Embodied-Harness: Towards Recursive Self-Improvement for Robotic Manipulation

Zhenxuan Fan[1], **Zhuoya Wang**[2], **Yuqian Yuan**[1], **Wenqiao Zhang**[1†], **Juncheng Li**[1], **Jun Xiao**[1], **Siliang Tang**[1], **Yueting Zhuang**[1]

[1]Zhejiang University, [2]Huazhong University of Science and Technology
[†]Corresponding author

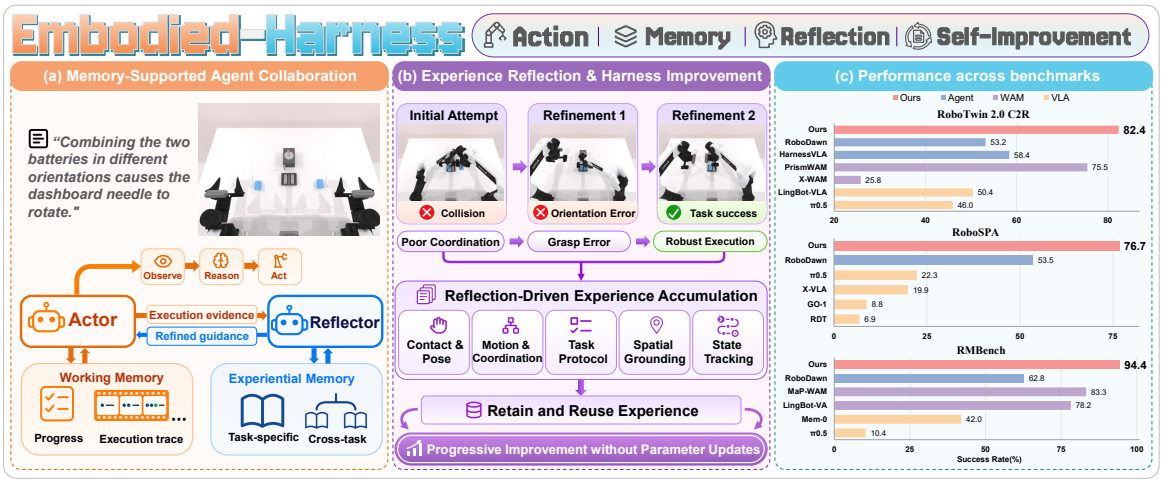


Figure 1 Overview of Embodied-Harness. **(a)** Memory-supported collaboration between the Actor and Reflector for task execution, experience analysis, and guidance generation. **(b)** Experience-driven reflection and recursive harness self-improvement, progressively correcting execution failures through evidence-based guidance without model parameter updates. **(c)** Performance comparison on RoboTwin 2.0, RoboSPA, and RMBench, demonstrating the effectiveness of Embodied-Harness across diverse manipulation benchmarks.

Abstract

Vision-language model (VLM)-based agents offer a promising approach to robotic manipulation, yet remain limited by inadequate memory organization, uncertain failure attribution, and limited reuse of execution experience. We propose **Embodied-Harness**, a dual-agent framework that improves manipulation by refining the execution harness without updating model parameters. The Actor interacts with the environment through a robot-control interface, while the Reflector converts execution evidence into reusable operational guidance. Hierarchical Memory Architecture maintains short-term execution context and long-term experiential knowledge to support execution and reflection. Evidence-based Experience Reflection grounds failure diagnosis in observed interactions, while Recursive Harness Self-Improvement refines operational guidance across successive attempts. Experiments on RoboTwin 2.0, RoboSPA, and RMBench demonstrate the effectiveness of Embodied-Harness in generalization under environmental variations, spatial reasoning, long-horizon planning, and memory-dependent manipulation. Our method achieves success rates of **82.4%**, **76.7%**, and **94.4%** on the three benchmarks, respectively. Real-world evaluation further yields an average success rate of **96.7%** across three manipulation tasks. These results highlight the potential of Embodied-Harness to improve robotic manipulation through experience-driven harness self-improvement.

1 Introduction

Robotic manipulation is a fundamental capability of embodied intelligence [1–3], enabling robots to interact with the physical world and perform diverse tasks across household assistance [4, 5], industrial manufacturing [6, 7], and healthcare [8, 9]. Developing general manipulation capabilities is essential for enabling versatile and autonomous robotic operation in real-world environments, representing a critical step toward general embodied intelligence (GEI) [10–12].

Towards this goal, different paradigms have emerged for robotic manipulation. End-to-end frameworks, including vision-language-action (VLA) models [13–19] and world-action models (WAMs) [20–25], have demonstrated remarkable progress in learning robotic policies from multimodal inputs. However, these approaches often rely on large-scale robotic datasets, face challenges in generalizing to unseen environments [26–28], and struggle with reliable long-horizon manipulation [29–31]. In contrast, agentic frameworks leverage the reasoning and planning capabilities of vision-language models (VLMs) to decompose complex tasks and coordinate robot actions [32–34]. Such frameworks support flexible action planning and feedback-driven execution, offering a promising alternative without requiring task-specific updates to the underlying model parameters [35–37].

Despite these advantages, agentic frameworks still face several challenges in achieving reliable robotic manipulation. **(i) Inadequate Memory Organization.** Many real-world manipulation tasks are inherently long-horizon, requiring agents to track task progress and detailed execution histories while retaining experience across attempts and tasks [38–41]. Managing such information across different temporal scales and levels of granularity remains challenging. **(ii) Uncertain Failure Attribution.** Manipulation failures may stem from inaccurate action grounding, unsatisfied preconditions, or violated state constraints [42–44]. However, execution feedback often reveals only observable outcomes rather than underlying causes, making it difficult to identify critical errors and derive reliable operational guidance. **(iii) Limited Self-Improvement Capability.** Even when agents identify execution errors and extract useful insights, these reflections may remain isolated from subsequent executions. Without mechanisms to accumulate, validate, and refine operational guidance across attempts, agents may repeatedly encounter similar failures [45–47].

To address these challenges, we propose **Embodied-Harness**, a dual-agent framework for robotic manipulation that enables recursive harness self-improvement through accumulated execution experience (Fig. 1). The framework comprises two complementary roles, an **Actor** and a **Reflector**. The Actor interacts with the environment through control interfaces, while the Reflector analyzes execution trajectories and environmental feedback to improve subsequent task execution. Specifically, Embodied-Harness incorporates three complementary mechanisms: **(i) Hierarchical Memory Architecture** organizes short-term execution context and long-term experiential knowledge across different temporal scales; **(ii) Evidence-based Experience Reflection** identifies potential failure causes from execution evidence and derives actionable operational guidance; and **(iii) Recursive Harness Self-Improvement** refines the execution harness using accumulated guidance, with subsequent execution feedback informing further reflection and refinement. Together, these mechanisms enable continual improvement in robotic manipulation without updating the underlying model parameters.

We evaluate Embodied-Harness on three robotic manipulation benchmarks, RoboTwin 2.0, RoboSPA, and RMBench, covering generalization under environmental variations, fine-grained spatial reasoning, long-horizon procedural planning, and memory-dependent manipulation. Embodied-Harness outperforms baselines spanning VLA models, WAMs, and agentic frameworks, achieving average success rates of **82.4%**, **76.7%**, and **94.4%** on the three benchmarks, respectively. Real-world evaluation further yields an average success rate of **96.7%** across three manipulation tasks. Ablation studies and further analyses support the effectiveness of the proposed components and illustrate how accumulated experience enables more reliable manipulation.

Our main contributions are summarized as follows:

- We propose Embodied-Harness, an experience-driven dual-agent framework that improves robotic manipulation through complementary execution and reflection without updating model parameters.
- We design a Hierarchical Memory Architecture that integrates short-term working memory and long-term experiential memory to support coherent task execution and effective experience reuse.

- We introduce Evidence-based Experience Reflection and Recursive Harness Self-Improvement to derive actionable guidance from execution feedback and refine the harness across successive attempts.
- We conduct extensive experiments on three robotic manipulation benchmarks and real-world tasks, demonstrating the effectiveness of Embodied-Harness across diverse manipulation settings.

2 Related Work

2.1 End-to-End Models for Robotic Manipulation

End-to-end policies map visual observations and language instructions directly to robot actions. Vision-language-action (VLA) models build this mapping on large-scale robot data and pretrained vision-language backbones [13, 14, 48, 49], with later works adopting diffusion or flow-matching action generators [15, 16, 50, 51] and extending them with cross-embodiment training, efficiency, or visual reasoning [17–19, 52, 53]. World action models (WAMs) instead exploit video generative priors, either guiding actions with predicted futures [20, 54, 55] or jointly modeling video and actions [21–25, 56]. Despite their progress, these models typically require large-scale robotic training data and remain limited in generalizing to unseen environments and task variations.

2.2 Agentic Frameworks for Robotic Manipulation

Agentic frameworks connect foundation models to robots through perception, planning, and control interfaces. Language models sequence skills, generate plans or programs, and incorporate environmental feedback [32, 33, 35, 57, 58], while embodied VLMs predict plans, affordances, and spatial targets [12, 59–61]. Other methods translate VLM reasoning into value maps, keypoints, or geometric constraints for motion generation [34, 36, 62–65], or pair high-level reasoners with low-level policies [15, 66–68]. Closest to our setting, RoboDawn [69] controls robots through VLM-issued commands, while Harness VLA [37] combines analytic primitives with a frozen VLA through memory-guided planning. Embodied-Harness advances these frameworks through coordinated execution and reflection, enabling evidence-based failure diagnosis, experience reuse, and progressive refinement of execution strategies across attempts.

2.3 Reflection and Self-Improvement in Agents

Reflection lets agents improve from feedback without updating model parameters. Language agents reason while acting [70], revise their own outputs [71], or store verbal reflections across trials [45], while others distill experience into insights, memories, skills, or rules [38, 39, 46, 47]. Prompt and agent-design optimization further improves systems built on fixed models [72–76]. In robotics, prior work explains, detects, or corrects manipulation failures [42–44], and SenseNova-RoboRSI [68] revises the agent harness and retains changes validated by re-evaluation. Embodied-Harness emphasizes the link between failure diagnosis and corrective guidance, using observed execution evidence to identify potential failure causes and specify actionable execution conditions for subsequent robotic manipulation attempts.

3 Method

3.1 Framework Overview

We propose **Embodied-Harness**, a dual-agent framework that improves robotic manipulation through accumulated execution experience. As illustrated in Fig. 2, Embodied-Harness consists of two complementary agents, an **Actor** and a **Reflector**, which collaborate through a recursive execution–reflection process. The framework integrates Hierarchical Memory Architecture, Evidence-based Experience Reflection, and Recursive Harness Self-Improvement to enable progressive improvement without updating the underlying model parameters.

Agent Role Specialization. Embodied-Harness separates task execution from experience reflection. The Actor interacts with the environment through control interfaces to carry out the task, while the Reflector examines execution outcomes to improve subsequent attempts. Drawing on execution evidence and accumulated experience, the Reflector diagnoses potential failure causes and generates corrective guidance to refine the

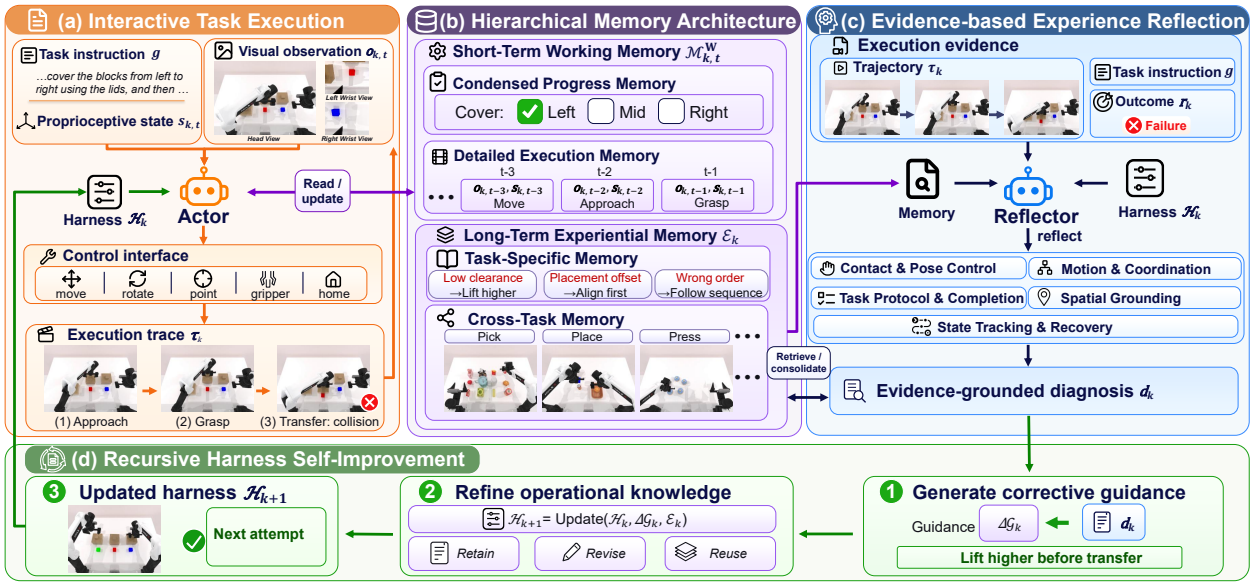


Figure 2 Framework of Embodied-Harness. **(a)** The Actor performs task execution through a perception–action loop guided by the execution harness. **(b)** Hierarchical memory maintains short-term execution context and long-term experience across attempts and tasks. **(c)** The Reflector analyzes execution evidence and experience to diagnose failures and identify improvements. **(d)** Recursive harness self-improvement generates guidance and refines operational knowledge to improve subsequent execution without updating model parameters.

execution harness. The updated harness then guides the Actor, forming a closed loop of execution, reflection, and self-improvement.

Interactive Action Execution. Given a task instruction g, the Actor selects a control command and grounded arguments based on visual observations, proprioceptive states, the current harness, and working memory:

$$a_{k,t} \sim \pi_\theta \left(\cdot \mid g, o_{k,t}, s_{k,t}, \mathcal{H}_k, \mathcal{M}^{\mathrm{W}}_{k,t} \right), \tag{1}$$

where $o_{k,t}$ and $s_{k,t}$ denote the visual observation and proprioceptive state at step t of execution attempt k, respectively. Here, $a_{k,t}$ denotes the selected action, π_θ is the frozen Actor model, $\mathcal{H}_k$ denotes the current execution harness, and $\mathcal{M}^{\mathrm{W}}_{k,t}$ represents the Actor's short-term working memory, detailed in Sec. 3.2. The action interface supports five types of control commands [69]:

$$\mathcal{A}_{\mathrm{type}} = \{\texttt{move}, \texttt{rotate}, \texttt{point}, \texttt{gripper}, \texttt{home}\}.$$

These commands support end-effector translation, rotation, predefined orientation control, gripper actuation, and returning to the home pose, respectively.

For commands involving end-effector motion, the harness converts the action arguments into a target pose $\hat{X}_{k,t} \in SE(3)$. The corresponding inverse kinematics (IK) problem is formulated as

$$\mathbf{q}^\star_{k,t} \in \arg \min_{\mathbf{q} \in \mathcal{Q}_{\mathrm{valid}}} D\Big(\mathrm{FK}(\mathbf{q}), \hat{X}_{k,t}\Big), \tag{2}$$

where FK denotes forward kinematics, D measures translational and rotational pose discrepancies, and $\mathcal{Q}_{\mathrm{valid}}$ denotes the set of joint configurations satisfying joint limits and collision constraints. The resulting joint configuration is used to plan a trajectory for robot execution. After each action, the Actor uses feedback $f_{k,t}$ and updated observations and states to track progress and select subsequent actions. The trajectory $\tau_k = \{(o_{k,t}, s_{k,t}, a_{k,t}, f_{k,t})\}_{t=1}^{T_k}$ provides evidence for reflection and harness refinement.

3.2 Hierarchical Memory Architecture

Robotic manipulation requires maintaining execution context within individual episodes while preserving reusable knowledge across attempts and tasks. To support these different temporal scales and information

granularities, we introduce a Hierarchical Memory Architecture comprising Short-Term Working Memory and Long-Term Experiential Memory.

Short-Term Working Memory. Short-Term Working Memory supports the Actor's decision-making within a single execution episode by continuously maintaining task progress and execution context. It comprises *Condensed Progress Memory* and *Detailed Execution Memory*, which capture high-level task states and fine-grained interaction histories, respectively.

Condensed Progress Memory. Long-horizon manipulation involves interdependent subgoals whose progress is often scattered across interaction histories, making it difficult to maintain coherent task states. To address this, *Condensed Progress Memory* consolidates completed subgoals, pending operations, environmental changes, and execution constraints into a compact representation:

$$\mathcal{M}^{c}_{k,t+1} = \text{Summarize}\left(\mathcal{M}^{c}_{k,t}, e_{k,t}\right), \tag{3}$$

where $e_{k,t} = (o_{k,t}, s_{k,t}, a_{k,t}, f_{k,t})$ denotes the interaction record at step t of attempt k. This enables the Actor to track task progress without repeatedly reconstructing it from historical interactions.

Detailed Execution Memory. Although condensed representations maintain task progress, they may omit critical action-level details needed to identify and avoid previous mistakes. *Detailed Execution Memory* therefore preserves chronological records of visual observations, proprioceptive states, actions, and execution feedback:

$$\mathcal{M}^{d}_{k,t+1} = \mathcal{M}^{d}_{k,t} \oplus e_{k,t}, \tag{4}$$

where $\oplus$ denotes the append operation. This memory supports feedback-driven action adjustment and provides execution evidence for the Reflector's failure analysis.

Together, these complementary components constitute the Actor's short-term working memory:

$$\mathcal{M}^{\mathrm{W}}_{k,t} = \left(\mathcal{M}^{c}_{k,t}, \mathcal{M}^{d}_{k,t}\right). \tag{5}$$

Long-Term Experiential Memory. Long-Term Experiential Memory primarily serves the Reflector by preserving experience across episodes, providing historical context for failure diagnosis and subsequent self-improvement. It comprises *Task-Specific Memory* and *Cross-Task Memory*, supporting task-dependent refinement and transferable knowledge reuse, respectively.

Task-Specific Memory. Repeated attempts at the same task may encounter similar failures, motivating the retention of previous execution experience. *Task-Specific Memory* records historical outcomes, diagnosed errors, and corrective guidance associated with each task:

$$\mathcal{E}^{s}_{k+1}(i) = \text{Update}\left(\mathcal{E}^{s}_{k}(i), r_k, d_k, \Delta\mathcal{G}_k\right), \tag{6}$$

where i denotes the task identity, shared across episodes with potentially different instructions, and r_k, d_k, and $\Delta\mathcal{G}_k$ denote the task outcome, execution diagnosis, and generated guidance, respectively. This memory helps identify recurring failure patterns and supports targeted improvement across attempts.

Cross-Task Memory. Different manipulation tasks often share transferable operational principles. *Cross-Task Memory* captures such knowledge from task-specific experience, including spatial grounding, contact control, action coordination, and feedback-based recovery. Its update is formulated as

$$\mathcal{E}^{x}_{k+1} = \text{Generalize}\left(\mathcal{E}^{x}_{k}, \mathcal{E}^{s}_{k+1}(i)\right), \tag{7}$$

where Generalize$(\cdot)$ abstracts transferable knowledge from task-specific experience. This enables the Reflector to draw on accumulated operational knowledge when analyzing new but related manipulation tasks.

Together, these complementary components constitute the long-term experiential memory available for task i:

$$\mathcal{E}_k = \left(\mathcal{E}^{s}_{k}(i), \mathcal{E}^{x}_{k}\right). \tag{8}$$

3.3 Evidence-based Experience Reflection

Robotic manipulation failures may arise from different underlying causes, while similar task outcomes often provide limited evidence for identifying execution errors. Moreover, unconstrained reflection may produce explanations unsupported by observed interactions. To address these challenges, we introduce Evidence-based Experience Reflection, which grounds failure diagnosis in concrete execution evidence and accumulated experience.

After each execution attempt, the Reflector analyzes the task instruction g, trajectory τ_k, task outcome r_k, current harness $\mathcal{H}_k$, and experiential memory $\mathcal{E}_k$:

$$d_k = \mathcal{R}_\phi(g, \tau_k, r_k, \mathcal{H}_k, \mathcal{E}_k), \tag{9}$$

where $\mathcal{R}_\phi$ denotes the frozen Reflector model and d_k represents the execution diagnosis.

Each reflection outcome is assigned to exactly one of the following five operational categories:

- **Contact & Pose Control:** Examines errors in physical interactions, including grasp configurations, contact positions, gripper operations, and object orientations.
- **Motion & Coordination:** Focuses on improper action sequencing, arm coordination, and motion planning that may lead to unsuccessful or unsafe manipulation.
- **Task Protocol & Completion:** Captures failures to satisfy task requirements, execution restrictions, success conditions, or termination criteria.
- **Spatial Grounding:** Addresses incorrect target identification, spatial relationship interpretation, and coordinate grounding that result in actions at unintended locations.
- **State Tracking & Recovery:** Identifies deficiencies in tracking task progress, verifying action outcomes, interpreting feedback, and recovering from execution failures.

The Reflector identifies discrepancies between intended and observed outcomes to produce evidence-grounded diagnoses for subsequent harness self-improvement.

3.4 Recursive Harness Self-Improvement

Following evidence-based reflection, the framework converts execution diagnosis into reusable corrective guidance and refines the execution harness.

Given the diagnosis d_k produced by the Reflector, the framework identifies operations requiring adjustment and specifies their execution conditions, conditioned on the current harness and accumulated experience:

$$\Delta\mathcal{G}_k = \text{Generate}(d_k, \mathcal{H}_k, \mathcal{E}_k), \tag{10}$$

where $\Delta\mathcal{G}_k$ denotes candidate guidance that translates identified execution issues into actionable instructions for subsequent attempts.

The candidate guidance is integrated with existing operational knowledge to update the harness:

$$\mathcal{H}_{k+1} = \text{Update}(\mathcal{H}_k, \Delta\mathcal{G}_k, \mathcal{E}_k). \tag{11}$$

The update uses accumulated experience to retain effective guidance, resolve inconsistencies, and refine instructions based on execution evidence. The refined harness $\mathcal{H}_{k+1}$ guides subsequent Actor attempts, whose trajectories and feedback support further reflection. This cycle of execution, diagnosis, guidance generation, and harness refinement enables progressive improvement without updating model parameters.

4 Experiments

4.1 Experimental Setup

Evaluation Benchmarks. We evaluate Embodied-Harness on three representative simulation benchmarks: RoboTwin 2.0 [85], RoboSPA [31], and RMBench [41], alongside three real-world manipulation tasks. RoboTwin

Table 1 Success rates (SR, %) on RoboTwin 2.0 under the Clean-to-Randomized (C2R) setting. Methods are grouped into VLA models, WAMs, and agentic frameworks. Shots denotes the number of task-specific demonstrations used. Our method is highlighted in blue, and the best result is boldfaced.

Method	Shots	SR (%)
Vision-Language-Action Models		
StarVLA [77]	Full Set	10.6
Xiaomi-Robotics-0 [78]	Full Set	18.2
GR00T-N1.7 [15]	Full Set	20.7
X-VLA [18]	Full Set	20.9
Spatial Forcing [79]	Full Set	26.7
$\pi_{0.5}$ [16]	Full Set	46.0
LingBot-VLA [19]	Full Set	50.4
GigaBrain-0.7 [67]	Full Set	67.9
World-Action Models		
FastWAM [25]	Full Set	1.9
X-WAM [80]	Full Set	25.8
PatchWAM [81]	Full Set	66.7
UniWAM [82]	Full Set	68.3
PICA-WAM [83]	Full Set	74.7
PrismWAM* [84]	Full Set	75.5
Agentic Frameworks		
HarnessVLA (Codex) [37]	Full Set	58.0
HarnessVLA (CC) [37]	Full Set	58.4
RoboDawn (GPT-6 Astra) [69]	0	53.2
RoboDawn (GPT-6 Astra) [69]	1	73.6
Ours (GPT-6 Astra)	0	**82.4**

2.0 includes 50 tasks evaluated under the Clean-to-Randomized (C2R) setting. RoboSPA comprises 24 fine-grained spatial reasoning tasks and 32 long-horizon procedural planning tasks, both evaluated at the highest difficulty level (L5). We additionally evaluate spatial reasoning across L1–L4 to examine performance under varying task difficulty. RMBench contains 9 memory-dependent manipulation tasks requiring information retention across an episode. Finally, three real-world tasks assess the effectiveness of Embodied-Harness in physical environments.

Evaluation Protocol. We use GPT-6 Astra as the backbone model, with medium reasoning effort as the default setting. Each evaluation episode is limited to 45 interaction rounds. For real-world experiments, we use the Piper robotic arm as the manipulation platform. We conduct 10 evaluation episodes per task on RoboTwin 2.0, RoboSPA, and real-world tasks, and 20 episodes per task on RMBench. We report success rates (SR) for each benchmark, and additionally report progress scores (PS) on RoboSPA long-horizon procedural planning tasks to measure the degree of partial task completion.

4.2 Main Results

RoboTwin 2.0 C2R. Table 1 compares Embodied-Harness with representative vision-language-action models and world-action models, as well as agentic frameworks, on RoboTwin 2.0 C2R. Our method achieves an overall success rate of **82.4%**, outperforming all compared methods. Under the same zero-shot setting and with the same GPT-6 Astra backbone, Embodied-Harness surpasses RoboDawn by 29.2 percentage points. This comparison supports the effectiveness of our framework in improving manipulation performance beyond the underlying model's capabilities.

Our method also exceeds the strongest VLA and WAM baselines, GigaBrain-0.7 and PrismWAM, by 14.5 and 6.9 percentage points, respectively, and outperforms the best HarnessVLA variant by 24.0 percentage points. These methods use the full demonstration set, whereas Embodied-Harness operates with zero demonstrations.

Table 2 Success rates (SR, %) on RoboSPA fine-grained spatial reasoning at the highest difficulty level (L5). Avg. denotes the unweighted average across five categories.

Method	GAC	SDE	CPI	RRR	CVR	**Avg.**
RDT [50]	4.8	9.4	8.6	4.4	3.4	6.1
GO-1 [52]	11.5	11.6	7.2	6.6	16.2	10.6
$\pi_{0.5}$ [16]	12.5	23.4	23.6	9.8	37.6	21.4
X-VLA [18]	25.3	34.6	8.8	14.2	36.6	23.9
RoboDawn [69]	80.0	68.0	74.0	82.0	24.0	65.6
Ours	**100.0**	**94.0**	**92.0**	**86.0**	**98.0**	**94.0**

Table 3 Success rates (SR, %) and progress scores (PS, %) on RoboSPA long-horizon procedural planning tasks at L5.

Method	**SR (%)**						**PS (%)**					
	RPF	OFE	OCE	CAC	MIP	Avg.	RPF	OFE	OCE	CAC	MIP	Avg.
RDT [50]	35.0	0.8	0.1	2.4	0.0	7.7	44.5	24.9	14.7	26.5	3.6	22.8
GO-1 [52]	23.8	5.6	3.1	2.6	0.0	7.0	32.0	25.7	14.8	22.7	6.0	20.2
$\pi_{0.5}$ [16]	**65.0**	26.1	12.0	12.5	0.0	23.1	**77.3**	62.1	28.0	48.0	11.4	45.4
X-VLA [18]	22.5	43.6	7.7	5.9	0.0	15.9	31.9	73.9	25.1	40.1	7.3	35.7
RoboDawn [69]	47.5	60.0	48.6	**35.0**	16.0	41.4	53.0	75.0	63.3	63.5	27.2	56.4
Ours	57.5	**66.2**	**58.6**	**35.0**	**80.0**	**59.5**	60.5	**81.0**	**72.6**	**68.5**	**92.0**	**74.9**

Together, these results demonstrate the effectiveness of Embodied-Harness on RoboTwin 2.0 C2R without task demonstrations or model parameter updates.

RoboSPA-Spatial. Table 2 presents the results on fine-grained spatial reasoning at the highest difficulty level (L5). Embodied-Harness achieves an average SR of 94.0%, outperforming the strongest baseline, RoboDawn, by 28.4 percentage points. Our method leads across all five categories, with a particularly substantial improvement on CVR from 24.0% to 98.0%. Furthermore, Fig. 3 shows consistently high performance across all five difficulty levels, with only minor degradation as task complexity increases. In contrast, competing methods exhibit more pronounced declines, highlighting the robustness of Embodied-Harness in challenging spatial reasoning scenarios.

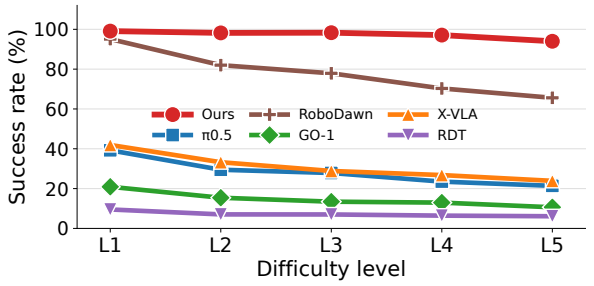


Figure 3 Success rates across five difficulty levels on RoboSPA fine-grained spatial reasoning tasks. Embodied-Harness maintains 94–99% success rates from L1 to L5, whereas RoboDawn and VLA baselines generally decline with increasing difficulty.

RoboSPA-Long. Long-horizon procedural planning requires agents to track task progress and coordinate interdependent operations over extended sequences. As shown in Table 3, Embodied-Harness achieves the highest average SR of 59.5% and PS of 74.9%, surpassing RoboDawn by 18.1 and 18.5 percentage points, respectively. Our method ranks first in four of five categories for both metrics, including a tie in CAC SR. The largest gains occur in MIP, with improvements of 64.0 and 64.8 percentage points in SR and PS over the strongest baseline. These results support the effectiveness of Embodied-Harness in sustaining progress and completing long-horizon manipulation tasks.

RMBench. Table 4 reports results on nine memory-dependent manipulation tasks. Embodied-Harness achieves the highest average success rate of 94.4%, surpassing the strongest baseline, MaP-WAM, by 11.1 percentage points. It achieves 100% success on five tasks and at least 85% on the remaining four, demonstrating consistent performance across diverse scenarios. These results support our hierarchical memory design: Short-Term Working Memory preserves task states and interaction histories for action selection, while Long-Term Experiential Memory retains task-specific and transferable knowledge for reflection and iterative improvement.

Table 4 Success rates (%) on the nine memory-dependent manipulation tasks in RMBench.

Task	*VLA Models*			*WAMs*			*Agentic Frameworks*	
	$\pi_{0.5}$ [16]	X-VLA [18]	Mem-0 [41]	LingBot-VA [23]	MemoryWAM [86]	MaP-WAM [87]	RoboDawn [69]	**Ours**
Observe and Pick Up	9	9	4	13	27	19	**100**	**100**
Rearrange Blocks	13	13	89	**100**	**100**	66	**100**	**100**
Put Back Block	11	18	90	**100**	**100**	**100**	25	**100**
Swap Blocks	24	16	67	99	**100**	97	0	90
Swap T	15	3	14	88	94	**96**	80	90
Battery Try	16	26	28	41	41	82	35	**100**
Blocks Ranking Try	6	1	18	**100**	**100**	94	**100**	**100**
Cover Blocks	0	2	68	79	98	**100**	60	85
Press Button	0	0	0	84	87	**96**	65	85
Average	10.4	9.8	42.0	78.2	83.0	83.3	62.8	**94.4**

Table 5 Success rates (%) on three real-world robotic manipulation tasks. Embodied-Harness outperforms $\pi_{0.5}$ on all three tasks, with an average success rate of 96.7% versus 40.0%.

Method	Arrangement	Placement	Stacking	Avg.
$\pi_{0.5}$	40.0	50.0	30.0	40.0
Ours	**90.0**	**100.0**	**100.0**	**96.7**

Table 6 Ablation results on RoboSPA-Long (L5).

Method	SR (%)	PS (%)
Embodied-Harness	**59.5**	**74.9**
w/o Reflector	43.0	61.2
w/o Short-Term Working Memory	32.1	49.5
w/o Long-Term Experiential Memory	48.5	65.6
w/o Evidence Grounding	52.3	67.9

Real-World Experiments. We evaluate Embodied-Harness on three real-world manipulation tasks requiring multi-step reasoning and long-horizon execution (Fig. 4). As reported in Table 5, Embodied-Harness achieves an average success rate of 96.7%, outperforming $\pi_{0.5}$ by 56.7 percentage points. Our method achieves 90.0% success on Letter Arrangement and 100.0% on both Block Placement and Block Stacking, while $\pi_{0.5}$ achieves 40.0% success on average. These results demonstrate the effectiveness of Embodied-Harness in real-world manipulation requiring complex reasoning and coordinated multi-step execution.

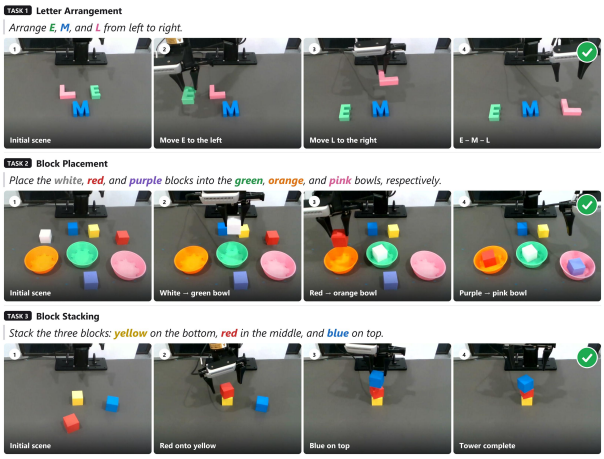
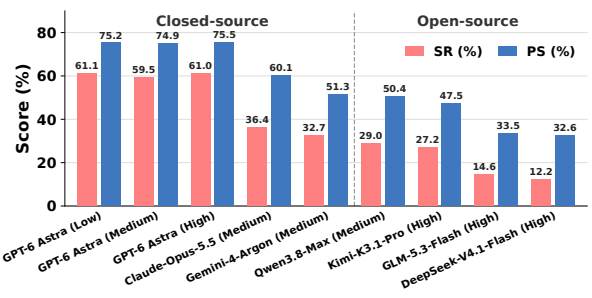


Figure 4 Execution examples on three real-world tasks.

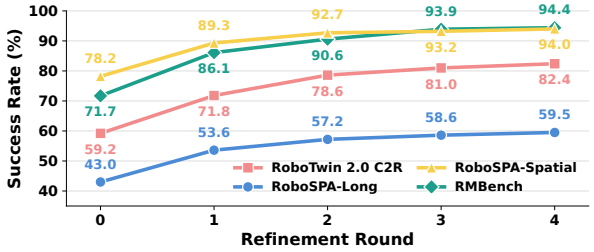
4.3 Ablation Studies

Effect of individual components. Table 6 reports ablation results on RoboSPA-Long (L5), where all ablations degrade performance. Without Short-Term Working Memory, SR drops from 59.5% to 32.1% and PS from 74.9% to 49.5%, marking the largest decline and highlighting its role in maintaining execution context. Ablating the Reflector reduces SR to 43.0%, while excluding Long-Term Experiential Memory yields 48.5% SR, supporting the value of reflection and accumulated experience. Without Evidence Grounding, SR and PS decrease to 52.3% and 67.9%, respectively, underscoring the importance of execution evidence for reliable corrective guidance.

Effect of VLM Backbone and Reasoning Effort. Figure 5a compares the performance of different VLM backbones and reasoning effort levels on RoboSPA-Long. GPT-6 Astra consistently achieves the strongest performance, with SRs of 59.5–61.1% and PSs of 74.9–75.5%, substantially outperforming other backbones. Closed-source models generally outperform open-source alternatives, with the strongest open-source model, Qwen3.8-Max, achieving only 29.0% SR and 50.4% PS. Notably, increasing the reasoning effort of GPT-6 Astra from low to



(a) Backbone and reasoning effort.



(b) Recursive harness self-improvement.

Figure 5 **(a)** Effect of VLM backbone and reasoning effort on RoboSPA-Long. **(b)** Success rates across recursive harness self-improvement rounds.

high yields no consistent improvement, suggesting that backbone capability plays a more critical role than reasoning effort.

4.4 In-Depth Analysis

Recursive Harness Self-Improvement. Figure 5b shows consistent performance gains across successive harness refinement rounds. From round 0 to round 4, success rates increase from 59.2% to 82.4% on RoboTwin 2.0 C2R, 78.2% to 94.0% on RoboSPA-Spatial, 43.0% to 59.5% on RoboSPA-Long, and 71.7% to 94.4% on RMBench. Improvements are largest in the early rounds and gradually diminish. These results support the effectiveness of repeated reflection and harness refinement in improving manipulation performance without updating model parameters.

Distribution of reflection outcomes. Figure 6 summarizes the reflection outcomes across five categories. Contact and pose control accounts for the largest share (34.5%), followed by motion and coordination (22.7%) and task protocol and completion (21.8%). Spatial grounding (10.1%) and state tracking and recovery (10.9%) comprise the remaining outcomes. This distribution suggests that reflection addresses both low-level execution issues and higher-level task requirements, providing corrective guidance across multiple aspects of manipulation.

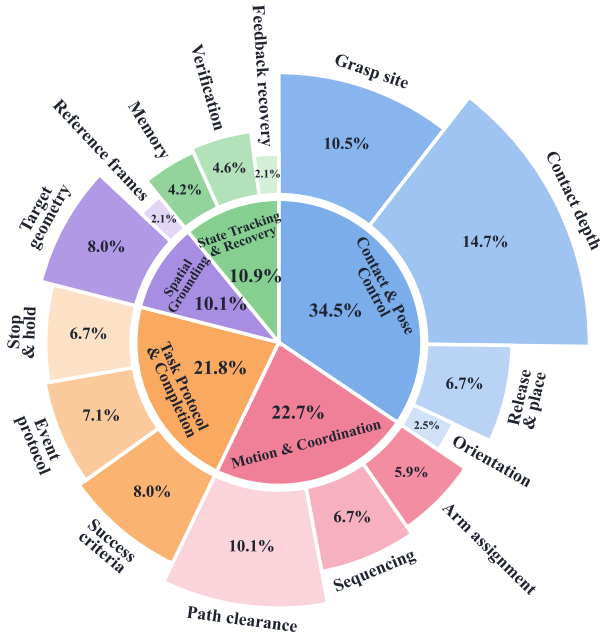


Figure 6 Distribution of reflection outcomes across operational categories.

5 Conclusion

We presented Embodied-Harness, a dual-agent framework that converts execution experience into reusable operational guidance for robotic manipulation without updating model parameters. Hierarchical memory supports context-aware execution by the Actor and evidence-based reflection by the Reflector, enabling iterative harness improvement. Experiments on RoboTwin 2.0, RoboSPA, RMBench, and real-world tasks demonstrate its effectiveness, while ablations and iterative analyses substantiate the roles of memory, reflection, and evidence grounding. These findings establish harness refinement as a promising direction for improving embodied agents. Future work will investigate experience transfer across tasks and embodiments and more efficient execution and reflection.

References

[1] Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446):eaat8414, 2019. doi: 10.1126/science.aat8414. URL https://www.science.org/doi/abs/10.1126/science.aat8414.

[2] Oliver Kroemer, Scott Niekum, and George Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *Journal of Machine Learning Research*, 22(30):1–82, 2021. URL http://jmlr.org/papers/v22/19-804.html.

[3] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022. doi: 10.1109/TETCI.2022.3141105.

[4] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, Mona Anvari, Minjune Hwang, Manasi Sharma, Arman Aydin, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Silvio Savarese, Hyowon Gweon, Karen Liu, Jiajun Wu, and Li Fei-Fei. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 80–93. PMLR, 14–18 Dec 2023. URL https://proceedings.mlr.press/v205/li23a.html.

[5] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation using low-cost whole-body teleoperation. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 4066–4083. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/fu25b.html.

[6] Yashraj Narang, Kier Storey, Iretiayo Akinola, Miles Macklin, Philipp Reist, Lukasz Wawrzyniak, Yunrong Guo, Adam Moravanszky, Gavriel State, Michelle Lu, Ankur Handa, and Dieter Fox. Factory: Fast Contact for Robotic Assembly. In *Proceedings of Robotics: Science and Systems*, New York City, NY, USA, June 2022. doi: 10.15607/RSS.2022.XVIII.035.

[7] Bingjie Tang, Michael A Lin, Iretiayo A Akinola, Ankur Handa, Gaurav S Sukhatme, Fabio Ramos, Dieter Fox, and Yashraj S Narang. IndustReal: Transferring Contact-Rich Assembly Tasks from Simulation to Reality. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.039.

[8] Pierre E. Dupont, Bradley J. Nelson, Michael Goldfarb, Blake Hannaford, Arianna Menciassi, Marcia K. O'Malley, Nabil Simaan, Pietro Valdastri, and Guang-Zhong Yang. A decade retrospective of medical robotics research from 2010 to 2020. *Science Robotics*, 6(60):eabi8017, 2021. doi: 10.1126/scirobotics.abi8017. URL https://www.science.org/doi/abs/10.1126/scirobotics.abi8017.

[9] Ji Woong Kim, Tony Z. Zhao, Samuel Schmidgall, Anton Deguet, Marin Kobilarov, Chelsea Finn, and Axel Krieger. Surgical robot transformer (srt): Imitation learning for surgical tasks. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 130–144. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/kim25a.html.

[10] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, Brian Ichter, Danny Driess, Jiajun Wu, Cewu Lu, and Mac Schwager. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, 44(5):701–739, 2025. doi: 10.1177/02783649241281508. URL https://doi.org/10.1177/02783649241281508.

[11] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics*, 30(6):7253–7274, 2025. doi: 10.1109/TMECH.2025.3574943.

[12] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Oscar Chang, Jose Enrique Chen, Xi Chen, Hao-Tien Lewis Chiang, Krzysztof Choromanski, David D'Ambrosio, Sudeep Dasari, Todor Davchev, Coline Devin, Norman Di Palo, Tianli Ding, Adil Dostmohamed, Danny Driess, Yilun Du, Debidatta Dwibedi, Michael Elabd, Claudio Fantacci, Cody Fong, Erik Frey, Chuyuan Fu, Marissa Giustina, Keerthana Gopalakrishnan, Laura Graesser, Leonard Hasenclever, Nicolas Heess, Brandon

Hernaez, Alexander Herzog, R. Alex Hofer, Jan Humplik, Atil Iscen, Mithun George Jacob, Deepali Jain, Ryan Julian, Dmitry Kalashnikov, M. Emre Karagozler, Stefani Karp, Chase Kew, Jerad Kirkland, Sean Kirmani, Yuheng Kuang, Thomas Lampe, Antoine Laurens, Isabel Leal, Alex X. Lee, Tsang-Wei Edward Lee, Jacky Liang, Yixin Lin, Sharath Maddineni, Anirudha Majumdar, Assaf Hurwitz Michaely, Robert Moreno, Michael Neunert, Francesco Nori, Carolina Parada, Emilio Parisotto, Peter Pastor, Acorn Pooley, Kanishka Rao, Krista Reymann, Dorsa Sadigh, Stefano Saliceti, Pannag Sanketi, Pierre Sermanet, Dhruv Shah, Mohit Sharma, Kathryn Shea, Charles Shu, Vikas Sindhwani, Sumeet Singh, Radu Soricut, Jost Tobias Springenberg, Rachel Sterneck, Razvan Surdulescu, Jie Tan, Jonathan Tompson, Vincent Vanhoucke, Jake Varley, Grace Vesom, Giulia Vezzani, Oriol Vinyals, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Fei Xia, Ted Xiao, Annie Xie, Jinyu Xie, Peng Xu, Sichun Xu, Ying Xu, Zhuo Xu, Yuxiang Yang, Rui Yao, Sergey Yaroshenko, Wenhao Yu, Wentao Yuan, Jingwei Zhang, Tingnan Zhang, Allan Zhou, and Yuxiang Zhou. Gemini robotics: Bringing ai into the physical world, 2025. URL https://arxiv.org/abs/2503.20020.

[13] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

[14] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 2679–2713. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/kim25c.html.

[15] NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots, 2025. URL https://arxiv.org/abs/2503.14734.

[16] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, brian ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In Joseph Lim, Shuran Song, and Hae-Won Park, editors, *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 17–40. PMLR, 27–30 Sep 2025. URL https://proceedings.mlr.press/v305/black25a.html.

[17] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, Simon Alibert, Matthieu Cord, Thomas Wolf, and Remi Cadene. Smolvla: A vision-language-action model for affordable and efficient robotics, 2025. URL https://arxiv.org/abs/2506.01844.

[18] Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, Tai Wang, Ya-Qin Zhang, Jingjing Liu, and Xianyuan Zhan. X-VLA: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kt51kZH4aG.

[19] Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, Yiyu Ren, Kejia Zhang, Hui Yu, Jingmei Zhao, Shuai Zhou, Zhenqi Qiu, Houlong Xiong, Ziyu Wang, Zechen Wang, Ran Cheng, Yong-Lu Li, Yongtao Huang, Xing Zhu, Yujun Shen, and Kecheng Zheng. A pragmatic vla foundation model, 2026. URL https://arxiv.org/abs/2601.18692.

[20] 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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 24328–24346. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/hu25g.html.

[21] Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 35101–35113, June 2026.

[22] Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos policy: Fine-tuning video models for visuomotor control and planning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 71531–71552, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/748becc400a57c0e31cfe6a2e7951467-Paper-Conference.pdf.

[23] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Luyao Zhang, Mingrui Yu, Zelin Gao, Nan Xue, Boyu Zhou, Xing Zhu, Mingyu Ding, Yujun Shen, and Yinghao Xu. Causal World Modeling for Robot Control. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.016.

[24] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies, 2026. URL https://arxiv.org/abs/2602.15922.

[25] Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination?, 2026. URL https://arxiv.org/abs/2603.16666.

[26] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi Jim Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick Tree Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist,

Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477.

[27] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.120.

[28] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision–language–action models for embodied ai. *IEEE Transactions on Neural Networks and Learning Systems*, 37(7):3031–3051, July 2026. ISSN 2162-2388. doi: 10.1109/tnnls.2025.3650584. URL http://dx.doi.org/10.1109/TNNLS.2025.3650584.

[29] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7 (3):7327–7334, 2022. doi: 10.1109/LRA.2022.3180108.

[30] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 44776–44791. Curran Associates, Inc., 2023. doi: 10.52202/075280-1939. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8c3c666820ea055a77726d66fc7d447f-Paper-Datasets_and_Benchmarks.pdf.

[31] Zhenxuan Fan, Bo Zhang, Yutong Lin, Yuqian Yuan, Juekai Lin, Liang Liang, Zhuoyi Huang, Wenqiao Zhang, Juncheng Li, Siliang Tang, Jun Xiao, and Yueting Zhuang. RoboSPA: Can VLA models go beyond simple scenes and short-horizon tasks? In Hung-Yi Lee, Sunipa Dev, Junyi Jessy Li, and Desmond Elliott, editors, *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 23457–23492, Budapest, Hungary, October 2026. Association for Computational Linguistics. ISBN 979-8-89176-459-0. doi: 10.18653/2026.emnlp-main.1021. URL https://aclanthology.org/2026.emnlp-main.1021/.

[32] brian ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as i can, not as i say: Grounding language in robotic affordances. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 287–318. PMLR, 14–18 Dec 2023. URL https://proceedings.mlr.press/v205/ichter23a.html.

[33] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng.

Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500, 2023. doi: 10.1109/ICRA48891.2023.10160591.

[34] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 540–562. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/huang23b.html.

[35] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and brian ichter. Inner monologue: Embodied reasoning through planning with language models. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 1769–1782. PMLR, 14–18 Dec 2023. URL https://proceedings.mlr.press/v205/huang23c.html.

[36] Jiafei Duan, Wentao Yuan, Wilbert Pumacay, Yi Ru Wang, Kiana Ehsani, Dieter Fox, and Ranjay Krishna. Manipulate-anything: Automating real-world robots using vision-language models. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 5326–5350. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/duan25a.html.

[37] Yixian Zhang, Huanming Zhang, Feng Gao, Xiao Li, Zhihao Liu, Yi Nie, Chunyang Zhu, Jiaxing Qiu, Yuchen Yan, Jiyuan Liu, Wenhao Tang, Jiaji Rao, Zhengru Fang, Changxu Wei, Yu Wang, Wenbo Ding, and Chao Yu. Harness vla: Steering frozen vlas into reliable manipulation primitives via memory-guided agents, 2026. URL https://arxiv.org/abs/2607.08448.

[38] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

[39] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=ehfRiF0R3a.

[40] Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Trans. Inf. Syst.*, 43(6), September 2025. ISSN 1046-8188. doi: 10.1145/3748302. URL https://doi.org/10.1145/3748302.

[41] Tianxing Chen, Yuran Wang, Mingleyang Li, Yan Qin, Hao Shi, Zixuan Li, Yifan Hu, Yingsheng Zhang, Kaixuan Wang, Yue Chen, Hongcheng Wang, Junjie Wang, Tianhang Yang, Renjing Xu, Ruihai Wu, Yao Mu, Yaodong Yang, Hao Dong, and Ping Luo. Rmbench: Memory-dependent robotic manipulation benchmark with insights into policy design, 2026. URL https://arxiv.org/abs/2603.01229.

[42] Zeyi Liu, Arpit Bahety, and Shuran Song. Reflect: Summarizing robot experiences for failure explanation and correction. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 3468–3484. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/liu23g.html.

[43] Jiafei Duan, Wilbert Pumacay, Nishanth Kumar, Yi Ru Wang, Shulin Tian, Wentao Yuan, Ranjay Krishna, Dieter Fox, Ajay Mandlekar, and Yijie Guo. Aha: A vision-language-model for detecting and reasoning over failures in robotic manipulation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 45493–45517, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/70a06501001e1820fd1eb9ee821302d2-Paper-Conference.pdf.

[44] Wenke Xia, Ruoxuan Feng, Dong Wang, and Di Hu. Phoenix: A motion-based self-reflection framework for fine-grained robotic action correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6981–6990, June 2025.

[45] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc., 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf.

[46] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19632–19642, Mar. 2024. doi: 10.1609/aaai.v38i17.29936. URL https://ojs.aaai.org/index.php/AAAI/article/view/29936.

[47] Minghao Chen, Yihang Li, Yanting Yang, Shiyu Yu, Binbin Lin, and Xiaofei He. Automanual: Constructing instruction manuals by llm agents via interactive environmental learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 589–631. Curran Associates, Inc., 2024. doi: 10.52202/079017-0019. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0142921fad7ef9192bd87229cdafa9d4-Paper-Conference.pdf.

[48] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S Ryoo, Grecia Salazar, Pannag R Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics Transformer for Real-World Control at Scale. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.025.

[49] Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An Open-Source Generalist Robot Policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.090.

[50] 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 *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yAzN4tz7oI.

[51] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action Flow Model for General Robot Control. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.010.

[52] AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, Shu Jiang, Yuxin Jiang, Cheng Jing, Hongyang Li, Jialu Li, Chiming Liu, Yi Liu, Yuxiang Lu, Jianlan Luo, Ping Luo, Yao Mu, Yuehan Niu, Yixuan Pan, Jiangmiao Pang, Yu Qiao, Guanghui Ren, Cheng Ruan, Jiaqi Shan, Yongjian Shen, Chengshi Shi, Mingkang Shi, Modi Shi, Chonghao Sima, Jianheng Song, Huijie Wang, Wenhao Wang, Dafeng Wei, Chengen Xie, Guo Xu, Junchi Yan, Cunbiao Yang, Lei Yang, Shukai Yang, Maoqing Yao, Jia Zeng, Chi Zhang, Qinglin Zhang, Bin Zhao, Chengyue Zhao, Jiaqi Zhao, and Jianchao Zhu. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems, 2025. URL https://arxiv.org/abs/2503.06669.

[53] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1702–1713, June 2025.

[54] 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 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 9156–9172. Curran Associates, Inc., 2023. doi: 10.52202/075280-0403. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1d5b9233ad716a43be5c0d3023cb82d0-Paper-Conference.pdf.

[55] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 10641–10662, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/2c37c5bcef24b9541550261dcd63261b-Paper-Conference.pdf.

[56] Yuran Wang, Siqiao Huang, Mingleyang Li, Chenhao Zhang, Jiaqi Liang, Weiyang Jin, Yue Chen, Xuemin Chi, Donghao Zhou, Qize Yu, Yu-Kai Wang, Yuhan Rui, Shenzhe Yao, Zhen Yuan, Zhenhao Shen, Kefei

Zhu, Zijie Zhu, Ning Gao, Xiaowei Chi, Guanqi He, Shanghang Zhang, Hao Dong, Lin Shao, and Hang Zhao. Openwam: An open, modular exploration towards systematic world-action model pretraining, 2026. URL https://arxiv.org/abs/2609.07398.

[57] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530, 2023. doi: 10.1109/ICRA48891.2023.10161317.

[58] Andy Zeng, Maria Attarian, brian ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=G2Q2Mh3avow.

[59] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-e: An embodied multimodal language model. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8469–8488. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/driess23a.html.

[60] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction in robotics. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 4005–4020. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/yuan25c.html.

[61] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, Xinda Xue, Qinghang Su, Huaihai Lyu, Xiaolong Zheng, Jiaming Liu, Zhongyuan Wang, and Shanghang Zhang. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1724–1734, June 2025.

[62] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 4573–4602. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/huang25g.html.

[63] Kuan Fang, Fangchen Liu, Pieter Abbeel, and Sergey Levine. MOKA: Open-World Robotic Manipulation through Mark-Based Visual Prompting. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.062.

[64] Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9488–9495, 2024. doi: 10.1109/IROS58592.2024.10801352.

[65] Mingjie Pan, Jiyao Zhang, Tianshu Wu, Yinghao Zhao, Wenlong Gao, and Hao Dong. Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17359–17369, June 2025.

[66] Lucy Xiaoyang Shi, Brian Ichter, Michael Robert Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, Adrian Li-Bell, Danny Driess, Lachy Groom, Sergey Levine, and Chelsea Finn. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 54919–54933. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/shi25d.html.

[67] GigaBrain Team, Angen Ye, Axiang Sun, Can Jin, Chenxi Cheng, Chong Shi, Dengke Shang, Dingqian Zhang, Guan Huang, Guangqiang Wang, Guangqing Ding, Guo Li, Hangcong Li, Hengyu Zhong, Hongtao Lu, Jianbo Qin, Jiming Mao, Jing Zhu, Jindi Lv, Jingzhi Cui, Junjie Xie, Junyi Bao, Kai Liu, Lei Yuan, Limin Long, Lv Feng, Mingming Yu, Peng Li, Pengfei Yi, Qi Li, Qianli Zhang, Qingfang Li, Qitang Hu, Rui Zhang, Shaoyan Sun, Shibo Sun, Shiying Duan, Tenghui Chen, Tianze Liu, Weijie Ke, Wenyao Xue, Xiaofeng Wang, Xiaoyu Tian, Xinyu Liu, Xinze Chen, Yang Wang, Yankai Wang, Yejun Zeng, Yifan Li, Yifei Nie, Yilong Li, Yilong Liu,

Yongchao Feng, Yumeng Wang, Yun Ye, Zhichao Liu, Ziheng He, Zonghai Yang, and Zheng Zhu. Gigabrain-0.7: Scaling embodied foundation models to emergent capabilities with a three-system architecture, 2026. URL https://arxiv.org/abs/2608.15875.

[68] SenseNova, Joint Research Team, Zhongang Cai, Ziang Cao, Tailai Chen, Han Chong Chua, Linjun Dai, Hanming Deng, Hokin Deng, Haiwen Diao, Chenjing Ding, Weichen Fan, Xiao Fu, Zerui Gong, Chenyang Gu, Zexin He, Zhi Hou, Bo Li, Qinyue Li, Quanzhou Li, Xiaowan Li, Zhichao Liao, Dahua Lin, Zilong Lin, Bo Liu, Chunxiao Liu, Pengfei Liu, Tianqi Liu, Ziwei Liu, Chen Change Loy, Lewei Lu, Jiapeng Luo, Yuexiao Ma, Liang Pan, Xingang Pan, Oscar Qian, Ziyi Shang, Yuyao Shao, Dawei Su, Dacheng Tao, Qingyi Tao, Quan Wang, Ruisi Wang, Xiaogang Wang, Ziyi Wang, Beichen Wen, Donghao Wu, Jiaxu Wu, Silei Wu, Xiaoming Wu, Zhonghua Wu, Hongcan Xiao, Weiye Xu, Atsushi Yamashita, Lei Yang, Zhe Yang, Baiqiao Yin, Wanqi Yin, Hanfei Zhao, Tianbo Zhong, Long Zhuo, and Jing Zuo. Has gpt-6-astra achieved embodied intelligence?, 2026. URL https://arxiv.org/abs/2610.22471.

[69] Meng-Hao Guo, Zhe-Han Mo, Jia-Jun Wang, Yi Zhang, Kejin Wang, Yi-Xuan Deng, Jia-Peng Zhang, Yongming Rao, and Shi-Min Hu. Transferring the intelligence of vlms to robotic control, 2026. URL https://arxiv.org/abs/2609.22966.

[70] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

[71] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc., 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

[72] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 12028–12068, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/3339f19c5fcee3ad74502947a32be9e6-Paper-Conference.pdf.

[73] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan A, Saiful Haq, Ashutosh Sharma, Thomas Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. Dspy: Compiling declarative language model calls into state-of-the-art pipelines. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 54928–54958, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/f1cf02ce09757f57c3b93c0db83181e0-Paper-Conference.pdf.

[74] Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. Optimizing generative ai by backpropagating language model feedback. *Nature*, 639(8055):609–616, Mar 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08661-4. URL https://doi.org/10.1038/s41586-025-08661-4.

[75] Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 21344–21377, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/36b7acf6f6010652b3f2a433774a66fe-Paper-Conference.pdf.

[76] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, XiongHui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. Aflow: Automating agentic workflow generation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 34040–34077, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5492ecbce4439401798dcd2c90be94cd-Paper-Conference.pdf.

[77] StarVLA Community. Starvla: A lego-like codebase for vision-language-action model developing, 2026. URL https://arxiv.org/abs/2604.05014.

[78] Rui Cai, Jun Guo, Xinze He, Piaopiao Jin, Jie Li, Bingxuan Lin, Futeng Liu, Wei Liu, Fei Ma, Kun Ma, Feng Qiu, Heng Qu, Yifei Su, Qiao Sun, Dong Wang, Donghao Wang, Yunhong Wang, Rujie Wu, Diyun Xiang, Yu Yang, Hangjun Ye, Yuan Zhang, and Quanyun Zhou. Xiaomi-robotics-0: An open-sourced vision-language-action model with real-time execution, 2026. URL https://arxiv.org/abs/2602.12684.

[79] Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long ZENG, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=euMVC1DO4k.

[80] Jun Guo, Qiwei Li, Peiyan Li, Zilong Chen, Nan Sun, Yifei Su, Heyun Wang, Yuan Zhang, Xinghang Li, and Huaping Liu. Unified 4d world action modeling from video priors with asynchronous denoising, 2026. URL https://arxiv.org/abs/2604.26694.

[81] Tianheng Wang, Zhou Xie, Heng Jia, Jianhua Xu, Tong Zhang, and Kaicheng Yu. An action is worth one patch: Unified world-action modeling with patchwam, 2026. URL https://arxiv.org/abs/2609.25961.

[82] Wenxuan Song, Jiayi Chen, Jingbo Wang, Shuai Zhou, Xicheng Gong, Zehua Fan, Ziyang Zhou, Junwu E, Haodong Yan, Fuhao Li, Qize Yu, Xu Huang, Pengwei Wang, Wen Chen, Shunbo Zhou, and Haoang Li. Uniwam: Unified world-action model, 2026. URL https://arxiv.org/abs/2610.02054.

[83] Yude Zou, Yifei Yao, Guanbo Jiao, Jianming Ma, Guanjie Zheng, and Xing Gao. Pica-wam: Visual perturbation invariance via cross-appearance self-alignment in world action models, 2026. URL https://arxiv.org/abs/2610.26813.

[84] Tianyu Xin, Wanqi Zhong, Bowen Hu, Zhangchi Guo, Lijie Wang, Yiming Wang, Haoran Liao, Heyang Yu, Xiaoyu Tian, Zhiding Yu, and Yiming Li. Prismwam: Decoupling state representations in world action models, 2026. URL https://arxiv.org/abs/2610.26564.

[85] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan-Ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. In Tong Zhang, Miroslav Dudik, Martin Jaggi, Alekh Agarwal, Sharon Li, Dale Schuurmans, Jerry Zhu, Felix Berkenkamp, Hanze Dong, and Alberto Bietti, editors, *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 13673–13699. PMLR, 06–11 Jul 2026. URL https://proceedings.mlr.press/v306/chen26j.html.

[86] Sizhe Yang, Juncheng Mu, Tianming Wei, Chenhao Lu, Xiaofan Li, Linning Xu, Zhengrong Xue, Zhecheng Yuan, Dahua Lin, Jiangmiao Pang, and Huazhe Xu. Memorywam: Efficient world action modeling with persistent memory, 2026. URL https://arxiv.org/abs/2606.20562.

[87] Sizhe Zhao, Haozhe Xie, Weiyu Zhao, Chenchu Zhang, Huan Wang, Chenyang Wang, Qinglin Liu, and Shengping Zhang. Memory as plans: World-action modeling with memory-grounded planning, 2026. URL https://arxiv.org/abs/2609.11561.