Mendonca, and Deepak Pathak. Bimanual dexterity for complex tasks. *arXiv preprint arXiv:2411.13677*, 2024.

[127] Han Zhang, Songbo Hu, Zhecheng Yuan, and Huazhe Xu. Doglove: Dexterous manipulation with a low-cost open-source haptic force feedback glove. *arXiv preprint arXiv:2502.07730*, 2025.

[128] Jiabao Gan, Shihui Guo, Zhijun Li, and Xiangren Shi. Telemotion: A realtime humanoid teleoperation system with motion capture. In *International Conference on Extended Reality*, pages 31–45. Springer, 2024.

[129] Yunfan Jiang, Ruohan Zhang, Josiah Wong, Chen Wang, Yanjie Ze, Hang Yin, Cem Gokmen, Shuran Song, Jiajun Wu, and Li Fei-Fei. Behavior robot suite: Streamlining real-world whole-body manipulation for everyday household activities. *arXiv preprint arXiv:2503.05652*, 2025.

[130] Yaru Niu, Yunzhe Zhang, Mingyang Yu, Changyi Lin, Chenhao Li, Yikai Wang, Yuxiang Yang, Wenhao Yu, Tingnan Zhang, Zhenzhen Li, et al. Human2locoman: Learning versatile quadrupedal manipulation with human pretraining. In *ICRA 2025 Workshop on Foundation Models and Neuro-Symbolic AI for Robotics*.

[131] Aadhithya Iyer, Zhuoran Peng, Yinlong Dai, Irmak Guzey, Siddhant Haldar, Soumith Chintala, and Lerrel Pinto. Open teach: A versatile teleoperation system for robotic manipulation. *arXiv preprint arXiv:2403.07870*, 2024.

[132] Shivin Dass, Wensi Ai, Yuqian Jiang, Samik Singh, Jiaheng Hu, Ruohan Zhang, Peter Stone, Ben Abbatematteo, and Roberto Martín-Martín. Telemoma: A modular and versatile teleoperation system for mobile manipulation. *arXiv preprint arXiv:2403.07869*, 2024.

[133] Younghyo Park, Jagdeep Singh Bhatia, Lars Lien Ankile, and Pulkit Agrawal. Dexhub and dart: Towards internet scale robot data collection. *ArXiv*, abs/2411.02214, 2024.

[134] Daniel Honerkamp, Harsh Mahesheka, Jan Ole von Hartz, Tim Welschehold, and Abhinav Valada. Whole-body teleoperation for mobile manipulation at zero added cost. *IEEE Robotics and Automation Letters*, 2025.

[135] Jianshu Zhou, Boyuan Liang, Junda Huang, Ian Zhang, Pieter Abbeel, and Masayoshi Tomizuka. Global-local interface for on-demand teleoperation. *arXiv preprint arXiv:2502.09960*, 2025.

[136] Shivin Dass, Karl Pertsch, Hejia Zhang, Youngwoon Lee, Joseph J Lim, and Stefanos Nikolaidis. Pato: Policy assisted teleoperation for scalable robot data collection. *arXiv preprint arXiv:2212.04708*, 2022.

[137] Kosei Tanada, Yuka Iwanaga, Masayoshi Tsuchinaga, Yuji Nakamura, Takemitsu Mori, Remi Sakai, and Takashi Yamamoto. Sketch-moma: Teleoperation for mobile manipulator via interpretation of hand-drawn sketches. *arXiv preprint arXiv:2412.19153*, 2024.

[138] Shengcheng Luo, Quanquan Peng, Jun Lv, Kaiwen Hong, Katherine Rose Driggs-Campbell, Cewu Lu, and Yong-Lu Li. Human-agent joint learning for efficient robot manipulation skill acquisition. *arXiv preprint arXiv:2407.00299*, 2024.

[139] Yunfan Jiang, Chen Wang, Ruohan Zhang, Jiajun Wu, and Li Fei-Fei. Transic: Sim-to-real policy transfer by learning from online correction. *arXiv preprint arXiv:2405.10315*, 2024.

[140] Michael Hagenow and Julie A Shah. Realm: Real-time estimates of assistance for learned models in human-robot interaction. *IEEE Robotics and Automation Letters*, 2025.

[141] Huihan Liu, Shivin Dass, Roberto Martín-Martín, and Yuke Zhu. Model-based runtime monitoring with interactive imitation learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4154–4161. IEEE, 2024.

[142] Zhiyuan Xu, Yinuo Zhao, Kun Wu, Ning Liu, Junjie Ji, Zhengping Che, Chi Harold Liu, and Jian Tang. Hacts: a human-as-copilot teleoperation system for robot learning. *arXiv preprint arXiv:2503.24070*, 2025.

[143] Cheng Pan, Hung Hon Cheng, and Josie Hughes. Online imitation learning for manipulation via decaying relative correction through teleoperation. *arXiv preprint arXiv:2503.15368*, 2025.

[144] Philipp Wu, Yide Shentu, Qiayuan Liao, Ding Jin, Menglong Guo, Koushil Sreenath, Xingyu Lin, and Pieter Abbeel. Robocopilot: Human-in-the-loop interactive imitation learning for robot manipulation. *arXiv preprint arXiv:2503.07771*, 2025.

[145] Ajay Mandlekar, Caelan Reed Garrett, Danfei Xu, and Dieter Fox. Human-in-the-loop task and motion planning for imitation learning. In *Conference on Robot Learning*, pages 3030–3060. PMLR, 2023.

[146] Yansong Wu, Xiao Chen, Yu Chen, Hamid Sadeghian, Fan Wu, Zhenshan Bing, Sami Haddadin, Alexander König, and Alois Knoll. Sharedassembly: A data collection approach via shared tele-assembly. *arXiv preprint arXiv:2503.12287*, 2025.

[147] Siyuan Huang, Yue Liao, Siyuan Feng, Shu Jiang, Si Liu, Hongsheng Li, Maoqing Yao, and Guanghui Ren. Adversarial data collection: Human-collaborative perturbations for efficient and robust robotic imitation learning. *arXiv preprint arXiv:2503.11646*, 2025.

[148] Hongjie Fang, Chenxi Wang, Yiming Wang, Jingjing Chen, Shangning Xia, Jun Lv, Zihao He, Xiyan Yi, Yunhan Guo, Xinyu Zhan, et al. Airexo-2: Scaling up generalizable robotic imitation learning with low-cost exoskeletons. *arXiv preprint arXiv:2503.03081*, 2025.

[149] Rui Zhong, Chuang Cheng, Junpeng Xu, Yantong Wei, Ce Guo, Daoxun Zhang, Wei Dai, and Huimin Lu. Nuexo: A wearable exoskeleton covering all upper limb rom for outdoor data collection and teleoperation of humanoid robots. *arXiv preprint arXiv:2503.10554*, 2025.

[150] Xintao Chao, Shilong Mu, Yushan Liu, Shoujie Li, Chuqiao Lyu, Xiao-Ping Zhang, and Wenbo Ding. Exo-viha: A cross-platform exoskeleton system with visual and haptic feedback for efficient dexterous skill learning. *arXiv preprint arXiv:2503.01543*, 2025.

[151] Michael Hagenow, Dimosthenis Kontogiorgos, Yanwei Wang, and Julie Shah. Versatile demonstration interface: Toward more flexible robot demonstration collection. *arXiv preprint arXiv:2410.19141*, 2024.

[152] Wenhai Liu, Junbo Wang, Yiming Wang, Weiming Wang, and Cewu Lu. Forcemimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation. *arXiv preprint arXiv:2410.07554*, 2024.

[153] Kelin Yu, Yunhai Han, Qixian Wang, Vaibhav Saxena, Danfei Xu, and Ye Zhao. Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation. *arXiv preprint arXiv:2310.16917*, 2023.

[154] Fangchen Liu, Chuanyu Li, Yihua Qin, Ankit Shaw, Jing Xu, Pieter Abbeel, and Rui Chen. Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface. *arXiv preprint arXiv:2504.06156*, 2025.

[155] Samuel Clarke, Suzannah Wistreich, Yanjie Ze, and Jiajun Wu. X-capture: An open-source portable device for multi-sensory learning. *arXiv preprint arXiv:2504.02318*, 2025.

[156] Tony Tao, Mohan Kumar Srirama, Jason Jingzhou Liu, Kenneth Shaw, and Deepak Pathak. Dexwild: Dexterous human interactions for in-the-wild robot policies. *arXiv preprint arXiv:2505.07813*, 2025.

[157] Chengyi Xing, Hao Li, Yi-Lin Wei, Tian-Ao Ren, Tianyu Tu, Yuhao Lin, Elizabeth Schumann, Wei-Shi Zheng, and Mark R Cutkosky. Taccap: A wearable fbg-based tactile sensor for seamless human-to-robot skill transfer. *arXiv preprint arXiv:2503.01789*, 2025.

[158] Kei Takahashi, Hikaru Sasaki, and Takamitsu Matsubara. Feasibility-aware imitation learning from observations through a hand-mounted demonstration interface. *arXiv preprint arXiv:2503.09018*, 2025.

[159] Huy Ha, Yihuai Gao, Zipeng Fu, Jie Tan, and Shuran Song. Umi on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers. *arXiv preprint arXiv:2407.10353*, 2024.

[160] Mingyo Seo, H Andy Park, Shenli Yuan, Yuke Zhu, and Luis Sentis. Legato: Cross-embodiment imitation using a grasping tool. *IEEE Robotics and Automation Letters*, 2025.

[161] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. *arXiv preprint arXiv:2410.24221*, 2024.

[162] Haonan Chen, Cheng Zhu, Yunzhu Li, and Katherine Rose Driggs-Campbell. Tool-as-interface: Learning robot tool use from human play through imitation learning. In *ICRA 2025 Workshop: Beyond Pick and Place*, 2025.

[163] Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. *arXiv preprint arXiv:2505.03729*, 2025.

[164] Nataliya Nechyporenko, Ryan Hoque, Christopher Webb, Mouli Sivapurapu, and Jian Zhang. Armada: Augmented reality for robot manipulation and robot-free data acquisition. *arXiv preprint arXiv:2412.10631*, 2024.

[165] Sirui Chen, Chen Wang, Kaden Nguyen, Li Fei-Fei, and C Karen Liu. Arcap: Collecting high-quality human demonstrations for robot learning with augmented reality feedback. *arXiv preprint arXiv:2410.08464*, 2024.

[166] Jun Wang, Chun-Cheng Chang, Jiafei Duan, Dieter Fox, and Ranjay Krishna. Eve: Enabling anyone to train robots using augmented reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–13, 2024.

[167] Christoph Bartneck, Marius Soucy, Kevin Fleuret, and Eduardo B Sandoval. The robot engine—making the unity 3d game engine work for hri. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 431–437. IEEE, 2015.

[168] Xingjian Li, Jeremy Park, Chris Reberg-Horton, Steven Mirsky, Edgar Lobaton, and Lirong Xiang. Photorealistic arm robot simulation for 3d plant reconstruction and automatic annotation using unreal engine 5. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5488, 2024.

[169] Ekim Otan Karaoğlu, Dilek Tükel, and Bagus Arthaya. Vr based visualization of robotic workcells using cryengine. In *2019 International Conference on Mechatronics, Robotics and Systems Engineering (MoRSE)*, pages 118–121. IEEE, 2019.

[170] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[171] Anderson Maciel, Tansel Halic, Zhonghua Lu, Luciana P Nedel, and Suvranu De. Using the physx engine for physics-based virtual surgery with force feedback. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 5(3):341–353, 2009.

[172] Nathan Koenig and Andrew Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ international conference on intelligent robots and systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 3, pages 2149–2154. Ieee, 2004.

[173] Christopher Mower, Theodoros Stouraitis, Joao Moura, Christian Rauch, Lei Yan, Nazanin Zamani Behabadi, Michael Gienger, Tom Vercauteren, Christos Bergeles, and Sethu Vijayakumar. Ros-pybullet interface: A framework for reliable contact simulation and human-robot interaction. In *Conference on robot learning*, pages 1411–1423. PMLR, 2023.

[174] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

[175] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020.

[176] Xian Zhou, Yiling Qiao, Zhenjia Xu, TH Wang, Z Chen, J Zheng, Z Xiong, Y Wang, M Zhang, P Ma, et al. Genesis: A generative and universal physics engine for robotics and beyond. *arXiv preprint arXiv:2401.01454*, 2024.

[177] Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. *ArXiv*, abs/2410.24185, 2024.

[178] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf. *Communications of the ACM*, 65:99 – 106, 2020.

[179] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42:1 – 14, 2023.

[180] Xinhai Li, Jialin Li, Ziheng Zhang, Rui Zhang, Fan Jia, Tiancai Wang, Haoqiang Fan, Kuo-Kun Tseng, and Ruiping Wang. Robogsim: A real2sim2real robotic gaussian splatting simulator. *ArXiv*, abs/2411.11839, 2024.

[181] Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Min Xu, Lunkai Lin, Zhiqiang Xie, Mingyu Ding, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins. 2025.

[182] Yan Ding, Xiaohan Zhang, S. Amiri, Nieqing Cao, Hao Yang, Andy Kaminski, Chad Esselink, and Shiqi Zhang. Integrating action knowledge and llms for task planning and situation handling in open worlds. *Autonomous Robots*, 47:981 – 997, 2023.

[183] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, Weiyu Liu, Percy Liang, Fei-Fei Li, Jiayuan Mao, and Jiajun Wu. Embodied agent interface: Benchmarking llms for embodied decision making. *ArXiv*, abs/2410.07166, 2024.

[184] Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. Gensim: Generating robotic simulation tasks via large language models. *ArXiv*, abs/2310.01361, 2023.

[185] Yao Mu, Junting Chen, Qinglong Zhang, Shoufa Chen, Qiaojun Yu, Chongjian Ge, Runjian Chen, Zhixuan Liang, Mengkang Hu, Chaofan Tao, Peize Sun, Haibao Yu, Chao Yang, Wenqi Shao, Wenhai Wang, Jifeng Dai, Yu Qiao, Mingyu Ding, and Ping Luo. Robocodex: Multimodal code generation for robotic behavior synthesis. *ArXiv*, abs/2402.16117, 2024.

[186] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *ArXiv*, abs/2409.16283, 2024.

[187] Lirui Wang, Kevin Zhao, Chaoqi Liu, and Xinlei Chen. Learning real-world action-video dynamics with heterogeneous masked autoregression. *ArXiv*, abs/2502.04296, 2025.

[188] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal M. P. Behbahani, Stephanie Chan, Nicolas Manfred Otto Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktaschel. Genie: Generative interactive environments. *ArXiv*, abs/2402.15391, 2024.

[189] Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025.

[190] Caelan Reed Garrett, Ajay Mandlekar, Bowen Wen, and Dieter Fox. Skillgen: Automated demonstration generation for efficient skill learning and deployment. In *2nd CoRL Workshop on Learning Effective Abstractions for Planning*, 2024.

[191] Caelan Garrett, Ajay Mandlekar, Bowen Wen, and Dieter Fox. Skillmimicgen: Automated demonstration generation for efficient skill learning and deployment. *arXiv preprint arXiv:2410.18907*, 2024.

[192] Zhengrong Xue, Shuying Deng, Zhenyang Chen, Yixuan Wang, Zhecheng Yuan, and Huazhe Xu. Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning. *arXiv preprint arXiv:2502.16932*, 2025.

[193] Martin Spitznagel, Jan Vaillant, and Janis Keuper. Physicsgen: Can generative models learn from images to predict complex physical relations? *arXiv preprint arXiv:2503.05333*, 2025.

[194] Runyi Yu, Yinhuai Wang, Qihan Zhao, Hok Wai Tsui, Jingbo Wang, Ping Tan, and Qifeng Chen. Skillmimic-v2: Learning robust and generalizable interaction skills from sparse and noisy demonstrations. *arXiv preprint arXiv:2505.02094*, 2025.

[195] Yiran Qin, Li Kang, Xiufeng Song, Zhenfei Yin, Xiaohong Liu, Xihui Liu, Ruimao Zhang, and Lei Bai. Robofactory: Exploring embodied agent collaboration with compositional constraints. *arXiv preprint arXiv:2503.16408*, 2025.

[196] Wensheng Wang and Ning Tan. Hybridgen: Vlm-guided hybrid planning for scalable data generation of imitation learning. *arXiv preprint arXiv:2503.13171*, 2025.

[197] Sizhe Yang, Wenye Yu, Jia Zeng, Jun Lv, Kerui Ren, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Novel demonstration generation with gaussian splatting enables robust one-shot manipulation. *arXiv preprint arXiv:2504.13175*, 2025.

[198] Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, Lan Xu, Wei Yang, Jiayuan Gu, and Jingyi Yu Cast. Component-aligned 3d scene reconstruction from an rgb image. *arXiv preprint arXiv:2502.12894*, 8, 2025.

[199] Jane Wu, Georgios Pavlakos, Georgia Gkioxari, and Jitendra Malik. Reconstructing hand-held objects in 3d. *arXiv preprint arXiv:2404.06507*, 2024.

[200] Junzhe Zhu, Yuanchen Ju, Junyi Zhang, Muhan Wang, Zhecheng Yuan, Kaizhe Hu, and Huazhe Xu. Densematcher: Learning 3d semantic correspondence for category-level manipulation from a single demo. *arXiv preprint arXiv:2412.05268*, 2024.

[201] Justin Yu, Letian Fu, Huang Huang, Karim El-Refai, Rares Andrei Ambrus, Richard Cheng, Muhammad Zubair Irshad, and Ken Goldberg. Real2render2real: Scaling robot data without dynamics simulation or robot hardware. *arXiv preprint arXiv:2505.09601*, 2025.

[202] Jianhua Sun and Cewu Lu. Digital gene: Learning about the physical world through analytic concepts. *arXiv e-prints*, pages arXiv–2504, 2025.

[203] Alisson Azzolini, Hannah Brandon, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Francesco Ferroni, Rama Govindaraju, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

[204] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. *arXiv preprint arXiv:2503.22020*, 2025.

[205] Liming Zheng, Feng Yan, Fanfan Liu, Chengjian Feng, Yufeng Zhong, Yiyang Huang, and Lin Ma. Dataplatter: Boosting robotic manipulation generalization with minimal costly data. *arXiv preprint arXiv:2503.19516*, 2025.

[206] Nick Heppert, Minh Quang Nguyen, and Abhinav Valada. Real2gen: Imitation learning from a single human demonstration with generative foundational models. In *ICRA 2025 Workshop on Foundation Models and Neuro-Symbolic AI for Robotics*, 2025.

[207] Tyler Ga Wei Lum, Olivia Y Lee, C Karen Liu, and Jeannette Bohg. Crossing the human-robot embodiment gap with sim-to-real rl using one human demonstration. *arXiv preprint arXiv:2504.12609*, 2025.

[208] Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. *arXiv preprint arXiv:2406.16862*, 2024.

[209] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos. *arXiv preprint arXiv:2503.00779*, 2025.

[210] Xinpeng Liu, Junxuan Liang, Zili Lin, Haowen Hou, Yong-Lu Li, and Cewu Lu. Imdy: Human inverse dynamics from imitated observations. *arXiv preprint arXiv:2410.17610*, 2024.

[211] Zifan Wang, Ziqing Chen, Junyu Chen, Jilong Wang, Yuxin Yang, Yunze Liu, Xueyi Liu, He Wang, and Li Yi. Mobileh2r: Learning generalizable human to mobile robot handover exclusively from scalable and diverse synthetic data. *arXiv preprint arXiv:2501.04595*, 2025.

[212] Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *European Conference on Computer Vision*, pages 329–345. Springer, 2024.

[213] Yurui Chen, Junge Zhang, Ziyang Xie, Wenye Li, Feihu Zhang, Jiachen Lu, and Li Zhang. S-nerf++: Autonomous driving simulation via neural reconstruction and generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[214] Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control. *arXiv preprint arXiv:2412.03568*, 2024.

[215] Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, et al. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025.

[216] Zishen Wan, Yiming Gan, Bo Yu, Shaoshan Liu, Arijit Raychowdhury, and Yuhao Zhu. The vulnerability-adaptive protection paradigm. *Communications of the ACM*, 67(9):66–77, 2024.