Machine Intelligence Research
www.mi-research.net

22(4), August 2025, 588-626
DOI: 10.1007/s11633-025-1542-8

A Survey of Embodied Learning for Object-centric Robotic Manipulation

Ying Zheng[1†] Lei Yao[1†] Yuejiao Su[1] Yi Zhang[1] Yi Wang[1]
Sicheng Zhao[2] Yiyi Zhang[3] Lap-Pui Chau[1]

[1] Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China

[2] Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China

[3] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong 999077, China

Abstract: Embodied learning for object-centric robotic manipulation is a rapidly developing and challenging area in embodied AI. It is crucial for advancing next-generation intelligent robots and has garnered significant interest recently. Unlike data-driven machine learning methods, embodied learning focuses on robot learning through physical interaction with the environment and perceptual feedback, making it especially suitable for robotic manipulation. In this paper, we provide a comprehensive survey of the latest advancements in this field and categorize the existing work into three main branches: 1) Embodied perceptual learning, which aims to predict object pose and affordance through various data representations; 2) Embodied policy learning, which focuses on generating optimal robotic decisions using methods such as reinforcement learning and imitation learning; 3) Embodied task-oriented learning, designed to optimize the robot's performance based on the characteristics of different tasks in object grasping and manipulation. In addition, we offer an overview and discussion of public datasets, evaluation metrics, representative applications, current challenges, and potential future research directions. A project associated with this survey has been established at https://github.com/RayYoh/OCRM_survey.

Keywords: Embodied learning, robotic manipulation, pose estimation, affordance learning, policy learning.

Citation: Y. Zheng, L. Yao, Y. Su, Y. Zhang, Y. Wang, S. Zhao, Y. Zhang, L. P. Chau. A survey of embodied learning for object-centric robotic manipulation. *Machine Intelligence Research*, vol.22, no.4, pp.588–626, 2025. http://doi.org/10.1007/s11633-025-1542-8

1 Introduction

During the previous decade, remarkable progress has been made in machine learning research centered on the field of deep learning, revolutionizing various applications such as computer vision[1–3] and natural language processing[4, 5]. Different from traditional machine learning methods that solely rely on pre-constructed datasets for pattern recognition and prediction, embodied learning, a cornerstone of embodied AI, aims to empower intelligent agents the capability of environment perception and decision making. Embodied learning allows robots to learn through physical interaction with the environment and feedback from sensors, enabling them to adapt to new situations. It emphasizes the importance of the robot's embodiment and knowledge acquisition through physical interactions and practical experiences[6, 7]. The data sources encompass a broad spectrum, including sensory inputs, bodily actions, and immediate environmental feedback. This learning mechanism is highly dynamic, continuously refining behaviors and manipulation strategies through real-time interactions and feedback loops. Embodied learning is essential in robotics as it equips robots with enhanced environmental adaptability, enabling them to handle changing conditions and undertake more intricate and complex tasks.

While a plethora of embodied learning methods have been proposed, this survey primarily focuses on the task of object-centric robotic manipulation. The inputs for this task are data collected from sensors, and the outputs are operational strategies and control signals for the robot to perform manipulation tasks. The objective is to enable the robot to efficiently and autonomously perform various object-centric manipulation tasks while enhancing its generality and flexibility across different environments and tasks. This task is highly challenging due to the diversity of objects and manipulation tasks, the complexity and uncertainty of the environment, and challenges such as noise, occlusion, and real-time constraints in real-world applications.

Fig. 1(a) illustrates a typical robotic manipulation system. It features a robotic arm equipped with sensors like

Review
Special Issue on Embodied Intelligence
Manuscript received on September 19, 2024; accepted on January 14, 2025; published online on June 20, 2025
Recommended by Associate Editor Wei He
Colored figures are available in the online version at https://link.springer.com/journal/11633
†These authors contributed equally to this work

cameras and end-effectors such as grippers, enabling it to manipulate a wide range of objects. The system's intelligence revolves around three key aspects, corresponding to the three types of embodied learning methods depicted in Fig. 1(b). 1) Advanced perception capabilities, which involve utilizing data captured by different sensors to understand the target object and external environment; 2) Precise policy generation, which entails analyzing the perceived information to make optimal decisions; 3) Task-orientation, which ensures the system can adapt to specific tasks by optimizing the execution process for maximum effectiveness.

In recent years, extensive research has been conducted around those above three key aspects, particularly with the flourishing of large language models (LLMs)[8], neural radiance fields (NeRFs)[9], diffusion models[10], and 3D Gaussian splatting[11], leading to a host of innovative solutions. However, there is a notable absence of a comprehensive survey that encapsulates the latest research in this rapidly evolving field. This motivates us to write this survey to systematically recap the cutting-edge advancements and summarize the encountered challenges, along with the prospective research directions.

1.1 Comparison with recent surveys

Over the past few years, many survey articles have emerged on embodied AI and robot learning, addressing various domains like navigation[12], planning[13], grasping[14], and manipulation[15]. In Table 1, we summarize and categorize some recent relevant surveys in this field and compare them with our work. To explicitly compare these survey papers, we utilized two key criteria: timeliness and systematicness. Timeliness assesses whether the reviewed papers are up-to-date and cover the latest research. Specifically, we consider review papers that include work from the past three years, i.e., those published in 2022 and later, as timely. Systematicness, on the other hand, applies specifically to surveys related to robotic manipulation (RM). Other types of surveys are not assessed for their systematic nature and are marked with a "–" in Table 1. If a survey only addresses certain aspects of RM like datasets[30] and imitation learning[32], it is deemed lacking in systematicness.

From Table 1, it can be observed that the number of surveys related to RM is the highest, indicating the significance of research in the RM field. In addition, although RG can be considered as the foundation of RM, it is often studied as an independent field due to its involvement in many subtasks and specific problems. These existing surveys primarily focus on specific aspects of robot manipulation, such as deep learning-based grasp synthesis[20] and manipulation policy learning[15, 21]. Additionally, some latest surveys delve into recent advancements in vision-language-action models[24] and large language model-based autonomous agents[25]. However, our survey is unique that it provides a comprehensive overview of embodied learning methods for object-centric robotic manipulation, encompassing embodied perceptual learning, policy learning, and task-oriented learning.

The most closely related work to ours is the survey paper by Cong et al. (2021)[35], which primarily reviews research on 3D vision-based robotic manipulation up to 2021. In contrast, our work is not limited to specific input modalities; we systematically summarize and categorize representation methods based on 2D images, 3D-

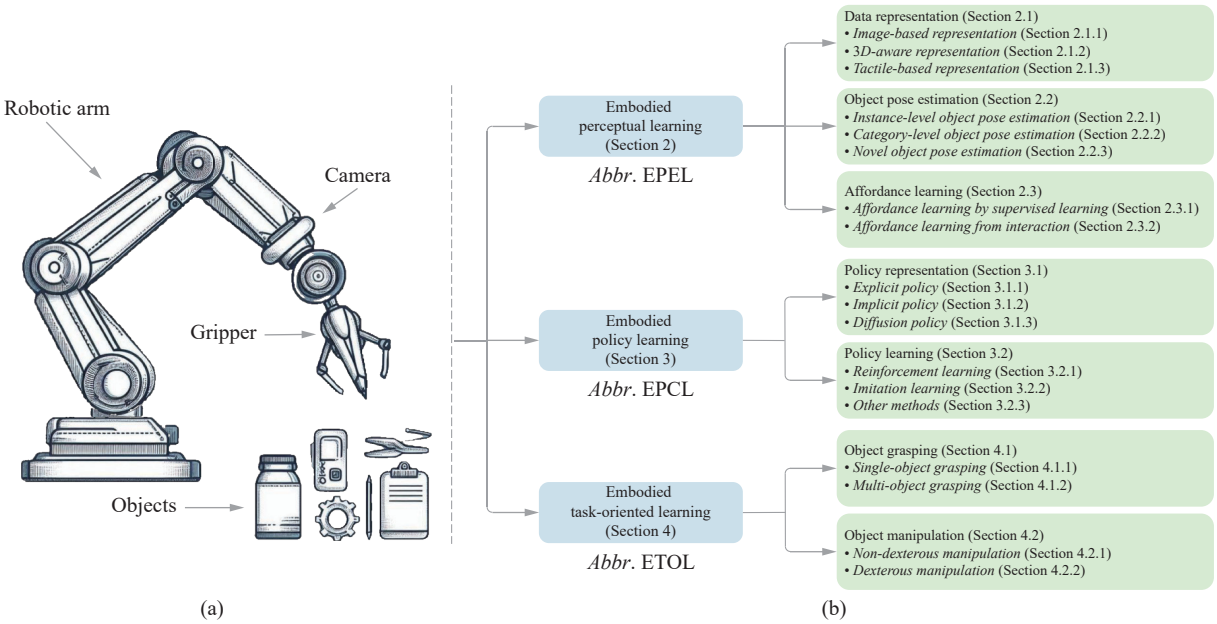

Fig. 1 An illustration of robotic manipulation system (left) and the typology of embodied learning methods for object-centric robotic manipulation (right).

Table 1 Summary of recent surveys related to embodied AI and robot learning. RM: Robotic manipulation; RG: Robotic grasping; RL: Reinforcement learning.

Authors	Reference	Year	Category	Timeliness	Systematicness	Short description
Jin et al.	[16]	2018	Control	×	–	Robot manipulator control using neural networks
Zhu et al.	[17]	2021	Navigation	×	–	Deep learning for embodied visual navigation
Duan et al.	[18]	2022	Simulator	✓	–	Simulators for embodied AI
Francis et al.	[19]	2022	Planning	✓	–	Embodied vision-language planning
Gervet et al.	[12]	2023	Navigation	✓	–	Real-world empirical study for robot navigation
Newbury et al.	[20]	2023	Grasp synthesis	✓	–	Deep learning approaches to grasp synthesis
Guo et al.	[13]	2023	Planning	✓	–	Task and motion planning for robotics
Xiao et al.	[22]	2023	Robot learning	✓	–	Foundation models for robot learning
Zare et al.	[21]	2024	Imitation learning	✓	–	Imitation learning
Chen et al.	[23]	2024	Policy learning	✓	–	Generative models for offline policy learning
Ma et al.	[24]	2024	Action	✓	–	Vision-language-action models for embodied AI
Xu et al.	[25]	2024	Planning and control	✓	–	Foundation models for robot planning and control
Kleeberger et al.	[26]	2020	RG	×	–	Machine learning for vision-based RG
Du et al.	[14]	2021	RG	×	–	Vision-based RG
Zhang et al.	[27]	2022	RG	✓	–	Traditional and recent methods for RG
Xie et al.	[28]	2023	RG	✓	–	Learning-based RG
Tian et al.	[29]	2023	RG	✓	–	RG for unknown objects
Huang et al.	[30]	2016	RM	×	×	Datasets of RM
Yamanobe et al.	[31]	2017	RM	×	×	Affordance in RM
Fang et al.	[32]	2019	RM	×	×	Imitation learning for RM
Billard and Kragic	[33]	2019	RM	×	×	Trends and challenges in RM
Kroemer et al.	[34]	2021	RM	×	×	Machine learning for RM
Cong et al.	[35]	2021	RM	×	×	3D vision-based RM
Cui and Trinkle	[36]	2021	RM	×	×	Adaptability of learned RM
Zhu et al.	[37]	2022	RM	×	×	RM of deformable objects
Mohammed et al.	[38]	2022	RM	✓	×	RL-based RM in cluttered environments
Suomalainen et al.	[39]	2022	RM	✓	×	RM in contact
Han et al.	[15]	2023	RM	✓	×	RL for RM
Weinberg et al.	[40]	2024	RM	✓	×	Learning approaches for in-hand RM
Ours		2024	RM	✓	✓	Embodied learning for object-centric RM

aware techniques, and tactile sensing. Moreover, we provide a comprehensive introduction to critical aspects of robotic manipulation, such as policy and task-oriented learning. Notably, our survey covers a wide range of the latest research achievements mainly published after 2021, offering a more cutting-edge and comprehensive perspective. Therefore, our work stands out as the only survey in the RM field that combines both timeliness and systematicness. We hope this survey will serve as a worthwhile reference for researchers and practitioners in the field of embodied learning for object-centric robotic manipulation.

1.2 Text organization

This paper presents a comprehensive survey of embodied learning methods for object-centric robotic manipulation, encompassing three main domains and seven sub-directions. The three domains are embodied perceptual learning (Section 2), embodied policy learning (Section 3), and embodied task-oriented learning (Section 4). The seven sub-directions include data representation (Section 2.1), object pose estimation (Section 2.2), affordance learning (Section 2.3), policy representation (Section 3.1), policy learning (Section 3.2), object grasping (Section 4.1), and object manipulation (Section 4.2). We also extensively cover the commonly used datasets and evaluation metrics (Section 5), along with several representative applications (Section 6). Additionally, we delve into the primary challenges and provide insights into potential future research directions (Section 7).

2 Embodied perceptual learning

To perform object-centric robotic manipulation, the robot must first learn to perceive the target object and its surrounding environment, which involves data representation, object pose estimation, and affordance learning. In this section, we will provide a comprehensive overview of these works.

2.1 Data representation

In object-centric robotic manipulation, robots utilize various sensors to perceive their surroundings. These encompass visual sensors like RGB and depth cameras, which capture color images and depth maps; LiDARs, which create high-resolution 3D point clouds through distance measurements; and tactile sensors, which detect forces during grasping and pressure distribution on contact surfaces. The data collected by these sensors come in different forms, leading to various representations tailored to specific solutions. Next, we will introduce three primary types of data representation approaches: image-based representation, 3D-aware representation, and tactile-based representation.

2.1.1 Image-based representation

This line of work primarily focuses on constructing effective representations solely from RGB images, thereby providing a robust foundation for subsequent tasks in robotic manipulation, such as object pose estimation. Depending on the number of input images and variations in network architecture, existing methods can be categorized into four types: single-image single-branch (SISB)[41], single-image multi-branch (SIMB)[42], multi-image single-branch (MISB)[43], and multi-image multi-branch (MIMB)[44], as illustrated in Fig. 2.

1) As depicted in Fig. 2(a), the SISB methods take a single RGB image as input, with a streamlined network architecture featuring a single main pathway. It conventionally employs deep learning models like convolutional neural networks (CNNs) to extract deep features from the source image, which are then fed into a pose estimator to generate the essential object pose information for robotic manipulation. SISB incorporates a typical approach to deep feature representation within an end-to-end network framework. Despite its speed and simplicity, the SISB's limitation in expressing objects' 3D geometric information may result in subsequently coarser object pose estimation.

2) To overcome the limitations of SISB, SIMB methods introduce extra network branches alongside the main pathway, as shown in Fig. 2(b). These additional branches are designed to capture richer auxiliary information. For instance, MonoGraspNet[42] combines a keypoint network and a normal network to produce keypoint heatmaps and normal maps, respectively. It provides a more robust intermediate representation, improving pose estimation accuracy. However, this method relies heavily on the prediction accuracy of the additional branches. Due to the inherent limitations of making predictions based on a single image, errors are inevitably introduced in the generated intermediate representations. These errors can amplify the adverse effect on subsequent processing steps and increase uncertainty in robotic manipulation tasks.

3) Owing to the lack of scale information in a single image, accurately estimating the 3D geometric information of objects is quite challenging. Therefore, a lot of research has focused on exploring methods that use multiple images to address this constraint. Among these approaches, the MISB framework has received significant attention. As shown in Fig. 2(c), this framework aims to use multiple images for 3D reconstruction to recover depth information of the scene[45, 46], which in turn facilitates the generation of efficient 3D representations. Specifically, the depth recovery can be achieved through advanced techniques such as NeRFs[9] or Gaussian splatting[11].

4) Unlike MISB, MIMB aims to directly generate

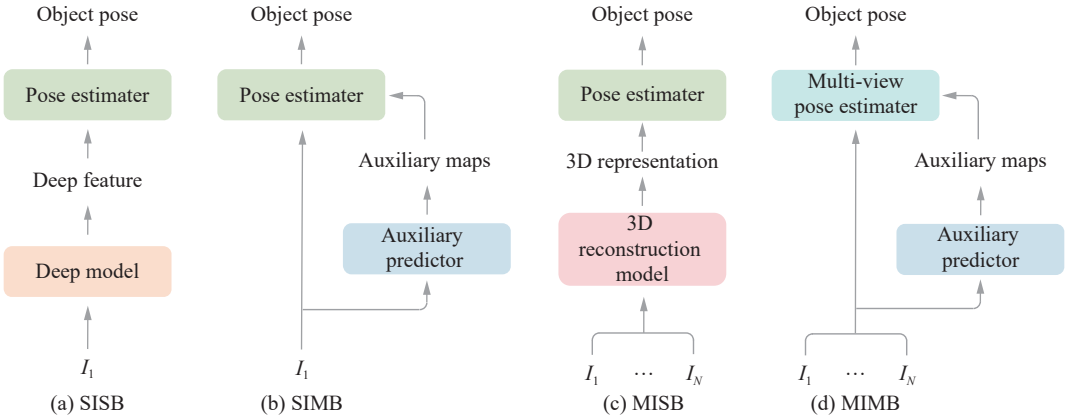

Fig. 2 Conceptual comparison of four image-based representation frameworks. SISB: Single-Image Single-Branch; SIMB: Single-Image Multi-Branch; MISB: Multi-Image Single-Branch; MIMB: Multi-Image Multi-Branch. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

multi-view image representations from images captured by a robot at multiple positions, bypassing the phase of 3D reconstruction. As illustrated in Fig. 2(d), the MIMB methods incorporate additional predictors to acquire extra information, compensating for the lack of 3D information and enhancing the robot's scene perception. For example, RGBManip[44] introduces a multi-view active learning method and utilizes the segmentation maps produced by the segment anything model (SAM) model[47] to provide enhanced representations for the multi-view pose estimator.

2.1.2 3D-aware representation

This section explores 3D-aware representation, which usually takes RGB-D images as input. Existing methods fall into three categories based on the representations they generate: depth-based representation (DR), point cloud-based representation (PR), and transition-based representation (TR), as shown in Fig. 3.

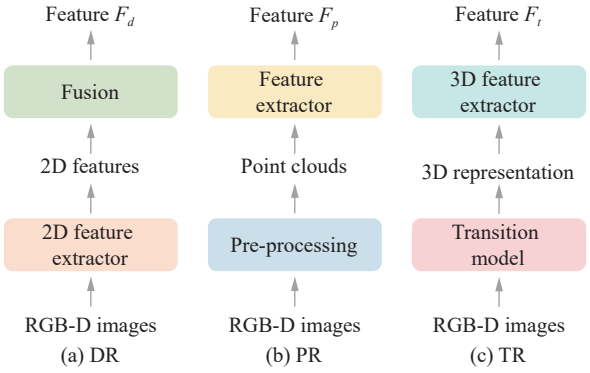

Fig. 3 Conceptual comparison of three 3D-aware representation frameworks. DR: Depth-based representation; PR: Point cloud-based representation; TR: Transition-based representation. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

1) The DR methods usually employ a network to extract 2D features from RGB-D images simultaneously, as illustrated in Fig. 3(a). Some use these extracted features directly for subsequent tasks[48, 49], which typically necessitate posterior refinement. For example, Lenz et al.[49] introduced a two-stage cascade network architecture, where the first network efficiently filters out numerous unlikely grasps generated upon extracted features, and the second network concentrates on evaluating the detections from the first network. Another line of studies[50, 51] utilizes a two-stream network to independently extract 2D features from RGB and depth images. Subsequently, these features are combined or fused to generate the final feature F_d for downstream tasks.

2) Instead of directly extracting features on RGB-D images, PR methods first create point clouds through preprocessing, as depicted in Fig. 3(b). Previous approaches for processing point clouds converted from RGB-D images[52] often involve voxelizing the point clouds and utilizing 3D convolutional neural networks to extract features. However, such approaches are inefficient in terms of memory usage. The introduction of PointNet[53], a network architecture designed explicitly for point clouds, has revolutionized the field. Many methods[54, 55] now prefer to leverage PointNet-like frameworks that enable direct feature extraction from individual points in the point cloud, followed by task-specific modules customized for different objectives.

3) Fig. 3(c) presents the framework of TR works[56] that focus on improving the model's understanding of 3D geometry by translating the input RGB-D data into 3D representations such as occupancy fields, NeRFs, or 3D Gaussians. For example, [57] involves converting RGB-D data into a voxel representation, using a voxel encoder to create a 3D feature volume. This volume is then employed to construct a neural radiance field to model the 3D space and predict robot actions. [58, 59] project RGB-D data into dense point clouds or voxelized point clouds, which are the foundation for placing 3D Gaussians within the scene and enhancing support for robotic manipulation tasks.

2.1.3 Tactile-based representation

Tactile sensing acquires crucial force and positional information, allowing the robot to perceive contact with objects and subtle surface changes sensitively. This information is vital for enhancing the robot's capacity to perform complicated tasks and improving its operational accuracy and adaptability.

The field of tactile sensing technologies is diverse, with examples such as Gelsight[60], DIGIT[61], and AllSight[62]. These sensors can capture various tactile information such as contact positions, normal forces, tangential forces, and torques. The representation methods for this data also vary. One common representation is time sequences obtained through multiple samplings of tactile feedback within a specific time window[63]. These sequences can be converted into feature vectors using neural networks like long short-term memory (LSTM)[64], which simplifies the processing in subsequent models. Another form of representation is the tactile image[65, 66], which presents tactile information visually in an intuitive format similar to a standard RGB image and can be directly processed using CNN for feature extraction. Additionally, tactile data can be integrated with other modalities, such as vision and audio, to create a multimodal representation[67, 68], providing a comprehensive understanding of the environment and objects.

Furthermore, creating high-quality tactile representations often requires extensive training data. However, gathering tactile data is more time-consuming than visual data. To overcome this challenge, researchers have proposed leveraging technologies like NeRF or GANs to generate tactile data[69, 70] or building simulation environments to imitate tactile experiences[71, 72]. With the continuous development of these techniques, we anticipate that tactile-based representations will play an even more

significant role in robotic manipulation.

2.1.4 Discussion

Image-based representation minimizes sensor requirements but is limited by relying solely on RGB image information. 3D-aware representation leverages both image and depth data to provide a more robust representation for learning tasks. Tactile-based representation serves as a supplementary method, further enhancing the robot's perception abilities. Future research should focus on combining these methods to fully exploit their respective strengths.

2.2 Object pose estimation

Grasp detection, an essential component of robotic manipulation, relies on accurate object pose estimation as a crucial step[73]. The precision of pose estimation significantly affects the robot's ability to successfully grasp target objects, emphasizing the need to develop robust and efficient pose estimation algorithms. Based on the type of predicted output, there are two main categories of object pose estimation methods: 2D planar pose estimation[74] and 6D pose estimation in 3D space[75]. The former predicts the object's position in the 2D plane and a 1D rotation angle, primarily employed for manipulating objects within a 2D plane. An example application for this method is product sorting in industrial assembly lines, where robotic arm grippers are typically positioned above the sorting platform and utilize a vertical downward angle to grasp target objects. The latter predicts the object's 6DoF (degrees of freedom), including 3D rotation and 3D translation, which can fully describe the object's position and orientation in 3D space. Compared to 2D planar pose estimation, 6D pose estimation has a broader range of applications, allowing robotic arms to manipulate objects from any angle.

Most existing work focuses on the 6D object pose estimation, which can be divided into three categories: instance-level, category-level, and novel object pose estimation.

2.2.1 Instance-level object pose estimation (ILOPE)

It refers to estimating the pose of a specific instance of an object, such as a particular cup. Existing methods typically require detailed prior knowledge of the object's shape and appearance, which a textured CAD model can furnish. Since these methods conduct training on specific samples of target objects, the trained models are object-specific.

The ILOPE problem can be formulated as (1): Given a set of N_o objects $\boldsymbol{O} = \{\boldsymbol{o}_i \mid i = 1, 2, \cdots, N_o\}$, along with their corresponding 3D models $\boldsymbol{M} = \{\boldsymbol{m}_i \mid i = 1, 2, \cdots, N_o\}$, the objective is to learn a model $\boldsymbol{\Phi}$ to estimate the transformation matrix $\boldsymbol{T}$ for each object instance S that is present in a given RGB or RGB-D image I. This transformation $\boldsymbol{T}$ consists of a 3D rotation $\boldsymbol{R} \in SO(3)$ and a translation component $\boldsymbol{t} \in \mathbb{R}^3$, which can map the target S to the camera coordinate system.

$$\boldsymbol{T} \leftarrow \boldsymbol{\Phi}(I \mid \boldsymbol{O}, \boldsymbol{M}). \tag{1}$$

Significant research has been conducted to estimate the pose of objects at the instance level. Some methods utilize deep neural networks to directly regress the 6D pose of objects, such as PoseCNN[76] and CDPN[77]. However, these methods may still require post-processing optimization[78, 79] to achieve better prediction results, as they are relatively simple. Another class of methods involves learning 2D-3D or 3D-3D correspondences using keypoints[80] and then employing a RANSAC-based PnP (Perspective-n-Point) algorithm[81, 82] to generate pose estimation results. Furthermore, template matching[83] or feature point voting[84] are promising approaches for 6D object pose estimation.

The above methods have the advantage of yielding highly accurate pose estimation results. However, they require training for each instance, which makes them unsuitable for handling large-scale and diverse sets of objects.

2.2.2 Category-level object pose estimation (CLOPE)

It involves estimating the pose of objects belonging to predefined categories, such as cups. Existing methods for this task generally do not rely on training on specific instances of objects. Instead, they perform pose estimation using certain features within or across object classes. These methods do not require a 3D model for each instance, which is particularly beneficial when the exact shape and appearance of the objects are not known in advance.

Formally, the CLOPE problem can be stated as (2): Given a set of N_c object categories $\boldsymbol{C} = \{\boldsymbol{c}_i \mid i = 1, 2, \cdots, N_c\}$ and a set of objects $\boldsymbol{O}$ belonging to different categories, the goal is to learn a model $\boldsymbol{\Phi}$ to estimate the transformation matrix $\boldsymbol{T}$ for each object instance s that appears in the observed RGB or RGB-D image I and belongs to category $\boldsymbol{c}_k$. In this case, the 3D model of each object is not available.

$$\boldsymbol{T} \leftarrow \boldsymbol{\Phi}(I \mid \boldsymbol{O}, \boldsymbol{C}). \tag{2}$$

To estimate object pose at the category level, Wang et al.[85] introduced normalized object coordinate space (NOCS), a coordinate system based on the object category. NOCS encodes the pose and size of the object as a normalized coordinate vector, and then the correspondence between observed pixels and NOCS can be directly inferred with a neural network. Chen and Dou[86] utilized the structured prior of the object category to guide pose adaptation and employed a transformer-based network to model the global structural similarity between the object

instance and the prior. These methods are mainly suitable for the pose estimation of rigid objects[87, 88]. However, they are not effectively generalized for articulated objects due to the complexity of articulated object poses, which involve not only translation and rotation but also various joint movements. To address category-level articulation pose estimation (CAPE), Li et al.[89] expanded upon NOCS and introduced articulation-aware normalized coordinate space hierarchy (ANCSH), a category-level representation method tailored for articulated objects. Additionally, Liu et al.[90] proposed a real-world task setting called CAPER (CAPE-Real), which can handle multiple instances and diverse kinematic structures.

The aforementioned methods all estimate an object's pose under the assumption that the object category is known. They typically train models using datasets of known object categories and then perform pose estimation on new instances of the object. These methods enable generalization within the predefined object categories, but they are not capable of handling new object categories.

2.2.3 Novel object pose estimation (NOPE)

It has emerged as a highly active research area in recent years to estimate the pose of novel objects from previously unseen categories during training. In this case, instance-level 3D models and category-level prior information are unavailable, but we can take reference images of the target object as an aid. This problem can be formalized as (3): Given one or multiple test images I along with several reference images I_r associated with the target object, the objective is to learn a model $\boldsymbol{\Phi}$ to estimate the transformation matrix $\boldsymbol{T}$ within the test images by leveraging the visual information from the reference images.

$$\boldsymbol{T} \leftarrow \boldsymbol{\Phi}(I \mid I_r). \tag{3}$$

In this field, classic methods usually employ image matching[75, 91] or feature matching[92, 93] techniques and subsequently perform pose estimation on new object instances. For example, Liu et al.[91] developed Gen6D, a novel 6D pose estimation method that integrates an object detector, viewpoint selector and pose refiner, enabling the inference of the 6D pose of unseen objects without relying on 3D models. Goodwin et al.[93] proposed a method based on a self-supervised vision transformer and semantic correspondence to achieve zero-shot object pose estimation.

Recently, the research community has been increasingly focused on utilizing large models to enhance the generalization capability of deep models for the NOPE task. Lin et al.[94] introduced the SAM-6D approach, which employs the powerful semantic segmentation capabilities of SAM[47] to generate potential object proposals. Simultaneously, Wen et al.[95] investigated methods to integrate large language models (LLMs) with contrastive learning, significantly improving model generalization by training on large-scale synthetic datasets. The primary advantage of these methods is that they can handle new object categories, thereby enhancing their generalizability and applicability in a broader range of real-world scenarios. However, it should be noted that large models usually require more training data and computational resources, which could be a potential limitation.

2.2.4 Discussion

These three types of pose estimation methods each have specific application scenarios and advantages and disadvantages: ILOPE offers high accuracy but is only suitable for known objects; CLOPE has a wide range of applicability but relatively lower accuracy; NOPE is highly flexible but faces significant challenges in accuracy and robustness.

2.3 Affordance learning

Once the estimated object pose is obtained, the next step involves identifying potential interactive regions of the object as shown in Fig. 4, a process known as affordance learning[97]. As a crucial component of robotic manipulation, affordance learning enables robots to comprehend the object's functionality and potential actions. Based on the data source, affordance learning can be categorized into two types: affordance learning by supervised learning and affordance learning from interaction.

2.3.1 Affordance learning by supervised learning

In order to make robots understand object manipula-

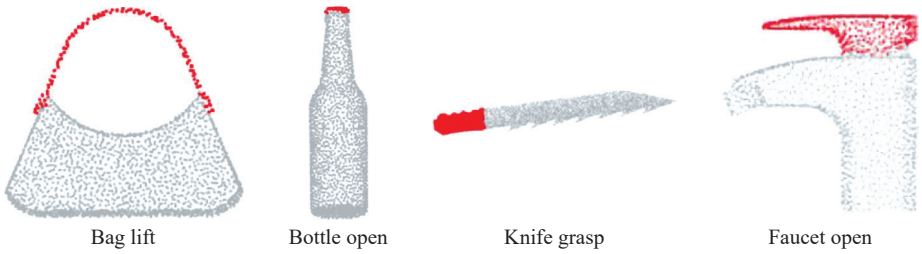

Fig. 4 Visualization of four representative affordance prediction examples from the dataset provided by [96], including bag lift, bottle open, knife grasp, and faucet open. The affordance ground truth labels are highlighted in red. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

tion, various methods have been proposed that utilize static data to learn affordances[98]. For example, AffordanceNet[98] considered human-annotated RGB images from public datasets as input and simultaneously performed object localization and affordance prediction through two distinct branches. Specifically, this method assigned a probable affordance label to each pixel within the predicted object, effectively making it a part of a semantic segmentation task[99]. Additionally, Nagarajan et al.[100] utilized interaction hotspot maps to depict object affordances and trained their model on large-scale human-object interaction video datasets.

While the aforementioned methods have shown promising results on static datasets, they have not explored applying learned affordances to robotic manipulation tasks. To bridge this gap, vision-robotics bridge (VRB)[101] incorporated a trajectory prediction model to extract affordances from egocentric videos and integrated the resulting model into various robot learning frameworks. To improve generalization to unseen objects, Robo-ABC[102] emphasized semantic correspondence and has successfully implemented its model on real-world platforms for grasping novel objects. Additionally, Kuang et al.[103] developed a retrieval-based architecture that lifts 2D affordances to 3D, enabling embodiment-agnostic robotic manipulation. This framework introduced a hierarchical retrieval pipeline to transfer actionable knowledge from out-of-domain data to specific target domains. To overcome the constraints of closed-set affordance learning, OpenAD[104] measured the similarity between language-based affordance labels and point-wise high-dimensional features and extended affordance learning to an open-vocabulary context.

2.3.2 Affordance learning from interaction

Rather than relying on supervised learning from static data, affordance learning from interaction framework seeks to gather training data through simulations. This method allows the system to learn from interactions, providing it with essential prior knowledge for real-world deployment. As a pioneer in this field, Where2Act[105] employed self-supervised interaction for articulated 3D objects, which uses single-frame images or partial point clouds as observations in the SAPIEN[106] simulator. But AdaAfford[107] identified that this paradigm ignores hidden kinematic uncertainties that lead to inaccurate affordances. To address this, AdaAfford proposed a method that involves sampling multiple test-time interactions to facilitate rapid adaptation. Building on similar concepts, DualAfford[108] expanded the interactive learning framework to dual-gripper manipulation to broaden the robot′s manipulation capabilities. Nevertheless, relying on random interactions for data collection makes these methods sample-inefficient. ActAIM[109] tackled this with a clustering-based strategy and a generative model to improve interaction diversity and data quality. Additionally, IDA[110] put forward an information-driven method for affordance discovery to boost interaction efficiency. Where2Explore[111] generalized affordance recognition to novel instances and even various object categories by leveraging local geometries for actionable parts.

It is important to note that all the aforementioned methods operate under the assumption of noiseless visual information, which is often unrealistic. In response, Ling et al.[112] introduced a coarse-to-fine architecture to reduce point cloud noise and improve the affordance learning performance. Beyond focusing solely on single-object affordances, Wu et al.[113] incorporated realistic physical constraints within environments and employed a data-efficient contrastive learning method to acquire environment-aware affordances, even under occlusions. RLAfford[114], in contrast to prior work limited by predefined affordance primitives, integrated reinforcement learning to facilitate end-to-end affordance learning. Specifically, they considered contact maps of interest during the RL process as visual affordances and seamlessly adapted the architecture to various manipulation tasks.

2.3.3 Discussion

Current supervised affordance learning methods are limited by their focus on specific domain data or tasks, while interaction-based approaches are constrained by sample inefficiency. Future research should investigate how to design effective frameworks that can harness the vast potential of internet-scale data and rapidly adapt to specific tasks.

3 Embodied policy learning

Embodied policy learning aims to empower robots with the sophisticated decision-making capabilities required to perform manipulation tasks efficiently. This section will delineate the process of embodied policy learning into two fundamental phases: policy representation and policy learning, elucidating how these techniques enable robots to accomplish predefined objectives. We present the overview of embodied policy learning in Fig. 5 and summarize the key works in Table 2.

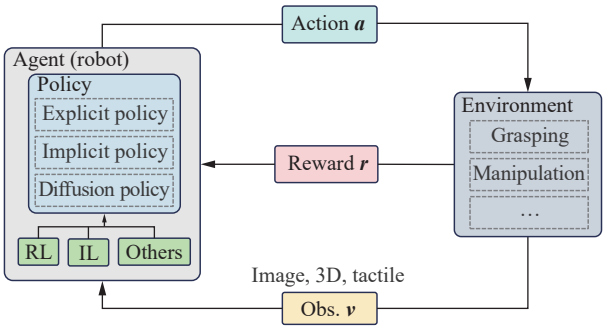

Fig. 5 Illustration of embodied policy learning architecture. The robot agent interacts with the environment to learn a policy that maps observations to actions.

Table 2 Summary of embodied policy learning. RL: Reinforcement learning; IL: Imitation learning.

Task	Type	Subfields & references
Policy representation	Explicit	Deterministic policy[115], Stochastic policy[116]
	Implicit	EBMs[117], Implicit behavioral cloning[118], IDAC[119], EBIL[120]
	Diffusion	Diffusion policy[121], Decision diffuser[122], Diffusion-QL[123], HDP[124], UniDexFPM[125], BESO[126]
		Incorporating language instructions: MDT[127], Lan-o3dp[128]
Policy learning	RL	ViSkill[129], RMA^2[116], SAM-RL[130], Offline RL[131, 132], Demonstration-guided RL[133]
		Rewards function learning: Text2reward[134], EUREKA[135]
	IL	DMPs[136], DAgger[137], SpawnNet[138], ACT[139]
		Scaling up demonstration data: MimicGen[140], Bridge Data[141], Open X-embodiment[142]
		Learning from human videos: Vid2Robot[143], Ag2Manip[144], MPI[145]
		Equivariant models: NDFs[146], L-NDF[147], EDFs[148], EDGI[149], Diffusion-EDFs[150], SE(3)-DiffusionFields[151]
	Others	Combination of RL & IL: UniDexGrasp[152], UniDexGrasp++[153]
		LLM- or VLM-driven: VILA[154], Grounding-RL[155], OpenVLA[156], 3D-VLA[157]

3.1 Policy representation

The role of policy is to model the robot's behavior by taking its observation as input and determining the corresponding action to execute. This process is mathematically represented as $\pi : \mathcal{O} \mapsto \mathcal{A}$, where $\mathcal{O}$ and $\mathcal{A}$ represent the observation space and action space, respectively. Policy representation is critical in embodied policy learning, as it significantly affects the robot's decision-making ability. Depending on the modeling options, policy representation is classified into explicit, implicit, and diffusion policies, regardless of whether the action space is discrete or continuous.

3.1.1 Explicit policy

Explicit policies utilize a parameterized function to map a robot's current observation $\boldsymbol{v} \in \mathcal{O}$ directly to an action $\boldsymbol{a} \in \mathcal{A}$. Typically, explicit policies are parameterized using feed-forward models like neural networks and can be either deterministic[115] or stochastic[116]. A deterministic policy directly predicts an action $\boldsymbol{a} = \pi_\theta(\boldsymbol{v})$ to execute, while a stochastic policy samples actions from an estimated distribution $\boldsymbol{a} \sim \pi_\theta(\cdot \mid \boldsymbol{v})$, where θ indicates parameters of the policy. Stochastic policies enhance an agent's exploration capabilities and provide greater robustness in complex, uncertain environments compared to deterministic policies.

In a discrete action space, policy representation can be transformed into an optimal action selection process from a finite set of actions. The categorical distribution is commonly used to calculate action probabilities, from which actions are sampled based on the estimated results. For instance, Zhang et al.[158] conceptualized the robot assembly manipulation policy as translation, rotation, and insertion primitives, with RL subsequently optimizing the policy. In continuous action spaces, a diagonal Gaussian distribution is often chosen to represent the action distribution, guided by regression losses such as mean squared error (MSE) or RL-based objectives. The policy outputs both the mean $\mu_\theta(\boldsymbol{v})$ and the standard deviation $\sigma_\theta(\boldsymbol{v})$, and actions are sampled from the resulting distribution as follows:

$$\boldsymbol{a} = \mu_\theta(\boldsymbol{v}) + \sigma_\theta(\boldsymbol{v}) \odot \boldsymbol{\xi}. \tag{4}$$

Here, $\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{I})$ represents a vector of Gaussian noise, and $\odot$ signifies the Hadamard product. It should be noted that, in practical applications, the logarithm of the standard deviation $\log \sigma_\theta(\boldsymbol{v})$ is typically used to prevent the standard deviation from taking on negative values.

3.1.2 Implicit policy

Unlike explicit policy models, implicit policies attempt to assign value to each action by leveraging energy-based models (EBMs)[117, 118], which are also recognized as action-value functions[119]. This paradigm learns the policy by optimizing a continuous function to find the action with minimal energy:

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a} \in \mathcal{A}} \mathcal{E}_\theta(\boldsymbol{v}, \boldsymbol{a}) \tag{5}$$

where θ denotes the parameters of the energy function $\mathcal{E}_\theta$. Consequently, the problem of action prediction is effectively reformulated as an optimization problem.

Generally, given a series of expert demonstrations or online-collected trajectories denoted as $\{(\boldsymbol{v}_t, \boldsymbol{a}_t)\}_{t=0}^{T}$, implicit policies are trained by an InfoNCE-style loss[159]. Once trained, stochastic optimization will be applied to identify the optimal action for implicit inference. Energy-based imitation learning (EBIL)[120] incorporated EBMs into the inverse RL architecture, utilizing the estimated expert energy as a surrogate reward. Florence et al.[118] further proposed an implicit behavioral cloning approach grounded in this framework and assessed its performance

across various robotic task domains such as simulated pushing and bi-manual sweeping.

3.1.3 Diffusion policy

Drawing inspiration from denoising diffusion probabilities models (DDPMs)[10], which gradually denoise random inputs to generate data samples, diffusion policies model the policy as a conditional generative model[123]. This approach approximates the action distribution $\pi(\cdot \mid \boldsymbol{v})$ considering the observation $\boldsymbol{v}$ as a condition for producing the corresponding action $\boldsymbol{a}$:

$$\boldsymbol{a}^{k-1} = \alpha(\boldsymbol{a}^k - \beta\epsilon_\theta(\boldsymbol{a}^k, \boldsymbol{v}, k)) + \sigma\mathcal{N}(0, \boldsymbol{I}) \tag{6}$$

where $k = 1, 2, \cdots, K$, representing the denosing iterations, α, β, σ are functions that rely on the noise schedule, ϵ_θ denotes the denosing network with parameters θ, and $\mathcal{N}(0, \boldsymbol{I})$ represents the standard Gaussian noise. The diffusion policy architecture is illustrated in Fig. 6.

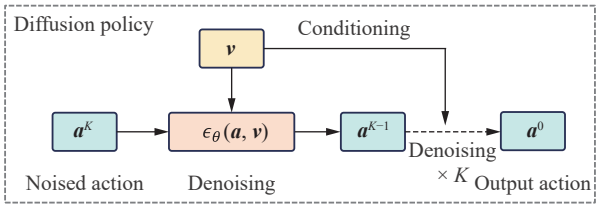

Fig. 6 Illustration of diffusion policy architecture. The policy is modeled as a conditional generative model that gradually denoises random inputs to generate actions. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

As concurrent work, decision diffuser[122] and diffusion-QL[123] have pioneered the integration of diffusion policies into offline RL. These studies revealed that this approach yields highly expressive policy representation that surpasses traditional policy formats. While decision diffuser[122] suggested extending diffusion policies to handle high-dimensional visual observations, its current focus remains on state-based benchmarks. In contrast, Chi et al.[121] proposed a novel diffusion policy tailored for vision-based robotic manipulation tasks. Their experimental findings highlighted the efficacy of diffusion policies in visuomotor policies and their superiority in managing behavioral multimodality in imitation learning. They also incorporated techniques such as receding horizon control and time-series diffusion transformers to adapt the policy for high-dimensional action spaces, resulting in more stable training. Hierarchical diffusion policy (HDP)[124] integrated diffusion policies into a high-level planning agent for multi-task robotic manipulation, whereas UniDexFPM[125] applied diffusion policies to pre-grasp manipulation. By leveraging the conditional generative paradigm, diffusion policies are well-suited for multimodal policy learning. For example, multimodal diffusion transformer (MDT)[127] and Lan-o3dp[128] advanced multimodal policy learning by incorporating language instructions. Differently, BESO[126] facilitated rapid inference in diffusion policies by decoupling the score model learning from the sampling process.

3.1.4 Discussion

Explicit policies are straightforward to implement but struggle with complex tasks, while implicit policies face challenges in training stability and computational costs. Diffusion policies offer a promising alternative to provide a more expressive and robust policy representation, but how to accelerate the sampling process remains to be explored.

3.2 Policy learning

After establishing a suitable policy representation, the next critical task is to train the policy π to execute specific manipulation tasks effectively. Policy learning methods can be broadly categorized into several approaches, including reinforcement learning (RL)[116, 129], imitation learning (IL)[160, 161], and other methods[153, 154] that combine elements of both or introduce entirely different learning paradigms. The choice of policy learning method depends on factors such as the availability of demonstration data, task complexity, and computational resources. Each method has its advantages and challenges, and the field of embodied policy learning continues to evolve with new techniques and insights.

3.2.1 Reinforcement learning

By modeling the policy learning procedure as a Markov decision process (MDP), RL aims to discover the optimal policy π^* that can maximize expected cumulative discounted reward, formulated as

$$\mathcal{J}(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T}\gamma^t r_t(\boldsymbol{v}_t, \boldsymbol{a}_t)\right] \tag{7}$$

where $\tau = \{(\boldsymbol{v}_t, \boldsymbol{a}_t)\}_{t=0}^{T}$ denotes a trajectory, with $\boldsymbol{v}_t$ and $\boldsymbol{a}_t$ representing the observation and action at time step t, respectively. The function r_t corresponds to the reward provided as feedback from the environment after each action is taken. Here, $\gamma \in [0, 1]$ is a discount factor used to balance the importance of immediate and future rewards. Therefore, the objective of RL can be expressed as

$$\pi^* = \arg\max_{\pi} \mathcal{J}(\pi). \tag{8}$$

As a pivotal element for decision-making, RL has been extensively investigated in robotic manipulation. Researchers from OpenAI[162] developed a sim-to-real training pipeline to enable a physical five-fingered robot hand to perform vision-based object reorientation. This pipeline initially trained the policy in simulation using proximal policy optimization (PPO)[163] and then adapted it to physical hardware through domain randomization. It should be highlighted that PPO is a widely used on-

policy RL algorithm in robotic manipulation, valued for its simplicity and effectiveness. For long-horizon surgical robot tasks, ViSkill[129] introduced a novel mechanism named value-informed skill chaining to learn smooth subtask policies. To create generalizable manipulation policies adaptable to various object shapes, Liang et al.[116] presented a two-stage training framework with an extra adapter training phase within PPO, enhancing the policy's robustness across diverse objects. Inspired by model-based RL[164], SAM-RL[130] proposed a sensing-aware architecture that renders images from different viewpoints and refines the learned world model by aligning these generated images with actual raw observations, demonstrating significant real-world performance. Mandlekar et al.[131] explored the impact of various design choices in offline RL[132] and made their dataset publicly available for further research. To overcome exploration challenges in RL, Huang et al.[133] proposed demonstration-guided RL, which assigns high values to expert-preferred actions using non-parametric regression.

Beyond algorithmic enhancements, crafting the reward function remains a significant challenge in RL due to the need for domain-specific knowledge to accurately capture the task objectives. Recently, research has increasingly explored the capabilities of LLMs for reward learning. For instance, Text2reward[134] and EUREKA[135] leveraged the understanding and generation capabilities of LLMs to convert natural language descriptions of goals into dense and interpretable reward codes, which can be iteratively refined by human feedback. This iterative process is crucial as it allows the reward function to evolve in response to new insights or changes in the task requirements. Consequently, this method streamlined the resolution of complex manipulation tasks, reducing reliance on manually crafted reward functions and potentially increasing the effectiveness of the learning process.

3.2.2 Imitation learning

Instead of learning in a trial-and-error manner as RL, the objective of IL is to mimic the expert behavior. Typically, IL encompasses three primary methodologies: behavioral cloning (BC)[160], inverse reinforcement learning (IRL)[165], and generative adversarial imitation learning (GAIL)[161]. BC is a straightforward yet effective approach that learns the policy by minimizing the mean squared error between the expert's action and the policy's predictions through supervised learning. IRL operates in a two-stage loop, initially deducing a reward function from the demonstrations, followed by policy optimization using RL techniques. GAIL is a generative model-based method that relies on adversarial learning to develop a discriminator and an action generator simultaneously, distinguishing between the actions of an expert and those produced by the policy.

Earlier, differentiable nonlinear dynamic systems like dynamic movement primitives (DMPs)[136] were used to acquire skills from demonstrations at the trajectory level. The essence of DMPs lies in incorporating a forcing term, comprised of a set of weighted basis functions, into the system dynamics. These weights are determined through regression analysis of the desired trajectory. Despite using a limited number of parameters, the effectiveness of DMPs is constrained by the choice of basis functions. Instead, DAgger[137] incrementally aggregated current policy interaction data with expert policy demonstrations to augment the training data. SpawnNet[138] incorporated a pre-trained visual model to develop a generalizable policy for diverse manipulation tasks. Kim et al.[166] introduced a self-attention mechanism to filter out irrelevant information, while action chunking with transformers (ACT)[139] directly trained a generative transformer model on action sequences specifically for dual-arm manipulation on real-world collected data.

Given the high cost of collecting human demonstrations, there is a focus on scaling up the demonstration data. MimicGen[140] designed a system that inputs a few expert demonstrations and created an augmented dataset by integrating various scenes and segmented objects. Instead, initiatives like Bridge Data[141] and Open X-embodiment[142] strove to compile extensive human demonstration datasets across diverse domains. In addition, some researchers explored the potential of in-the-wild data for IL, capitalizing on the availability of extensive egocentric human activity videos. Vid2Robot[143] proposed an end-to-end policy learning framework by training a unified model on human video data. Recent efforts, such as Ag2Manip[144] and MPI[145], also adopted this approach to extract skills from human videos, demonstrating substantial performance in multi-task robotic manipulation.

Equivariant models have garnered attention for their advantages of enhancing sample efficiency and generalization in IL. A notable example is the work by Simeonov et al.[167], which introduces neural descriptor fields (NDFs). These fields leverage SE(3)-equivariance to represent manipulated objects and facilitate IL by searching matched poses within demonstration data. Building on this foundation, local neural descriptor fields (L-NDF)[147] extended the concept by introducing shared local geometric features between objects. However, NDFs face inherent limitations that restrict their generalization to non-fixed targets. To address this, equivariant descriptor fields (EDFs)[148] reformulated NDFs within a probabilistic learning framework, enhancing its flexibility. Further advancements include the integration of diffusion models into EDFs, as seen in diffusion-EDFs[150], equivariant diffuser for generating interactions (EDGI)[149], and SE(3)-DiffusionFields[151]. These approaches aim to improve the model's ability to generalize across a broader range of scenarios.

3.2.3 Other methods

In the domain of embodied policy learning, several in-

novative methods have emerged that combine the strengths of RL and IL. As a series of work, UniDexGrasp[152] and UniDexGrasp++[153] perpetuated the paradigm of teacher-student learning, aiming to develop a universal grasp policy that can effectively generalize across diverse objects and scenarios. Initially, these methods employed model-free RL algorithms to cultivate a teacher model that takes oracle states as input. Subsequently, the skills acquired by the teacher model are distilled to a student policy via IL, where the student policy solely has access to realistic observations, such as those obtained through vision.

Recent breakthroughs in LLMs and vision-language models (VLMs) have sparked interests in their applications for policy learning in robotics, leveraging their capabilities in perception, reasoning, and decision-making. These models took current visual observations and language instructions as inputs to generate corresponding action sequences through trainable adapters, enabling robots to perform complex tasks and adapt to new situations[154, 155]. Notable examples include VILA[154] and Grounding-RL[155], which use pre-trained LLMs in their policy learning methods. In contrast, OpenVLA[156] employed pre-trained visual encoders to extract visual features, subsequently mapping them into a language embedding space. This method utilized a low-rank adaptation fine-tuning strategy to customize LLMs for robotic manipulation tasks. 3D-VLA[157] further extended this concept by incorporating 3D spatial observations and integrating a diffusion model for goal-aware state generation, resulting in a 3D generative world model.

3.2.4 Discussion

Although RL and IL have shown remarkable progress in embodied policy learning, challenges remain in terms of sample efficiency, domain adaptation, and generalization. Future research needs to unleash the potential of LLMs and VLMs for building generalizable and versatile agents for robotic manipulation tasks.

4 Embodied task-oriented learning

Embodied task-oriented learning not only involves strategic planning through powerful perception but also necessitates robots to understand how their physical attributes influence decision-making and task execution. It helps robots develop the ability to make decisions in complex and dynamic scenarios. Specifically, existing work of embodied task-oriented learning centers on two domains: object grasping and object manipulation. As shown in Table 3, this section will introduce methods tailored for these two tasks, revealing how embodied learning improves the efficiency and precision of robots.

4.1 Object grasping

Object grasping is the fundamental cornerstone of object manipulation. It encapsulates a robot's ability to capture targets reliably using end-effectors such as grippers or suction cups. This process requires analyzing object attributes like location, shape, size, and material to formulate grasp strategies that ensure steadfast control while preserving the object's intactness. Grasping methods are further differentiated into single-object grasping[225] and multi-object grasping[203], each presenting its own set of complexities. Fig. 7 illustrates examples of these two types of methods.

4.1.1 Single-object grasping (SOG)

Prior research has defined SOG as the configuration of

Table 3 Summary of embodied task-oriented learning methods. SOG: Single-object grasping; MOG: Multi-object grasping; NDM: Non-dexterous manipulation; DM: Dexterous manipulation; H2R: Human-to-robot.

Task	Type	Subfields & references
Object grasping	SOG	Open-loop grasping (STEM-CaRFs[168], FANet[169], AnyGrasp[170]), closed-loop grasping (adaptive grasping[171], GG-CNN[172], VFAS-Grasp[173])
		Transparent object grasping: DFNet[174], Dex-NeRF[175], GraspNeRF[176], NFL[177], TRansPose[178], TGF-Net[179]
		Grasping in clutter: Collision-free grasping (Contact-GraspNet[180], CaTGrasp[181], CollisionNet[182], DDGC[183], GSNet[184], DAL[185]), reposition-based grasping (push-grasping synergy[186], object singulation[187], grasping invisible[188], vision-language grasping[189])
		Dynamic object grasping: H2R handover (wearable sensing[190], TLP[191], reactive handover[192], flexible handover[193], GenH2R[194]), human-free moving object grasping (velocity decomposition[195], adaptive motion generation[196], Moving GraspNet[197])
	MOG	Holistic grasping (MOG in the Plane[198], MOG-Net[199], experience forest[200], Push-MOG[201])
		Independent grasping (MOG by exploiting kinematic redundancy[202], MultiGrasp[203])
Object manipulation	NDM	Pick-and-place[204], object rearrangement[205], kit assembly[206], deformable object manipulation (clothing[207], ropes[208], and fluids[209]), articulated object manipulation (GAPartNet[210], UniDoorManip[211], PartManip[212])
	DM	Trajectory optimization[213], kinodynamic planning[214], PDDM[215], in-hand object reorientation[216], DIME[217], DexDeform[218]
		Tool Manipulation: KETO[219], TOG-Net[220], DiffSkill[221], tool cognition[222], ATLA[223], RoboTool[224]

Fig. 7 Illustration of single-object grasping (top row) and multi-object grasping (bottom row). The examples are respectively from the ARNOLD benchmark[226] and Grasp′ Em dataset[203]. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

an end-effector designed to achieve partial or complete form-closure or force-closure of a targeted object[49]. Achieving stability and robustness in single-object grasping involves accurately determining object positions and identifying the appropriate grasping pose, which has a wide range of applications in fields such as industrial manufacturing[227] and medical assistant[228].

The typical and direct SOG involves three steps: grasp detection, trajectory planning, and execution. In this pipeline, the robot first captures the local scene using external cameras and plans a set of candidate configurations for the target object. Some methods executed the optimal grasp in an open-loop manner, where the grasp is performed directly without further sensor feedback after selecting the optimal grasp. In open-loop grasping, grasp detection is critical as subsequent steps rely on the coordinates generated during this phase. Consequently, various studies have endeavored to enhance the precision of grasp detection to facilitate effective grasping procedures. For example, Asif et al.[168] proposed hierarchical cascaded forests to infer object class and grasp-pose probabilities at both patch and object levels. More recently, Zhai et al.[169] designed FANet, which leverages grasp keypoints to enhance the grasp detection accuracy while maintaining real-time efficiency. In AnyGrasp[170], the center-of-mass of objects is incorporated into target detection, and an open-loop strategy is employed throughout the grasping process.

Although open-loop grasping has been extensively studied, it might fail due to inadequate pose estimation and other perception artifacts. To address these issues, closed-loop grasping has been proposed, leveraging real-time feedback to correct perception errors and handle object disturbances. Specifically, object tracking and visual servoing are two primary methods for achieving closed-loop grasping. For instance, Marturi et al.[171] explicitly tracked the 6DoF object pose and combined it with pre-computed grasp poses to enable adaptive grasp planning and execution. Furthermore, Morrison et al.[172] proposed GG-CNN to perform close-loop object-independent grasping, using a lightweight CNN to predict pixel-wise grasp quality. After that, Piacenza et al.[173] presented VFAS-Grasp, which uses visual feedback from point clouds and an uncertainty-aware adaptive sampling strategy to maintain a closed-loop system.

In addition to the general SOG methods mentioned above, three specific tasks, as illustrated in Fig. 8, have garnered significant attention due to their high level of challenge: transparent object grasping, grasping in clutter, and dynamic object grasping.

Transparent object grasping. Transparent objects are items through which light can pass without significant scattering or reflection, such as glass containers and plastic bottles commonly found in daily life. Research into embodied learning technology for grasping transparent objects has a profound impact on robotic applications[230]. However, grasping transparent objects presents significant challenges. Firstly, the lack of distinctive texture and appearance features, combined with light reflection and refraction, prevents most sensors from accurately capturing surface information, making it difficult for traditional vision systems to recognize and locate these objects. Secondly, the low friction of transparent objects complicates stable manipulation during the grasping process.

Most grasping methods heavily rely on depth images, necessitating precise depth information for transparent objects. Fang et al.[174] developed DFNet, an end-to-end depth completion network using RGB images and inaccurate depth maps to produce refined depth maps. Some other approaches utilized NeRFs to generate the depth information of transparent objects directly. For instance, Ichnowski et al.[175] augmented the specular reflection of transparent objects by placing additional lighting and used NeRFs for transparency-aware depth rendering. However, this method requires several hours of computation per grasp. To speed up the grasping process, Dai et al.[176] introduced GraspNeRF, which uses six sparse multi-view RGB images for zero-shot NeRF construction

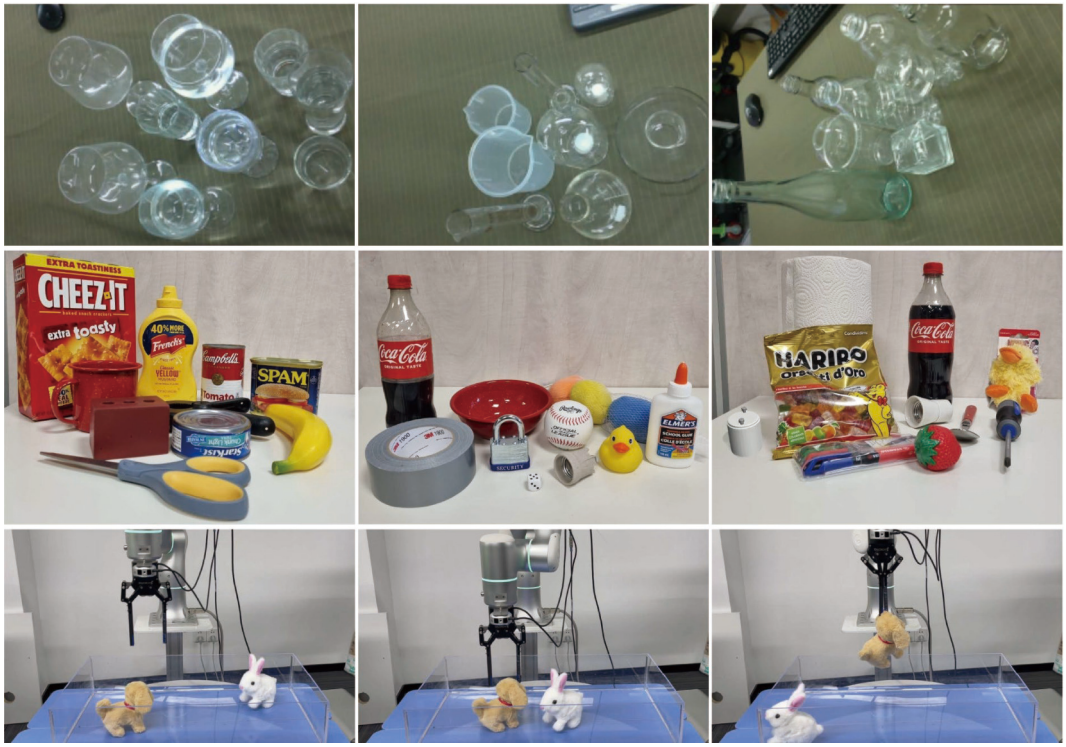

Fig. 8 Illustration of transparent objects (the first row), cluttered environment (the second row), and dynamic object grasping (the last row). The examples are respectively from TRansPose dataset[178], CEPB benchmark[229], and Moving GraspNet[197].

and grasp detection, achieving material-agnostic grasp detection in 90 ms. In contrast, Lee et al.[177] proposed normal field learning (NFL) to train neural volume from per-pixel surface normal estimation instead of RGB images and employed segmentation to identify transparent objects, requiring only 40 seconds of training time.

In addition to depth information, some researchers have employed other information to detect transparent objects in the scene. For example, Kim et al.[178] created TRansPose, the first large-scale multispectral dataset combining stereo RGB-D, thermal infrared images, and object poses, to promote the study of transparent object grasping. Moreover, Yu et al.[179] proposed using TGF-Net to learn surface fragments, edge features, and geometric features for 6D pose estimation of transparent objects, aiming to enhance robustness to variations in appearance. Additionally, they introduced a low-cost dataset generation scheme for obtaining a high-fidelity, large-scale dataset of transparent objects.

Grasping in clutter. It refers to a robot's ability to grasp target objects in crowded and cluttered environments precisely, which is crucial for applications like automated home services, manufacturing parts picking, and waste sorting. For instance, in domestic settings, robots must pick up items from a cluttered desk or cabinet for organization or delivery. Compared to tasks in an organized environment, grasping in clutter is more complex and challenging. This is because target objects may be hidden or overlapping, making them hard to identify and locate, and robots must also avoid collisions with surrounding objects to ensure safety[231].

Current research primarily focuses on collision-free object grasping[180, 181], with the goal of planning a safe and efficient grasping and execution path for the robot to ensure a smooth and unobstructed process. Murali et al.[182] proposed a grasp learning method, which uses a deep network called CollisionNet to assess the collision risk of generated grasps in cluttered scenes. Lundell et al.[183] introduced DDGC, a fast method for generating multi-finger collision-free grasp samples, addressing the issues of long computation times and challenges in obstacle avoidance. Wang et al.[184] introduced the concept of graspness, which is a quality metric combining geometric cues and collision labels to evaluate graspable regions in a cluttered scene. To alleviate the reliance on extensive labeled data, Wei et al.[185] introduced a discriminative active learning framework, which employs a discriminator to assess the informational value of unlabeled samples and intelligently select samples for annotation.

Furthermore, a body of research has expanded beyond collision-free object grasping and explored strategies involving grasping and pushing to reposition surrounding objects[186], which is particularly crucial when the target object is occluded or not directly accessible. Kiatos and Malassiotis[187] addressed the challenge of collision avoidance and proposed using pushing actions to isolate the target object from surrounding clutter. Yang et al.[188] further explored the problem of grasping invisible objects in clutter and integrated deep Q-learning with domain knowledge to devise optimal pushing and grasping mo-

tions. Recently, Xu et al.[189] employed a visual-language model to grasp objects in a cluttered environment based on language instructions and utilized a series of obstacle-removal actions to guide the robot to grasp the target object. Although substantial progress has been made, grasping in clutter continues to present significant challenges. Language-conditioned grasping has emerged as a novel and promising research field, increasingly attracting attention for future exploration.

Dynamic object grasping. It is a highly challenging research area focusing on enhancing a robot's ability to grasp moving objects stably. Its applications span from picking up product parts on factory assembly lines to delivering items in household services and even accurately grasping and transferring surgical instruments during surgery procedures. Compared to grasping stationary objects, dynamic object grasping is much more difficult. It requires the robot to quickly adjust its grasp to match the object's movement for precise docking and to predict continuous motion to handle the object's potentially nonlinear and unpredictable trajectories. This presents significant challenges for the robot's real-time processing, adaptability, and advanced motion prediction capabilities.

One key focus of existing work is on human-to-robot (H2R) handover, which aims to enable robots to receive objects from humans. Some studies have been conducted to improve the success rate of H2R handovers by understanding human intention[190, 191]. Wang et al.[191] explored the integration of multimodal inputs, e.g., vision and language, and proposed a multimodal learning framework to predict human behavioral intentions. An alternative research direction involves dynamic motion planning[192, 193] for H2R handover. Yang et al.[192] tackled the challenges posed by object variability in dynamic environments by integrating a closed-loop motion planning strategy with grasp generation. To further enhance the generalizability of H2R handover methods, Wang et al.[194] employed large-scale simulated demonstrations and imitation learning, enabling the robot to pick up objects of any shape transferred by humans in complex trajectories.

Another line of work is human-free moving object grasping, which does not involve human participation. Early studies simplified this problem by assuming prior knowledge of object motion. For example, Ye and Liu[195] focused on the top-down grasping strategy and proposed a planning algorithm based on analyzing velocity components to predict the trajectory of moving objects. Subsequently, Akinola et al.[196] expanded on this assumption by developing a dynamic grasping approach that combines reachability and motion awareness, thereby improving the overall success rate without relying on prior knowledge of object motion or constraining the grasping direction. In recent years, the robotics community has expressed significant interest in reactive grasping due to its autonomous adaptability in complex and dynamic environments. Specifically, Liu et al.[197] introduced a target-referenced reactive grasping method that emphasizes both the temporal smoothness and the semantic consistency of the predicted grasp poses. Overall, the technology for grasping moving objects is still evolving, and further in-depth research will enable robots to develop more robust and universally applicable grasping capabilities.

4.1.2 Multi-object grasping (MOG)

It embodies the advanced performance of robots in efficiently capturing two or more objects within a single operational cycle. This capability holds tremendous potential for various robotic applications, including logistics automation, product packaging, and home services, significantly boosting the efficiency of task execution. Compared to SOG, MOG imposes stricter demands on a robot's comprehensive abilities, encompassing acute perception, sophisticated strategy formulation, and precise execution coordination. Robots must adeptly identify and localize multiple target objects while mastering intricate planning and coordination mechanisms to ensure simultaneous grasping of multiple objects in various environments, achieving an impeccable blend of high efficiency and stability.

Within the domain of MOG, some methods treat it as a holistic grasping problem, in which multiple target objects are regarded as a whole entity for manipulation[198]. These methods intend to encompass all objects with a gripper or fingers in a single action, regardless of their individual placement or stacking configurations. In practice, the contact points are confined to the collective periphery of the objects, and the robot needs to apply precise grasping forces to ensure the ensemble's stability and the grasp's reliability. Sun et al.[232] developed a comprehensive taxonomy comprising twelve different types of MOG, which incorporate considerations of shape and function. Agboh et al.[199] integrated the factor of inter-object friction into their method, significantly improving the robot's ability to grasp multiple objects in a single motion. Many existing methods are grounded in the assumption that the target objects are closely adjacent in space[200]. However, this assumption does not always hold in practical, real-world situations. To tackle this challenge, Aeron et al.[201] introduced the Push-MOG method, which utilizes pushing maneuvers to systematically arrange a disordered ensemble of polygonal objects into compact and easily graspable clusters.

Another type of method involves considering each object as an independent unit, which significantly enhances the robot's flexibility and adaptability. Yao and Billard[202] proposed an algorithm that imitates human dexterous grasping, enabling a robotic hand to utilize the cooperative spaces between fingers for efficient and sequential grasping of multiple objects. Li et al.[203] introduced the multigrasp framework, focusing on maintaining the ability to independently manipulate each object while systematically enhancing the overall grasp effi-

ciency in complex MOG scenarios. Despite significant progress in this field, there is an urgent need for more exhaustive attention and further in-depth investigation.

4.1.3 Discussion

SOG, owing to its relative simplicity, has been extensively studied and has achieved significant progress, gradually moving into practical applications. In contrast, MOG, due to its high complexity, has experienced slower progress and still demands further efforts and breakthroughs.

4.2 Object manipulation

Object manipulation involves a wide range of control activities robots perform, from object grasping and utilization to environmental interactions. These capabilities are crucial in various applications, including product assembly, household services, and precision medical surgeries. Currently, the methodologies in this field are divided conceptually into two main categories: non-dexterous manipulation and dexterous manipulation, as depicted in Fig. 9. Next, we will introduce some representative methods of these two manipulation types.

4.2.1 Non-dexterous manipulation (NDM)

It refers to using simple end-effectors such as grippers, suction cups, or pushers by robots during task execution instead of relying on delicate finger manipulation or complex hand coordination. This type of manipulation typically has fewer degrees of freedom. It is well-suited for tasks that do not demand high precision or complexity, such as basic gripping, pushing, and pulling. While it may not be as flexible or adaptable as dexterous manipulation, its simplicity and efficiency make it highly promising in fields characterized by repetitive tasks, including industrial assembly, logistics sorting, and agricultural picking.

Pick-and-place is a fundamental task of NDM that has been extensively researched in recent years. It involves a robot picking up objects from one location and placing them in another specified location. Early studies primarily concentrated on estimating the poses of known objects[233] in structured environments or relied on scripted planning and motion control[234]. However, there has been a recent shift towards creating universal pick-and-place policies[204] for novel objects to enhance adaptability across broader scenarios. Furthermore, some research has expanded on the basic pick-and-place capabilities to perform higher-level tasks, such as object rearrangement[205] and kit assembly[206]. These advancements represent progress towards more advanced manipulation skills and signify the next steps in complex robotic operations.

Another line of research focuses on improving the intelligence of robots to handle more complex tasks, such as manipulating deformable and articulated objects. For deformable object manipulation, the variability in physical properties across different materials and the complex deformation behaviors under external forces introduce unpredictability in manipulation processes and heighten control complexity. Researchers in this field are drawing insights from human-object interactions in daily life to develop specific manipulation strategies for materials like clothing[207], ropes[208], and fluids[209]. For articulated object manipulation, the core challenge lies in precisely perceiving and controlling each joint part's angle and position while also necessitating a deep understanding of their kinematic properties and dynamic interactions. Typical articulated objects, such as doors, drawers, and buckets, form the nucleus of research interest. The current research frontiers center on establishing benchmarks[210, 211] for articulated object manipulation at the part level and developing universal manipulation policies[212] that can effectively handle previously unseen shapes and categories.

Although existing methods have demonstrated proficiency in various NDM tasks, they still confront several challenges. For example, in prolonged operations, ensuring high stability and seamless operational continuity is

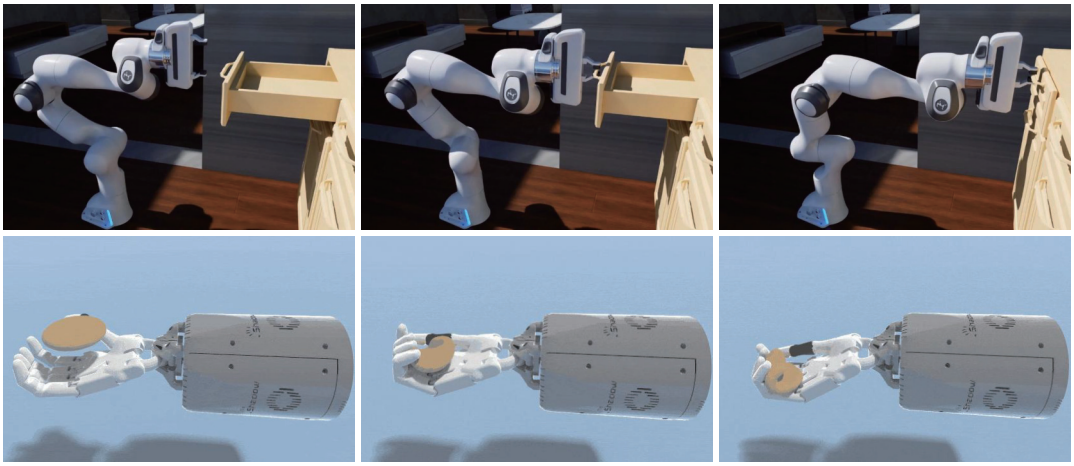

Fig. 9 Illustration of non-dexterous manipulation (top row: close drawer) and dexterous manipulation (bottom row: in-hand manipulation). The examples are respectively from the ARNOLD benchmark[226] and DexDeform[218].

paramount, which requires systems with robust endurance for long-horizon task execution[235]; in dynamic environments, the robot needs to adapt its position and orientation based on environmental changes or the instantaneous state of objects, which necessitates the integration of active visual adaptation and learning mechanisms[236]; when deviations or errors occur during robot operations, the immediate and precise identification of errors coupled with autonomous corrective actions become essential to ensure uninterrupted task completion[237]. Future research efforts must address these challenges to advance the development of NDM technologies.

4.2.2 Dexterous manipulation (DM)

It aims to replicate subtle human actions, such as unscrewing bottle caps or handling tools. It relies on sophisticated robotic hands[238], distinct from commonly used parallel grippers in NDM. Typically, these robotic hands emulate the structure of human hands, featuring multiple fingers and exhibiting exceptional flexibility[239], specializing in precise grasping and manipulation tasks.

Early methods for DM relied on analytical kinematic and dynamic models, using trajectory optimization[213] and kinodynamic planning[214] to establish robotic control policies and motion trajectories. However, these approaches had a significant limitation as they heavily relied on precise knowledge of dynamic properties and simplified assumptions on object geometries, which are often hard to obtain in complex real-world applications. In recent years, model-based[215] and model-free[216] RL approaches have increasingly become more prevalent in DM. The former aims to train a model from collected data that can predict state transitions and rewards to guide policy optimization. In contrast, the latter does not involve explicit model construction of the environment; instead, it learns directly from experiences gained through interaction with the environment. Another line of work lies in imitation learning, where optimal control strategies are learned from demonstrations[217], sometimes integrated with RL to enhance the effectiveness of DM[218]. These methods have shown effectiveness in executing DM tasks; nonetheless, they are primarily designed and optimized for specific categories of tasks. Consequently, developing universal and broadly adaptable DM frameworks remains an area for further exploration.

Tool manipulation. As a universal and fundamental human skill, tool manipulation has emerged as a pivotal focus in the field of DM, which is dedicated to enabling robots to proficiently manipulate a wide range of tools using intricate dexterous hands or specialized end-effectors[240]. Its applicability spans from industrial automation to surgical interventions and even space exploration, empowering robots to undertake tasks of remarkable complexity and specificity. In contrast to conventional object manipulation, tool manipulation poses a more stringent challenge to robots. It entails not merely the precise grasping of tools but also the intricate use of tactile feedback to accurately discern the contact status and effects of tool-workpiece interactions[241]. Considering the wide variety of tools in the real world, with their differing shapes, materials, and usage, robots need to demonstrate robust perception and decision-making capabilities to adapt flexibly and handle the specific physical properties and operational requirements of each tool[242, 243].

Current research in tool manipulation often revolves around learning task-specific skills, encompassing movement strategies and manipulation techniques for using tools. This is similar to learning approaches for manipulating non-tool objects but emphasizes integrating tools into robotic actions. Qin et al.[219] proposed the KETO framework, which employs deep neural networks to predict task-relevant keypoints from point clouds. Fang et al.[220] introduced task-oriented grasping network (TOG-Net), which utilizes large-scale simulation for self-supervised learning to optimize tool grasping and manipulation strategies. Lin et al.[221] advanced the field with the DiffSkill framework, leveraging a differentiable physics simulator to learn skill abstraction for tool-based long-horizon manipulation of deformable objects. Distinguished from these methods that rely on prior tool learning, Tee et al.[222] introduced a framework inspired by neuroscience principles, enabling robots to recognize and deftly apply novel tools to perform a variety of tasks without prior learning. Recently, research trends have ventured into leveraging LLMs to enhance robots′ tool manipulation capabilities[223, 224], highlighting a new direction for novel approaches that enable more flexible and efficient robotic tool manipulation.

4.2.3 Discussion

Both NDM and DM involve diverse and complex tasks. Existing methods are typically designed for several specific tasks and still fall significantly short of achieving truly general object manipulation.

4.3 Analysis of different end-effectors

No matter for object grasping or manipulation, a crucial component for the embodied system is the end-effector, which directly interacts with the objects and surrounding environments. Generally, end-effectors can be categorized into two types: parallel-jaw grippers and multi-fingered grippers. Parallel-jaw grippers are simple and widely used in industrial scenarios due to their low cost and ease of control. Nevertheless, the lack of adaptability and dexterity restricts their application in complex tasks. In contrast, multi-fingered grippers, inspired by human hands, have more degrees of freedom and are more flexible and versatile, enabling robots to perform a wide range of manipulation tasks. Despite their advantages, they are more complex and expensive, which hinders their widespread adoption in practical scenarios.

In the context of object-centric robotic manipulation, the choice of end-effector is depending on the specific ma-

nipulation tasks and object properties to some extent. For instance, parallel-jaw grippers are more suitable for single-object grasping[225] and simple object manipulation tasks[205], due to their simplicity and efficiency. Instead, complicated multi-fingered grippers are more appropriate for dexterous manipulation tasks[218] and tasks involving multiple objects[203], where precise control and dexterity are required.

However, from the perspective of embodied learning, different types of end-effectors are not the primary concern and can be considered into one pipeline. In particular, the embodied agent just needs different action spaces to adapt to various end-effectors, and the learning algorithms can be shared across different end-effectors. Therefore, we do not classify the embodied learning methods based on different end-effectors in this survey.

5 Datasets and evaluation metrics

In this section, we will introduce some primary datasets and evaluation metrics in the area of robotic manipulation.

5.1 Datasets

Existing datasets can be divided into two categories based on the differences in specific manipulation tasks: object grasping and object manipulation. Table 4 presents an overview of widely used datasets. Most of them are from simulated environments and exhibit considerable variation in categories, objects, data domains, sizes, and modalities. For a detailed description of each dataset, please see Appendix A and Appendix B.

5.2 Evaluation metrics

Here, we will introduce some typical evaluation metrics for object grasping and manipulation, as shown in Table 5.

5.2.1 Object grasping

Accuracy (Acc) is a classic metric for evaluating object grasping, which measures the percentage of correct predictions out of all predicted outputs. Acc can be expressed as follows:

$$Acc = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathbb{1}_{\{T(i)\}} \tag{9}$$

where N_p refers the number of all predicted output, $\mathbb{1}_{\{T(i)\}}$ is the indicator function, which equals 1 if the condition T of correctness is satisfied for the i-th prediction and 0 otherwise. There are two critical metrics for determining the correctness of a prediction: the "point" metric[258] and the "rectangle" metric[244]. In the "point" metric, a prediction is considered correct if the center point of the predicted rectangle falls within a

Table 4 Summary of the widely used datasets for robotic manipulation. "Domain" indicates whether the dataset is derived from real-world environments or generated through simulation, with "sim" being short for "simulation". "–" denotes that the corresponding quantity is unavailable.

Task	Dataset	#Categories	#Objects	Domain	Size	Modality
Object grasping	Cornell[244]	–	240	Real	885 images, 8 019 grasps	RGB-D
	Multi-Object[245]	–	∼400	Real	96 images, 2 904 grasps	RGB-D
	Jacquard[246]	–	11 K	Sim	54 K images, 1.1 M grasps	RGB-D
	VR-Grasping-101[247]	7	101	Sim	150 K grasping demonstrations	RGB-D
	ACRONYM[248]	262	8 872	Sim	17.7 M parallel-jaw grasps	PointCloud
	EGAD[249]	–	2 331	Sim	233 K antipodal grasps	PointCloud
	GraspNet-1Billion[250]	–	88	Real	97 280 images, ∼1.2 B grasps	RGB-D
	Grasp-Anything[251]	236	∼3 M	Sim	∼1 M samples, ∼600 M grasps	Text/Image
Object manipulation	YCB[252]	5	77	Real	600 RGB-D images for each object	RGB-D
	AKB-48[253]	48	2 037	Real	100 K generated RGB-D images	RGB-D
	PartNet-Mobility[106]	46	2 346	Sim	14 068 articulated parts	PointCloud/RGB-D
	GAPartNet[210]	27	1 166	Sim	8 489 part instances	PointCloud/RGB-D
	ManiSkill2[254]	20	2 000+	Sim	4 M demonstration frames	PointCloud/RGB-D
	ARNOLD[226]	8	1 078	Sim	10 080 demonstrations	Text/RGB-D
	Bi-DexHands[255]	20	–	Sim	1 638 400 step demonstrations	Force/PointCloud/RGB-D
	DexArt[256]	4	82	Sim	6 K point clouds for each object	PointCloud
	PartManip[212]	6	494	Sim	11 object categories, 1 432 tasks	PointCloud
	BEHAVIOR-1K[257]	1 000	9 318	Sim	50 scenes, 1 949 object categories	RGB-D

Table 5 Some typical evaluation metrics for object grasping and manipulation

Task	Metric	Formula	Short description
Object grasping	Accuracy	$\frac{1}{N_p}\sum_{i=1}^{N_p}\mathbb{1}_{\{T(i)\}}$	The percentage of correct predictions out of all predicted output
	Grasp success rate	$\frac{S_g}{N_g}$	The proportion of successful grasps out of the total number of grasp attempts
Object manipulation	Task success rate	$\frac{S_t}{N_t}$	The proportion of successful executions out of the total number of task executions

certain threshold distance from the ground truth grasp point. However, this metric does not account for grasp orientation, which may lead to overestimating actual performance. On the other hand, the "rectangle" metric is designed explicitly for rectangles and incorporates orientation error into the evaluation criteria. It first filters out predicted rectangles with an angular deviation from the ground truth G that surpasses 30 degrees. Then, among the remaining set, it calculates the intersection over union (IoU) between the predicted rectangle $\hat{G}$ and G:

$$J(\hat{G}, G) = \frac{|\hat{G} \cap G|}{|\hat{G} \cup G|}. \tag{10}$$

Finally, the prediction $\hat{G}$ is considered correct if $J(\hat{G}, G)$ is greater than a certain threshold τ.

Besides, grasp success rate (GSR) is commonly adopted as an evaluation metric in real-world robotic experiments. Assuming that a robot performs successful grasp S_g times out of N_g grasp attempts, the GSR is formulated as

$$GSR = \frac{S_g}{N_g}. \tag{11}$$

Moreover, several specially tailored evaluation metrics have been proposed, such as completion rate[42] and AP[250].

5.2.2 Object manipulation

The most commonly used evaluation metric for object manipulation is the task success rate (TSR). A task is considered successful when it satisfies specific conditions. Generally, each task is performed multiple times using different random seeds to reduce the impact of random variations on assessment results, and the mean value and variance are then reported. The following formula formally defines the TSR:

$$TSR = \frac{S_t}{N_t} \tag{12}$$

in which S_t and N_t are the numbers of successful and total executions, respectively. Notably, the conditions for success differ across various types of manipulation tasks. Taking the task of opening a cabinet drawer as an example, the condition for success is that the target drawer has been opened to at least 90% of its maximum opening range and must remain in a static state[254]. There are some extra metrics to evaluate models from various perspectives, such as the simulated time and kinematic object disarrangement[257].

6 Applications

With the continuous advancement of artificial intelligence, machine learning, and robotics technology, intelligent robots will be applied more extensively and deeply across various fields. Table 6 shows the applications of embodied learning for object-centric manipulations, i.e., industrial robots, agricultural robots, domestic robots, surgical robots, and other promising applications.

6.1 Industrial robots

Traditional industries, represented by manufacturing, rely heavily on human resources. However, intelligent robots are expected to revolutionize the conventional industrial production model, achieving goals such as increasing production efficiency, reducing labor costs, and enhancing safety. Typical application scenarios include: 1) Assembly line operations[259], where robots can perform tasks such as parts installation and circuit board welding; 2) Packaging and sorting operations[260], where robots can provide fast and accurate packaging and sorting services for industries like retail and food; 3) Maintenance operations[261], where robots can perform equipment maintenance tasks in hazardous environments.

To better demonstrate the practical application value of the embodied intelligence methods introduced in this paper, we use a factory assembly line as an example to illustrate how these methods can be combined with real-world applications. Here, the robot automation system can be divided into three stages: 1) Input RGB-D images captured by a depth camera and use object detection technology to identify the category and location of the parts. 2) Use pose estimation technology to determine the 6DOF grasping pose of each part so that the robot can accurately grasp the parts. 3) Use reinforcement learning technology to optimize the actions during the grasping and assembly process to achieve efficient assembly based on dexterous hands. The detailed case study is shown in Table 7. The hardware, software, and algorithms men-

Table 6 Applications of embodied learning for object-centric robotic manipulation

Area	Example	Short description
Industrial robots	Assembly line operations[259]	Performing tasks such as parts installation and circuit board welding
	Packaging and sorting operations[260]	Providing fast and accurate packaging and sorting services for industries like retail and food
	Maintenance operations[261]	Performing equipment maintenance tasks in hazardous environments
Agricultural robots	FarmBot	Accomplishing intelligent planting through monitoring plant growth, fertilization, and cultivation control
	FAR	Identify fruits and their ripeness, and performing tasks such as pruning
	Rowbot	Designing for crops like corn, capable of fertilizing, seeding, and weed removal
Domestic robots	Household assistants[262]	Helping with tasks like organizing desks, performing cleaning duties, and folding clothes
	Smart caregiving[263]	Offering daily care and health monitoring for people in need of special care
	Cooking[264]	Automatically completing specific cooking tasks, such as frying eggs, stir-fry, and toasting bread
Surgical robots	Cutting and suturing[265]	Utilizing specialized surgical tools to perform precise tissue cutting and suturing
	Automated control[266]	Performing intelligent adjustment of the speed and direction of surgical instruments
	Surgical collaboration[267]	Collaborating in real-time with doctors during surgery, providing real-time process monitoring and data-driven decision support
Other applications		Space exploration[268], education[269], research[270]

Table 7 Detailed case study of embodied learning methods in real-world applications

Section	Details
Background	On a factory assembly line, robots perform precise tasks by assembling multiple parts into a complete product.
Task description	The robot must complete assembly tasks, which involve recognizing parts, estimating their poses, detecting 6DoF positions, predicting affordances, and optimizing its operational strategy in a dynamic environment.
Implementation	– Hardware: A UR5 robotic arm equipped with a multi-functional gripper, an Intel RealSense camera for capturing RGB-D data, and a high-performance GPU computer. –Software: ROS for control, PyTorch for model training and inference, Gazebo for simulation. –Algorithm: PointNet++[53] for feature extraction, 6-DOF GraspNet[271] for 6DoF grasp pose detection, RLAfford[114] for affordance prediction, PPO[163] for reinforcement learning, behavioral cloning[160] for imitation learning.
Experimental setup	Initial training in a simulated environment, followed by validation in a real-world setting using various parts to assess generalization.
Results analysis	The system′s accuracy and efficiency are measured using task success rate and execution time.

tioned in this table serve as examples for the case study. In real-world applications, the most suitable options can be selected based on specific requirements.

In the process of robots becoming increasingly intelligent but not yet fully autonomous, collaboration between robots and humans is essential in many complex and high-precision task scenarios. Additionally, robots may encounter operational anomalies or make mistakes, which requires the implementation of intelligent detection and fault diagnosis methods[272]. This is crucial for ensuring intelligent robots′ stable and reliable application within the industrial field.

6.2 Agricultural robots

In modern agriculture, intelligent robots are crucial in completing various tasks within farms and orchards, such as planting, nurturing, and harvesting crops. This promotes high-quality and sustainable agricultural development[273]. Some representative agricultural intelligent robots include 1) FarmBot[1], which accomplishes intelligent planting through monitoring plant growth, fertilization, and cultivation control; 2) FAR[2], which utilizes artificial intelligence and computer vision technology to identify fruits and their ripeness, and can perform tasks such as pruning; 3) Rowbot[3], designed for crops like corn, capable of fertilizing, seeding, and weed removal.

In agricultural settings, goods are often delicate and prone to damage[274], as seen in tomato harvesting. Applying too much force can harm the tomatoes, while too little force can cause them to slip and fall. This makes it challenging to achieve precise control and gentle handling. Additionally, robots may encounter obstacles such as branches and leaves in the open agricultural environment, requiring high positional accuracy and flexibility

[1] https://farm.bot

[2] https://www.tevel-tech.com

[3] https://www.rowbot.com

during operation.

6.3 Domestic robots

Domestic intelligent robots have promising applications in areas such as family assistance, caregiving, and household chores. They are capable of enhancing living quality and convenience and providing assistance to individuals with particular needs. Specifically, some valuable application scenarios include: 1) Household assistants[262], which help with tasks like organizing desks, performing cleaning duties, and folding clothes, thereby alleviating the daily household workload; 2) Smart caregiving[263], which offers daily care and health monitoring for people in need of special care; 3) Cooking[264], capable of automatically completing specific cooking tasks, such as frying eggs, stir-fry, and toasting bread.

Due to the vast diversity in home environments and the high complexity of tasks, it is essential for robots to quickly adapt to different household settings and handle potentially unexpected situations as well as new types of tasks. Taking cooking as an example, it involves a wide variety of ingredients and vastly different cooking steps, and the process also requires continuous identification of the food's state[275], which poses a significant challenge for robots. Another critical aspect is that such robots must be sufficiently safe and reliable, especially around children. Furthermore, cost is one of the significant factors affecting the widespread adoption of domestic robots. Consequently, identifying strategies to reduce the expense of robots while ensuring the quality of service is also a crucial issue that must be considered.

6.4 Surgical robots

Research on robots in the surgical field is rapidly advancing. These robots have the potential to serve as intelligent assistant tools to help doctors improve the quality and efficiency of surgery. Some typical application scenarios include: 1) Cutting and suturing[265], where robots utilize specialized surgical tools to perform precise tissue cutting and suturing, thereby reducing surgical trauma and medical staff workload; 2) Automated control[266], where intelligent adjustment of the speed and direction of surgical instruments is performed, enabling precise control over surgical progress and outcomes; 3) Surgical collaboration[267], where robots collaborate in real-time with doctors during surgery, providing real-time process monitoring and data-driven decision support.

The privacy of surgical data makes obtaining large-scale real-world data challenging. Current methods often rely on simulation environments to enhance surgical machine learning[276]. However, there is a significant gap between simulation and reality, challenging intelligent robots to make real progress in practical applications. Therefore, while intelligent robots can perform certain parts of surgical procedures, it will take considerable time before they can replace doctors. In particular, complex and enduring surgical tasks still require doctors' oversight to ensure the surgery's safety and effectiveness.

6.5 Other applications

In addition to the applications above, intelligent robots can also be used in fields such as space exploration[268], education[269], and research[270]. The skills required in these fields are different, and most of the existing work involves designing specialized intelligent models tailored to domain expert knowledge. While these methods perform well on specific tasks, their generalization capability could be improved. Fortunately, recent explorations based on general large foundation models offer a promising solution.

7 Challenges and future directions

In the past few years, there has been a significant increase in research on embodied learning methods for object-centric robot manipulation tasks, leading to rapid development in this field. However, current technology still faces some highly challenging issues. Further exploration of these issues will be crucial in promoting the widespread application of intelligent robots in various fields. This section will discuss several challenges and potential future research directions.

7.1 Sim-to-real generalization

Collecting real-world data for robotic manipulation is difficult, making creating a large-scale dataset challenging. To address this issue, current research primarily focuses on training models within simulation environments[277] which offer safe, controllable, and cost-effective learning scenarios, and the ability to generate virtually unlimited simulated training data[278]. However, transferring robotic skills learned in simulations to real-world scenarios poses a significant challenge in robotic manipulation. This difficulty arises primarily due to the differences between simulated and real environments[279], which leads to a mismatch between the data used in simulations and that encountered in reality. For instance, a simulated environment might not accurately reflect the physical properties of the real world, such as friction and gravity. Furthermore, simulated environments may struggle to account for unforeseen events in real-world situations. Another challenge is the difficulty in accurately modeling various types of robot sensors and actuators within a simulated environment. An unseen sensor or actuator may have different characteristics and limitations. This discrepancy can result in variations in robot behavior and performance when transitioning from simulation to the real world.

The above issues make it difficult for robots to gener-

alize their learned capabilities to new situations, thus limiting their deployment and applications. However, this challenge also presents new opportunities for robotic manipulation. The training data and iterations in simulated environments can be unlimited, which can help improve the robot′s understanding and adaptation to the complexities and uncertainties of real-world environments.

Recent research has focused on reducing this sim-to-real gap by using methods like domain randomization[280], physical constraint regularization[281], and iterative self-training[282]. We propose further research on domain adaptation methods suitable for robotic manipulation, which involve in-depth exploration of large-scale pre-training in virtual environments and rapid adaptation in real-world settings. Additionally, methods based on adversarial training and contrastive learning are worth further investigation. Studies in these areas can help improve the adaptation of robotic manipulation methods to real-world environments and enhance their performance in practical situations.

7.2 Multimodal embodied LLMs

Humans possess rich perceptual abilities like sight, hearing, and touch, which help them gather detailed information about their surroundings. Besides, humans can utilize learned experiences to perform various tasks. This versatility is also the ultimate goal of general-purpose intelligent robots. To achieve this, robots must be equipped with multiple sensors to perceive the environment and collect multimodal data. Additionally, robots must quickly learn and adapt to new environments and tasks to perform efficient actions.

However, enabling robots to handle multi-modal data and master diverse manipulation tasks is extremely challenging. This is reflected in two aspects: 1) Robots must integrate and understand information from multiple sources to comprehend task instructions and their surrounding environment. Different data modalities have unique characteristics and structures, often containing lots of redundant information, which complicates the understanding of multi-modal data. 2) Robots are expected to perform effectively across a wide range of manipulation tasks. Nevertheless, models often exhibit varying performance levels depending on the task, and they may struggle with certain tasks. Furthermore, multi-task learning increases the model′s complexity, making training and optimization more difficult.

Most existing research on robotic manipulation is based on single modalities such as images, 3D, or tactile data, with relatively little research on general robotic manipulation models that simultaneously master multiple modalities. Enhancing robots′ multi-modal understanding and multi-task execution capabilities can improve their ability to interpret human intentions and respond accordingly, leading to more flexible and adaptive robots.

In recent years, LLMs have demonstrated remarkable capabilities in the fields of natural language processing and computer vision. Meanwhile, in the domain of robotic manipulation, researchers are gradually exploring how to leverage LLMs to enhance robots′ perception, reasoning, and action-generation abilities[283–286]. Xu et al.[285] introduced a method for tuning reasoning that generates accurate numerical outputs for robotic grasping, leveraging the extensive prior knowledge of LLMs. SMART-LLM[287] employed LLMs for multi-agent robot task planning, converting high-level task instructions into multi-robot task plans through processes like task decomposition, coalition formation, and task allocation. Huang et al.[286] integrated affordance and physical concepts into LLMs beyond regular image and text modalities, resulting in better performance in robotic manipulation. RobotGPT[288] leveraged ChatGPT′s problem-solving capabilities to train a more reliable agent for robotic manipulation. These LLM-based methods integrate inputs from multiple modalities and are usually referred to as multimodal LLMs (MLLMs). The success of MLLMs is closely linked to the input prompts[289]. Cheng et al.[290] introduced a framework named LLM+A(ffordance), which employs an LLM as both a sub-task planner and a motion controller. Additionally, they devised an affordance prompting technique to allow the language model to generate affordance values for relevant objects. Xiong et al.[291] proposed an autonomous interactive correction (AIC) MLLM, which refines the SE(3) pose prediction of articulated objects by leveraging low-level interaction experiences. For interactions with objects, they designed two types of prompt instructions, i.e., visual masks and textual descriptions, to optimize the output of the MLLM. Liu et al.[292] suggested improving the autonomous manipulation abilities of large language models (LLMs) by leveraging human-robot collaboration, which utilizes a prompted GPT-4 model to translate language instructions into action sequences that a robot can perform.

The above-mentioned works have promoted the development of multimodal embodied LLMs, but overall, the field is still in its early stages and necessitates further extensive and in-depth research. We recommend starting with the development of large multimodal LLMs that integrate vision, language, and tactile feedback to facilitate general robotic manipulation in open-world environments, with a particular focus on improving the efficiency of training LLMs. Furthermore, leveraging vast amounts of videos capturing human activities, in conjunction with general multimodal LLMs, represents a promising direction for further research.

7.3 Human-robot collaboration

Intelligent robots can potentially revolutionize industries such as manufacturing, healthcare, and services. To

fully realize this potential, human-robot collaboration is crucial[293]. By working together, robots can assist humans, enhancing efficiency and reducing human workload and safety risks. Meanwhile, humans can guide and monitor robot operations to improve accuracy. Nevertheless, achieving perfect human-robot collaboration is challenging due to communication and coordination barriers, over-reliance, and safety issues.

Specifically, the challenges of human-robot collaboration include: 1) Robots must be able to accurately and in real-time perceive human behavior and understand the intentions behind it, which is very complex and subtle. 2) Robots must quickly adapt to constantly changing situations during interactions, but unforeseen events often lead to erroneous decisions by the robot. 3) Human-robot collaboration requires robots to have clear safety decision-making capabilities, which presents a series of moral and social-ethical challenges. These challenges make it difficult for robots to effectively interact with humans, thus limiting their ability to adapt to different user needs and environmental conditions, reducing their adaptability in various scenarios.

The research community has already achieved some progress in addressing the challenges of human-robot collaboration. For instance, Jin et al.[294] proposed a two-level hierarchical control framework based on deep RL to establish an optimal human-robot cooperation policy. Wang et al.[295] introduced a policy training method called Co-GAIL, which is based on human-human collaboration demonstrations and co-optimization in an interactive learning process. However, these methods are implemented in simulation environments or can only perform a limited number of tasks, making them unsuitable for practical applications. In the future, human-robot collaboration will remain a crucial research area, requiring continuous exploration to enhance the efficiency and safety. We suggest developing a unified framework for human-robot collaboration to enable researchers to conduct research and testing more efficiently. This framework should be flexible and scalable, supporting various interaction modes and application scenarios. Additionally, establishing a set of universal evaluation standards for human-robot collaboration methods is crucial. These standards will help ensure the comparability and consistency of different methods.

7.4 Model compression and robot acceleration

In applications such as embedded systems, mobile devices, and edge computing, robots with embodied intelligent systems usually have minimal computational resources[296]. This makes it essential to optimize and compress the deep models to meet the requirements of storage space, real-time, and accuracy. While LLMs-based methods have made significant advancements in embodied AI, they have also led to increased computational resource demands, posing challenges for implementation on devices with limited computing capabilities. Specifically, this challenge mainly arises from two reasons: 1) Limited computational resources can slow down the model's inference speed, affecting the robot's execution efficiency. 2) In resource-constrained environments, storage space is often highly restricted, preventing the storage of large amounts of model parameters and affecting the model's performance. Therefore, future research on model compression is expected to facilitate the practical application of intelligent robots.

In real-world applications, long waiting times often result in a poor user experience. Thus, it is expected that robots should be able to complete tasks quickly. However, many current mainstream models have low operating frequencies. For instance, Google's RT-2 model[297] has a decision frequency ranging from 1–5 Hz, depending on the parameter scale of the used VLMs, indicating that there is still a substantial gap before it becomes practical. Recently, the humanoid robot Figure 01[4] can generate action instructions at a frequency of 200 Hz, which benefits from OpenAI's LLMs and an efficient end-to-end network architecture. This achievement brings greater optimism for future research on robot acceleration. We recommend intensifying research efforts on more efficient perception and control algorithms, particularly in exploring unified architectures capable of handling multiple tasks. Such an architecture should seamlessly integrate all stages between sensor data input and robotic execution control, creating a fast and efficient pathway. By optimizing data processing and decision-making processes, this architecture will enhance the robot's response speed and task execution capabilities, enabling real-time operations in dynamic and complex environments.

7.5 Model interpretability and application safety

Deep learning-based methods are commonly referred to as "black boxes"[298]. For intelligent robots based on deep learning, this black-box characteristic can lead to suspicion and mistrust from users. However, understanding how deep models make decisions is a challenging task. On one hand, this is due to the lack of transparency in the models. On the other hand, the model complexity makes it hard to determine which features are the most significant for making predictions. Additionally, models trained on imbalanced data may exhibit potential biases, leading to inaccurate predictions. These issues make it extremely difficult to identify and resolve potential problems, which can compromise the safety and reliability of robots[299], ultimately limiting the widespread adoption of robotic manipulation systems. Therefore, research on the

[4] https://www.figure.ai

interpretability of embodied learning methods is crucial, which can help people understand the model's decision-making process and increase user trust in robots.

In addition to model interpretability, the safety of intelligent robots needs to be guaranteed from other perspectives, including implementing more reliable online learning and control techniques to prevent potential harm caused by the robot's motion[300]. It is also essential to employ adversarial training to protect robots against attacks[301] and to design robust safety monitoring methods for detecting possible security risks[302]. We propose establishing comprehensive robotic motion constraints in real-world settings, as this is essential for ensuring safety. Additionally, with the growing application of LLM-based methods in robotic manipulation, future research into the interpretability of instructions generated by LLMs will be highly valuable. Enhancing the interpretability of LLMs will help us achieve a better understanding of the decision-making processes of robots, thereby increasing the transparency and reliability of the system.

8 Conclusions

In this paper, we present a comprehensive survey of the existing methods for embodied learning in object-centric robotic manipulation. We begin by introducing the concept of this task and its essential components and then compare it with related survey articles. Next, we systematically present the main works across three categories. We then explore the commonly used datasets and evaluation metrics, highlighting some representative applications. Finally, we discuss the challenges and suggest promising directions for future research. We hope this survey will provide researchers with a comprehensive understanding and new insights in this emerging field.

Acknowledgements

The research work was partly conducted in the JC STEM Lab of Machine Learning and Computer Vision funded by The Hong Kong Jockey Club Charities Trust. This work was supported in part by the National Natural Science Foundation of China (No. 62106236). Open access funding provided by The Hong Kong Polytechnic University, China.

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

Appendix A. Datasets for object grasping

The objective of the object grasping task is for the robotic arm to be able to successfully and stably grasp objects. Therefore, the related datasets are mainly constructed around aspects such as enriching the categories of objects and the grasping poses.

The **Cornell** dataset[244] encompasses 240 different objects, for which 885 RGB-D images have been captured with a top-down imaging perspective. It provides 8 019 manually annotated positive and negative grasp samples. This dataset was the first to introduce oriented rectangles to represent grasps, enabling efficient training of grasp detection network models on images.

The **Multi-Object** dataset[245] comprises 96 real-world images, each containing 3–5 objects, amounting to approximately 400 objects in total. Multiple grasps have been annotated for each object, with a total of 2 904 grasps included, making it suitable for evaluating multi-object/multi-grasp tasks.

The **Jacquard** dataset[246] contains 54 K images of 11 K objects and provides 1.1 million grasping positions, which are represented by 2D rectangles on the images. All data are generated through simulations on CAD models and can be leveraged for image-based grasping position estimation.

The **VR-Grasping-101** dataset[247] includes 101 daily objects across 7 categories and synthesizes approximately 150 K grasping samples based on human demonstrations collected in virtual reality (VR). This dataset is particularly suitable for 6DoF parallel-jaw grasping tasks.

The **ACRONYM** dataset[248] is constructed based on simulation, comprising 8 872 objects across 262 categories from the ShapeNetSem dataset[303], and provides data on 17.7 million parallel-jaw grasping poses. As a large-scale grasping dataset, it can be used to train learning-based grasp detection algorithms.

The **EGAD** dataset[249] contains 2 331 objects represented by 3D meshes, with each object annotated with 100 antipodal grasps. Furthermore, a curated selection of 49 objects amenable to 3D printing has been chosen to facilitate the testing of robotic grasping tasks in real-world scenarios.

The **GraspNet-1Billion** dataset[250] encompasses 88 daily objects and has collected 97 280 RGB-D images of these objects from 190 cluttered scenes, along with providing accurate 3D mesh models. All data were acquired using real-world sensors and cameras. Furthermore, over one billion grasp poses have been annotated through analytic computation, offering a large-scale benchmark for the advancement of robotic object grasping techniques.

The **Grasp-Anything** dataset[251] is constructed based on foundational models and includes 236 object categories from the LVIS dataset[304]. It provides one million scene descriptions generated by ChatGPT and corresponding images produced by Stable Diffusion[305], with a total object count of approximately 3 million. For each object, there is an average of 200 grasps represented by 2D rectangles, amounting to roughly 6 million grasps in total. This dataset has the potential to substantially bolster research in the domain of zero-shot grasp detection.

B. Datasets for object manipulation

Compared to grasp detection, object manipulation is a more complex and general task, encompassing both the action of grasping and the various operations performed on objects after grasping. Therefore, datasets in this field focus more on covering a wider variety of task types, scenarios, and skills.

The **YCB** dataset[252] consists of 5 types of object sets, totaling 77 sub-classes of objects. All data were captured using 5 RGB-D sensors and 5 high-resolution RGB cameras. For each object, it provides 600 RGB-D images, 600 high-resolution RGB images, segmentation masks, calibration information, and 3D mesh models with texture mapping. This dataset is primarily useful for research related to robotic grasping and manipulation.

The **AKB-48** dataset[253] contains 48 types of objects, encompassing a total of 2 037 articulated object instances. Each instance has been scanned from the real world and manually refined by humans. Based on these models, 100 K RGB-D images have been generated for network training. In addition, 10 K real-world images have been collected, with 50% used for fine-tuning and the other 50% for testing. This dataset can be utilized to facilitate the generalization of robotic manipulation methods from simulation to reality.

The **PartNet-Mobility** dataset[106] encompasses 46 categories of common indoor objects, totaling 2 346 articulated object models, and provides 14 068 annotations of moving parts. Specifically, the motion of parts is categorized into three types: hinge, slider, and screw. This dataset serves as an evaluation benchmark for robotic perception and part-based manipulation tasks.

The **GAPartNet** dataset[210] is a large-scale part-centric dataset for object manipulation, featuring 1 166 articulated object models across 27 categories, all sourced from the AKB-48 dataset and PartNet-Mobility dataset. It defines 9 classes of cross-category GAPart and provides annotations for functional parts on each object, amounting to a total of 8 489 GAPart instances. This dataset is suitable for domain-generalizable tasks on object perception and manipulation.

The **ManiSkill2** dataset[254] is developed on the OpenAI Gym simulator[306] and consists of 20 manipulation tasks that span a diverse range of task types, including stationary/mobile, single/dual-arm, and rigid/soft. It contains over 2 000 object models and offers more than 4 million demonstration frames. This dataset is particularly well-suited for advancing and evaluating research in the field of learning generalizable manipulation skills.

The **ARNOLD** dataset[226] is built on the Isaac Gym simulator[307] and encompasses 8 language-grounded manipulation tasks, each presenting four goal states articulated through human language. This dataset provides a comprehensive collection of 10 080 learning demonstrations, with each demonstration consisting of 4–6 keyframes. It consists of 1 078 3D object models across 40 distinct categories and 1 114 scenes spanning 20 different types. This dataset establishes an evaluation benchmark for methods of language-conditioned manipulation and the generalization of learned skills.

The **Bi-DexHands** dataset[255] includes 20 types of bimanual manipulation tasks, features thousands of target objects, and provides multi-modal information like contact force, RGB image, RGB-D image, and point cloud. This dataset offers a comprehensive benchmark for general reinforcement learning approaches aimed at dexterous two-handed manipulation tasks.

The **DexArt** dataset[256] comprises 4 categories of objects, i.e., faucet, bucket, laptop, and toilet, with a total of 82 objects included. For each object, it provides 6 k point clouds, which encompass both the actual observed and imagined points. The primary focus of this dataset is on the generalizable dexterous manipulation oriented towards articulated objects.

The **PartManip** dataset[212] is constructed based on the GAPartNet dataset and comprises 494 objects across 11 different types. It consists of 1 432 sub-tasks that fall under 6 major task categories, respectively: OpenDoor, OpenDrawer, CloseDoor, CloseDrawer, PressButton, and GraspHandle. This dataset is designed to advance the research into part-based cross-category object manipulation methods.

The **BEHAVIOR-1K** dataset[257] encompasses a comprehensive collection of 1 000 behavioral categories derived from 50 everyday scenarios, incorporating 1 949 object classes with a total of 9 318 individual object models. Additionally, it provides rich physical and semantic annotations for each object. All data within BEHAVIOR-1K are generated within a simulation environment named OMNIGIBSON, which is particularly tailored for the evaluation of diverse approaches to complex robotic manipulation tasks.

Open Access

References

[1] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[3] Y. Zheng, H. Yao, X. Sun, S. Zhang, S. Zhao, F. Porikli. Sketch-specific data augmentation for freehand sketch recognition. *Neurocomputing*, vol. 456, pp. 528–539, 2021. DOI: 10.1016/j.neucom.2020.05.124.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ACM, Long Beach, USA, pp. 6000–6010, 2017.

[5] J. Devlin, M. W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, Minneapolis, USA, vol. 1, pp. 4171–4186, 2019. DOI: 10.18653/v1/N19-1423.

[6] A. Gupta, S. Savarese, S. Ganguli, Li F. F. Embodied intelligence via learning and evolution. *Nature Communications*, vol. 12, no. 1, Article number 5721, 2021. DOI: 10.1038/s41467-021-25874-z.

[7] N. Roy, I. Posner, T. Barfoot, P. Beaudoin, Y. Bengio, J. Bohg, O. Brock, I. Depatie, D. Fox, D. Koditschek, T. Lozano-Perez, V. Mansinghka, C. Pal, B. Richards, D. Sadigh, S. Schaal, G. Sukhatme, D. Therien, M. Toussaint, M. Van De Panne. From machine learning to robotics: Challenges and opportunities for embodied intelligence, [Online], Available: https://arxiv.org/abs/2110.15245, 2021.

[8] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Y. Nie, J. R. Wen. A survey of large language models, [Online], Available: https://arxiv.org/abs/2303.18223, 2023.

[9] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2022. DOI: 10.1145/3503250.

[10] J. Ho, A. Jain, P. Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, Article number 574, 2020.

[11] B. Kerbl, G. Kopanas, T. Leimkuhler, G. Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, vol. 42, no. 4, Article number 139, 2023. DOI: 10.1145/3592433.

[12] T. Gervet, S. Chintala, D. Batra, J. Malik, D. S. Chaplot. Navigating to objects in the real world. *Science Robotics*, vol. 8, no. 79, Article number eadf6991, 2023. DOI: 10.1126/scirobotics.adf6991.

[13] H. Guo, F. Wu, Y. Qin, R. Li, K. Li, K. Li. Recent trends in task and motion planning for robotics: A survey. *ACM Computing Surveys*, vol. 55, no. 13s, Article number 289, 2023. DOI: 10.1145/3583136.

[14] G. Du, K. Wang, S. Lian, K. Zhao. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: A review. *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1677–1734, 2021. DOI: 10.1007/s10462-020-09888-5.

[15] D. Han, B. Mulyana, V. Stankovic, S. Cheng. A survey on deep reinforcement learning algorithms for robotic manipulation. *Sensors*, vol. 23, no. 7, Article number 3762, 2023. DOI: 10.3390/s23073762.

[16] L. Jin, S. Li, J. G. Yu, J. B. He. Robot manipulator control using neural networks: A survey. *Neurocomputing*, vol. 285, pp. 23–34, 2018. DOI: 10.1016/j.neucom.2018.01.002.

[17] F. Zhu, Y. Zhu, V. C. S. Lee, X. Liang, X. Chang. Deep learning for embodied vision navigation: A survey, [Online], Available: https://arxiv.org/abs/2108.04097, 2021.

[18] J. Duan, S. Yu, H. L. Tan, H. Zhu, C. Tan. A survey of embodied AI: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 2, pp. 230–244, 2022. DOI: 10.1109/TETCI.2022.3141105.

[19] J. Francis, N. Kitamura, F. Labelle, X. Lu, I. Navarro, J. Oh. Core challenges in embodied vision-language planning. *Journal of Artificial Intelligence Research*, vol. 74, pp. 459–515, 2022. DOI: 10.1613/jair.1.13646.

[20] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic, D. Fox, A. Cosgun. Deep learning approaches to grasp synthesis: A review. *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3994–4015, 2023. DOI: 10.1109/TRO.2023.3280597.

[21] M. Zare, P. M. Kebria, A. Khosravi, S. Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, vol. 54, no. 12, pp. 7173–7186, 2024. DOI: 10.1109/TCYB.2024.3395626.

[22] X. Xiao, J. Liu, Z. Wang, Y. Zhou, Y. Qi, Q. Cheng, B. He, S. Jiang. Robot learning in the era of foundation models: A survey, [Online], Available: https://arxiv.org/abs/2311.14379, 2023.

[23] J. Chen, B. Ganguly, Y. Xu, Y. Mei, T. Lan, V. Aggarwal. Deep generative models for offline policy learning: Tutorial, survey, and perspectives on future directions, [Online], Available: https://arxiv.org/abs/2402.13777, 2024.

[24] Y. E. Ma, Z. Song, Y. Zhuang, J. Hao, I. King. A survey on vision-language-action models for embodied AI, [Online], Available: https://arxiv.org/abs/2405.14093, 2024.

[25] Z. Xu, K. Wu, J. Wen, J. Li, N. Liu, Z. Che, J. Tang. A survey on robotics with foundation models: Toward embodied AI, [Online], Available: https://arxiv.org/abs/2402.02385, 2024.

[26] K. Kleeberger, R. Bormann, W. Kraus, M. F. Huber. A survey on learning-based robotic grasping. *Current Robotics Reports*, vol. 1, no. 4, pp. 239–249, 2020. DOI: 10.1007/s43154-020-00021-6.

[27] H. Zhang, J. Tang, S. Sun, X. Lan. Robotic grasping from

classical to modern: A survey, [Online], Available: https://arxiv.org/abs/2202.03631, 2022.

[28] Z. Xie, X. Liang, C. Roberto. Learning-based robotic grasping: A review. *Frontiers in Robotics and AI*, vol. 10, pp. 1–14, 2023. DOI: 10.3389/frobt.2023.1038658.

[29] H. K. Tian, K. C. Song, S. Li, S. Ma, J. Xu, Y. H. Yan. Data-driven robotic visual grasping detection for unknown objects: A problem-oriented review. *Expert Systems with Applications*, vol. 211, Article number 118624, 2023. DOI: 10.1016/j.eswa.2022.118624.

[30] Y. Q. Huang, M. Bianchi, M. Liarokapis, Y. Sun. Recent data sets on object manipulation: A survey. *Big Data*, vol. 4, no. 4, pp. 197–216, 2016. DOI: 10.1089/big.2016.0042.

[31] N. Yamanobe, W. Wan, I. G. Ramirez-Alpizar, D. Petit, T. Tsuji, S. Akizuki, M. Hashimoto, K. Nagata, K. Harada. A brief review of affordance in robotic manipulation research. *Advanced Robotics*, vol. 31, no. 19–20, pp. 1086–1101, 2017. DOI: 10.1080/01691864.2017.1394912.

[32] B. Fang, S. Jia, D. Guo, M. Xu, S. Wen, F. Sun. Survey of imitation learning for robotic manipulation. *International Journal of Intelligent Robotics and Applications*, vol. 3, pp. 362–369, 2019. DOI: 10.1007/s41315-019-00103-5.

[33] A. Billard, D. Kragic. Trends and challenges in robot manipulation. *Science*, vol. 364, no. 6446, Article number eaat8414, 2019. DOI: 10.1126/science.aat8414.

[34] O. Kroemer, S. Niekum, G. Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *Journal of Machine Learning Research*, vol. 22, no. 30, pp. 1–82, 2021

[35] Y. Cong, R. Chen, B. Ma, H. Liu, D. Hou, C. Yang. A comprehensive study of 3-D vision-based robot manipulation. *IEEE Transactions on Cybernetics*, 2021. DOI: 10.1109/TCYB.2021.3108165.

[36] J. Cui, J. Trinkle. Toward next-generation learned robot manipulation. *Science Robotics*, vol. 6, no. 54, Article number eabd9461, 2021. DOI: 10.1126/scirobotics.abd9461.

[37] J. Zhu, A. Cherubini, C. Dune, D. Navarro-Alarcon, F. Alambeigi, D. Berenson, F. Ficuciello, K. Harada, J. Kober, X. Li, J. Pan, W. Yuan, M. Gienger. Challenges and outlook in robotic manipulation of deformable objects. *IEEE Robotics & Automation Magazine*, vol. 29, no. 3, pp. 67–77, 2022. DOI: 10.1109/MRA.2022.3147415.

[38] M. Q. Mohammed, L. C. Kwek, S. C. Chua, A. Al-Dhaqm, S. Nahavandi, T. A. E. Eisa, M. F. Miskon, M. N. Al-Mhiqani, A. Ali, M. Abaker, E. A. Alandoli. Review of learning-based robotic manipulation in cluttered environments. *Sensors*, vol. 22, no. 20, Article number 7938, 2022. DOI: 10.3390/s22207938.

[39] M. Suomalainen, Y. Karayiannidis, V. Kyrki. A survey of robot manipulation in contact. *Robotics and Autonomous Systems*, vol. 156, Article number 104224, 2022. DOI: 10.1016/j.robot.2022.104224.

[40] A. I. Weinberg, A. Shirizly, O. Azulay, A. Sintov. Survey of learning approaches for robotic in-hand manipulation, [Online], Available: https://arxiv.org/abs/2401.07915, 2024.

[41] B. Tekin, S. N. Sinha, P. Fua. Real-time seamless single shot 6D object pose prediction. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 292–301, 2018. DOI: 10.1109/CVPR.2018.00038.

[42] G. Zhai, D. Huang, S. C. Wu, H. Jung, Y. Di, F. Manhardt, F. Tombari, N. Navab, B. Busam. MonoGraspNet: 6-DoF grasping with a single RGB image. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, London, UK, pp. 1708–1714, 2023. DOI: 10.1109/ICRA48891.2023.10160779.

[43] C. Liu, K. Shi, K. Zhou, H. Wang, J. Zhang, H. Dong. RGBGrasp: Image-based object grasping by capturing multiple views during robot arm movement with neural radiance fields. *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 6012–6019, 2023. DOI: 10.1109/LRA.2024.3396101.

[44] B. An, Y. Geng, K. Chen, X. Li, Q. Dou, H. Dong. RGB-Manip: Monocular image-based robotic manipulation through active object pose estimation. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Yokohama, Japan, pp. 7748–7755, 2024. DOI: 10.1109/ICRA57147.2024.10610690.

[45] J. Kerr, L. Fu, H. Huang, Y. Avigal, M. Tancik, J. Ichnowski, A. Kanazawa, K. Goldberg. Evo-NeRF: Evolving NeRF for sequential robot grasping of transparent objects. In *Proceedings of the Conference on Robot Learning*, Auckland, New Zealand, pp. 353–367, 2023.

[46] A. Agrawal, R. Roy, B. P. Duisterhof, K. B. Hekkadka, H. Chen, J. Ichnowski. Clear-splatting: Learning residual Gaussian splats for transparent object manipulation. In *the 1st Workshop on Neural Fielda in Robotics at ICRA, Yukohama, Japan*, 2024.

[47] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Y. Lo, P. Dollár, R. Girshick. Segment anything. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Paris, France, pp. 3992–4003, 2023. DOI: 10.1109/ICCV51070.2023.00371.

[48] J. Redmon, A. Angelova. Real-time grasp detection using convolutional neural networks. In *Proceedings of the IEEE International Conference on Robotics and Automation*, IEEE, Seattle, USA, pp. 1316–1322, 2015. DOI: 10.1109/ICRA.2015.7139361.

[49] I. Lenz, H. Lee, A. Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, vol. 34, no. 4–5, pp. 705–724, 2015. DOI: 10.1177/0278364914549607.

[50] M. Schwarz, A. Milan, A. S. Periyasamy, S. Behnke. RGB-D object detection and semantic segmentation for autonomous manipulation in clutter. *The International Journal of Robotics Research*, vol. 37, no. 4–5, pp. 437–451, 2018. DOI: 10.1177/0278364917713117.

[51] S. Kumra, C. Kanan. Robotic grasp detection using deep convolutional neural networks. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Vancouver, Canada, pp. 769–776, 2017. DOI: 10.1109/IROS.2017.8202237.

[52] J. Varley, C. DeChant, A. Richardson, J. Ruales, P. Allen. Shape completion enabled robotic grasping. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Vancouver, Canada, pp. 2442–2447, 2017. DOI: 10.1109/IROS.2017.8206060.

[53] C. R. Qi, H. Su, K. C. Mo, L. J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of IEEE Conference on Computer*

Vision and Pattern Recognition, IEEE, Honolulu, USA, pp. 77–85, 2017. DOI: 10.1109/CVPR.2017.16.

[54] H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, J. Zhang. PointNetGPD: Detecting grasp configurations from point sets. In *Proceedings of International Conference on Robotics and Automation*, IEEE, Montreal, Canada, pp. 3629–3635, 2019. DOI: 10.1109/ICRA.2019.8794435.

[55] C. Zhong, Y. Zheng, Y. Zheng, H. Zhao, L. Yi, X. Mu, L. Wang, P. Li, G. Zhou, C. Yang, X. Zhang, J. Zhao. 3D implicit transporter for temporally consistent keypoint discovery. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Paris, France, pp. 3846–3857, 2023. DOI: 10.1109/ICCV51070.2023.00358.

[56] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y. W. Chao, D. Fox. RVT: Robotic view transformer for 3D object manipulation. In *Proceedings of the 7th Conference on Robot Learning*, Atlanta, USA, pp. 694–710, 2023.

[57] Y. Ze, G. Yan, Y. H. Wu, A. Macaluso, Y. Ge, J. Ye, N. Hansen, L. E. Li, X. Wang. GNFactor: Multi-task real robot learning with generalizable neural feature fields. In *Proceedings of the 7th Conference on Robot Learning*, Atlanta, USA, pp. 284–301, 2023.

[58] G. Lu, S. Zhang, Z. Wang, C. Liu, J. Lu, Y. Tang. ManiGaussian: Dynamic Gaussian splatting for multi-task robotic manipulation. In *Proceedings of the 18th European Conference on Computer Vision*, Springer, Milan, Italy, pp. 349–366, 2025. DOI: 10.1007/978-3-031-72761-0_20.

[59] Y. Li, D. Pathak. Object-aware Gaussian splatting for robotic manipulation. In *LCRA 2024 Workshop on 3D Visual Representations for Robot Manipulation, Yokohama, Japan*, 2024.

[60] W. Yuan, S. Dong, E. H. Adelson. GelSight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, vol. 17, no. 12, Article number 2762, 2017. DOI: 10.3390/s17122762.

[61] M. Lambeta, P. W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, R. Calandra. DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020. DOI: 10.1109/LRA.2020.2977257.

[62] O. Azulay, N. Curtis, R. Sokolovsky, G. Levitski, D. Slomovik, G. Lilling, A. Sintov. AllSight: A low-cost and high-resolution round tactile sensor with zero-shot learning capability. *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 483–490, 2024. DOI: 10.1109/LRA.2023.3333701.

[63] M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, Li F. F., A. Garg, J. Bohg. Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 582–596, 2020. DOI: 10.1109/TRO.2019.2959445.

[64] S. Hochreiter, J. Schmidhuber. Long short-term memory. *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[65] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, S. Levine. The feeling of success: Does touch sensing help predict grasp outcomes? In *Proceedings of the 1st Annual Conference on Robot Learning*, Mountain View, USA, pp. 314–323, 2017.

[66] I. Guzey, B. Evans, S. Chintala, L. Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play. In *Proceedings of the 7th Conference on Robot Learning*, Atlanta, USA, pp. 3142–3166, 2023.

[67] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, S. Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300–3307, 2018. DOI: 10.1109/LRA.2018.2852779.

[68] R. Gao, Y. Y. Chang, S. Mall, Li F. F., J. Wu. ObjectFolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *Proceedings of the 5th Conference on Robot Learning*, London, UK, pp. 466–476, 2021.

[69] S. Wang, M. Lambeta, P. W. Chou, R. Calandra. TACTO: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors. *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3930–3937, 2022. DOI: 10.1109/LRA.2022.3146945.

[70] J. Xu, S. Kim, T. Chen, A. R. Garcia, P. Agrawal, W. Matusik, S. Sueda. Efficient tactile simulation with differentiability for robotic manipulation. In *Proceedings of the 6th Conference on Robot Learning*, Auckland, New Zealand, pp. 1488–1498, 2022.

[71] S. Zhong, A. Albini, O. P. Jones, P. Maiolino, I. Posner. Touching a NeRF: Leveraging neural radiance fields for tactile sensory data generation. In *Proceedings of the 6th Conference on Robot Learning*, Auckland, New Zealand, pp. 1618–1628, 2022.

[72] Y. Dou, F. Yang, Y. Liu, A. Loquercio, A. Owens. Tactile-augmented radiance fields. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 26519–26529, 2024. DOI: 10.1109/CVPR52733.2024.02505.

[73] J. Liu, W. Sun, H. Yang, Z. Zeng, C. Liu, J. Zheng, X. Liu, H. Rahmani, N. Sebe, A. Mian. Deep learning-based object pose estimation: A comprehensive survey, [Online], Available: https://arxiv.org/abs/2405.07801, 2024.

[74] D. De Gregorio, R. Zanella, G. Palli, L. Di Stefano. Effective deployment of CNNs for 3DoF pose estimation and grasping in industrial settings. In *Proceedings of the 25th International Conference on Pattern Recognition*, IEEE, Milan, Italy, pp. 7419–7426, 2021. DOI: 10.1109/ICPR48806.2021.9411912.

[75] V. N. Nguyen, T. Groueix, M. Salzmann, V. Lepetit. GigaPose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 9903–9913, 2024. DOI: 10.1109/CVPR52733.2024.00945.

[76] Y. Xiang, T. Schmidt, V. Narayanan, D. Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In *Proceedings of the Robotics: Science and Systems*, Pittsburgh, USA, 2018. DOI: 10.15607/RSS.2018.XIV.019.

[77] Z. Li, G. Wang, X. Ji. CDPN: Coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 7677–7686, 2019. DOI: 10.1109/

ICCV.2019.00777.

[78] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, Li F. F. DenseFusion: 6D object pose estimation by iterative dense fusion. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 3338–3347, 2019. DOI: 10.1109/CVPR.2019.00346.

[79] S. Iwase, X. Y. Liu, R. Khirodkar, R. Yokota, K. M. Kitani. RePOSE: Fast 6D object pose refinement via deep texture rendering. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 3283–3292, 2021. DOI: 10.1109/ICCV48922.2021.00329.

[80] L. Xu, H. Qu, Y. Cai, J. Liu. 6D-Diff: A keypoint diffusion framework for 6D object pose estimation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 9676–9686, 2024. DOI: 10.1109/CVPR52733.2024.00924.

[81] X. S. Gao, X. R. Hou, J. Tang, H. F. Cheng. Complete solution classification for the perspective-three-point problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 930–943, 2003. DOI: 10.1109/TPAMI.2003.1217599.

[82] V. Lepetit, F. Moreno-Noguer, P. Fua. EPnP: An accurate $O(n)$ solution to the PnP problem. *International Journal of Computer Vision*, vol. 81, no. 2, pp. 155–166, 2009. DOI: 10.1007/s11263-008-0152-6.

[83] Z. Dang, L. Wang, Y. Guo, M. Salzmann. Match normalization: Learning-based point cloud registration for 6D object pose estimation in the real world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4489–4503, 2024. DOI: 10.1109/TPAMI.2024.3355198.

[84] J. Zhou, K. Chen, L. Xu, Q. Dou, J. Qin. Deep fusion transformer network with weighted vector-wise keypoints voting for robust 6D object pose estimation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Paris, France, pp. 13921–13931, 2023. DOI: 10.1109/ICCV51070.2023.01284.

[85] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. R. Song, L. J. Guibas. Normalized object coordinate space for category-level 6D object pose and size estimation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 2637–2646, 2019. DOI: 10.1109/CVPR.2019.00275.

[86] K. Chen, Q. Dou. SGPA: Structure-guided prior adaptation for category-level 6D object pose estimation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 2753–2762, 2021. DOI: 10.1109/ICCV48922.2021.00277.

[87] Y. Hai, R. Song, J. Li, M. Salzmann, Y. Hu. Rigidity-aware detection for 6D object pose estimation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Vancouver, Canada, pp. 8927–8936, 2023. DOI: 10.1109/CVPR52729.2023.00862.

[88] J. Corsetti, D. Boscaini, F. Poiesi. Revisiting fully convolutional geometric features for object 6D pose estimation. In *Proceedings of IEEE/CVF International Conference on Computer Vision Workshops*, IEEE, Paris, France, pp. 2095–2104, 2023. DOI: 10.1109/ICCVW60793.2023.00224.

[89] X. Li, H. Wang, L. Yi, L. J. Guibas, A. L. Abbott, S. R. Song. Category-level articulated object pose estimation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 3703–3712, 2020. DOI: 10.1109/CVPR42600.2020.00376.

[90] L. Liu, H. Xue, W. Xu, H. Fu, C. Lu. Toward real-world category-level articulation pose estimation. *IEEE Transactions on Image Processing*, vol. 31, pp. 1072–1083, 2022. DOI: 10.1109/TIP.2021.3138644.

[91] Y. Liu, Y. Wen, S. Peng, C. Lin, X. Long, T. Komura, W. Wang. Gen6D: Generalizable model-free 6-DoF object pose estimation from RGB images. In *Proceedings of the 17th European Conference on Computer Vision*, Springer, Tel Aviv, Israel, pp. 298–315, 2022. DOI: 10.1007/978-3-031-19824-3_18.

[92] J. Sun, Z. Wang, S. Zhang, X. He, H. C. Zhao, G. Zhang. OnePose: One-shot object pose estimation without cad models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 6815–6824, 2022. DOI: 10.1109/CVPR52688.2022.00670.

[93] W. Goodwin, S. Vaze, I. Havoutis, I. Posner. Zero-shot category-level object pose estimation. In *Proceedings of the 17th European Conference on Computer Vision*, Springer, Tel Aviv, Israel, pp. 516–532, 2022. DOI: 10.1007/978-3-031-19842-7_30.

[94] J. Lin, L. Liu, D. Lu, K. Jia. SAM-6D: Segment anything model meets zero-shot 6D object pose estimation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 27906–27916, 2024. DOI: 10.1109/CVPR52733.2024.02636.

[95] B. Wen, W. Yang, J. Kautz, S. Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 17868–17879, 2024. DOI: 10.1109/CVPR52733.2024.01692.

[96] Y. Li, N. Zhao, J. Xiao, C. Feng, X. Wang, T. S. Chua. LASO: Language-guided affordance segmentation on 3D object. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 14251–14260, 2024. DOI: 10.1109/CVPR52733.2024.01351.

[97] J. J. Gibson. The theory of affordances. *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, R. E Shaw, J. Bransford, Eds., Hilldale, USA: Lawrence Erlbaum Associates, pp. 67–82, 1977.

[98] T. T. Do, A. Nguyen, I. Reid. AffordanceNet: An end-to-end deep learning approach for object affordance detection. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Brisbane, Australia, pp. 5882–5889, 2018. DOI: 10.1109/ICRA.2018.8460902.

[99] Y. Zheng, H. Yao, X. Sun. Deep semantic parsing of freehand sketches with homogeneous transformation, soft-weighted loss, and staged learning. *IEEE Transactions on Multimedia*, vol. 23, pp. 3590–3602, 2021. DOI: 10.1109/TMM.2020.3028466.

[100] T. Nagarajan, C. Feichtenhofer, K. Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 8687–8696, 2019. DOI: 10.1109/ICCV.2019.00878.

[101] S. Bahl, R. Mendonca, L. Chen, U. Jain, D. Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Vancouver, Canada, pp. 13778–13790, 2023. DOI: 10.1109/CVPR52729.2023.01324.

[102] Y. Ju, K. Hu, G. Zhang, G. Zhang, M. Jiang, H. Xu. Robo-ABC: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *Proceedings of the 18th European Conference on Computer Vision*, Springer, Milan, Italy, pp. 222–239, 2025. DOI: 10.1007/978-3-031-72940-9_13.

[103] Y. Kuang, J. Ye, H. Geng, J. Mao, C. Deng, L. Guibas, H. Wang, Y. Wang. RAM: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation, [Online], Available: https://arxiv.org/abs/2407.04689, 2024.

[104] T. Nguyen, M. N. Vu, A. Vuong, D. Nguyen, T. Vo, N. Le, A. Nguyen. Open-vocabulary affordance detection in 3D point clouds. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Detroit, USA, pp. 5692–5698, 2023. DOI: 10.1109/IROS55552.2023.10341553.

[105] K. Mo, L. Guibas, M. Mukadam, A. Gupta, S. Tulsiani. Where2Act: From pixels to actions for articulated 3D objects. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 6793–6803, 2021. DOI: 10.1109/ICCV48922.2021.00674.

[106] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, L. Yi, A. X. Chang, L. J. Guibas, H. Su. SAPIEN: A simulated part-based interactive environment. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 11094–11104, 2020. DOI: 10.1109/CVPR42600.2020.01111.

[107] Y. Wang, R. Wu, K. Mo, J. Ke, Q. N. Fan, L. J. Guibas, H. Dong. AdaAfford: Learning to adapt manipulation affordance for 3D articulated objects via few-shot interactions. In *Proceedings of the 17th European Conference on Computer Vision*, Springer, Tel Aviv, Israel, pp. 90–107, 2022. DOI: 10.1007/978-3-031-19818-2_6.

[108] Y. Zhao, R. Wu, Z. Chen, Y. Zhang, Q. Fan, K. Mo, H. Dong. DualAfford: Learning collaborative visual affordance for dual-gripper manipulation. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[109] L. Wang, N. Dvornik, R. Dubeau, M. Mittal, A. Garg. Self-supervised learning of action affordances as interaction modes. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, London, UK, pp. 7279–7286, 2023. DOI: 10.1109/ICRA48891.2023.10161371.

[110] P. Mazzaglia, T. Cohen, D. Dijkman. Information-driven affordance discovery for efficient robotic manipulation. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Yokohama, Japan, pp. 7780–7787, 2024. DOI: 10.1109/ICRA57147.2024.10611170.

[111] C. Ning, R. Wu, H. Lu, K. Mo, H. Dong. Where2Explore: Few-shot affordance learning for unseen novel categories of articulated objects. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ACM, New Orleans, USA, Article number 203, 2023.

[112] S. Ling, Y. Wang, S. Wu, Y. Zhuang, T. Xu, Y. Li, C. Liu, H. Dong. Articulated object manipulation with coarse-to-fine affordance for mitigating the effect of point cloud noise. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Yokohama, Japan, pp. 10895–10901, 2024. DOI: 10.1109/ICRA57147.2024.10610593.

[113] R. Wu, K. Cheng, Y. Zhao, C. Ning, G. Zhan, H. Dong. Learning environment-aware affordance for 3D articulated object manipulation under occlusions. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ACM, New Orleans, USA, pp. 2664, 2023.

[114] Y. Geng, B. An, H. Geng, Y. Chen, Y. Yang, H. Dong. RLAfford: End-to-end affordance learning for robotic manipulation. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, London, UK, pp. 5880–5886, 2023. DOI: 10.1109/ICRA48891.2023.10161571.

[115] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, S. Levine. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Proceedings of the 2nd Annual Conference on Robot Learning*, Zürich, Switzerland, pp. 651–673, 2018.

[116] Y. Liang, K. Ellis, J. Henriques. Rapid motor adaptation for robotic manipulator arms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 16404–16413, 2024. DOI: 10.1109/CVPR52733.2024.01552.

[117] R. Boney, J. Kannala, A. Ilin. Regularizing model-based planning with energy-based models. In *Proceedings of the 3rd Annual Conference on Robot Learning*, Osaka, Japan, pp. 182–191, 2019.

[118] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, J. Tompson. Implicit behavioral cloning. In *Proceedings of the 5th Conference on Robot Learning*, London, UK, pp. 158–168, 2021.

[119] Y. Yue, Z. Wang, M. Zhou. Implicit distributional reinforcement learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ACM, Vancouver, Canada, Article number 599, 2020.

[120] M. Liu, T. He, M. Xu, W. Zhang. Energy-based imitation learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, ACM, pp. 809–817, 2021.

[121] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of the 19th Robotics: Science and Systems 2023*, Daegu, Republic of Korea, 2023.

[122] A. Ajay, Y. Du, A. Gupta, J. B. Tenenbaum, T. S. Jaakkola, P. Agrawal. Is conditional generative modeling all you need for decision making? In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[123] Z. Wang, J. J. Hunt, M. Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[124] X. Ma, S. Patidar, I. Haughton, S. James. Hierarchical diffusion policy for kinematics-aware multi-task robotic

manipulation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 18081–18090, 2024. DOI: 10.1109/CVPR52733.2024.01712.

[125] T. Wu, Y. Gan, M. Wu, J. B. Cheng, Y. Yang, Y. Zhu, H. Dong. UniDexFPM: Universal dexterous functional pregrasp manipulation via diffusion policy, [Online], Available: https://arxiv.org/abs/2403.12421, 2024.

[126] M. Reuss, M. Li, X. Jia, R. Lioutikov. Goal-conditioned imitation learning using score-based diffusion policies. In *Proceedings of the 19th Robotics: Science and Systems 2023*, Daegu, Republic of Korea, 2023.

[127] M. Reuss, Ö. E. Yağmurlu, F. Wenzel, R. Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals, [Online], Available: https://arxiv.org/abs/2407.05996, 2024.

[128] H. Li, Q. Feng, Z. Zheng, J. Feng, Z. Chen, A. Knoll. Language-guided object-centric diffusion policy for collision-aware robotic manipulation, [Online], Available: https://arxiv.org/abs/2407.00451, 2024.

[129] T. Huang, K. Chen, W. Wei, J. Li, Y. Long, Q. Dou. Value-informed skill chaining for policy learning of long-horizon tasks with surgical robot. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Detroit, USA, pp. 8495–8501, 2023. DOI: 10.1109/IROS55552.2023.10342180.

[130] J. Lv, Y. Feng, C. Zhang, S. Zhao, L. Shao, C. Lu. SAM-RL: Sensing-aware model-based reinforcement learning via differentiable physics-based simulation and rendering. In *Proceedings of the 19th Robotics: Science and Systems 2023*, Daegu, Republic of Korea, 2023.

[131] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, Li F. F., S. Savarese, Y. Zhu, R. Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Proceedings of the 5th Conference on Robot Learning*, London, UK, pp. 1678–1690, 2021.

[132] S. Levine, A. Kumar, G. Tucker, J. Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems, [Online], Available: https://arxiv.org/abs/2005.01643, 2020.

[133] T. Huang, K. Chen, B. Li, Y. H. Liu, Q. Dou. Demonstration-guided reinforcement learning with efficient exploration for task automation of surgical robot. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, London, UK, pp. 4640–4647, 2023. DOI: 10.1109/ICRA48891.2023.10160327.

[134] T. Xie, S. Zhao, C. H. Wu, Y. Liu, Q. Luo, V. Zhong, Y. Yang, T. Yu. Text2Reward: Automated dense reward function generation for reinforcement learning, [Online], Available: https://arxiv.org/abs/2309.11489, 2023.

[135] Y. J. Ma, W. Liang, G. Wang, D. A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, A. Anandkumar. Eureka: Human-level reward design via coding large language models. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, 2024.

[136] S. Schaal, J. Peters, J. Nakanishi, A. Ijspeert. Learning movement primitives. In *Proceedings of the 11th International Symposium on Robotics Research*, Springer, Berlin, Germany, pp. 561–572, 2005. DOI: 10.1007/11008941_60.

[137] S. Ross, G. J. Gordon, D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, USA, pp. 627–635, 2011.

[138] X. Lin, J. So, S. Mahalingam, F. Liu, P. Abbeel. SpawnNet: Learning generalizable visuomotor skills from pre-trained networks, [Online], Available: https://arxiv.org/abs/2307.03567, 2023.

[139] T. Z. Zhao, V. Kumar, S. Levine, C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of the 19th Robotics: Science and Systems 2023*, Daegu, Republic of Korea, 2023.

[140] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. S. Narang, L. Fan, Y. Zhu, D. Fox. MimicGen: A data generation system for scalable robot learning using human demonstrations. In *Proceedings of the 7th Conference on Robot Learning*, Atlanta, USA, pp. 1820–1864, 2023.

[141] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, S. Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. In *Proceedings of the 18th Robotics: Science and Systems 2022*, New York, USA, 2022.

[142] Open X-Embodiment Collaboration. Open X-embodiment: Robotic learning datasets and RT-X models, [Online], Available: https://arxiv.org/abs/2310.08864, 2023.

[143] V. Jain, M. Attarian, N. J. Joshi, A. Wahid, D. Driess, Q. Vuong, P. R. Sanketi, P. Sermanet, S. Welker, C. Chan, I. Gilitschenski, Y. Bisk, D. Dwibedi. Vid2Robot: End-to-end video-conditioned policy learning with cross-attention transformers, [Online], Available: https://arxiv.org/abs/2403.12943, 2024.

[144] P. Li, T. Liu, Y. Li, M. Han, H. Geng, S. Wang, Y. Zhu, S. C. Zhu, S. Huang. Ag2Manip: Learning novel manipulation skills with agent-agnostic visual and action representations. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Abu Dhabi, UAE, pp. 573–580, 2024. DOI: 10.1109/IROS58592.2024.10801835.

[145] J. Zeng, Q. Bu, B. Wang, W. Xia, L. Chen, H. Dong, H. Song, D. Wang, D. Hu, P. Luo, H. Cui, B. Zhao, X. Li, Y. Qiao, H. Li. Learning manipulation by predicting interaction, [Online], Available: https://arxiv.org/abs/2406.00439, 2024.

[146] A. Simeonov, Y. Du, Y. C. Lin, A. R. Garcia, L. P. Kaelbling, T. Lozano-Pérez, P. Agrawal. Se(3)-equivariant relational rearrangement with neural descriptor fields. In *Proceedings of the 6th Conference on Robot Learning*, Auckland, New Zealand, pp. 835–846, 2022.

[147] E. Chun, Y. Du, A. Simeonov, T. Lozano-Perez, L. Kaelbling. Local neural descriptor fields: Locally conditioned object representations for manipulation. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, London, UK, pp. 1830–1836, 2023. DOI: 10.1109/ICRA48891.2023.10160423.

[148] H. Ryu, H. I. Lee, J. H. Lee, J. Choi. Equivariant descriptor fields: Se(3)-equivariant energy-based models for end-to-end visual robotic manipulation learning. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[149] J. Brehmer, J. Bose, P. De Haan, T. Cohen. EDGI: Equivariant diffusion for planning with embodied agents. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ACM, New Or-

leans, USA, Article number 2787, 2023.

[150] H. Ryu, J. Kim, H. An, J. Chang, J. Seo, T. Kim, Y. Kim, C. Hwang, J. Choi, R. Horowitz. Diffusion-EDFs: Bi-equivariant denoising generative modeling on SE(3) for visual robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 18007–18018, 2024. DOI: 10.1109/CVPR52733.2024.01705.

[151] J. Urain, N. Funk, J. Peters, G. Chalvatzaki. Se(3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion. In *Proceedings of the IEEE International Conference on Robotics and Automation*, IEEE, London, UK, pp. 5923–5930, 2023. DOI: 10.1109/ICRA48891.2023.10161569.

[152] Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. C. Wang, H. R. Geng, Y. J. Weng, J. Y. Chen, T. Y. Liu, L. Yi, H. Wang. UniDexGrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Vancouver, Canada, pp. 4737–4746, 2023. DOI: 10.1109/CVPR52729.2023.00459.

[153] W. Wan, H. Geng, Y. Liu, Z. Shan, Y. Yang, L. Yi, H. Wang. UniDexGrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Paris, France, pp. 3868–3879, 2023. DOI: 10.1109/ICCV51070.2023.00360.

[154] Y. Hu, F. Lin, T. Zhang, L. Yi, Y. Gao. Look before you leap: Unveiling the power of GPT-4V in robotic vision-language planning, [Online], Available: https://arxiv.org/abs/2311.17842, 2023.

[155] A. Szot, B. Mazoure, H. Agrawal, D. Hjelm, Z. Kira, A. Toshev. Grounding multimodal large language models in actions, [Online], Available: https://arxiv.org/abs/2406.07904, 2024.

[156] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, C. Finn. OpenVLA: An open-source vision-language-action model, [Online], Available: https://arxiv.org/abs/2406.09246, 2024.

[157] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, C. Gan. 3D-VLA: A 3D vision-language-action generative world model. In *Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria, 2024.

[158] X. Zhang, S. Jin, C. Wang, X. Zhu, M. Tomizuka. Learning insertion primitives with discrete-continuous hybrid action space for robotic assembly tasks. In *Proceedings of the International Conference on Robotics and Automation*, IEEE, Philadelphia, USA, pp. 9881–9887, 2022. DOI: 10.1109/ICRA46639.2022.9811973.

[159] A. Van Den Oord, Y. Li, O. Vinyals. Representation learning with contrastive predictive coding, [Online], Available: https://arxiv.org/abs/1807.03748, 2018.

[160] F. Torabi, G. Warnell, P. Stone. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, pp. 4950–4957, 2018.

[161] J. Ho, S. Ermon. Generative adversarial imitation learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ACM, Barcelona, Spain, pp. 4572–4580, 2016.

[162] M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, W. Zaremba. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020. DOI: 10.1177/0278364919887447.

[163] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov. Proximal policy optimization algorithms, [Online], Available: https://arxiv.org/abs/1707.06347, 2017.

[164] J. Fu, S. Levine, P. Abbeel. One-shot learning of manipulation skills with online dynamics adaptation and neural network priors. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Daejeon, Republic of Korea, pp. 4019–4026, 2016. DOI: 10.1109/IROS.2016.7759592.

[165] B. D. Ziebart, A. Maas, J. A. Bagnell, A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence*, Chicago, USA, pp. 1433–1438, 2008.

[166] H. Kim, Y. Ohmura, Y. Kuniyoshi. Transformer-based deep imitation learning for dual-arm robot manipulation. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Prague, Czech Republic, pp. 8965–8972, 2021. DOI: 10.1109/IROS51168.2021.9636301.

[167] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, V. Sitzmann. Neural descriptor fields: Se(3)-equivariant object representations for manipulation. In *Proceedings of the International Conference on Robotics and Automation*, IEEE, Philadelphia, USA, pp. 6394–6400, 2022. DOI: 10.1109/ICRA46639.2022.9812146.

[168] U. Asif, M. Bennamoun, F. A. Sohel. RGB-D object recognition and grasp detection using hierarchical cascaded forests. *IEEE Transactions on Robotics*, vol. 33, no. 3, pp. 547–564, 2017. DOI: 10.1109/TRO.2016.2638453.

[169] D. H. Zhai, S. Yu, Y. Xia. FANet: Fast and accurate robotic grasp detection based on keypoints. *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 3, pp. 2974–2986, 2024. DOI: 10.1109/TASE.2023.3272664.

[170] H. S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, C. Lu. AnyGrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3929–3945, 2023. DOI: 10.1109/TRO.2023.3281153.

[171] N. Marturi, M. Kopicki, A. Rastegarpanah, V. Rajasekaran, M. Adjigble, R. Stolkin, A. Leonardis, Y. Bekiroglu. Dynamic grasp and trajectory planning for moving objects. *Autonomous Robots*, vol. 43, no. 5, pp. 1241–1256, 2019. DOI: 10.1007/s10514-018-9799-1.

[172] D. Morrison, J. Leitner, P. Corke. Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. In *Proceedings of the 14th Robotics: Science and Systems Conference*, Pittsburgh, USA, 2018.

[173] P. Piacenza, J. Yuan, J. Huh, V. Isler. Vfas-grasp: Closed loop grasping with visual feedback and adaptive sampling. In *Proceedings of IEEE International Conference on Rubotics and Automation*, Yokhama, Japan,

pp. 4126–4132, 2024. DOI: 10.1109/ICRA57147.2024.10611183.

[174] H. Fang, H. S. Fang, S. Xu, C. Lu. TransCG: A large-scale real-world dataset for transparent object depth completion and a grasping baseline. *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7383–7390, 2022. DOI: 10.1109/LRA.2022.3183256.

[175] J. Ichnowski, Y. Avigal, J. Kerr, K. Goldberg. Dex-neRF: Using a neural radiance field to grasp transparent objects. In *Proceedings of the 5th Conference on Robot Learning*, London, UK, pp. 526–536, 2021.

[176] Q. Dai, Y. Zhu, Y. Geng, C. Ruan, J. Zhang, H. Wang. GraspNeRF: Multiview-based 6-DoF grasp detection for transparent and specular objects using generalizable NeRF. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, London, UK, pp. 1757–1763, 2023. DOI: 10.1109/ICRA48891.2023.10160842.

[177] J. Lee, S. M. Kim, Y. Lee, Y. M. Kim. NFL: Normal field learning for 6-DoF grasping of transparent objects. *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 819–826, 2024. DOI: 10.1109/LRA.2023.3336108.

[178] J. Kim, M. H. Jeon, S. Jung, W. Yang, M. Jung, J. Shin, A. Kim. Transpose: Large-scale multispectral dataset for transparent object. *The International Journal of Robotics Research*, vol. 43, no. 6, pp. 731–738, 2024. DOI: 10.1177/02783649231213117.

[179] H. Yu, S. Li, H. Liu, C. Xia, W. Ding, B. Liang. TGF-Net: Sim2Real transparent object 6D pose estimation based on geometric fusion. *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3868–3875, 2023. DOI: 10.1109/LRA.2023.3268041.

[180] M. Sundermeyer, A. Mousavian, R. Triebel, D. Fox. Contact-graspNet: Efficient 6-DoF grasp generation in cluttered scenes. In *Proceedings of the IEEE International Conference on Robotics and Automation*, IEEE, Xi'an, China, pp. 13438–13444, 2021. DOI: 10.1109/ICRA48506.2021.9561877.

[181] B. Wen, W. Lian, K. Bekris, S. Schaal. CaTGrasp: Learning category-level task-relevant grasping in clutter from simulation. In *Proceedings of the International Conference on Robotics and Automation*, IEEE, Philadelphia, USA, pp. 6401–6408, 2022. DOI: 10.1109/ICRA46639.2022.9811568.

[182] A. Murali, A. Mousavian, C. Eppner, C. Paxton, D. Fox. 6-DOF grasping for target-driven object manipulation in clutter. In *Proceedings of the IEEE International Conference on Robotics and Automation*, IEEE, Paris, France, pp. 6232–6238, 2020. DOI: 10.1109/ICRA40945.2020.9197318.

[183] J. Lundell, F. Verdoja, V. Kyrki. DDGC: Generative deep dexterous grasping in clutter. *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6899–6906, 2021. DOI: 10.1109/LRA.2021.3096239.

[184] C. Wang, H. S. Fang, M. Gou, H. Fang, J. Gao, C. Lu. Graspness discovery in clutters for fast and accurate grasp detection. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada pp. 15944–15953, 2021. DOI: 10.1109/ICCV48922.2021.01566.

[185] B. Wei, X. Ye, C. Long, Z. Du, B. Li, B. Yin, X. Yang. Discriminative active learning for robotic grasping in cluttered scene. *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1858–1865, 2023. DOI: 10.1109/LRA.2023.3243474.

[186] K. Xu, H. Yu, Q. Lai, Y. Wang, R. Xiong. Efficient learning of goal-oriented push-grasping synergy in clutter. *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6337–6344, 2021. DOI: 10.1109/LRA.2021.3092640.

[187] M. Kiatos, S. Malassiotis. Robust object grasping in clutter via singulation. In *Proceedings of the International Conference on Robotics and Automation*, IEEE, Montreal, Canada, pp. 1596–1600, 2019. DOI: 10.1109/ICRA.2019.8793972.

[188] Y. Yang, H. Liang, C. Choi. A deep learning approach to grasping the invisible. *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2232–2239, 2020. DOI: 10.1109/LRA.2020.2970622.

[189] K. Xu, S. Zhao, Z. Zhou, Z. Li, H. Pi, Y. Zhu, Y. Wang, R. Xiong. A joint modeling of vision-language-action for target-oriented grasping in clutter. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, London, UK, pp. 11597–11604, 2023. DOI: 10.1109/ICRA48891.2023.10161041.

[190] W. Wang, R. Li, Z. M. Diekel, Y. Chen, Z. Zhang, Y. Jia. Controlling object hand-over in human–robot collaboration via natural wearable sensing. *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 1, pp. 59–71, 2019. DOI: 10.1109/THMS.2018.2883176.

[191] W. Wang, R. Li, Y. Chen, Y. Sun, Y. Jia. Predicting human intentions in human–robot hand-over tasks through multimodal learning. *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 3, pp. 2339–2353, 2022. DOI: 10.1109/TASE.2021.3074873.

[192] W. Yang, C. Paxton, A. Mousavian, Y. W. Chao, M. Cakmak, D. Fox. Reactive human-to-robot handovers of arbitrary objects. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Xi'an, China, pp. 3118–3124, 2021. DOI: 10.1109/ICRA48506.2021.9561170.

[193] G. Zhang, H. S. Fang, H. J. Fang, C. W. Lu. Flexible handover with real-time robust dynamic grasp trajectory generation. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Detroit, USA, pp. 3192–3199, 2023. DOI: 10.1109/IROS55552.2023.10341777.

[194] Z. Wang, J. Chen, Z. Chen, P. Xie, R. Chen, L. Yi. GenH2R: Learning generalizable human-to-robot handover via scalable simulation, demonstration, and imitation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 16362–16372, 2024. DOI: 10.1109/CVPR52733.2024.01548.

[195] X. Ye, S. Liu. Velocity decomposition based planning algorithm for grasping moving object. In *Proceedings of the 7th Data Driven Control and Learning Systems Conference*, IEEE, Enshi, China, pp. 644–649, 2018. DOI: 10.1109/DDCLS.2018.8516083.

[196] I. Akinola, J. Xu, S. Song, P. K. Allen. Dynamic grasping with reachability and motion awareness. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Prague, Czech Republic, pp. 9422–9429, 2021. DOI: 10.1109/IROS51168.2021.9636057.

[197] J. Liu, R. Zhang, H. S. Fang, M. Gou, H. Fang, C. Wang, S. Xu, H. Yan, C. Lu. Target-referenced reactive grasp-

ing for dynamic objects. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Vancouver, Canada, pp. 8824–8833, 2023. DOI: 10.1109/CVPR52729.2023.00852.

[198] W. C. Agboh, J. Ichnowski, K. Goldberg, M. R. Dogar. Multi-object grasping in the plane. In *Proceedings of the 20th International Symposium on Robotics Research*, Springer, Cham, Germany, pp. 222–238, 2023. DOI: 10.1007/978-3-031-25555-7_15.

[199] W. C. Agboh, S. Sharma, K. Srinivas, M. Parulekar, G. Datta, T. Qiu, J. Ichnowski, E. Solowjow, M. Dogar, K. Goldberg. Learning to efficiently plan robust frictional multi-object grasps. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Detroit, USA, pp. 10660–10667, 2023. DOI: 10.1109/IROS55552.2023.10341895.

[200] T. Chen, Y. Sun. Multi-object grasping–experience forest for robotic finger movement strategies. *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5222–5229, 2024. DOI: 10.1109/LRA.2024.3389815.

[201] S. Aeron, E. Llontop, A. Adler, W. C. Agboh, M. Dogar, K. Goldberg. Push-MOG: Efficient pushing to consolidate polygonal objects for multi-object grasping. In *Proceedings of the 19th International Conference on Automation Science and Engineering*, IEEE, Auckland, New Zealand, 2023. DOI: 10.1109/CASE56687.2023.10260295.

[202] K. Yao, A. Billard. Exploiting kinematic redundancy for robotic grasping of multiple objects. *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 1982–2002, 2023. DOI: 10.1109/TRO.2023.3253249.

[203] Y. Li, B. Liu, Y. Geng, P. Li, Y. Yang, Y. Zhu, T. Liu, S. Huang. Grasp multiple objects with one hand. *IEEE Robotics and Automation Letters*, vol. 9, no. 5, pp. 4027–4034, 2024. DOI: 10.1109/LRA.2024.3374190.

[204] L. Berscheid, P. Meißner, T. Kröger. Self-supervised learning for precise pick-and-place without object model. *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4828–4835, 2020. DOI: 10.1109/LRA.2020.3003865.

[205] G. Zhai, X. Cai, D. Huang, Y. Di, F. Manhardt, F. Tombari, N. Navab, B. Busam. SG-Bot: Object rearrangement via coarse-to-fine robotic imagination on scene graphs. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Yokohama, Japan, pp. 4303–4310, 2024. DOI: 10.1109/ICRA57147.2024.10610792.

[206] K. Zakka, A. Zeng, J. Lee, S. Song. Form2Fit: Learning shape priors for generalizable assembly from disassembly. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Paris, France, pp. 9404–9410, 2020. DOI: 10.1109/ICRA40945.2020.9196733.

[207] Z. Huang, X. Lin, D. Held. Mesh-based dynamics with occlusion reasoning for cloth manipulation. In *Proceedings of the 18th Robotics: Science and Systems Conference*, New York, USA, 2022.

[208] P. Mitrano, D. McConachie, D. Berenson. Learning where to trust unreliable models in an unstructured world for deformable object manipulation. *Science Robotics*, vol. 6, no. 54, Article number eabd8170, 2021. DOI: 10.1126/scirobotics.abd8170.

[209] Y. Li, S. Li, V. Sitzmann, P. Agrawal, A. Torralba. 3D neural scene representations for visuomotor control. In *Proceedings of the 5th Conference on Robot Learning*, London, UK, pp. 112–123, 2021.

[210] H. Geng, H. Xu, C. Zhao, C. Xu, L. Yi, S. Huang, H. Wang. GAPartNet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Vancouver, Canada, pp. 7081–7091, 2023. DOI: 10.1109/CVPR52729.2023.00684.

[211] Y. Li, X. Zhang, R. Wu, Z. Zhang, Y. Geng, H. Dong, Z. F. He. UniDoorManip: Learning universal door manipulation policy over large-scale and diverse door manipulation environments, [Online], Available: https://arxiv.org/abs/2403.02604, 2024.

[212] H. Geng, Z. Li, Y. Geng, J. Chen, H. Dong, H. Wang. PartManip: Learning cross-category generalizable part manipulation policy from point cloud observations. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Vancouver, Canada, pp. 2978–2988, 2023. DOI: 10.1109/CVPR52729.2023.00291.

[213] B. Sundaralingam, T. Hermans. Relaxed-rigidity constraints: Kinematic trajectory optimization and collision avoidance for in-grasp manipulation. *Autonomous Robots*, vol. 43, no. 2, pp. 469–483, 2019. DOI: 10.1007/s10514-018-9772-z.

[214] D. Rus. In-hand dexterous manipulation of piecewise-smooth 3-D objects. *International Journal of Robotics Research*, vol. 18, no. 4, pp. 355–381, 1999. DOI: 10.1177/02783649922066268.

[215] A. Nagabandi, K. Konolige, S. Levine, V. Kumar. Deep dynamics models for learning dexterous manipulation. In *Proceedings of the 3rd Annual Conference on Robot Learning*, Osaka, Japan, pp. 1101–1112, 2019.

[216] T. Chen, J. Xu, P. Agrawal. A system for general in-hand object re-orientation. In *Proceedings of the 5th Conference on Robot Learning*, London, UK, pp. 297–307, 2021.

[217] S. P. Arunachalam, S. Silwal, B. Evans, L. Pinto. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, London, UK, pp. 5954–5961, 2023. DOI: 10.1109/ICRA48891.2023.10160275.

[218] S. Li, Z. Huang, T. Chen, T. Du, H. Su, J. B. Tenenbaum, C. Gan. DexDeform: Dexterous deformable object manipulation with human demonstrations and differentiable physics. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[219] Z. Qin, K. Fang, Y. Zhu, Li F. F., S. Savarese. KETO: Learning keypoint representations for tool manipulation. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Paris, France, pp. 7278–7285, 2020. DOI: 10.1109/ICRA40945.2020.9196971.

[220] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, Li F. F., S. Savarese. Learning task-oriented grasping for tool manipulation from simulated self-supervision. *The International Journal of Robotics Research*, vol. 39, no. 2–3, pp. 202–216, 2020. DOI: 10.1177/0278364919872545.

[221] X. Lin, Z. Huang, Y. Li, J. B. Tenenbaum, D. Held, C. Gan. Diffskill: Skill abstraction from differentiable physics for deformable object manipulations with tools. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[222] K. P. Tee, S. Cheong, J. Li, G. Ganesh. A framework for tool cognition in robots without prior tool learning or ob-

servation. *Nature Machine Intelligence*, vol. 4, no. 6, pp. 533–543, 2022. DOI: 10.1038/s42256-022-00500-9.

[223] A. Z. Ren, B. Govil, T. Y. Yang, K. R. Narasimhan, A. Majumdar. Leveraging language for accelerated learning of tool manipulation. In *Proceedings of the 6th Conference on Robot Learning*, Auckland, New Zealand, pp. 1531–1541, 2022.

[224] M. Xu, P. Huang, W. Yu, S. Liu, X. Zhang, Y. Niu, T. Zhang, F. Xia, J. Tan, D. Zhao. Creative robot tool use with large language models, [Online], Available: https://arxiv.org/abs/2310.13065, 2023.

[225] E. Jang, S. Vijayanarasimhan, P. Pastor, J. Ibarz, S. Levine. End-to-end learning of semantic grasping. In *Proceedings of the 1st Annual Conference on Robot Learning*, Mountain View, USA, pp. 119–132, 2017.

[226] R. Gong, J. Huang, Y. Zhao, H. Geng, X. Gao, Q. Wu, W. Ai, Z. Zhou, D. Terzopoulos, S. C. Zhu, B. Jia, S. Huang. ARNOLD: A benchmark for language-grounded task learning with continuous states in realistic 3D scenes. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Paris, France, pp. 20426–20438, 2023. DOI: 10.1109/ICCV51070.2023.01873.

[227] H. Sun, Z. Zhang, H. Wang, Y. Wang, Q. Cao. A novel robotic grasp detection framework using low-cost rgb-d camera for industrial bin picking. *IEEE Transactions on Instrumentation and Measurement*, vol. 73, Article number 2513212, 2024. DOI: 10.1109/TIM.2023.3346531.

[228] S. Ge, B. Hou, W. Zhu, Y. Zhu, S. Lu, Y. Zheng. Pixel-level collision-free grasp prediction network for medical test tube sorting on cluttered trays. *IEEE Robotics and Automation Letters*, vol. 8, no. 12, pp. 7897–7904, 2023. DOI: 10.1109/LRA.2023.3322896.

[229] S. D'Avella, M. Bianchi, A. M. Sundaram, C. A. Avizzano, M. A. Roa, P. Tripicchio. The cluttered environment picking benchmark (CEPB) for advanced warehouse automation: Evaluating the perception, planning, control, and grasping of manipulation systems. *IEEE Robotics & Automation Magazine*, vol. 31, no. 4, pp. 45–58, 2023. DOI: 10.1109/MRA.2023.3310861.

[230] J. Jiang, G. Cao, J. Deng, T. T. Do, S. Luo. Robotic perception of transparent objects: A review. *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 6, pp. 2547–2567, 2024. DOI: 10.1109/TAI.2023.3326120.

[231] D. Morrison, P. Corke, J. Leitner. Learning robust, real-time, reactive robotic grasping. *The International Journal of Robotics Research*, vol. 39, no. 2–3, pp. 183–201, 2020. DOI: 10.1177/0278364919859066.

[232] Y. Sun, E. Amatova, T. Chen. Multi-object grasping-types and taxonomy. In *Proceedings of the International Conference on Robotics and Automation*, IEEE, Philadelphia, USA, pp. 777–783, 2022. DOI: 10.1109/ICRA46639.2022.9812388.

[233] M. Zhu, K. G. Derpanis, Y. Yang, S. Brahmbhatt, M. Zhang, C. Phillips, M. Lecce, K. Daniilidis. Single image 3D object detection and pose estimation for grasping. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Hong Kong, China, pp. 3936–3943, 2014. DOI: 10.1109/ICRA.2014.6907430.

[234] E. Frazzoli, M. A. Dahleh, E. Feron. Maneuver-based motion planning for nonlinear systems with symmetries. *IEEE Transactions on Robotics*, vol. 21, no. 6, pp. 1077–1091, 2005. DOI: 10.1109/TRO.2005.852260.

[235] Y. Zhu, P. Stone, Y. Zhu. Bottom-up skill discovery from unsegmented demonstrations for long-horizon robot manipulation. *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4126–4133, 2022. DOI: 10.1109/LRA.2022.3146589.

[236] R. Jangir, N. Hansen, S. Ghosal, M. Jain, X. Wang. Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3046–3053, 2022. DOI: 10.1109/LRA.2022.3144512.

[237] J. Liu, C. Li, G. Wang, L. Lee, K. Zhou, S. Chen, C. Xiong, J. Ge, R. Zhang, S. Zhang. Self-corrected multimodal large language model for end-to-end robot manipulation, [Online], Available: https://arxiv.org/abs/2405.17418, 2024.

[238] A. M. Okamura, N. Smaby, M. R. Cutkosky. An overview of dexterous manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, IEEE, San Francisco, USA, pp. 255–262, 2000. DOI: 10.1109/ROBOT.2000.844067.

[239] H. Zhang, S. Christen, Z. Fan, O. Hilliges, J. Song. GraspXL: Generating grasping motions for diverse objects at scale. In *Proceedings of the 18th IEEE/CVF Conference on Computer Vision*, Springer, Milan, Italy, pp. 386–403, 2025. DOI: 10.1007/978-3-031-73347-5_22.

[240] M. Qin, J. Brawer, B. Scassellati. Robot tool use: A survey. *Frontiers in Robotics and AI*, vol. 9, Article number 1009488, 2023. DOI: 10.3389/frobt.2022.1009488.

[241] Y. Shirai, D. K. Jha, A. U. Raghunathan, D. Hong. Tactile tool manipulation. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, London, UK, pp. 12597–12603, 2023. DOI: 10.1109/ICRA48891.2023.10160480.

[242] Y. Zhu, Y. Zhao, S. C. Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 2855–2864, 2015. DOI: 10.1109/CVPR.2015.7298903.

[243] N. Saito, T. Ogata, S. Funabashi, H. Mori, S. Sugano. How to select and use tools?: Active perception of target objects using multimodal deep learning. *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2517–2524, 2021. DOI: 10.1109/LRA.2021.3062004.

[244] Y. Jiang, S. Moseson, A. Saxena. Efficient grasping from RGBD images: Learning using a new rectangle representation. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Shanghai, China, pp. 3304–3311, 2011. DOI: 10.1109/ICRA.2011.5980145.

[245] F. J. Chu, R. Xu, P. A. Vela. Real-world multiobject, multigrasp detection. *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3355–3362, 2018. DOI: 10.1109/LRA.2018.2852777.

[246] A. Depierre, E. Dellandréa, L. Chen. Jacquard: A large scale dataset for robotic grasp detection. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Madrid, Spain, pp. 3511–3516, 2018. DOI: 10.1109/IROS.2018.8593950.

[247] X. Yan, J. Hsu, M. Khansari, Y. F. Bai, A. Pathak, A. Gupta, J. Davidson, H. Lee. Learning 6-DOF grasping interaction via deep geometry-aware 3D representations. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Brisbane, Australia, pp. 3766–3773, 2018. DOI: 10.1109/ICRA.2018.8460609.

[248] C. Eppner, A. Mousavian, D. Fox. ACRONYM: A large-scale grasp dataset based on simulation. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Xi'an, China, pp. 6222–6227, 2021. DOI: 10.1109/ICRA48506.2021.9560844.

[249] D. Morrison, P. Corke, J. Leitner. EGAD! An evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation. *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4368–4375, 2020. DOI: 10.1109/LRA.2020.2992195.

[250] H. S. Fang, C. Wang, M. Gou, C. Lu. GraspNet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 11441–11450, 2020. DOI: 10.1109/CVPR42600.2020.01146.

[251] A. D. Vuong, M. N. Vu, H. Le, B. Huang, H. T. T. Binh, T. Vo, A. Kugi, A. Nguyen. Grasp-anything: Large-scale grasp dataset from foundation models. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Yokohama, Japan, pp. 14030–14037, 2024. DOI: 10.1109/ICRA57147.2024.10611277.

[252] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, A. M. Dollar. Yale-CMU-Berkeley dataset for robotic manipulation research. *The International Journal of Robotics Research*, vol. 36, no. 3, pp. 261–268, 2017. DOI: 10.1177/0278364917700714.

[253] L. Liu, W. Xu, H. Fu, S. Qian, Q. Yu, Y. Han, C. Lu. AKB-48: A real-world articulated object knowledge base. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 14789–14798, 2022. DOI: 10.1109/CVPR52688.2022.01439.

[254] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao, X. Yuan, P. Xie, Z. Huang, R. Chen, H. Su. ManiSkill2: A unified benchmark for generalizable manipulation skills. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[255] Y. Chen, Y. Geng, F. Zhong, J. Ji, J. Jiang, Z. Lu, H. Dong, Y. Yang. Bi-DexHands: Towards human-level bimanual dexterous manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2804–2818, 2024. DOI: 10.1109/TPAMI.2023.3339515.

[256] C. Bao, H. Xu, Y. Qin, X. Wang. DexArt: Benchmarking generalizable dexterous manipulation with articulated objects. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Vancouver, Canada, pp. 21190–21200, 2023. DOI: 10.1109/CVPR52729.2023.02030.

[257] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, W. Ai, B. Martinez, H. Yin, M. Lingelbach, M. Hwang, A. Hiranaka, S. Garlanka, A. Aydin, S. Lee, J. Sun, M. Anvari, M. Sharma, D. Bansal, S. Hunter, K. Y. Kim, A. Lou, C. R. Matthews, I. Villa-Renteria, J. H. Tang, C. Tang, F. Xia, Y. Z. Li, S. Savarese, H. Gweon, C. K. Liu, J. Wu, Li F. F. BEHAVIOR-1K: A human-centered, embodied AI benchmark with 1,000 everyday activities and realistic simulation, [Online], Available: https://arxiv.org/abs/2403.09227, 2024.

[258] A. Saxena, J. Driemeyer, A. Y. Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 157–173, 2008. DOI: 10.1177/0278364907087172.

[259] M. Kyrarini, M. A. Haseeb, D. Ristić-Durrant, A. Gräser. Robot learning of industrial assembly task via human demonstrations. *Autonomous Robots*, vol. 43, no. 1, pp. 239–257, 2019. DOI: 10.1007/s10514-018-9725-6.

[260] O. Koca, O. T. Kaymakci, M. Mercimek. Advanced predictive maintenance with machine learning failure estimation in industrial packaging robots. In *Proceedings of the International Conference on Development and Application Systems*, IEEE, Suceava, Romania, 2020. DOI: 10.1109/DAS49615.2020.9108913.

[261] J. Oyekan, M. Farnsworth, W. Hutabarat, D. Miller, A. Tiwari. Applying a 6 dof robotic arm and digital twin to automate fan-blade reconditioning for aerospace maintenance, repair, and overhaul. *Sensors*, vol. 20, no. 16, Article number 4637, 2020. DOI: 10.3390/s20164637.

[262] R. Lee, D. Ward, V. Dasagi, A. Cosgun, J. Leitner, P. Corke. Learning arbitrary-goal fabric folding with one hour of real robot experience. In *Proceedings of the 4th Conference on Robot Learning*, Cambridge, USA, pp. 2317–2327, 2020.

[263] R. Ye, W. Xu, H. Fu, R. K. Jenamani, V. Nguyen, C. Lu, K. Dimitropoulou, T. Bhattacharjee. RCare world: A human-centric simulation world for caregiving robots. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Kyoto, Japan, pp. 33–40, 2022. DOI: 10.1109/IROS47612.2022.9982244.

[264] J. Liu, Y. Chen, Z. Dong, S. Wang, S. Calinon, M. Li, F. Chen. Robot cooking with stir-fry: Bimanual non-prehensile manipulation of semi-fluid objects. *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5159–5166, 2022. DOI: 10.1109/LRA.2022.3153728.

[265] J. Lu, A. Jayakumari, F. Richter, Y. Li, M. C. Yip. SuPer deep: A surgical perception framework for robotic tissue manipulation using deep learning for feature extraction. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Xi'an, China, pp. 4783–4789, 2021. DOI: 10.1109/ICRA48506.2021.9561249.

[266] S. Krishnan, A. Garg, S. Patil, C. Lea, G. Hager, P. Abbeel, K. Goldberg. Transition state clustering: Unsupervised surgical trajectory segmentation for robot learning. *The International Journal of Robotics Research*, vol. 36, no. 13–14, pp. 1595–1618, 2017. DOI: 10.1177/0278364917743319.

[267] Y. Long, W. Wei, T. Huang, Y. Wang, Q. Dou. Human-in-the-loop embodied intelligence with interactive simulation environment for surgical robot learning. *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4441–4448, 2023. DOI: 10.1109/LRA.2023.3284380.

[268] X. Wang, B. Xu, Y. Cheng, H. Wang, F. Sun. Robust adaptive learning control of space robot for target capturing using neural network. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 10, pp. 7567–7577, 2023. DOI: 10.1109/TNNLS.2022.3144569.

[269] B. C. Zhong, L. Y. Xia. A systematic review on exploring the potential of educational robotics in mathematics education. *International Journal of Science and Mathematics Education*, vol. 18, no. 1, pp. 79–101, 2020. DOI: 10.1007/s10763-018-09939-y.

[270] S. James, Z. Ma, D. R. Arrojo, A. J. Davison. RLBench: The robot learning benchmark & learning environment.

IEEE Robotics and Automation Letters, vol. 5, no. 2, pp. 3019–3026, 2020. DOI: 10.1109/LRA.2020.2974707.

[271] A. Mousavian, C. Eppner, D. Fox. 6-DOF GraspNet: Variational grasp generation for object manipulation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Republic of Korea, pp. 2901–2910, 2019. DOI: 10.1109/ICCV.2019.00299.

[272] C. Chen, C. Liu, T. Wang, A. Zhang, W. Wu, L. Cheng. Compound fault diagnosis for industrial robots based on dual-transformer networks. *Journal of Manufacturing Systems*, vol. 66, pp. 163–178, 2023. DOI: 10.1016/j.jmsy.2022.12.006.

[273] S. Yury, M. Dmitry, B. Maria, Y. Alexander, P. Tatyana. Robotics in agriculture: Advanced technologies in livestock farming and crop cultivation. *E3S Web of Conferences*, vol. 480, Article number 03024, 2024. DOI: 10.1051/e3sconf/202448003024.

[274] S. H. Van Delden, M. SharathKumar, M. Butturini, L. J. A. Graamans, E. Heuvelink, M. Kacira, E. Kaiser, R. S. Klamer, L. Klerkx, G. Kootstra, A. Loeber, R. E. Schouten, C. Stanghellini, W. Van Ieperen, J. C. Verdonk, S. Vialet-Chabrand, E. J. Woltering, R. Van De zedde, Y. Zhang, L. F. M. Marcelis. Current status and future challenges in implementing and upscaling vertical farming systems. *Nature Food*, vol. 2, no. 12, pp. 944–956, 2021. DOI: 10.1038/s43016-021-00402-w.

[275] K. Kawaharazuka, N. Kanazawa, Y. Obinata, K. Okada, M. Inaba. Continuous object state recognition for cooking robots using pre-trained vision-language models and black-box optimization. *IEEE Robotics and Automation Letters*, vol. 9, no. 5, pp. 4059–4066, 2024. DOI: 10.1109/LRA.2024.3375257.

[276] Q. Yu, M. Moghani, K. Dharmarajan, V. Schorp, W. C. H. Panitch, J. Liu, K. Hari, H. Huang, M. Mittal, K. Goldberg, A. Garg. Orbit-surgical: An open-simulation framework for learning surgical augmented dexterity. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Yokohama, Japan, pp. 15509–15516, 2024. DOI: 10.1109/ICRA57147.2024.10611637.

[277] X. B. Peng, M. Andrychowicz, W. Zaremba, P. Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Brisbane, Australia, pp. 3803–3810, 2018. DOI: 10.1109/ICRA.2018.8460528.

[278] E. Aljalbout, F. Frank, M. Karl, P. van der Smagt. On the role of the action space in robot manipulation learning and sim-to-real transfer. *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5895–5902, 2024. DOI: 10.1109/LRA.2024.3398428.

[279] Y. Jiang, C. Wang, R. Zhang, J. Wu, Li F. F. TRANSIC: Sim-to-real policy transfer by learning from online correction, [Online], Available: https://arxiv.org/abs/2405.10315, 2024.

[280] F. Muratore, M. Gienger, J. Peters. Assessing transferability from simulation to reality for reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1172–1183, 2021. DOI: 10.1109/TPAMI.2019.2952353.

[281] H. Ma, M. Shi, B. Gao, D. Huang. Generalizing 6-DoF grasp detection via domain prior knowledge. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 18102–18111, 2024. DOI: 10.1109/CVPR52733.2024.01714.

[282] K. Chen, R. Cao, S. James, Y. Li, Y. H. Liu, P. Abbeel, Q. Dou. Sim-to-real 6D object pose estimation via iterative self-training for robotic bin picking. In *Proceedings of the 17th European Conference on Computer Vision*, Springer, Tel Aviv, Israel, pp. 533–550, 2022. DOI: 10.1007/978-3-031-19842-7_31.

[283] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, P. Florence. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, USA, pp. 8469–8488, 2023.

[284] X. Li, M. Zhang, Y. Geng, H. Geng, Y. Long, Y. Shen, R. Zhang, J. Liu, H. Dong. ManipLLM: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 18061–18070, 2024. DOI: 10.1109/CVPR52733.2024.01710.

[285] J. Xu, S. Jin, Y. Lei, Y. Zhang, L. Zhang. Reasoning tuning grasp: Adapting multi-modal large language models for robotic grasping. In *the 2nd Workshop on Language and Robot Learning: Language as Grounding*, 2023.

[286] S. Huang, I. Ponomarenko, Z. Jiang, X. Li, X. Hu, P. Gao, H. Li, H. Dong. ManipVQA: Injecting robotic affordance and physically grounded information into multi-modal large language models. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Abu Dhabi, UAE, pp. 7580–7587, 2024. DOI: 10.1109/IROS58592.2024.10801993.

[287] S. S. Kannan, V. L. N. Venkatesh, B. C. Min. SMART-LLM: Smart multi-agent robot task planning using large language models. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Abu Dhabi, UAE, pp. 12140–12147, 2024. DOI: 10.1109/IROS58592.2024.10802322.

[288] Y. Jin, D. Li, A. Yong, J. Shi, P. Hao, F. Sun, J. Zhang, B. Fang. RobotGPT: Robot manipulation learning from ChatGPT. *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2543–2550, 2024. DOI: 10.1109/LRA.2024.3357432.

[289] M. G. Arenas, T. Xiao, S. Singh, V. Jain, A. Ren, Q. Vuong, J. Varley, A. Herzog, I. Leal, S. Kirmani, M. Prats, D. Sadigh, V. Sindhwani, K. Rao, J. Liang, A. Zeng. How to prompt your robot: A PromptBook for manipulation skills with code as policies. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Yokohama, Japan, pp. 4340–4348, 2024. DOI: 10.1109/ICRA57147.2024.10610784.

[290] G. Cheng, C. Zhang, W. Cai, L. Zhao, C. Sun, J. Bian. Empowering large language models on robotic manipulation with affordance prompting, [Online], Available: https://arxiv.org/abs/2404.11027, 2024.

[291] C. Xiong, C. Shen, X. Li, K. Zhou, J. Liu, R. Wang, H. Dong. Autonomous interactive correction MLLM for robust robotic manipulation, [Online], Available: https://arxiv.org/abs/2406.11548v6, 2024.

[292] H. Liu, Y. Zhu, K. Kato, A. Tsukahara, I. Kondo, T. Aoyama, Y. Hasegawa. Enhancing the LLM-based robot ma-

nipulation through human-robot collaboration. *IEEE Robotics and Automation Letters*, vol. 9, no. 8, pp. 6904–6911, 2024. DOI: 10.1109/LRA.2024.3415931.

[293] A. Ajoudani, A. M. Zanchettin, S. Ivaldi, A. Albu-Schäffer, K. Kosuge, O. Khatib. Progress and prospects of the human–robot collaboration. *Autonomous Robots*, vol. 42, no. 5, pp. 957–975, 2018. DOI: 10.1007/s10514-017-9677-2.

[294] Z. Jin, A. Liu, W. A. Zhang, L. Yu, C. Y. Su. A learning based hierarchical control framework for human–robot collaboration. *IEEE Transactions on Automation Science and Engineering*, vol. 20, no. 1, pp. 506–517, 2023. DOI: 10.1109/TASE.2022.3161993.

[295] C. Wang, C. Perez-D'Arpino, D. Xu, Li F. F., C. K. Liu, S. Savarese. Co-GAIL: Learning diverse strategies for human-robot collaboration. In *Proceedings of the 5th Conference on Robot Learning*, London, UK, pp. 1279–1290, 2021.

[296] T. W. Chin, R. Ding, C. Zhang, D. Marculescu. Towards efficient model compression via learned global ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 1515–1525, 2020. DOI: 10.1109/CVPR42600.2020.00159.

[297] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, Q. Vuong, V. Vanhoucke, H. T. Tran, R. Soricut, A. Singh, J. Singh, P. Sermanet, P. R. Sanketi, G. Salazar, M. S. Ryoo, K. Reymann, K. Rao, K. Pertsch, I. Mordatch, H. Michalewski, Y. Lu, S. Levine, L. Lee, T. W. E. Lee, I. Leal, Y. Kuang, D. Kalashnikov, R. Julian, N. J. Joshi, A. Irpan, B. Ichter, J. Hsu, A. Herzog, K. Hausman, K. Gopalakrishnan, C. Fu, P. Florence, C. Finn, K. A. Dubey, D. Driess, T. Ding, K. M. Choromanski, X. Chen, Y. Chebotar, J. Carbajal, N. Brown, A. Brohan, M. G. Arenas, K. Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, Atlanta, USA, pp. 2165–2183, 2023.

[298] T. H. Wang, W. Xiao, T. Seyde, R. M. Hasani, D. Rus. Measuring interpretability of neural policies of robots with disentangled representation. In *Proceedings of the 7th Conference on Robot Learning*, Atlanta, USA, pp. 602–641, 2023.

[299] X. Li, Z. Serlin, G. Yang, C. Belta. A formal methods approach to interpretable reinforcement learning for robotic planning. *Science Robotics*, vol. 4, no. 37, Article number eaay6276, 2019. DOI: 10.1126/scirobotics.aay6276.

[300] J. Ji, B. Zhang, J. Zhou, X. Pan, W. Huang, R. Sun, Y. Geng, Y. Zhong, J. Dai, Y. Yang. Safety-gymnasium: A unified safe reinforcemei learning benchmark. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 831, 2023.

[301] Y. Jia, C. M. Poskitt, J. Sun, S. Chattopadhyay. Physical adversarial attack on a robotic arm. *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9334–9341, 2022. DOI: 10.1109/LRA.2022.3189783.

[302] M. Machin, J. Guiochet, H. Waeselynck, J. P. Blanquart, M. Roy, L. Masson. SMOF: A safety monitoring framework for autonomous systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 5, pp. 702–715, 2018. DOI: 10.1109/TSMC.2016.2633291.

[303] M. Savva, A. X. Chang, P. Hanrahan. Semantically-enriched 3D models for common-sense knowledge. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Boston, USA, pp. 24–31, 2015. DOI: 10.1109/CVPRW.2015.7301289.

[304] A. Gupta, P. Dollár, R. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 5351–5359, 2019. DOI: 10.1109/CVPR.2019.00550.

[305] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, USA, pp. 10674–10685, 2022. DOI: 10.1109/CVPR52688.2022.01042.

[306] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, W. Zaremba. OpenAI Gym, [Online], Available: https://arxiv.org/abs/1606.01540, 2016.

[307] V. Makoviychuk, L. Wawrzyniak, Y. R. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, G. State. Isaac Gym: High performance GPU based physics simulation for robot learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, 2021.

Ying Zheng received the Ph. D. degree in computer science from Harbin Institute of Technology, China in 2019. He is currently a research fellow at The Hong Kong Polytechnic University, China.

His research interests include computer vision and embodied AI.

E-mail: ying1.zheng@polyu.edu.hk
ORCID iD: 0000-0002-8042-6196

Lei Yao received the B. Eng. degree in measurement & control technology and instruments and M. Eng. degree in mechanical engineering from Huazhong University of Science and Technology, China in 2020 and 2023, respectively. He is currently a Ph. D. degree candidate with The Hong Kong Polytechnic University, China.

His research interests include 3D scene understanding and embodied AI.

E-mail: rayyoh.yao@connect.polyu.hk
ORCID iD: 0009-0007-0304-3056

Yuejiao Su received the B. Sc. and M. Sc. degrees in computer science and engineering from the Northwestern Polytechnical University, China in 2020 and 2023 respectively. She is currently a Ph. D. degree candidate with The Hong Kong Polytechnic University, China.

Her research interests include egocentric analysis, image segmentation, and embodied AI.

E-mail: yuejiao.su@connect.polyu.hk
ORCID iD: 0009-0006-9118-9217

Yi Zhang received the B. Sc. degree in computer science from The Hong Kong University of Science and Technology, China in 2020, the M. Sc. degree in data and artificial intelligence form Polytechnic Institute of Paris, France in 2022. She is currently a Ph. D. degree candidate with The Hong Kong Polytechnic University, China.

Her research interests include 3D scene understanding and embodied AI.

E-mail: yi-eee.zhang@connect.polyu.hk

ORCID iD: 0009-0009-8242-1581

Yi Wang received the B. Eng. degree in electronic information engineering and the M. Eng. degree in information and signal processing from the School of Electronics and Information, Northwestern Polytechnical University, China in 2013 and 2016, respectively, and the Ph. D. degree in information processing from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore in 2021. He is currently a research assistant professor at the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, China.

His research interests include image/video processing, computer vision, intelligent transport systems, and digital forensics.

E-mail: yi-eie.wang@polyu.edu.hk

ORCID iD: 0000-0001-8659-4724

Sicheng Zhao received the Ph. D. degree in computer science from the Harbin Institute of Technology, China in 2016. He was a visiting scholar with the National University of Singapore, Singapore from 2013 to 2014, a research fellow with Tsinghua University, China from 2016 to 2017, a posteoctorial research fellow with the University of California at Berkeley, USA from 2017 to 2020, and a posteoctorial research scientist with Columbia University, USA from 2020 to 2022. He is currently a research associate professor with Tsinghua University. He is an associate editor of IEEE TIP and IEEE TAFFC.

His research interests include affective computing, multimedia, and computer vision.

E-mail: schzhao@tsinghua.edu.cn

ORCID iD: 0000-0001-5843-6411

Yiyi Zhang received the B. Eng. degree in digital media technology from Zhejiang University, China in 2015, and the M. Sc. degree in computer science from the Institut Polytechnique de Paris – Télécom Paris, France in 2017. She is currently a Ph. D. degree candidate with The Chinese University of Hong Kong, China.

Her research interests include computer vision and medical AI.

E-mail: yyzhang24@cse.cuhk.edu.hk

ORCID iD: 0009-0000-6491-8513

Lap-Pui Chau received the Ph. D. degree in electronic engineering from The Hong Kong Polytechnic University, China in 1997. He was with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore from 1997 to 2022. He is currently a professor in the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, China. He is an IEEE Fellow. He was the Chair of Technical Committee on Circuits & Systems for Communications of IEEE Circuits and Systems Society from 2010 to 2012. He was General Chairs and Program Chairs for some international conferences. Besides, he served as Associate Editors for several IEEE journals and Distinguished Lecturer for IEEE BTS.

His research interests include large language model, perception for autonomous driving, egocentric computer vision, and 3D computer vision.

E-mail: lap-pui.chau@polyu.edu.hk (Corresponding author)

ORCID iD: 0000-0003-4932-0593