EchoVLA: Robotic Vision-Language-Action Model with Synergistic Declarative Memory for Mobile Manipulation

Min Lin[1], Xiwen Liang[1], Bingqian Lin[2], Jingzhi Liu[1], Zijian Jiao[1], Kehan Li[1], Ziang Yan[1], Yu Sun[1], Weijia Liufu[1], Yuhan Ma[1], Jiarui Hu[1], Yuecheng Liu[3], Shen Zhao[1], Yuzheng Zhuang[3], and Xiaodan Liang[1,4⋆]

[1] Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] Huawei Noah's Ark Lab, China
[4] Yinwang Intelligent Technology Co., Ltd., China
linm57@mail2.sysu.edu.cn, xdliang32@gmail.com

Abstract. Recent progress in Vision–Language–Action (VLA) models has enabled embodied agents to interpret multimodal instructions and perform complex tasks. However, existing VLAs are mostly confined to short-horizon, table-top manipulation, lacking the memory and reasoning capability required for mobile manipulation, where agents must coordinate navigation and manipulation under changing spatial contexts. In this work, we present EchoVLA, a memory-aware VLA model for mobile manipulation. EchoVLA incorporates a synergistic declarative memory inspired by the human brain, consisting of a scene memory that maintains a collection of spatial–semantic maps and an episodic memory that stores task-level experiences with multimodal contextual features. The two memories are individually stored, updated, and retrieved based on current observations, task history, and instructions, and their retrieved representations are fused via coarse- and fine-grained attention to guide base–arm diffusion policies. To support large-scale training, we further introduce MoMani, an automated benchmark that generates expert-level trajectories through multimodal large language model (MLLM)–guided planning and feedback-driven refinement, supplemented with real-robot demonstrations. Comprehensive simulated and real-world results demonstrate that EchoVLA substantially improves overall performance, e.g., it achieves the highest success rates of 0.52 on manipulation/navigation tasks and 0.31 on mobile manipulation tasks in simulation, exceeding the strong baseline $\pi_{0.5}$ by +0.20 and +0.11, respectively.

1 Introduction

Recent advances in Vision–Language–Action (VLA) models [2,4,13,20,22,37,41, 44,53] have shown strong potential for general-purpose robot learning, enabling

⋆ Corresponding author.

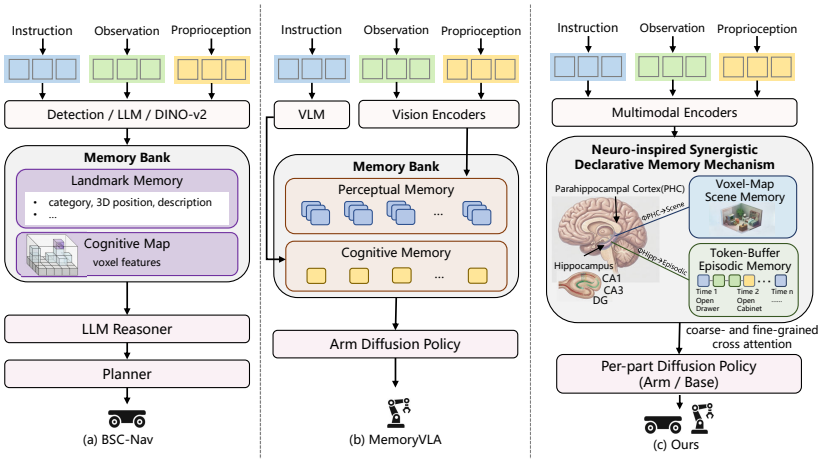

Fig. 1: Comparison of memory designs for mobile manipulation control. (a) BSC-Nav [39] uses a memory bank with landmark memory and a cognitive map to support LLM-based reasoning and planning. (b) MemoryVLA [40] builds a memory bank with perceptual memory and cognitive memory to condition a diffusion policy. (c) EchoVLA (ours) introduces dual memories: a persistent scene memory (voxel map) and a time-indexed episodic memory (token buffer), fused via coarse- and fine-grained cross attention for context-aware per-part diffusion control (arm/base).

embodied agents to interpret multimodal inputs and perform diverse manipulation tasks. Models such as RT-2 [54], OpenVLA [24], and ManipLLM [28] demonstrate that large-scale vision–language pretraining can be extended to robot control, achieving impressive zero-shot generalization across unseen objects and scenes. However, most existing VLAs are restricted to short-horizon, tabletop manipulation, and rely on Markovian control, where each decision depends solely on the current observation. This limitation prevents consistent reasoning over extended task sequences and hinders long-term spatial understanding.

To overcome this limitation, we draw inspiration from the declarative memory system in the human brain, which comprises complementary subsystems for spatial and experiential memory encoding, respectively. The parahippocampal cortex (PHC) and related retrosplenial regions represent the spatial and semantic structure of scenes, providing contextual information about environmental layouts and object relations [1, 15, 16]. These regions project to the hippocampus, which integrates contextual inputs into temporally organized episodic traces that capture individual experiences and task outcomes [14, 36]. Such a dual-system interaction allows humans to remember where things are and how tasks were accomplished—offering a useful analogy for memory-guided embodied reasoning.

Building on this insight, we propose **EchoVLA**, a memory-aware Vision–Language–Action model for mobile manipulation. EchoVLA introduces two complementary memories: a **scene memory** parallels the PHC by maintaining a persistent spatial representation as a voxel map, and an **episodic memory**

mirrors hippocampal processing by storing time-indexed multimodal tokens to track task progress. As shown in Figure 1, prior work such as BSC-Nav [39] organizes memory as landmark descriptors and a cognitive map for LLM-based reasoning and planning, while MemoryVLA [40] builds perceptual and cognitive memories to condition diffusion control. In contrast, EchoVLA retrieves from scene memory and episodic memory with separate coarse- and fine-grained cross attention, and fuses them to guide per-part diffusion policies (arm/base). This neuro-inspired dual-memory with attention-guided formulation helps maintain coherent spatial context and consistent task execution over long horizons.

We further introduce MoMani, an automated benchmark for mobile manipulation that generates expert-level trajectories via MLLM-guided planning and feedback-based refinement. Compared with existing mobile manipulation benchmarks—which remain limited in scale and task diversity—as well as datasets like RoboTwin 2.0 [8] and RoboCasa [32], MoMani offers richer task coverage and collects real-robot demonstrations on a holonomic mobile platform [48] to support large scale training.

Both simulation and real-world validation reveal that our EchoVLA maintains strong robustness in long-horizon mobile manipulation tasks by leveraging its synergistic declarative memory. On the RoboCasa [32] simulator, EchoVLA brings significant performance gain in SR over strong baseline $\pi_{0.5}$ [21] on manipulation/navigation tasks and mobile manipulation tasks by 0.20 and 0.11, respectively. Real-world experiments on the TidyBot++ platform [48] across a 7m $\times$ 7m arena shows that EchoVLA achieves the highest SR of 0.44, outperforming $\pi_{0.5}$ [21] (0.33) and Diffusion Policy [9] (0.32).

Our contributions are summarized as follows:

- We propose EchoVLA, a neuro-inspired memory-aware VLA model equipped with synergistic scene and episodic memory for mobile manipulation.
- We introduce MoMani, an automated benchmark that provides expert-level multimodal trajectories with real-robot demonstrations for scalable embodied data generation.
- Extensive simulated and real-world results demonstrate that EchoVLA consistently outperforms strong baselines on various mobile manipulation tasks.

2 Related Work

2.1 Vision-Language-Action Model

Driven by advances in visual–language foundation models [6, 30, 35] and large-scale robot datasets [5, 33], Vision–Language–Action (VLA) models unify perception, reasoning, and control within a single multimodal policy. Representative works such as RT-2 [54], OpenVLA [24], and ManipLLM [28] discretize continuous actions into token sequences and perform autoregressive prediction, enabling large-scale pretraining but limiting motion continuity. Recent approaches (e.g., CogACT [27], DexVLA [46], HybridVLA [29]) introduce diffusion- or regression-based action heads to capture the multimodal distribution of robot trajectories,

while AC-DiT [7] explicitly models coordination between the mobile base and the manipulator through adaptive diffusion. Notable works $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_{0.5}$ [3, 21] address hierarchical trajectory diffusion for mobile manipulation. $\boldsymbol{\pi}_0$ models the full robot action space, while $\boldsymbol{\pi}_{0.5}$ introduces per-part decomposition for mobile base and manipulator, adding partial episodic memory. Despite improved trajectory fidelity, both lack explicit scene-level memory for planning. Meanwhile, AnywhereVLA [18] adopts a modular design that combines classical SLAM with a fine-tuned VLA head for robust real-world pick-and-place, highlighting the practicality of hybrid systems. However, most existing VLAs remain Markovian, relying only on current observations without structured memory for reasoning.

In this work, we take inspiration from the memory system of human brain and propose EchoVLA, which constructs a synergistic declarative memory composed of scene and episodic memories. The scene memory explicitly builds a spatial–semantic map of object layouts and environmental structure, while the episodic memory records task-specific experience traces for contextual retrieval. This dual-memory mechanism enables reasoning in mobile-manipulation tasks beyond the capability of prior single-memory VLAs. A concurrent work MemoryVLA [40] augments VLAs with a perceptual–cognitive memory that caches visual features to enhance manipulation stability. However, its memory remains an implicit perceptual cache without an explicit spatial representation or task-level experience storage, limiting its ability for long-horizon task reasoning.

2.2 Mobile Manipulation Benchmark

Early embodied benchmarks such as RoboTHOR [11] for sim-to-real indoor navigation and iGibson 2.0 [25] for object-centric household tasks mainly offer mobile perception and basic interaction. Habitat 2.0 [43] further supports interactive rearrangement with articulated objects, but the mobile–manipulation coupling is still limited to predefined skills. Large-scale suites like BEHAVIOR-100 / 1K [26, 42] and ManiSkill2 [17] broaden the task families to 100–1000 household activities or 20+ manipulation families, yet most tasks assume a fixed or simplified manipulation setting rather than full mobile manipulation. More recent, robotics-oriented simulators move closer to real mobile manipulation: RoboCasa [32] targets realistic kitchen scenes with 120+ layouts and 2,500+ assets but the task spectrum is still mostly kitchen rearrangement; MoMa-Kitchen [52] provides 100K+ affordance-grounded samples for "last-mile" base positioning but is domain-specific to kitchens; RoboTwin 2.0 [8] automates data generation for 50 dual-arm tasks in 731-object settings but does not involve base navigation; TidyBot [47] shows LLM-guided, user-conditioned tidy-up on a mobile manipulator, while TidyBot++ [48] mainly contributes an open-source holonomic mobile-manipulation platform for data collection rather than a broad, automatic task generator. Different from existing benchmarks, our constructed MoMani provides a unified, automated pipeline that produces expert-level trajectories for diverse mobile-manipulation tasks and augments them with real-robot demonstrations.

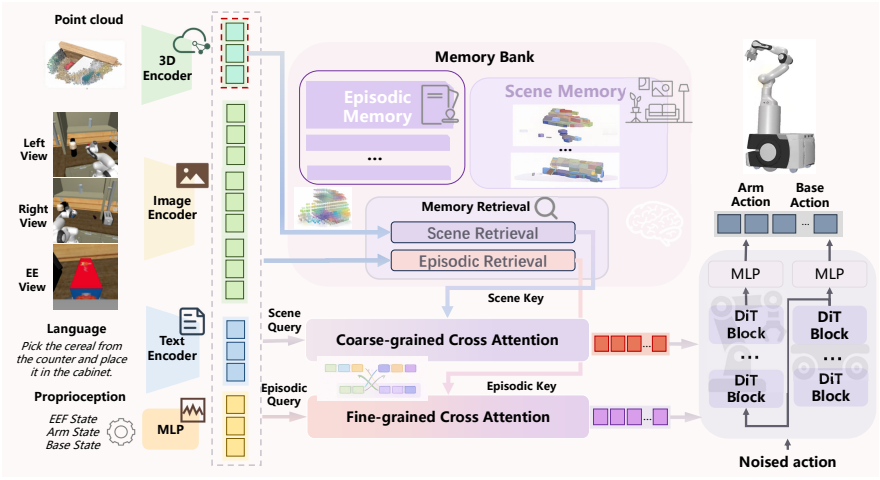

Fig. 2: Overview of EchoVLA. Multi-modal observations (RGB views, point clouds, language, and proprioception) are encoded into a unified token sequence. The model retrieves relevant information from both episodic and scene memory through coarse- and fine-grained cross-attention. The retrieved memory augments the diffusion policy, which generates base and arm actions through a per-part denoising process.

3 EchoVLA

EchoVLA is a memory-augmented vision–language–action framework designed for mobile manipulation. Its key idea is to maintain two complementary memory systems—scene memory and episodic memory—and integrate them into a coarse-to-fine retrieval hierarchy that improves spatial consistency, temporal reasoning, and manipulation precision. The overall architecture as illustrated in Figure 2 consists of: (1) multimodal state representation (Sec. **??**), (2) memory retrieval and interaction (Sec. 3.3), and (3) diffusion-based action generation (Sec. 3.4). We first present the problem formulation of the mobile manipulation task in Sec. 3.1, then we describe each component of EchoVLA in detail.

3.1 Problem Formulation

At timestep t, the agent observes multi-view RGB-D images $\mathcal{O}_t$, proprioceptive states $\mathbf{s}_t$, and a natural-language instruction $\mathcal{I}$. The goal is to output continuous arm and base actions:

$$(a_t^{\text{arm}}, a_t^{\text{base}}) = \pi_\theta(\mathcal{I}, \mathcal{O}_{1:t}, \mathbf{s}_{1:t}). \tag{1}$$

We focus on non-Markovian tasks— two visually similar frames may correspond to completely different progress states (e.g., “cabinet opened” vs. “about to open”). Thus, a memory mechanism is needed.

3.2 Multimodal State Representation

EchoVLA encodes language, appearance, 3D structure, and robot configuration into a unified token sequence to drive memory retrieval and action generation. Natural language instructions are encoded via the text tower of a frozen SigLIP model [51] to produce language tokens $\mathbf{L}$. Concurrently, multi-view RGB images from three fixed cameras are processed independently by the frozen SigLIP vision tower to yield per-view features $\mathbf{V}_t^{(i)}$, which are concatenated and projected into a shared embedding space as $\mathbf{V}_t$ to maintain cross-modal alignment.

Depth observations are fused into a point cloud and encoded by a trainable PointAttn backbone [45]. The resulting tokens $\mathbf{P}_t$ provide fine-grained geometric cues regarding free space and object boundaries while adapting to the specific robot embodiment. Lastly, the proprioceptive state $\mathbf{s}_t$ is transformed into configuration tokens $\mathbf{R}_t$ using a small MLP. We then form the unified token sequence:

$$\mathbf{S}_t = [\mathbf{L},\ \mathbf{V}_t,\ \mathbf{P}_t,\ \mathbf{R}_t], \tag{2}$$

which serves as the query for hierarchical memory retrieval and conditions the diffusion policy.

3.3 Memory Retrieval and Interaction

EchoVLA maintains two complementary memory banks—scene memory and episodic memory—and retrieves information from them using a two-level hierarchical (coarse-to-fine) attention mechanism. The hierarchy comes from the semantic granularity of the two memories: scene memory provides slowly varying spatial structure, while episodic memory captures fine-grained, time-indexed task progress. The retrieved memory features are later fused with the current tokens and used to condition the diffusion policy.

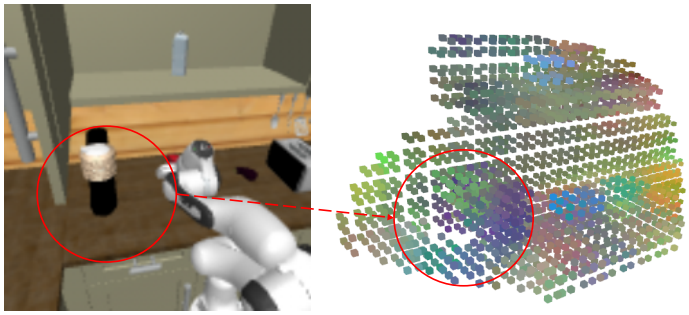

Fig. 3: Visualization of the voxelized 3D feature map, which encodes local geometric structure from depth observations.

Scene Memory Scene memory $\mathcal{M}^{\text{scene}}$ is maintained as a Key-Value bank, where Keys are 3D spatial indices (x, y, z) and Values are aggregated PointAttn features. It represents a persistent, environment-specific 3D structure refined as new episodes unfold. At the beginning of training in a new environment, $\mathcal{M}^{\text{scene}}$ is initialized as an empty voxel grid. As the agent interacts with the environment, depth observations are encoded by the PointAttn network, producing a local 3D feature volume $\mathbf{V}_t^{3D} \in \mathbb{R}^{X \times Y \times Z \times C}$ (see Figure 3). While $\mathcal{M}^{\text{scene}}$ stores a global environment-specific representation, feature retrieval for the downstream policy is executed via a spatial-query process, which dynamically selects the top-k relevant entries within the current local camera frustum to maintain efficiency.

To incorporate new observations into the global structure while avoiding redundant updates, a discrepancy-driven rule is formally defined by a reconstruction error δ:

$$\delta = \|\mathbf{V}_t^{3D} - \mathcal{D}(\mathcal{M}^{\text{scene}}[x, y, z])\|^2, \tag{3}$$

where $\mathcal{D}$ is an MLP decoder. Regions whose reconstruction error exceeds a threshold ($\delta > \tau$, where $\tau = 0.5$) are updated via exponential moving average (EMA):

$$\mathcal{M}_{\text{new}}^{\text{scene}} = (1 - \alpha)\mathcal{M}_{\text{old}}^{\text{scene}} + \alpha \mathbf{V}_t^{3D}, \tag{4}$$

with $\alpha = 0.2$, while unchanged areas retain their previous values. This enables $\mathcal{M}^{\text{scene}}$ to gradually converge to a stable representation of the environment's geometry.

Episodic Memory Episodic memory stores a short horizon of recent token sequences, each associated with its timestep. Formally, we maintain a time-indexed buffer:

$$\mathcal{M}^{\text{epi}} = \{(\mathbf{S}_{t-k}, t-k), \ldots, (\mathbf{S}_{t-1}, t-1)\}, \tag{5}$$

where the window size k is determined by the memory capacity.

Unlike scene memory, which evolves slowly and captures stable 3D structure, episodic memory preserves fine-grained temporal information tied to recent interactions—such as whether a drawer has been opened, whether an object has already been grasped, or the precise end-effector configuration in the immediate past. The memory is maintained as a fixed-size FIFO buffer updated at each timestep, ensuring that temporally coherent information is retained while avoiding unbounded growth. By storing the original encoded tokens instead of compressing them into abstract summaries, the model preserves detailed temporal cues that are essential for resolving non-Markov ambiguities and maintaining consistent task execution across visually similar states.

Memory Matching and Attention EchoVLA accesses its two memory banks in two steps: it first matches the current representations to memory entries using cosine similarity, and then interacts with the selected entries via cross-attention.

For scene memory, the query is the current voxelized 3D feature map $\mathbf{V}_t^{3D}$. We compute cosine similarity between $\mathbf{V}_t^{3D}$ and stored scene maps in $\mathcal{M}^{\text{scene}}$,

select the top-k matches $\mathcal{M}_{\text{sel}}^{\text{scene}}$, and apply coarse cross-attention:

$$\mathbf{Z}_t^{\text{scene}} = \text{CrossAttn}\big(\text{q} = \mathbf{V}_t^{3D},\ \text{k / v} = \mathcal{M}_{\text{sel}}^{\text{scene}}\big). \tag{6}$$

For episodic memory, the query is the current multi-modal state tokens $\mathbf{S}_t$. We match $\mathbf{S}_t$ to historical states in $\mathcal{M}^{\text{epi}}$, select the most relevant subset $\mathcal{M}_{\text{sel}}^{\text{epi}}$, and apply fine cross-attention:

$$\mathbf{Z}_t^{\text{epi}} = \text{CrossAttn}\big(\text{q} = \mathbf{S}_t,\ \text{k / v} = \mathcal{M}_{\text{sel}}^{\text{epi}}\big). \tag{7}$$

The outputs of the two interactions are combined into a memory-augmented representation:

$$\mathbf{H}_t = \big[\mathbf{Z}_t^{\text{scene}},\ \mathbf{Z}_t^{\text{epi}}\big], \tag{8}$$

which serves as the conditioning input for the diffusion-based action policy.

3.4 Diffusion-based Action Generation

To model the heterogeneous dynamics of mobile-base motion and arm manipulation, we propose a per-part diffusion policy conditioned on the memory-augmented representation $\mathbf{H}_t$.

Each action subspace $p \in \{\text{base, arm}\}$ is generated by an independent denoising diffusion process. For part p, the denoiser predicts the noise of a perturbed action sample:

$$\epsilon_\theta^{(p)} = \text{Denoiser}_p(\mathbf{z}_t, \mathbf{H}_t, t), \quad p \in \{\text{base}, \text{arm}\} \tag{9}$$

where $\mathbf{z}_t$ is the noisy action at diffusion step t, $\mathbf{H}_t$ captures the fused current observation with retrieved scene-level and episodic context.

During training, the objective minimizes the denoising loss for each part:

$$\mathcal{L} = \sum_{p \in \{\text{base,arm}\}} \mathbb{E}_{t, \mathbf{z}_t}\Big[\|\epsilon - \epsilon_\theta^{(p)}(\mathbf{z}_t, \mathbf{H}_t, t)\|^2\Big]. \tag{10}$$

This per-part design enables structured action generation, allowing coordinated yet decoupled learning of locomotion and manipulation behaviors, which improves scene generalization and cross-task transfer.

4 MoMani

We introduce **MoMani**, an automated benchmark for high-quality mobile manipulation data generation. As shown in Table 1, while frameworks like ManiSkill2 [17] and RoboCasa [32] focus on static tasks, MoMani provides a unified pipeline for integrated "navigation + manipulation" (Co-Gen) data.

Distinguishing itself from simulation-only platforms like BEHAVIOR-1K [26] and InfinitedWorld [38], MoMani uniquely incorporates both simulated (S) and real-robot (R) data sources. It stands as the only benchmark to support a full spectrum of automated generation modes across both Mobile-M (Sim) and Real-Robot platforms, delivering expert-level trajectories for complex, multi-stage embodied tasks.

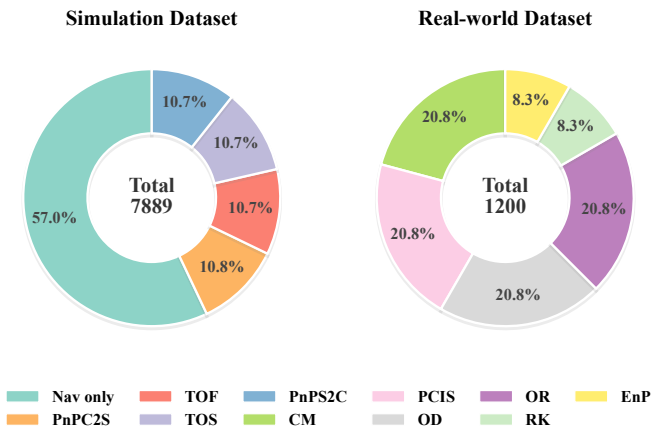

Fig. 4: Dataset composition. Simulation: pure navigation and four mobile manipulation tasks (TOF, PnPS2C, PnPC2S, TOS). Real-world: six mobile manipulation tasks (CM, OD, OR, PCIS, RK, EnP).

Table 1: Comparison of MoMani with representative embodied AI platforms. Sub-tasks indicates count per trajectory. Data Source: R (Real-world), S (Simulation); Generation: Navigation, Manipulation, Co-Gen; Robotic Platforms: Mobile-M (simulated), Real-Robot (physical); Action: N (Navigation), M (Manipulation).

Name	Sub-tasks	Data Source		Automated Generation			Robotic Platforms		Action
		Real (R)	Sim (S)	Navigation	Manipulation	Co-Gen	Mobile-M (Sim)	Real-Robot	
ManiSkill2 [17]	1-2	-	✓	-	✓	-	✓	-	*M*
Social Nav [34]	1	-	✓	✓	-	-	✓	-	*N, M*
HomeRobot [49]	3	-	✓	✓	✓	-	✓	✓	*N, M*
ProcTHOR [12]	3	-	✓	✓	✓	-	✓	-	*N, M*
Behavior-1K [26]	> 5	-	✓	✓	✓	-	✓	-	*N, M*
Open X-Embodiment [33]	1-3	✓	-	-	-	-	✓	✓	*M*
RoboCasa365 [19]	2-16	-	✓	✓	✓	-	✓	-	*N, M*
InfinitedWorld [38]	3	-	✓	✓	✓	✓	✓	-	*N, M*
MoMani (Ours)	2-3	✓	✓	✓	✓	✓	✓	✓	*N, M*

4.1 Simulation Data Construction

As shown in Fig. 5, MoMani employs a two-stage automated pipeline to generate high-quality long-horizon trajectories.

Stage I: Online Candidate Generation. Task specifications are processed via an MMLM prompter to execute sequential Target-Aligned Sampling (L1), Safety-Aware Navigation (L2), and Continuous Nav-Manip Stitching (L3). Crucially, our engine supports simultaneous base and arm execution, utilizing continuous collision detection to generate coordinated "nav-manip" trajectories. Candidates must pass Hard Quality Gates—verifying zero collisions, precise alignment ($\Delta\text{pos} < 0.05\text{m}, \Delta\text{ori} < 5°$), and 100% task success—before entering the feasible pool D_{pool}.

Stage II: Offline Selection & Audit. Candidates in D_{pool} undergo Lexicographic Ranking (L4) based on path length and planning cost to identify the

expert-level Top-K set D_{topK}. A final Scene-Camera Audit ensures the dataset is training-ready, enabling the scalable generation of $5,000+$ multimodal trajectories that bridge the gap between navigation and manipulation.

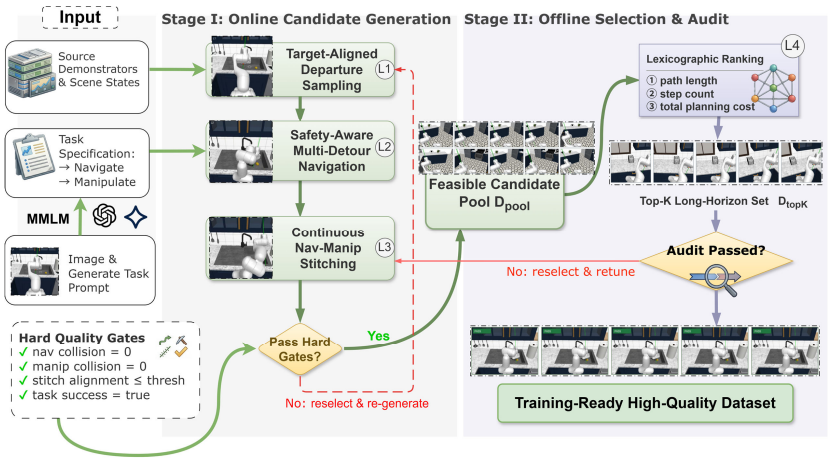

Fig. 5: Overview of the Quality-Controlled Data Engine. Online generation synthesizes trajectories filtered by quality gates (collision-free, task success), while offline audit ranks candidates via lexicographic criteria (path length, cost) for training-ready data.

4.2 Real-Robot Demonstration Collection

To bridge the simulation-to-real gap, we collect a real-world dataset using a hardware configuration based on TidyBot++ [48]. Our platform features a Kinova Gen3 7-DoF manipulator on a holonomic mobile base, equipped with front RGB-D and top-view stereo cameras synchronized via ROS. We utilize a web-based teleoperation interface to record multimodal data streams at 30 Hz, which are then segmented into motion primitives. Successful trajectories are verified via replay, while failed attempts are discarded.

4.3 Dataset Distribution

To provide a comprehensive evaluation for mobile manipulation, we construct a diverse and highly heterogeneous dataset that captures both broad environmental variations and intricate long-horizon task multi-modality. As visually summarized in Figure 4, our hierarchical data distribution comprises:

– **Simulation (7,889 episodes):** Primarily consists of a large navigation-only subset (57.0%) designed to prime base mobility, supplemented by di-

verse contact-rich mobile manipulation tasks including *PnPC2S* (PickPlaceCounterToStove, 10.8%), *PnPS2C* (PickPlaceSinkToCounter, 10.7%), *TOS* (TurnOnStove, 10.7%), and *TOF* (TurnOnSinkFaucet, 10.7%).
– **Real-world (1,200 episodes):** Features a balanced distribution (20.8% each) of interactive manipulation tasks requiring precise contact, including *OR* (Open refrigerator door), *CM* (Close the microwave), *OD* (Open the drawer), and *PCIS* (Place the cup from the rack into the sink), along with additional long-horizon tasks such as *EnP* (Enter the room and place the pears on the cabinet) and *RK* (Rotate the knob of a microwave) (8.3% each).

5 Experiment

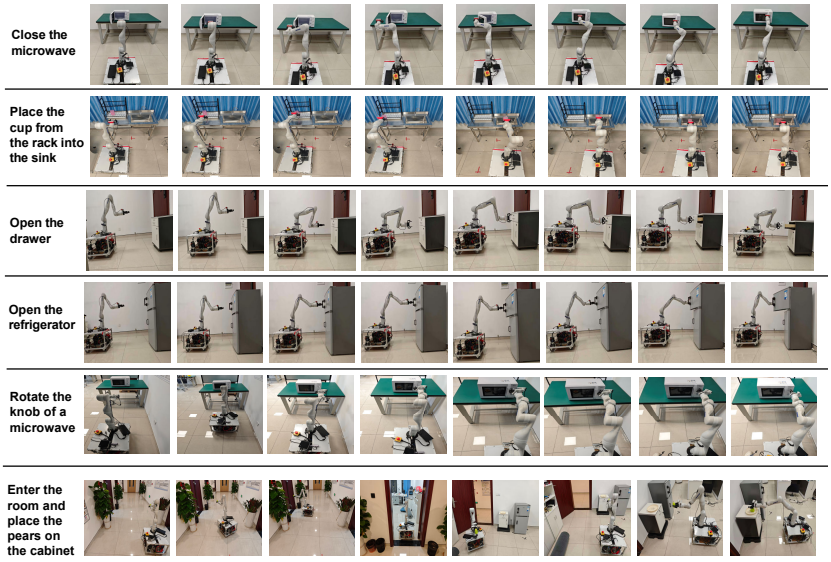

Fig. 6: Real-world rollouts generated by EchoVLA. The robot successfully completes four mobile manipulation tasks—turning off the microwave, placing the cup from the rack into the sink, opening the drawer, and opening the refrigerator.

5.1 Experimental Setup

We evaluate our proposed EchoVLA model in both simulation and real-world environments. In **simulation**, experiments are conducted in the RoboCasa simulator using multiple mobile manipulation tasks. For **real-world experiments**, we deploy EchoVLA using the *TidyBot++* platform [48] in a 7m $\times$ 7m arena to execute diverse household tasks, notably *EnP* (Enter and Pick).

Table 2: Success Rate (SR) comparison on Manipulation, Navigation, and Mobile Manipulation tasks. EchoVLA (Ours) achieves the highest performance in most scenarios.

Category	Method	PnP C2S	PnP S2C	TOF	TOS	Nav only	Avg SR
Manip. / Nav.	BC-T	0.04	0.46	0.33	0.45	0.10	0.28
	Diffusion Policy	0.00	0.00	0.03	0.00	0.03	0.01
	DP3	0.03	0.25	0.43	0.26	0.13	0.22
	WB-VIMA	0.00	0.12	0.11	0.19	0.31	0.15
	$\pi_{0.5}$	0.18	0.20	0.44	0.52	0.24	0.32
	EchoVLA (Ours)	**0.21**	**0.68**	**0.51**	**0.67**	**0.51**	**0.52**
Mobile Manip.	BC-T	0.00	0.11	0.04	0.04	–	0.05
	DP3	0.00	0.10	0.08	0.04	–	0.06
	WB-VIMA	0.00	0.12	0.11	0.19	–	0.11
	$\pi_{0.5}$	0.08	0.25	0.18	0.27	–	0.20
	EchoVLA (Ours)	**0.17**	**0.34**	**0.29**	**0.43**	–	**0.31**

Unlike other stationary or single-room tasks, *EnP* requires cross-room navigation and manipulation, moving from a starting corridor into a target room to retrieve objects. This validates the model's capability in handling long-horizon spatial transitions and multi-stage logic in complex environments.

We train EchoVLA on 8 NVIDIA A100 GPUs with observations including multi-view RGB-D images and robot states. And the action space is decomposed into mobile base and arm actions following the per-part diffusion design.

5.2 Comparison with Baselines in RoboCasa Simulation

We evaluate **EchoVLA** against representative baselines including BC-T [31], Diffusion Policy [10], DP3 [50], $\pi_{0.5}$ [21], and WB-VIMA [23]. The evaluation spans four core tasks in the RoboCasa simulator: PickPlaceCounterToStove, PickPlaceSinkToCounter, TurnOnSinkFaucet, and TurnOnStove. To assess **Mobile Manipulation**, we introduce additional challenges by initializing the robot at a distance from targets or requiring follow-up navigation post-manipulation.

Performance is analyzed across three categories: (1) **Manipulation** (arm-only), (2) **Navigation** (base precision), and (3) **Mobile Manipulation** (coordinated control). We report the **Success Rate (SR)** averaged over three random seeds and 50 evaluation episodes per task. These findings are consolidated in Table 2, highlighting the inherent difficulty of coordinated mobile manipulation. The results in Table 2 reveal several important observations:

1) Method Comparison. For Manipulation and Navigation tasks, traditional behavior cloning (BC-T) and Diffusion Policy exhibit low performance, with Diffusion Policy often achieving near-zero success rates. $\pi_{0.5}$ shows improved performance, while our proposed EchoVLA achieves the highest average success rate (Avg Success = 0.52), demonstrating robust task execution.

Table 3: Ablation results on the **PnP Counter to Stove** task. M: Mobile; S: Static.

	RGB	PC	EM	SM	M	S
(a)	✗	✓	✓	✓	0.02	0.13
(b)	✓	✗	✓	✓	0.08	0.15
(c)	✓	✓	✓	✗	0.09	0.16
(d)	✓	✓	✗	✓	0.14	0.13
Ours	✓	✓	✓	✓	**0.17**	**0.21**

2) Increased Difficulty in Mobile Manipulation. Coordinating base and arm control significantly increases complexity, causing success rates to drop across all methods. For instance, $\pi_{0.5}$'s average success decreases from 0.32 to 0.20, while BC-T exhibits a drastic decline from 0.28 to 0.05.

3) Superiority of EchoVLA. Despite the increased difficulty, EchoVLA maintains the highest success rate (0.31) in Mobile Manipulation tasks, outperforming all other methods. This highlights the method's effectiveness in handling coordinated base–arm motion, long-range navigation, and sequential manipulation tasks.

4) Task-specific Performance Variations. Tasks such as PickPlaceSinkToCounter and TurnOnSinkFaucet show larger performance variations across methods, reflecting higher requirements for spatial understanding and precise manipulation. EchoVLA consistently achieves strong performance on these challenging tasks, demonstrating its capability in multi-modal perception and action reasoning.

5.3 Ablation Study

Ablations are performed on two representative RoboCasa tasks, **PnP Counter to Stove (Mobile)** and **PnP Counter to Stove**, and results are reported as **Success Rate (SR)** averaged over 50 episodes.

1) Observation Ablation. We examine the effect of removing point cloud input from the observation space. When the agent relies solely on RGB features without 3D geometric cues, its spatial grounding ability decreases notably. As shown in Table 3, removing point cloud input results in consistent performance drops across both mobile and non-mobile variants, demonstrating the importance of geometric depth information for accurate object localization and grasping.

2) Memory Module Ablation. We further analyze the roles of our two memory mechanisms: **Episodic Memory (EM)** and **Scene Memory (SM)**. The ablation rows in Table 3 show that disabling either EM or SM leads to substantial degradation, and removing both RGB or PC signals further compounds these effects. Overall, the results highlight that observation diversity (RGB + PC) and hierarchical memory (EM + SM) are both essential for robust performance in manipulation and mobile manipulation tasks in RoboCasa.

Table 4: Sensitivity analysis of hyperparameters on the PnPC2S (Mobile) task. $L = 8$ and $\tau = 0.5$ are selected as optimal values for EchoVLA.

Window Size (L)	2	4	8	16
SR ↑	0.08	0.12	**0.17**	0.15
Threshold (τ)	0.1	0.3	0.5	0.7
SR ↑	0.11	0.14	**0.17**	0.13

3) Sensitivity Analysis. We evaluate EchoVLA's sensitivity to EM window size L and update threshold τ. As shown in Table 4, a small L (2 or 4) leads to "plan forgetting" and oscillations, while $L = 16$ introduces latency with diminishing returns. $L = 8$ provides the optimal performance-efficiency balance. Furthermore, a low threshold ($\tau = 0.1$) causes representation instability, whereas a high threshold ($\tau = 0.7$) results in outdated scene memory. $\tau = 0.5$ strikes the best balance, ensuring memory freshness and stable spatial grounding for long-horizon tasks.

5.4 Real Robot Experiment

We evaluate EchoVLA across a diverse set of tasks, including opening a drawer (*OD*), closing a microwave (*CM*), placing a cup into the sink (*PCIS*), opening a refrigerator (*OR*), rotating a knob (*RK*), and entering a room to place pears (*EnP*). For each task, we conduct 20 independent trials with randomized initial robot base positions to ensure statistical significance.

1) Real-World Performance. As shown in Table 5, EchoVLA achieves a 0.44 mean success rate, outperforming $\pi_{0.5}$ (0.33) and Diffusion Policy (0.32). While baselines suffice for shorter tasks like *OD* and *OR*, EchoVLA excels in complex ones like *CM* (0.70) and *RK* (0.50). Crucially, it leads in the longest-horizon task, *EnP* (362.9 frames), confirming that our synergistic memory stabilizes control and mitigates perceptual noise.

2) Robustness to Real-World Perceptual Noise. As shown in Table 5, EchoVLA achieves a mean success rate of 0.44, surpassing Diffusion Policy (0.32). EchoVLA excels in long-horizon precision, achieving 50% SR on *RK* while baselines fail (10%) due to memory drift. This gap suggests that temporal context from EM acts as a corrective anchor, compensating for spatial inaccuracies like voxel map "ghosting" in scene memory.

3) Failure Analysis on Dynamic Occlusion. Despite its robustness, performance drops in *OR* due to dynamic occlusion. As the fridge door opens, rapidly changing geometry degrades EchoVLA's explicit 3D Scene Memory. Conversely, baselines relying on implicit "muscle memory" are more robust to such blind spots where explicit geometry fails. This confirms that while synergistic memory ensures overall stability, explicit 3D representations remain sensitive to extreme structural dynamics.

Table 5: Success rates for real robot mobile manipulation tasks. We report performance on six basic actions. Avg SR is the mean success rate. (**Notes:** OD: Open the drawer, CM: Close the microwave, PCIS: Place the cup from the rack into the sink, OR: Open refrigerator door, RK: Rotate the knob of a microwave, EnP: Enter the room and place the pears on the cabinet.)

Method	OD	CM	PCIS	OR	RK	EnP	Avg SR
$\pi_{0.5}$	0.35	0.55	0.20	0.50	0.40	0.00	0.33
Diffusion Policy	0.40	0.60	0.50	0.30	0.10	0.03	0.32
EchoVLA (Ours)	**0.45**	**0.70**	**0.50**	**0.40**	**0.50**	**0.10**	**0.44**

6 Conclusion

We present EchoVLA, a memory-aware vision–language–action framework for mobile manipulation. By integrating a 3D Scene Memory with an Episodic Memory, EchoVLA effectively handles non-Markovian decision-making and mitigates perceptual noise. To validate our approach, we introduce MoMani, a benchmark spanning procedurally generated simulation tasks and real-robot evaluations. Extensive experiments demonstrate that EchoVLA consistently outperforms strong baselines across diverse scenarios.

Limitations & Future Work. EchoVLA's performance depends on high-quality depth and pose streams; cumulative odometry drift can cause spatial misalignments or "ghosting" in the Scene Memory. Future work will integrate loop-closure or visual-SLAM to improve tracking robustness. Furthermore, to address the cold-start problem in novel environments where the 3D memory is initially sparse, we plan to explore active exploration strategies or generative 3D priors. Lastly, we will investigate alternative encoder architectures—such as unified multi-modal transformers—to enhance scalability and cross-task generalization.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62506231, the National Key Research and Development Program of China (2024YFE0203100), the Scientific Research Innovation Capability Support Project for Young Faculty (No. ZYGXQNJSKYCXNLZCXM-I28), the Shenzhen Science and Technology Program (Grant No. QNXMA20250701093544048), and the General Embodied AI Center of Sun Yat-sen University.

References

1. Aminoff, E.M., Kveraga, K., Bar, M.: The role of the parahippocampal cortex in cognition. Trends in cognitive sciences **17**(8), 379–390 (2013)

2. Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al.: Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734 (2025)
3. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L.X., Tanner, J., Vuong, Q., Walling, A., Wang, H., Zhilinsky, U.: π_0: A vision-language-action flow model for general robot control (2024), https://arxiv.org/abs/2410.24164, accessed 25 June 2026
4. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022)
5. Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., Gao, S., He, X., Hu, X., Huang, X., et al.: Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. arXiv preprint arXiv:2503.06669 (2025)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
7. Chen, S., Liu, J., Qian, S., Jiang, H., Li, L., Zhang, R., Liu, Z., Gu, C., Hou, C., Wang, P., et al.: Ac-dit: Adaptive coordination diffusion transformer for mobile manipulation. arXiv preprint arXiv:2507.01961 (2025)
8. Chen, T., Chen, Z., Chen, B., Cai, Z., Liu, Y., Li, Z., Liang, Q., Lin, X., Ge, Y., Gu, Z., et al.: Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. arXiv preprint arXiv:2506.18088 (2025)
9. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion (2024), https://arxiv.org/abs/2303.04137, accessed 25 June 2026
10. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research **44**(10-11), 1684–1704 (2024). https://doi.org/10.1177/02783649241273668
11. Deitke, M., Han, W., Herrasti, A., Kembhavi, A., Kolve, E., Mottaghi, R., Salvador, J., Schwenk, D., VanderBilt, E., Wallingford, M., et al.: Robothor: An open simulation-to-real embodied ai platform. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3164–3174 (2020)
12. Deitke, M., VanderBilt, E., Herrasti, A., Weihs, L., Ehsani, K., Salvador, J., Han, W., Kolve, E., Kembhavi, A., Mottaghi, R.: Procthor: Large-scale embodied ai using procedural generation. Advances in Neural Information Processing Systems **35**, 5982–5994 (2022)
13. Ding, P., Ma, J., Tong, X., Zou, B., Luo, X., Fan, Y., Wang, T., Lu, H., Mo, P., Liu, J., et al.: Humanoid-vla: Towards universal humanoid control with visual integration. arXiv preprint arXiv:2502.14795 (2025)
14. Eichenbaum, H.: On the integration of space, time, and memory. Neuron **95**(5), 1007–1018 (2017)
15. Epstein, R., Kanwisher, N.: A cortical representation of the local visual environment. Nature **392**(6676), 598–601 (1998)
16. Epstein, R.A.: Parahippocampal and retrosplenial contributions to human spatial navigation. Trends in cognitive sciences **12**(10), 388–396 (2008)
17. Gu, J., Xiang, F., Li, X., Ling, Z., Liu, X., Mu, T., Tang, Y., Tao, S., Wei, X., Yao, Y., et al.: Maniskill2: A unified benchmark for generalizable manipulation skills. arXiv preprint arXiv:2302.04659 (2023)

18. Gubernatorov, K., Voronov, A., Voronov, R., Pasynkov, S., Perminov, S., Guo, Z., Tsetserukou, D.: Anywherevla: Language-conditioned exploration and mobile manipulation (2025), https://arxiv.org/abs/2509.21006, accessed 25 June 2026
19. He, Z., Zhang, Y., Zhou, Y., Tao, M., Li, H., Tian, Y., Zeng, J., Wang, T., Cai, W., Chen, Y., et al.: Nimbus: A unified embodied synthetic data generation framework. arXiv preprint arXiv:2601.21449 (2026)
20. Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., Fei-Fei, L.: Voxposer: Composable 3d value maps for robotic manipulation with language models. arXiv preprint arXiv:2307.05973 (2023)
21. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Galliker, M.Y., Ghosh, D., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., LeBlanc, D., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Ren, A.Z., Shi, L.X., Smith, L., Springenberg, J.T., Stachowicz, K., Tanner, J., Vuong, Q., Walke, H., Walling, A., Wang, H., Yu, L., Zhilinsky, U.: $\pi_{0.5}$: a vision-language-action model with open-world generalization (2025), https://arxiv.org/abs/2504.16054, accessed 25 June 2026
22. Jiang, Y., Gupta, A., Zhang, Z., Wang, G., Dou, Y., Chen, Y., Fei-Fei, L., Anandkumar, A., Zhu, Y., Fan, L.: Vima: General robot manipulation with multimodal prompts. arXiv preprint arXiv:2210.03094 **2**(3), 6 (2022)
23. Jiang, Y., Zhang, R., Wong, J., Wang, C., Ze, Y., Yin, H., Gokmen, C., Song, S., Wu, J., Fei-Fei, L.: Behavior robot suite: Streamlining real-world whole-body manipulation for everyday household activities (2025), https://arxiv.org/abs/2503.05652, accessed 25 June 2026
24. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
25. Li, C., Xia, F., Martín-Martín, R., Lingelbach, M., Srivastava, S., Shen, B., Vainio, K., Gokmen, C., Dharan, G., Jain, T., et al.: igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. arXiv preprint arXiv:2108.03272 (2021)
26. Li, C., Zhang, R., Wong, J., Gokmen, C., Srivastava, S., Martín-Martín, R., Wang, C., Levine, G., Ai, W., Martinez, B., et al.: Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. arXiv preprint arXiv:2403.09227 (2024)
27. Li, Q., Liang, Y., Wang, Z., Luo, L., Chen, X., Liao, M., Wei, F., Deng, Y., Xu, S., Zhang, Y., et al.: Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. arXiv preprint arXiv:2411.19650 (2024)
28. Li, X., Zhang, M., Geng, Y., Geng, H., Long, Y., Shen, Y., Zhang, R., Liu, J., Dong, H.: Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18061–18070 (2024)
29. Liu, J., Chen, H., An, P., Liu, Z., Zhang, R., Gu, C., Li, X., Guo, Z., Chen, S., Liu, M., et al.: Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. arXiv preprint arXiv:2503.10631 (2025)
30. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55 (2024)

31. Mandlekar, A., Xu, D., Wong, J., Nasiriany, S., Wang, C., Kulkarni, R., Fei-Fei, L., Savarese, S., Zhu, Y., Martín-Martín, R.: What matters in learning from offline human demonstrations for robot manipulation (2021), https://arxiv.org/abs/2108.03298, accessed 25 June 2026
32. Nasiriany, S., Maddukuri, A., Zhang, L., Parikh, A., Lo, A., Joshi, A., Mandlekar, A., Zhu, Y.: Robocasa: Large-scale simulation of everyday tasks for generalist robots. arXiv preprint arXiv:2406.02523 (2024)
33. O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 6892–6903 (2024)
34. Puig, X., Undersander, E., Szot, A., Cote, M.D., Yang, T.Y., Partsey, R., Desai, R., Clegg, A.W., Hlavac, M., Min, S.Y., et al.: Habitat 3.0: A co-habitat for humans, avatars and robots. arXiv preprint arXiv:2310.13724 (2023)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021)
36. Ranganath, C., Ritchey, M.: Two cortical systems for memory-guided behaviour. Nature reviews neuroscience **13**(10), 713–726 (2012)
37. Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S.G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J.T., et al.: A generalist agent. arXiv preprint arXiv:2205.06175 (2022)
38. Ren, P., Li, M., Luo, Z., Song, X., Chen, Z., Liufu, W., Yang, Y., Zheng, H., Xu, R., Huang, Z., et al.: Infiniteworld: A unified scalable simulation framework for general visual-language robot interaction. arXiv preprint arXiv:2412.05789 (2024)
39. Ruan, S., Wang, L., Kang, C., Zhu, Q., Liu, S., Wei, X., Su, H.: From reactive to cognitive: brain-inspired spatial intelligence for embodied agents. intelligence (AGI) **3**(9), 10 (2025)
40. Shi, H., Xie, B., Liu, Y., Sun, L., Liu, F., Wang, T., Zhou, E., Fan, H., Zhang, X., Huang, G.: Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. arXiv preprint arXiv:2508.19236 (2025)
41. Shridhar, M., Manuelli, L., Fox, D.: Cliport: What and where pathways for robotic manipulation. In: Conference on robot learning. pp. 894–906 (2022)
42. Srivastava, S., Li, C., Lingelbach, M., Martín-Martín, R., Xia, F., Vainio, K.E., Lian, Z., Gokmen, C., Buch, S., Liu, K., et al.: Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In: Conference on robot learning. pp. 477–490 (2022)
43. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D.S., Maksymets, O., et al.: Habitat 2.0: Training home assistants to rearrange their habitat. Advances in neural information processing systems **34**, 251–266 (2021)
44. Team, O.M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al.: Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213 (2024)
45. Wang, J., Guo, D., et al.: Pointattn: You only need attention for point cloud completion. In: Proceedings of the AAAI Conference on Artificial Intelligence (2023)
46. Wen, J., Zhu, Y., Li, J., Tang, Z., Shen, C., Feng, F.: Dexvla: Vision-language model with plug-in diffusion expert for general robot control. arXiv preprint arXiv:2502.05855 (2025)

47. Wu, J., Antonova, R., Kan, A., Lepert, M., Zeng, A., Song, S., Bohg, J., Rusinkiewicz, S., Funkhouser, T.: Tidybot: Personalized robot assistance with large language models. Autonomous Robots **47**(8), 1087–1102 (2023)
48. Wu, J., Chong, W., Holmberg, R., Prasad, A., Gao, Y., Khatib, O., Song, S., Rusinkiewicz, S., Bohg, J.: Tidybot++: An open-source holonomic mobile manipulator for robot learning. arXiv preprint arXiv:2412.10447 (2024)
49. Yenamandra, S., Ramachandran, A., Yadav, K., Wang, A., Khanna, M., Gervet, T., Yang, T.Y., Jain, V., Clegg, A.W., Turner, J., et al.: Homerobot: Open-vocabulary mobile manipulation. arXiv preprint arXiv:2306.11565 (2023)
50. Ze, Y., Zhang, G., Zhang, K., Hu, C., Wang, M., Xu, H.: 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations (2024), https://arxiv.org/abs/2403.03954, accessed 25 June 2026
51. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training (2023), https://arxiv.org/abs/2303.15343, accessed 25 June 2026
52. Zhang, P., Gao, X., Wu, Y., Liu, K., Wang, D., Wang, Z., Zhao, B., Ding, Y., Li, X.: Moma-kitchen: A 100k+ benchmark for affordance-grounded last-mile navigation in mobile manipulation. arXiv preprint arXiv:2503.11081 (2025)
53. Zhao, T.Z., Kumar, V., Levine, S., Finn, C.: Learning fine-grained bimanual manipulation with low-cost hardware. arXiv preprint arXiv:2304.13705 (2023)
54. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183 (2023)