Intelligent Robot Manipulation Requires Self-Directed Learning

Li Chen[1] **Chonghao Sima**[1] **Kashyap Chitta**[2]
Antonio Loquercio[3] **Ping Luo**[1] **Yi Ma**[1] **Hongyang Li**[1]

[1]The University of Hong Kong [2]NVIDIA Research [3]University of Pennsylvania

Abstract

The Embodied AI community has long aspired to create a robotic system with human-level intelligence and dexterity. Recent advances in vision and language models motivated researchers to follow a similar paradigm for robotics and scale up imitation learning from demonstrations. However, imitation learning lacks the mechanism to incorporate feedback from the agent's own experience during interaction with the environment. This perspective argues that enabling agents to learn from their own experience, which we term as *self-directed learning*, is indispensable for advancing the intelligence and dexterity of modern robotic systems. Despite possessing a similar concept and toolkit to existing reinforcement learning methods, self-directed learning imposes extra challenges that could completely alter the algorithmic landscape of robot learning: the lack of resets and an explicit and noise-free reward signal. To overcome this limitation, we argue that future endeavors in self-directed learning should be focused into three aspects: goal identification, skill acquisition, and performance evaluation. To improve the efficiency of each step, we are inspired by education theory, suggesting that learning is not confined to a single modality, but rather relies on shared mechanisms across visual, textual, and kinesthetic processes. The key challenges and prospective research avenues to self-directed learning are outlined. We further foster a discussion on alternatives to self-directed learning to train robots for physically dexterous tasks.

1 Introduction

The mind is not a vessel to be filled, but a fire to be kindled. Then it motivates one towards originality and instills the desire for truth.

— PLUTARCH, "ON LISTENING"

Embodied AI (EAI) aims to build intelligent robotic systems capable of perceiving their surroundings, interacting with the environment, and executing actions based on sensor inputs [92, 77]. Robotic manipulation presents a central challenge in attaining human-level intelligence in robots [55]. This complexity arises from the diversity and intricate nature of manipulation tasks and the objects being interacted with, along with the direct impact manipulation exerts on the robot's environment [55, 23].

In recent years, the emergence of foundation models, such as large language models (LLMs) [74] and vision language models (VLMs) [96], has catalyzed a surge in large-scale machine learning research. By scaling up data and model capacities for behavior cloning (BC), the robotics community has made significant strides in developing large-scale robot datasets to chase the goal of training a general robot policy [51, 16, 76, 99]. These efforts have yielded remarkable outcomes, such as manipulation policies exhibiting improved robustness to variations in object positions, lighting conditions, and background [35, 25, 10]. However, these advancements primarily rely on expert demonstrations and pre-trained large models, which, as recent studies suggest, may limit their adaptability to long-horizon or entirely new tasks [72, 11]. While we recognize the importance of scaling up data for supervised learning as part of the roadmap towards generalization, we argue that the mere scaling of behavior cloning will *not* be sufficient for robots to achieve general dexterous intelligence.

But if this hypothesis is true, what can we do? For insight, we take inspiration from the broader history of human intelligence [94, 101], particularly the Dartmouth workshop [69] where the term "artificial intelligence" was first coined. Artificial intelligence aims to empower machines to master all skills that humans have [8, 23]. Ideally, for a robotic manipulation system to demonstrate intelligence, it must exhibit adaptability across diverse scenarios and be able to evolve and handle new tasks without requiring prior labeled data. We refer to this capability as *self-directed learning*. In essence, this entails the agent's capability for adaptation to generalized goals through self-derived reasoning and guidance. While conceptually related to reinforcement learning, self-directed learning differs in two key ways: it lacks a controlled environment with resets and does not rely on noise-free, carefully engineered reward functions.

In this position paper, we advocate for the development of self-directed learning approaches to surmount the limitations of scaling BC and to fulfill the desiderata for an intelligent robotic system. Drawing on pedagogical theories [53, 38], we posit that self-directed learning in embodied agents should operate as a closed-loop feedback process, grounded in three core pillars: goal identification, skill acquisition, and performance monitoring or evaluation.

These components should also be developed through learning from unlabeled data and structured exploration, enabling adaptation to new tasks in a self-reflective and continually evolving manner. To bridge the gap between the principles of self-directed learning and real-world deployment, we propose leveraging three human learning strategies (VTK): visual, textual (including auditory and reading/writing), and kinesthetic.

In summary, **this perspective advocates for self-directed learning being the essential component for enabling an intelligent robotic manipulation system**. Although this work focuses on robot manipulation as a testbed for the self-directed learning, these principles might be universally applicable to diverse EAI domains, e.g., robot navigation.

The remainder of this paper is organized as follows. In Sec. 2, we first provide an overview of the task formulation and highlight the primary challenges in general-purpose manipulation (Sec. 2.1). We then propose key principles, processes, and necessities of self-directed learning to address the challenges towards intelligent manipulation systems (Sec. 2.2). We reveal how visual, textual, and kinesthetic learning can be applied for various self-directed learning stages in Sec. 3, while also discussing challenges and potential directions in developing an effective self-directed learning framework in Sec. 4. Finally, alternative views, including scaling up BC, are outlined in Sec. 5.

2 The Principles of Self-Directed Learning

2.1 Formulation and Challenges Involved in Robot Manipulation Tasks

In contrast to classical policy learning tasks in simple settings [95], generalized robot manipulation for real-world deployment presents a multitude of complexities [23]. Arguably, the problem can be formulated as a Goal-conditioned Markov Decision Process (GcMDP), which involves continuous and dynamic interactions with an environment while accommodating diverse objectives [105, 3].

Concretely, it can be represented by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{G}, r, \gamma, \phi, p_g \rangle$, where $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{G}$ define the state space, action space, and goal space, respectively. $\mathcal{T}$ is the dynamics transition function and γ denotes a scalar discount factor. $\phi : \mathcal{S} \rightarrow \mathcal{G}$ denotes a function mapping a state s to a specific goal $g \in \mathcal{G}$, and p_g represents the distribution of desired goals. The state and goal space typically involve objects for manipulation, environmental background, and some relational or behavioral descriptions about the current and desired status. Variations in $\mathcal{S}$ or $\mathcal{G}$ are often referred to as generalization problems [35, 11]. While the latter is general, we argue for an alternative task formulation that is more practical and clearly highlights pronounced deviations from the training data distribution.

A manipulation task begins at time step $t = 0$, and concludes at $t = T$ upon meeting a termination condition, which could involve reaching a time limit or receiving success or failure signals. The sequence of state-action pairs, $\{s_t, a_t\}_{t=0}^{T}$, is referred to as an episode. Typically, the robot receives a sparse reward indicating success, set by $r : \mathcal{S} \times \mathcal{A} \times \mathcal{G} \rightarrow \mathbb{R}$. The objective is optimizing the goal-conditioned policy $\pi : \mathcal{S} \times \mathcal{G} \rightarrow \mathcal{A}$. Next, we delineate the challenging settings for real-world manipulation, and introduce corresponding requisite elements to achieve dexterity in Sec. 2.2.

Long-Horizon Tasks. Current prevailing benchmarks focus on simplified laboratory settings with single or limited motions involved [107, 36, 70]. Meanwhile, there are substantially more complex long-horizon manipulation tasks which test a robot in the real world. For instance, a simple household task such as "`bring me a snack from the drawer`" involves at least four sub-goals: opening the drawer, picking up the snack, delivering it to the user, and closing the drawer at the end. Complex tasks like "`wash the dishes in the sink`" will yield more sub-processes to achieve the final target. Since the entire process is temporally dependent, even minor deviations in achieving any of these sub-goals can lead to overall task failure.

In particular, a *desired goal* is defined as the ultimate objective, with a binary reward $r \in \{0, 1\}$ awarded at episode completion. In addition, *behavior goals* are introduced to represent intermediate objectives throughout the episode. Real-world task goals can manifest in various forms, including textual instructions [1, 52], target images [50, 106], or specified locations and motion references [19].

Lack of Privileged Information. Though not explicitly formulated as a partially observable MDP with state estimations in many robot manipulation works, full and accurate observation and transition dynamics are often not available. Unlike tasks with full privileged information [95, 46], robots rely on limited sensory inputs, such as RGB-D images captured in first-person, third-person, or wrist views, and tactile sensors mounted on grippers and dexterous hands. However, even with multi-modal inputs, comprehensive environmental information remains elusive due to viewpoint constraints, occlusions, and unattainable object-specific knowledge like the gravity center and friction parameters [103, 71]. Most importantly, rewards for long-horizon tasks are very challenging to engineer.

In robot manipulation benchmarks, evaluation of tasks' success is typically conducted by humans or simulators with privileged knowledge of the environment's states [107, 70, 76], such as objects' contact status for the task "`close a laptop`". The evaluator assigns a reward of $r = 1$ if the task is completed at the end of the episode and sometimes partial scores for the completion of sub-goals, *e.g.*, $r = 0.5$ for achieving the first sub-goal in a two-sub-goals manipulation task [36]. However, the aforementioned partially observable nature of robotic perception severely undermines this requirement for sub-goals' verification, leading to failures in long-term tasks' generalization.

2.2 Principles of Self-Directed Learning

Conventional robot manipulation systems deployed in physical environments are typically human-directed in terms of goal setting, data collection, and progress monitoring. For instance, traditional robotic frameworks often divide perception and planning algorithms and integrate them into an engineered system, *e.g.*, a combination of detectors, a state machine, and A* motion planning. Hand-crafted rules are devised to cover most situations robots may encounter, with meticulously selected data employed to train independent modules. Although adequate for constrained applications such as industrial robots, this paradigm exhibits limitations in generalizing across different environments and embodiments [23].

Recent advancements in LLMs offer an alternative path towards generalization requiring reduced human intervention. Leveraging extensive textual data sourced from the internet, LLMs adopt a unified model encapsulating a vast array of world knowledge, demonstrating an impressive ability to interpolate and extrapolate for adaptation [74, 96]. Unfortunately, most contemporary robotic systems fail to exhibit a comparable level of generalization, especially in novel or long-horizon tasks requiring self-reflection and self-adaptation. In contrast to the unified text data, robot data is usually collected through teleoperation or simulations, with task specifications manually defined including goals and optimization targets. While popular end-to-end methodologies benefit from scaling techniques and building upon LLMs/VLMs to establish strong perception and memory foundations, they still struggle with complex physical interactions [27, 76, 17]. For example, systems like OpenVLA [52] often work only within the confines of the same environment and embodiment used during training, which indicates only interpolation capabilities rather than extrapolation and adaptation (we refer to Sec. 5.1 for detailed discussions on scaling behavior cloning).

EAI systems are generally composed of four key components: perception, action, memory, and learning [77]. While recent end-to-end approaches have made significant progress in scaling unified models that implicitly integrate perception, action, and memory [20, 27, 97, 52, 6], the learning component remains underdeveloped. Here, we define learning broadly as the ability to generalize—to make accurate predictions under novel input distributions [94, 37]. At the same time, there is growing recognition that open-endedness, i.e., the capacity for continual, unsupervised skill acquisition, is essential for the emergence of intelligence [44]. To foster such adaptive capabilities, especially once existing robotic datasets have been exhausted, advancing the learning module in robot systems is critical.

In this vein, **self-directed learning** from data and experience in the wild, instead of relying solely on pre-scripted paired demonstrations, emerges as a viable direction. To be specific, given the current state s, the desired goal g, and a base goal-conditioned policy π, where s and g significantly deviate from prior experience and render π inadequate for achieving g, the robot needs to autonomously adapt and evolve to the new task without pre-collected perception-action demonstrations. This learning paradigm mirrors the self-directed learning in human education [53, 38]. It can be divided into three main steps: **(1) set learning goals, (2) engage in learning to acquire skills, and (3) monitor and evaluate learning progress and outcomes**. One essential feature of this paradigm is the closed-loop structure which facilitates adaptive progress towards the ultimate goal [94]. Consequently, to address the challenges posed by long-horizon goals and the absence of privileged information in general-purpose manipulation, advancements in all three dimensions are desired.

2.2.1 Step 1: Goal Identification

Task Decomposition. As mentioned in Sec. 2.1, a complex long-horizon task usually involves multiple stages, with a specified desired goal. While behavioral goals can be generated either by the environment or human monitors given the desired goal, we emphasize the intrinsic goal setting due to the lack of privileged information and human labelers. Specifically, the robot must *autonomously generate sub-goals* for learning based on the desired goal provided in the instructions without external intervention. A primary advantage is that the identified sub-goals naturally guide the policy to achieve the ultimate target during both the learning phase and deployment.

State Space Abstraction. Though most existing robot manipulation benchmarks employ high-level language instructions to represent desired goals [51, 76, 16], they can be abstracted and provided in multiple forms. Text-based representations are widely favored due to their human-friendliness and universal format, akin to their utility for foundation models [74]. Besides, depicting goals in an image format is also feasible [2, 28, 29]. In limited manipulation tasks like object reorientation [81, 67], the target is a numeric position with which dense rewards can be calculated precisely. These various forms do not have substantial advantages over each other, and would be applicable in different learning strategies, as elaborated in Sec. 3.

For robotic manipulation in the real world, goals often highlight the most relevant content about task rewards. Thus, the goal space is typically a sub-space of the full state space (*i.e.*, $\mathcal{G} \subset \mathcal{S}$), and goal identification functions as a state abstraction [41, 79]. Besides, given that goals are the conditions for skill acquisition and reward evaluation in GcMDP, the clarity and propriety of identified goals are crucial to guide the learning process, affecting the efficiency and efficacy of self-directed learning.

2.2.2 Step 2: Skill Acquisition

We consider a setting in which no paired data between perception and robot actions is available. Even if the goals and evaluation mechanisms are predefined, the lack of labeled data demands that the robot acquire new knowledge autonomously.

Skill Transfer. Humans possess the ability to acquire new skills through demonstrations, such as mastering a dance routine, even in the absence of precise action annotations. Robots should reach similar competence by adapting to new skills by utilizing non-identical exemplar data formats. This skill transfer process is analogous to the concept of in-context learning, a trending and encouraging direction in foundation models, where pre-trained LLMs can be boosted based on augmented examples [13]. However, our target skill source diverges from the text expressions for LLMs. In the realm of robot manipulation, demonstrations can stem from heterogeneous data sources, including human videos, similar rollouts of robots with distinct embodiment configurations, or even instruction books [17, 89]. Though limited paired data is present at this stage, a vast amount of mixed data in the wild can be leveraged for this purpose.

Skill Acquisition from Scratch. When no valuable internal memory or external resources can be retrieved, the system has to search on its own. Yet, prior research in reinforcement learning (RL) areas has demonstrated that exploration in the vast state and action space is extremely inefficient [47, 61, 107], which raises further demands for an effective monitor to guide the learning direction.

2.2.3 Step 3: Monitoring and Evaluation

Self-directed learning is a closed-loop and dynamic process, where evaluation of the learning progress provides feedback to the agent to adjust the subsequent learning direction [53]. This can sometimes be referred to as *self-reflection* [63]. For robots, the core of self-reflection lies in their ability to evaluate their behavior process and causally determine their actions [2]. As discussed in Section 2.1, long-horizon tasks in robotics are typically decomposed into several stages, referred to as behavioral goals, and it is essential to enable the system to detect errors and ensure that each intermediate goal is reliably achieved before proceeding to the next sub-goal. We usually evaluate the task progress or success status with signals from the simulators or supervision from humans aside the robots. However, in a fully autonomous, self-directed setting, we need to define an explicit distance function $d(\phi(s), g)$ to measure the distance between the state and the goal in $\mathcal{G}$ and set a threshold ϵ to assess whether the agent has achieved its goals, without access to privileged information as in simulators.

Specifically, we express the distance function as a **value model** V, which quantifies the "goodness" of a given state s_t. While a state-action value function $Q^\pi(s_t, a_t)$ (or Q-function) is conditioned on a specific action a_t, our focus in this work is on learning the *state value* alone. Importantly, although value functions are often associated with reinforcement learning, our framework does *not* require the use of RL algorithms. The value estimate can be learned jointly with the policy or obtained independently. For example, it may be instantiated as a classical value function trained via reinforcement learning [73], or as a vision-language model that predicts task progress as a proxy for value [7, 45].

Thus, while we adopt the formalism of MDPs and emphasize the utility of value learning, our method deliberately decouples value estimation from reward-based interaction or trial-and-error learning. The value function serves as a flexible signal that can support both policy training and test-time monitoring, regardless of how it is acquired. To this end, value learning helps to use the value function as a metric to detect the intermediate failure modes and task progress during deployment. We refer to this as *closed-loop feedback* or error measurement in this work. This capability can be applied to two major perspectives: (1) verifying the reliability of the self-generated behavioral goals, (2) determining the appropriate time to transition to the next sub-goal by assessing the progress or completion of the task.

Sub-Goal Verification. Since the sub-goal sequence is directed toward the ultimate target, which receives a reward of $r = 1$, the intermediate values should, in principle, increase monotonically from 0 to 1 after reward shaping. With a reliable value estimate, it becomes straightforward to check if the value falls below a specified threshold or decreases significantly compared to previous sub-goals. A small threshold ensures that sub-goals are completed and that the transition to the next sub-goal or task termination occurs successfully. When these conditions are met, the agent can re-plan and generate a new set of behavioral goals to correct the course of action and manage long-term tasks.

Outcome Assessment with Closed-loop Feedback. Taking both the value and the goal, the distance between these two states can be expressed as $d(\phi(s), g) = V^\pi(s) - V^\pi(\phi^{-1}(g_k))$. Distinct from approaches involving fixed sub-goal action prediction and execution, this allows the agent to progressively advance toward the goal. This approach is consistent with the closed-loop feedback philosophy, where the error serves as guidance for the next sub-goal and benefits causal reasoning [15, 68].

3 Potential Strategies for Self-Directed Learning

Motivated by the principles outlined in Sec. 2.2, we demonstrate potential learning strategies to fulfill the proposed self-directed learning paradigm in this section. Fundamentally, any candidate that tackles all the three steps above would be an effective implementation. However, we focus on a set of strategies inspired from education theory. Specifically, we categorize such strategies into three classes (VTK): visual, textual (including auditory, reading/writing), and kinesthetic. The categorization follows a soft standard from pedagogy theories, as will be presented in Sec. 3.1. Then, we discuss in detail what each class of works has achieved, and what aspects they are still missing for enabling general robot manipulation.

3.1 Background of VTK Learning Strategies

Since the 1970s, educators have investigated how individuals differ in learning, which has led to the development of modern education theory. One widely acknowledged model of learning styles is categorized by learning modalities. In particular, Barbe *et al.* [5] proposed that there are mainly three types of learners: visual learners, auditory learners, and kinesthetic learners (VAK learning styles); and one style is often predominant for each individual's learning preference. Fleming *et al.* [34] further extended the theory with one additional modality, *i.e.*, reading/writing, leading to the VARK model.

For robotic systems, visual and kinesthetic skills can be reflected in the perception and action modules [77]. However, auditory and reading/writing abilities are not exactly developed like humans; but often exist in another form of the input and output of the system–text. For example, auditory instructions are usually transformed into text with a speech recognition model, and then processed by foundation models [82, 74]. Thus, in this paper, we merge these two learning strategies as a whole to be textual learning for better analyzing current advancements in robot manipulation. The resulting visual, textual, and kinesthetic learning strategies are then abbreviated as VTK learning. We compare below how VTK learning strategies differ from each other, along with their connection to the original VARK learning styles in human education:

- Visual learning: The agent imitates humans' or other experts' demonstrations via visual information (images, videos, or data from other visual sensors). This is similar to human visual learners who absorb information primarily by observation.
- Textual learning: Learners follow text-based demonstrations or involve text-based abstraction to accomplish a task (*e.g.*, aided by foundation models).
- Kinesthetic learning: The learning happens mostly through online physical interaction with the real world (usually implemented by reinforcement learning with more modalities such as tactility). This is analogous to kinesthetic learners who actively participate in hands-on activities or events.

3.2 Visual Learning

Most elemental human activities originate from the imitation of others' behaviors through the visual system. As incorporating visual imitation in robot learning is promising, many works focus on visual learning in robotics, aiming to extract valuable semantics from visual observations, potentially form a series of sub-goals to accomplish a task, and finally predict executions to reach the sub-goals/goal.

Numerous works design perceptual tasks to extracting semantic clues from visual demonstrations. Such cues are then used to compute robot actions. These demonstrations may involve hand trajectories [18, 87, 91] or poses [100], as well as semantic contact points [93] and affordances [4, 56, 87]. The extraction process leverages advanced vision models [78, 84] in an off-the-shelf fashion [56]. The extracted semantics are then utilized either as action labels for imitation learning within the policy network [111, 32], or as perception labels for fine-tuning the perception model followed by execution by a low-level controller [33, 11].

The construction of sub-goals has been an active topic in visual learning, aiming to predict possible future frames based on the current frame and the corresponding semantics. This predictive model is sometimes presented as a world model in the literature [54, 28, 42]. This is done by synthesizing a video of the imagined execution of the task using a video prediction model [40] conditioned on the initial frame. The predicted frames are either directly employed by the visuomotor policy [54, 15] or utilized for extracting sub-goal semantics, subsequently translated into sequential actions [28].

3.3 Textual Learning

Advanced human activities, such as assembling IKEA furniture following a user manual, demand the capacity to comprehend text-based instructions and execute them step-by-step to accomplish the task. With the advent of LLMs/VLMs, it is attractive to utilize them either directly or through fine-tuning to mimic the procedure of learning from text-based demonstrations. This approach typically involves converting the task completion process to text, constructing text-based demonstrations [79, 109], and employing LLMs/VLMs to finish the task via in-context learning or supervised finetuning.

The first stage is to generate the textual description of accomplishing a task (demonstrations) with the help of LLMs/VLMs [74, 96]. This involves image understanding and reasoning about the task completion [109, 110], both of which may or may not include the final executable action. The second stage is to utilize the in-context ability of the off-the-shelf LLMs/VLMs to infer on the new tasks with the generated demonstrations as a prompt, followed up with vision-language-action models (VLAs) [111, 52, 14] or other task executors to finish the task.

What is Missing in Visual Learning and Textual Learning? These works moderately address the Step 1 in the proposed self-directed learning (visual learning has predicted future frames [54] and textual learning has the text-based decomposition for accomplishing a task [109]). However, many of them neglect the monitoring and evaluation of their predicted actions when trying to reach sub-goals (Step 3) and thus may "blindly" execute the predicted sub-goals during inference. Recent works are dedicated to providing the proposed monitoring and evaluation by evaluating the intermediate sub-goals of the task completion via VLMs [66] or a video prediction model [31, 43], which is feasible but also introduces more challenges as discussed in Sec. 4.

3.4 Kinesthetic Learning

Indirect learning strategies such as visual and textual learning entail robots retrieving demonstrations from external sources for skill transfer. In contrast, kinesthetic learning represents a more direct approach, emphasizing learning through hands-on interaction with the physical environment, *e.g.*, manipulating the robot around physically to find the task's solution. Kinesthetic learning, particularly in the realm of reinforcement learning within simulated environments and subsequent policy transfer to real-world settings, encounters challenges in large-scale deployment due to issues inherent in simulation construction and the sim-to-real disparity [37, 23]. An emerging trend in addressing these challenges lies in real-world RL [64, 65, 58], where actions from the policy network are applied to the physical robot directly. This line of research, while promising, has not been as extensively explored as visual or textual learning strategies, regarding core issues such as safety protocols for human-robot-interaction, data sampling efficiency, and the variance in real-world feedback.

What is Missing in Kinesthetic Learning? RL frameworks can provide the utility of monitoring and evaluation (Step 3) of the predicted actions via rewards or similar mechanisms, yet many of them tend to overlook or underemphasize the crucial step of goal identification (Step 1). Alternatively, hierarchical RL frameworks, closely related to goal identification, have only been demonstrated in constrained environments and short-term tasks [57, 102]. These limitations diminish the transferability of acquired skills, as no intermediate atomic actions are shared explicitly across diverse tasks. Furthermore, even RL-based methods fall short in adequately addressing monitoring and evaluation aspects, with challenges persisting in learning a general reward function itself [47].

4 Challenges and Future Work

While the preliminary techniques in Sec. 3 for self-directed learning show promise, pivotal challenges still remain, and we summarize them below.

Ambiguity in Goal Abstraction. As illustrated in Sec. 2.1, the full state space includes a significant amount of information, including physical parameters, contact status and points, and 3D priors, which usually cannot be described clearly via text or limited images. It has become popular to represent sub-goals as similar forms of text instructions or target images, and generate them with foundation models and generative models due to their powerful reasoning ability. However, currently, these models often suffer from significant hallucination issues, which can lead to the generation of unreachable or unreasonable sub-goals. For example, a generated goal-state image that is inconsistent with the given instruction may mislead the policy, resulting in incorrect actions [2]. Additionally, certain motion targets such as rotating a tool may not even be expressed across images.

Ambiguity in Value Annotation and Estimation. The value function models the latent distance between goals and states. Therefore, analogous to the aforementioned goal abstraction issue, generating annotated data for supervised learning or reinforcement learning of value functions for complex tasks is non-trivial. For example, having a human labeler assign scalar value functions for robot data at scale is challenging, due to the inherent ambiguity in the definition (*i.e.*, sub-goals) of this task. One potential alleviation to obtain data with value estimates often requires a separate stage of online reinforcement learning or rule-based planning with known reward functions [85]. Annotating preferences or comparisons between states, as is common in reinforcement learning with human feedback for LLMs, is another interesting avenue to explore for this task [83].

Meanwhile, due to the ambiguity of goals, precise value estimation is challenging as well. Despite the extensive applications, existing value learning approaches for real-world robots are typically formulated as sparse success detectors based on image observations [30, 106, 29]. In many cases, these value functions can be insufficient, as they do not fully capture the three-dimensional and physical states and cannot indicate the task progress, making it challenging to evaluate nuanced state transitions in dexterous tasks. How to enable more precise value estimation and incorporate more informative factors such as multi-modal sensory data for value estimation remain underexplored.

Generalized Value Learning. The multi-goal setting demands the value function to be robust in diverse or even novel scenarios, necessitating a universal value function approximation [86]. This could raise similar challenges as learning the policy itself. Nonetheless, current value learning approaches are often established on simple or simulated tasks [67, 66, 15]. As a result, they may suffer from covariate shifts when applied to the real world. This discrepancy can lead to suboptimal executions, as the accuracy of the learned value functions may drop severely across varying conditions. This highlights the need for diverse data for value learning. Potential solutions may include leveraging the generalization of foundation models [7, 30, 68] and domain adaptation techniques.

Integrated VTK Learning. While different learning strategies (visual, textual and kinesthetic) have demonstrated promising results in certain perspectives of self-directed learning, how to effectively integrate them together to cover the full learning stages and enjoy the best of them remains puzzling. As visual and textual learning are mainly implemented as imitation learning, and kinesthetic learning usually employs reinforcement learning, a natural idea is to combine imitation learning with reinforcement learning in the training procedure. For instance, one potential direction is to learn a residual policy via RL using controlled exploration strategies [48]. The residual component learned with kinesthetic strategies may serve as an additional value estimator, and be integrated with the base policy cultivated with visual and textual learning. Yuan *et al.* [108] propose Policy Decorator as it functions similarly to Python decorators—wrapping with additional error correction the base policy learned from visual or textual learning. Such error correction ability initially comes from the goal identification and progress monitoring capability of the base policy while evolving during RL training, making the ultimate solution cover all three steps in the self-directed learning process.

5 Alternative Views

In this section, we outline two alternative paradigms to self-directed learning: "full" human involvement in deciding the goals, values, and learning by scaling up behavior cloning (Sec. 5.1), and "partial" human involvements, where humans focus on error correction and feedback (Sec. 5.2).

5.1 Scaling Behavior Cloning

Built upon the rapid advancements in both AI software and hardware infrastructure [98, 90], scaling laws have been demonstrated in domains such as LLMs/VLMs [74, 96] and vision generators [12, 80]. They exhibit generalization abilities such as instruction following [75] and few-shot learning [13]. The robotics community, pursuing the same goal of generalization, could potentially reproduce similar successes. The primary motivation for adopting scaling BC in robotics is that it circumvents a case-by-case sophisticated design for each task, or structured VTK learning paradigms to incorporate unannotated sources, and embraces the generalization learned from data at scale. Researchers have made substantial efforts in this direction, including collecting large-scale demonstrations [76, 51, 99, 16] and leveraging pre-trained foundation models to build VLA models [59, 62, 52, 9]. One might argue that extremely comprehensive data coverage could resolve issues such as goal identification and value modeling in new tasks [25], especially with the help of simulation and low-cost devices for crowdsourcing, while also revealing some limitations worthy of further exploration.

Memorization instead of Intelligence. Although recent works [62, 52, 25, 59] demonstrate that increasing the size of both the pretraining dataset and the model improves success rates and the ability of instruction following, they still fail to generalize across unseen tasks and environments. One possible explanation is that large-scale imitation learning in robotic data is merely memorizing the data distribution from the collected demonstrations [22, 39]. The lack of failure data in training also constrains models, resulting in an inability to self-correct and self-improve.

A Plateau with Scaling Behavior Cloning. Even when success rate is considered the primary metric, the typical 'scaling law' appears to be invalid for robotic models, as it reaches a plateau rather than following the expected power-law relationship [49] with respect to data and model size [72]. This plateau in performance has been observed in several models pretrained on the Open X-Embodiment dataset [76], including RDT-1B [62] and Octo [97]. Even when expanding generalization across multiple embodiments [26], a power-law relationship does not emerge as expected.

5.2 Human-in-the-loop Learning (Assistive Training)

Note that an important property of self-directed learning is that it is highly self-motivated and monitored, *without* external intervention, or with minimal interference. However, as in traditional classroom education, teachers are still occasionally necessary to provide feedback to the self-directed learners [53, 53]. Human-in-the-loop learning (HITL) has demonstrated remarkable success in LLM post-training, utilizing reinforcement learning to align with human preference [75] or select valuable training data to strengthen certain abilities [24]. Similarly, in robot learning, HITL also presents great promise for these benefits. This could make the learning process more efficient as humans inherently introduce privileged information that is previously inaccessible to robots.

With a base policy, monitors can take over or send real-time instructions to manipulation systems to correct mistakes and guide exploration directions [21, 60, 65, 104]. The error correction data involved may be helpful to enhance the causal reasoning ability of the robot. The resulting action sequences can be leveraged to continuously update the policy with RL [65] or BC [60]. Witnessing these advancements, an alternative opinion against self-directed learning might stand that we can employ HITL for adapting to complex tasks and learning value functions with the monitors' feedback efficiently. However, one crucial issue has inevitably obstructed its broader application.

Limited Scalability. In scaling behavior cloning (Sec. 5.1), researchers have addressed scalability challenges by leveraging low-cost hardware and large-scale simulation data. In contrast, human-in-the-loop (HITL) algorithms present several fundamental obstacles to scalability. First, HITL methods require real-time human intervention during robot execution, without halting rollouts. Most existing work relies on Franka robot arms to support this need [60, 65], as few other hardware platforms offer the reliability and integration ease needed. Second, effective human feedback requires accurate and responsive teleoperation systems to correct robot behavior, which is especially demanding for tasks involving end-effectors or dexterous hands [21, 104]. Third, recent approaches have explored incorporating high-level human feedback during task execution [88]. However, this assumes the presence of a strong low-level policy that can meaningfully interpret human intent—an assumption that often limits applicability to narrow, well-defined tasks rather than general settings. Across all these approaches, the need for skilled teleoperators or monitors introduces serious barriers to both scalability and safety, posing a major challenge to the widespread adoption of HITL methods for advancing robotic manipulation.

6 Conclusion and Outlook

In this work, we argue that an intelligent robot manipulation system should prioritize developing self-directed learning abilities, including autonomous goal identification, policy learning, and monitoring with a value model. Although scaling techniques have achieved significant success in numerous domains, we argue that scaling alone is insufficient to realize the goal. Instead, we propose that robots can achieve self-directed learning inspired by human learning styles, through visual, textual (auditory, reading/writing), and kinesthetic learning.

As the field of Embodied AI continues to evolve rapidly, it is clear that algorithmic breakthroughs do not occur in isolation. As discussed above, progress in areas such as hardware design, data collection, and ecosystem development will also be essential to unlock the full potential of self-directed learning. We anticipate that advances in self-directed learning for robots will not only drive innovation across robot manipulation, but may also inspire progress in adjacent domains beyond robotics.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do As I Can, Not As I Say: Grounding language in robotic affordances. In *CoRL*, 2022.

[2] Anurag Ajay, Seungwook Han, Yilun Du, Shuang Li, Abhi Gupta, Tommi Jaakkola, Josh Tenenbaum, Leslie Kaelbling, Akash Srivastava, and Pulkit Agrawal. Compositional foundation models for hierarchical planning. In *NeurIPS*, 2023.

[3] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *NeurIPS*, 2017.

[4] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *CVPR*, 2023.

[5] Walter Burke Barbe, Raymond H Swassing, and Michael N Milone. *Teaching Through Modality Strengths: Concepts and Practices*. Zaner-Bloser, 1979.

[6] Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, et al. A careful examination of large behavior models for multitask dexterous manipulation. *arXiv preprint arXiv:2507.05331*, 2025.

[7] Kate Baumli, Satinder Baveja, Feryal Behbahani, Harris Chan, Gheorghe Comanici, Sebastian Flennerhag, Maxime Gazeau, Kristian Holsheimer, Dan Horgan, Michael Laskin, et al. Vision-language models as a source of rewards. *arXiv preprint arXiv:2312.09187*, 2023.

[8] Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science*, 2019.

[9] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[10] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y Galliker, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *CoRL*, 2025.

[11] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, et al. RT-1: Robotics transformer for real-world control at scale. In *RSS*, 2023.

[12] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 2024.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[14] Qingwen Bu, Hongyang Li, Li Chen, Jisong Cai, Jia Zeng, Heming Cui, Maoqing Yao, and Yu Qiao. Towards synergistic, generalized, and efficient dual-system for robotic manipulation. *arXiv preprint arXiv:2410.08001*, 2024.

[15] Qingwen Bu, Jia Zeng, Li Chen, Yanchao Yang, Guyue Zhou, Junchi Yan, Ping Luo, Heming Cui, Yi Ma, and Hongyang Li. Closed-loop visuomotor control with generative expectation for robotic manipulation. In *NeurIPS*, 2024.

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

[17] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. UniVLA: Learning to act anywhere with task-centric latent actions. In *RSS*, 2025.

[18] Guangyan Chen, Meiling Wang, Te Cui, Yao Mu, Haoyang Lu, Tianxing Zhou, Zicai Peng, Mengxiao Hu, Haizhou Li, Li Yuan, et al. VLMimic: Vision language models are visual imitation learner for fine-grained actions. In *NeurIPS*, 2024.

[19] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-TeleVision: Teleoperation with immersive active visual feedback. In *CoRL*, 2024.

[20] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor policy learning via action diffusion. In *RSS*, 2023.

[21] Eugenio Chisari, Tim Welschehold, Joschka Boedecker, Wolfram Burgard, and Abhinav Valada. Correct Me if I am Wrong: Interactive learning for robotic manipulation. *RA-L*, 2022.

[22] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT Memorizes, RL Generalizes: A comparative study of foundation model post-training. In *CPAL*, 2025.

[23] Jinda Cui and Jeff Trinkle. Toward next-generation learned robot manipulation. *Science Robotics*, 2021.

[24] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, et al. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[25] Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Heming Cui, et al. GraspVLA: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*, 2025.

[26] Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling Cross-Embodied Learning: One policy for manipulation, navigation, locomotion and aviation. In *CoRL*, 2024.

[27] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An embodied multimodal language model. In *ICML*, 2023.

[28] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *NeurIPS*, 2023.

[29] Yilun Du, Sherry Yang, Pete Florence, Fei Xia, Ayzaan Wahid, brian ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Pack Kaelbling, Andy Zeng, and Jonathan Tompson. Video language planning. In *ICLR*, 2024.

[30] Yuqing Du, Ksenia Konyushkova, Misha Denil, Akhil Raju, Jessica Landon, Felix Hill, Nando de Freitas, and Serkan Cabi. Vision-language models as success detectors. *arXiv preprint arXiv:2303.07280*, 2023.

[31] Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Danijar Hafner, and Pieter Abbeel. Video prediction models as rewards for reinforcement learning. *NeurIPS*, 2023.

[32] Bin Fang, Shidong Jia, Di Guo, Muhua Xu, Shuhuan Wen, and Fuchun Sun. Survey of imitation learning for robotic manipulation. *IJIRA*, 2019.

[33] Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. AnyGrasp: Robust and efficient grasp perception in spatial and temporal domains. *TRO*, 2023.

[34] Neil Fleming, David Baume, et al. Learning Styles Again: Varking up the right tree! *Educational Developments*, 2006.

[35] Jensen Gao, Suneel Belkhale, Sudeep Dasari, Ashwin Balakrishna, Dhruv Shah, and Dorsa Sadigh. A taxonomy for evaluating generalist robot policies. *arXiv preprint arXiv:2503.01238*, 2025.

[36] Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay Policy Learning: Solving long-horizon tasks via imitation and reinforcement learning. In *CoRL*, 2020.

[37] Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. *Nature Communications*, 2021.

[38] Todd M Gureckis and Douglas B Markant. Self-Directed Learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, 2012.

[39] Chengyang He, Xu Liu, Gadiel Sznaier Camps, Guillaume Sartoretti, and Mac Schwager. Demystifying Diffusion Policies: Action memorization and simple lookup table alternatives. *arXiv preprint arXiv:2505.05787*, 2025.

[40] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022.

[41] Mark K Ho, Jonathan D Cohen, and Thomas L Griffiths. Rational simplification and rigidity in human planning. *Psychological Science*, 2023.

[42] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video Prediction Policy: A generalist robot policy with predictive visual representations. In *ICML*, 2025.

[43] Tao Huang, Guangqi Jiang, Yanjie Ze, and Huazhe Xu. Diffusion Reward: Learning rewards via conditional video diffusion. In *ECCV*, 2024.

[44] Edward Hughes, Michael D Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktäschel. Position: Open-Endedness is essential for artificial superhuman intelligence. In *ICML*, 2024.

[45] Kuo-Han Hung, Pang-Chi Lo, Jia-Fong Yeh, Han-Yuan Hsu, Yi-Ting Chen, and Winston H Hsu. VICtoR: Learning hierarchical vision-instruction correlation rewards for long-horizon manipulation. In *ICLR*, 2025.

[46] Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 2019.

[47] Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: Lessons we have learned. *IJRR*, 2021.

[48] Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *ICRA*, 2019.

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

[50] Alexander Khazatsky, Ashvin Nair, Daniel Jing, and Sergey Levine. What can i do here? learning new skills by imagining visual affordances. In *ICRA*, 2021.

[51] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *RSS*, 2024.

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

[53] Malcolm S Knowles. *Self-Directed Learning: A guide for learners and teachers.* ERIC, 1975.

[54] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to act from actionless videos through dense correspondences. *arXiv preprint arXiv:2310.08576*, 2023.

[55] Oliver Kroemer, Scott Niekum, and George Konidaris. A Review of Robot Learning for Manipulation: Challenges, representations, and algorithms. *JMLR*, 2021.

[56] Yuxuan Kuang, Junjie Ye, Haoran Geng, Jiageng Mao, Congyue Deng, Leonidas Guibas, He Wang, and Yue Wang. RAM: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation. In *CoRL*, 2024.

[57] Chengshu Li, Fei Xia, Roberto Martin-Martin, and Silvio Savarese. HRL4IN: Hierarchical reinforcement learning for interactive navigation with mobile manipulators. In *CoRL*, 2020.

[58] Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Zhaohui Yang, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, et al. SimpleVLA-RL: Scaling vla training via reinforcement learning. *arXiv preprint arXiv:2509.09674*, 2025.

[59] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. In *ICLR*, 2025.

[60] Huihan Liu, Soroush Nasiriany, Lance Zhang, Zhiyao Bao, and Yuke Zhu. Robot Learning on the Job: Human-in-the-loop autonomy and learning during deployment. *IJRR*, 2022.

[61] Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-Conditioned Reinforcement Learning: Problems and solutions. In *IJCAI*, 2022.

[62] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1B: a diffusion foundation model for bimanual manipulation. In *ICLR*, 2025.

[63] Sofie MM Loyens, Joshua Magda, and Remy MJP Rikers. Self-directed learning in problem-based learning and its relationships with self-regulated learning. *Educational Psychology Review*, 2008.

[64] Jianlan Luo, Zheyuan Hu, Charles Xu, You Liang Tan, Jacob Berg, Archit Sharma, Stefan Schaal, Chelsea Finn, Abhishek Gupta, and Sergey Levine. SERL: A software suite for sample-efficient robotic reinforcement learning. In *ICRA*, 2024.

[65] Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. *arXiv preprint arXiv:2410.21845*, 2024.

[66] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *ICLR*, 2023.

[67] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *ICLR*, 2024.

[68] Yecheng Jason Ma, Joey Hejna, Chuyuan Fu, Dhruv Shah, Jacky Liang, Zhuo Xu, Sean Kirmani, Peng Xu, Danny Driess, Ted Xiao, et al. Vision language models are in-context value learners. In *ICLR*, 2025.

[69] John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. Dartmouth workshop. https://en.wikipedia.org/wiki/Dartmouth_workshop, 1956.

[70] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *RA-L*, 2022.

[71] Marius Memmel, Andrew Wagenmaker, Chuning Zhu, Dieter Fox, and Abhishek Gupta. ASID: Active exploration for system identification in robotic manipulation. In *ICLR*, 2024.

[72] Suvir Mirchandani, Suneel Belkhale, Joey Hejna, Evelyn Choi, Md Sazzad Islam, and Dorsa Sadigh. So you think you can scale up autonomous robot data collection? In *CoRL*, 2024.

[73] Mitsuhiko Nakamoto, Oier Mees, Aviral Kumar, and Sergey Levine. Steering Your Generalists: Improving robotic foundation models via value guidance. In *CoRL*, 2024.

[74] OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

[75] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

[76] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In *ICRA*, 2024.

[77] Giuseppe Paolo, Jonas Gonzalez-Billandon, and Balázs Kégl. Position: A Call for Embodied AI. In *ICML*, 2024.

[78] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *CVPR*, 2024.

[79] Andi Peng, Ilia Sucholutsky, Belinda Z. Li, Theodore Sumers, Thomas L. Griffiths, Jacob Andreas, and Julie Shah. Learning with language-guided state abstractions. In *ICLR*, 2024.

[80] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

[81] Haozhi Qi, Brent Yi, Sudharshan Suresh, Mike Lambeta, Yi Ma, Roberto Calandra, and Jitendra Malik. General in-hand o=object rotation with vision and touch. In *CoRL*, 2023.

[82] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023.

[83] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your language model is secretly a reward model. In *NeurIPS*, 2024.

[84] Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment anything meets point tracking. In *WACV*, 2025.

[85] Anian Ruoss, Gregoire Deletang, Sourabh Medapati, Jordi Grau-Moya, Li Kevin Wenliang, Elliot Catt, John Reid, Cannada A Lewis, Joel Veness, and Tim Genewein. Amortized Planning with Large-Scale Transformers: A case study on chess. In *NeurIPS*, 2024.

[86] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *ICML*, 2015.

[87] Junyao Shi, Zhuolun Zhao, Tianyou Wang, Ian Pedroza, Amy Luo, Jie Wang, Jason Ma, and Dinesh Jayaraman. ZeroMimic: Distilling robotic manipulation skills from web videos. *arXiv preprint arXiv:2503.23877*, 2025.

[88] Lucy Xiaoyang Shi, Zheyuan Hu, Tony Z Zhao, Archit Sharma, Karl Pertsch, Jianlan Luo, Sergey Levine, and Chelsea Finn. Yell At Your Robot: Improving on-the-fly from language corrections. In *RSS*, 2024.

[89] Modi Shi, Li Chen, Jin Chen, Yuxiang Lu, Chiming Liu, Guanghui Ren, Ping Luo, Di Huang, Maoqing Yao, and Hongyang Li. Is diversity all you need for scalable robotic manipulation? *arXiv preprint arXiv:2507.06219*, 2025.

[90] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[91] Himanshu Gaurav Singh, Antonio Loquercio, Carmelo Sferrazza, Jane Wu, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Hand-object interaction pretraining from videos. *arXiv preprint arXiv:2409.08273*, 2024.

[92] Linda Smith and Michael Gasser. The Development of Embodied Cognition: Six lessons from babies. *Artificial Life*, 2005.

[93] Mohan Kumar Srirama, Sudeep Dasari, Shikhar Bahl, and Abhinav Gupta. HRP: Human affordances for robotic pre-training. In *RSS*, 2024.

[94] Robert J Sternberg. *Handbook of Human Intelligence*. Cambridge university press, 1982.

[95] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[96] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[97] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. In *RSS*, 2024.

[98] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[99] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. BridgeData v2: A dataset for robot learning at scale. In *CoRL*, 2023.

[100] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. MimicPlay: Long-horizon imitation learning by watching human play. In *CoRL*, 2023.

[101] Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press, 2019.

[102] Fei Xia, Chengshu Li, Roberto Martín-Martín, Or Litany, Alexander Toshev, and Silvio Savarese. ReLMoGen: Integrating motion generation in reinforcement learning for mobile manipulation. In *ICRA*, 2021.

[103] Zhenjia Xu, Jiajun Wu, Andy Zeng, Joshua B Tenenbaum, and Shuran Song. DensePhysNet: Learning dense physical object representations via multi-step dynamic interactions. In *RSS*, 2019.

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

[105] Rui Yang, Yiming Lu, Wenzhe Li, Hao Sun, Meng Fang, Yali Du, Xiu Li, Lei Han, and Chongjie Zhang. Rethinking goal-conditioned supervised learning and its connection to offline RL. In *ICLR*, 2022.

[106] Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *ICLR*, 2024.

[107] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *CoRL*, 2020.

[108] Xiu Yuan, Tongzhou Mu, Stone Tao, Yunhao Fang, Mengke Zhang, and Hao Su. Policy Decorator: Model-agnostic online refinement for large policy model. In *ICLR*, 2025.

[109] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. In *CoRL*, 2024.

[110] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. CoT-VLA: Visual chain-of-thought reasoning for vision-language-action models. In *CVPR*, 2025.

[111] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023.