Spatially Generalizable Mobile Manipulation via Adaptive Experience Selection and Dynamic Imagination

Ping Zhong[1,2], **Liangbai Liu**[1], **Bolei Chen**[1*], **Tao Wu**[1], **Jiazhi Xia**[1,2], **Chaoxu Mu**[3], **Jianxin Wang**[1]
[1]School of Computer Science and Engineering, Central South University
[2]Xiangjiang Laboratory
[3]School of Electrical and Information Engineering, Tianjin University
{liangbailiu, boleichen, 8102221321, ping.zhong, xiajiazhi}@csu.edu.cn, cxmu@tju.edu.cn, jxwang@mail.csu.edu.cn

Abstract

Mobile **M**anipulation (MM) involves long-horizon decision-making over multi-stage compositions of heterogeneous skills, such as navigation and picking up objects. Despite recent progress, existing MM methods still face two key limitations: (i) low sample efficiency, due to ineffective use of redundant data generated during long-term MM interactions; and (ii) poor spatial generalization, as policies trained on specific tasks struggle to transfer to new spatial layouts without additional training. In this paper, we address these challenges through **A**daptive **E**xperience **S**election (AES) and model-based dynamic imagination. In particular, AES makes MM agents pay more attention to critical experience fragments in long trajectories that affect task success, improving skill chain learning and mitigating skill forgetting. Based on AES, a **R**ecurrent **S**tate-**S**pace **M**odel (RSSM) is introduced for **M**odel-**P**redictive **F**orward **P**lanning (MPFP) by capturing the coupled dynamics between the mobile base and the manipulator and imagining the dynamics of future manipulations. RSSM-based MPFP can reinforce MM skill learning on the current task while enabling effective generalization to new spatial layouts. Comparative studies across different experimental configurations demonstrate that our method significantly outperforms existing MM policies. Real-world experiments further validate the feasibility and practicality of our method. The source code is available at https://csu-hero-lab.github.io/SG-MM_Web.

1 Introduction

Mobile **M**anipulation (MM) is a cornerstone of embodied AI, enabling agents to perceive, reason, and act in unstructured environments. Unlike isolated navigation or fixed-base manipulation, MM requires agents to coordinate locomotion and dexterous interaction over extended spatial and temporal horizons. For example, MM allows embodied agents to complete complex goal-directed tasks such as opening doors and fetching objects, as shown in Fig. 1 (c). MM is quite challenging, as it typically involves long-horizon decision-making over multi-stage compositions of heterogeneous skills [Huang *et al.*, 2023; Lin *et al.*, 2025b]. To address the complexity of MM, early methods decompose the problem into navigation [Chen *et al.*, 2024; Kang *et al.*, 2024] and stationary manipulation [Nguyen and La, 2019]. This factorized design simplifies modeling and learning, and has shown promise in structured environments. However, such approaches [Jauhri *et al.*, 2024; Arduengo *et al.*, 2021; Feng *et al.*, 2025] often optimize only short-term objectives and overlook the dynamic coupling between mobility and manipulation, making them vulnerable to error accumulation

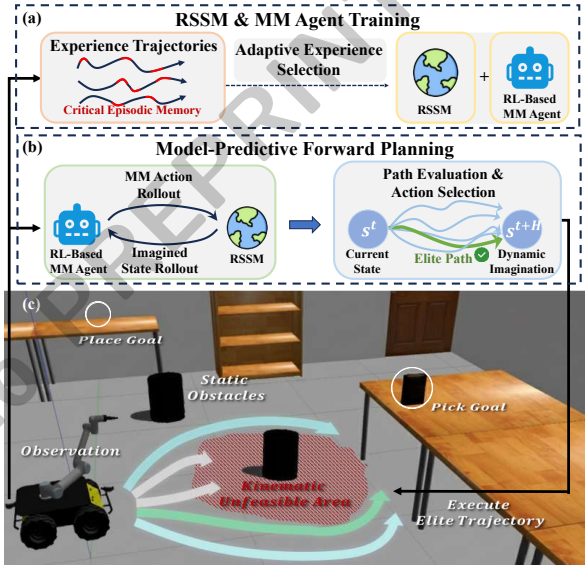

Figure 1: (a) Our AES enables RSSM to pay more attention to critical episodic memories in experience trajectories, which are used to improve and spatially generalize RL-based MM agents. (b) RSSM cooperates with RL-based MM skills for H-horizon dynamic imagination to achieve MPFP. The most promising elite path is selected for execution. (c) An example of MM tasks that consist of heterogeneous skills, i.e., navigation → picking → navigation → placing.

*Corresponding Author.

over long horizons.

Recent work [Honerkamp *et al.*, 2023a; Huang *et al.*, 2023] models the mobile base and manipulator as a unified system using whole-body control or joint policy learning. Despite encouraging progress, these solutions, especially model-free **R**einforcement **L**earning (RL)–based methods [Honerkamp *et al.*, 2023a; Huang *et al.*, 2023], usually suffer from inefficient exploration and low sample efficiency. The reason is that MM is essentially a long-horizon task consisting of skills with strict sequential dependencies, which introduces unique data distribution challenges that are different from those of standard RL benchmarking: **(1)** Failures in the early stages (navigation) can hinder the experience collection of all downstream tasks (picking and placing), thus hindering the learning of complete skill chains. **(2)** As the trajectory lengthens, the experience buffer becomes dominated by navigation-intensive fragments. This results in sparse but critical interaction moments in the trajectory being overwhelmed by redundant navigation experiences. **(3)** Such methods lack explicit mechanisms for learning reusable interaction patterns, which can easily lead to skill forgetting. Although recent data-driven or model-based methods [Chen *et al.*, 2025; Cen *et al.*, 2025; Lin *et al.*, 2025a] can help mitigate these issues, they are often sensitive to environmental variations and difficult to spatially generalize. The above limitations raise a fundamental question: **How can a MM policy efficiently exploit long-horizon experience to achieve robust decision-making and generalization across different spatial layouts?**

Recent advances [Sun *et al.*, 2023; Tafazoli *et al.*, 2025] in system neuroscience suggest that higher animals solve novel tasks by recalling task-relevant episodic fragments and composing them with previously acquired skills or habitual behaviors. Inspired by this, we aim to investigate how to memorize key and informative episodic fragments from rich experience trajectories, and use them to improve RL-learned MM policies so that they can adapt to new spatial layouts. To this end, we address the above limitations by recalling episodic memories and composing them with MM skills to produce robust and spatially generalizable MM behaviors.

As shown in Fig. 1 (a), we propose **A**daptive **E**xperience **S**election (AES) to make MM agents pay more attention to key experience fragments in long trajectories that affect task success, improving skill chain learning and mitigating skill forgetting. Based on the informative state transitions well-selected by AES, a **R**ecurrent **S**tate-**S**pace **M**odel (RSSM) is introduced for **M**odel-**P**redictive **F**orward **P**lanning (MPFP) by capturing the coupled dynamics between the mobile base and the manipulator and imagining the dynamics of future manipulations. RSSM is trained together with model-free RL-based MM skill learning, and they share a single experience buffer. The difference is that AES helps RSSM memorize critical episodic fragments in a sample-efficient manner, thereby identifying weaknesses of the current MM skills and further consolidating them. As shown in Fig. 1 (b), RSSM-based MPFP can prospectively imagine future state transitions and evaluate state goodness to support RL-based MM skill learning. As system neuroscience suggests, RSSM here acts as a function of memorizing and recalling episodic fragments, while RL-learned skills act as habitual behaviors specific to MM. Their combination can improve MM policies on the current task and generalize to new spatial layouts without performance loss and additional training.

We implement our above ideas based on the popular **E**nd-**E**ffector (EE)-centric MM framework [Honerkamp *et al.*, 2023a]. Our approach is evaluated through sufficient comparative studies in diverse experimental configurations, demonstrating significant performance gains over the **S**tate-**o**f-**T**he-**A**rt methods. Ablation studies further confirm the individual contributions of AES and RSSM-based MPFP to sample efficiency and spatial generalization. Finally, real-world robotic experiments highlight the practicality of our methods in realistic MM scenarios. In summary, our contributions are threefold: **(1)** We propose a system neuroscience-inspired method that leverages sample-efficient episodic recall to consolidate RL-based skill learning for robust and spatially generalizable MM. **(2)** An AES mechanism is introduced to make MM agents pay more attention to key experience fragments in long trajectories that affect task success, improving sample efficiency. **(3)** RSSM-based dynamic imagination is proposed for MPFP by capturing the coupled dynamics between the mobile base and the manipulator and imagining the dynamics of future manipulations.

2 Related Work

Model-Free RL-based Mobile Manipulation. Due to the difficulty of planning in the joint space of the mobile base and the manipulator, early work solves an MM task by decomposing it into sequential navigation and static manipulation tasks. Such a decomposition has been widely adopted in methods based on planning [Jauhri *et al.*, 2024], reachability analysis [Feng *et al.*, 2025], and RL [Gu *et al.*, 2022; Fang *et al.*, 2022]. Planning-based methods [Jauhri *et al.*, 2024] plan trajectories for robots in joint space to ensure the kinematic feasibility of MM tasks. Given explicit task constraints, inverse reachability maps [Vahrenkamp *et al.*, 2013] can also be exploited to determine favorable placements for the mobile base that facilitate subsequent manipulation. However, these classical methods are usually unable to instantly respond to diverse unstructured scenes and dynamic changes, and are thus less adaptable. Recently, RL-based approaches [Gu *et al.*, 2022; Huang *et al.*, 2023] mitigate these issues by using hierarchical policy learning and curriculum learning techniques. In addition, there are methods [Jauhri *et al.*, 2022] that use classical methods as behavioral priors to guide sample-efficient exploration. Most recently, some methods [He *et al.*, 2025; Honerkamp *et al.*, 2023a] investigate the EE-centric MM framework, which achieve strong autonomy and robustness by explicitly solving for base motion that cooperates with the manipulator during grasping.

Despite recent progress, existing MM policies still struggle with low sample efficiency, as they fail to exploit the sparse but informative transitions embedded in long-horizon trajectories. Moreover, policies learned for specific tasks generalize poorly to new spatial layouts due to the absence of mechanisms for memorizing and reusing task-relevant experience. In this work, we address these issues by proposing AES and

RSSM-based dynamic imagination.

Model-based Mobile Manipulation. **V**ision-**L**anguage-**A**ction (VLA) models [Liu *et al.*, 2025; Zhang *et al.*, 2025] have recently emerged as a powerful paradigm for grounding multimodal instruction and perception into direct robot control. HybridVLA [Liu *et al.*, 2025] jointly integrates autoregressive and diffusion-based policy heads to improve robustness across diverse manipulation tasks. EchoVLA [Lin *et al.*, 2025a] augments VLA with declarative and episodic memory to better handle long-horizon MM and spatial context reasoning. Diffusion-based action models [Wen *et al.*, 2025] introduce masked diffusion mechanisms tailored for structured action generation, setting new standards over autoregressive VLAs in multimodal control. In parallel, world-model-augmented frameworks [Cen *et al.*, 2025; Hafner *et al.*, 2025] seek to unify dynamics prediction with action planning, enabling joint learning of environment dynamics and policy, which can improve long-horizon foresight. Despite these advances, current model-based methods often require large multimodal datasets and still struggle with sample efficiency and reliable generalization across diverse spatial layouts. In this work, we employ RSSM-based MPFP to improve and generalize RL-based MM skills to address these issues.

3 Preliminaries

Problem Definition. The MM problem we consider in this work involves a robot that includes a mobile base and a manipulator. The MM tasks described in this paper involve skills such as navigation, collision avoidance, picking, placing, and opening. Such a problem is modeled as a goal-conditional **P**artially **O**bservable **M**arkov **D**ecision **P**rocess (POMDP) defined as a tuple $\mathcal{X} = (\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{G}, \rho, \gamma)$ with underlying state space $\mathcal{S}$, observation space $\mathcal{O}$, action space $\mathcal{A}$, reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, initial state distribution ρ and discount factor γ. The state transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}_+$ represents the probability density of the next state $\boldsymbol{s}^{t+1} \in \mathcal{S}$ given the current state $\boldsymbol{s}^t \in \mathcal{S}$ and action $\boldsymbol{a}^t \in \mathcal{A}$. For tasks like object rearrangement [Szot *et al.*, 2022], the goal space $\boldsymbol{g} \in \mathcal{G}$ is specified using the object's initial pickup pose or the goal pose where the object has to be placed. Our objective is to learn a policy $\pi_\theta(\boldsymbol{a}^t|\boldsymbol{o}^t, \boldsymbol{g})$ mapping an observation $\boldsymbol{o}^t \in \mathcal{O}$ and goal $\boldsymbol{g}$ to an action $\boldsymbol{a}^t$ that maximizes the expected return [Chen *et al.*, 2023]:

$$\pi^*_\theta(\boldsymbol{s}^t; \boldsymbol{g}) = \arg\max_{\boldsymbol{a}^t} \Big[\mathcal{R}(\boldsymbol{s}^t, \boldsymbol{a}^t; \boldsymbol{g}) + \gamma \int_{\boldsymbol{s}^{t+1}} \mathcal{P}(\boldsymbol{s}^{t+1} \mid \boldsymbol{s}^t, \boldsymbol{a}^t) V^*(\boldsymbol{s}^{t+1}; \boldsymbol{g})\, d\boldsymbol{s}^{t+1} \Big], \tag{1}$$

where $V^*(\boldsymbol{s}^t; \boldsymbol{g}) = \mathbb{E}_{\pi_\theta, \mathcal{P}} \Big[\sum_{k=t}^{T} \gamma^{k-t} \mathcal{R}(\boldsymbol{s}^k, \pi^*_\theta(\boldsymbol{s}^t; \boldsymbol{g})) \Big]$ is the optimal value function. In this setting, $\mathcal{P}(\boldsymbol{s}^{t+1} \mid \boldsymbol{s}^t, \boldsymbol{a}^t)$ represents the state transition over time, which is usually unknown to the MM agent. In this work, our MM policy recall key experience fragments and combine them with habitual skills to produce robust and spatially generalizable MM behaviors. Therefore, the state transition $\mathcal{P}$ and the optimal value function V^* are learned together through dynamic imagination and model-free RL, respectively.

EE-Centric MM. To instantiate the MM tasks, we use the **R**obot **O**perating **S**ystem (ROS) to control an robotic system consisting of a Husky mobile base and a 6-DoF UR5 manipulator. Following the EE-centric MM framework [He *et al.*, 2025; Honerkamp *et al.*, 2023a], given a sequence of 6D poses in $SE(3)$, the robot needs to find appropriate base placements in $SE(2)$ to allow the manipulator's EE to reach these 6D goal poses sequentially along dynamically feasible paths. Such a solution is natural, just as humans reach for objects through leg movements in conjunction with their arms.

In practice, we employ an **I**nverse **K**inematics (IK) solver to solve the manipulator's motion in joint space. On this basis, our objective is to learn the navigation of the mobile base to cooperate with the long-horizon manipulation. This is challenging because the mobile base must move in a collision-free manner while ensuring that the next 6D goal pose of the manipulator is reachable. Otherwise, the MM task is terminated due to collision of the mobile base or IK solving failure of the manipulator. To ensure the robot's kinematic feasibility, the mobile base's 2D action space $\mathcal{A} \in \mathbb{R}^2$ is continuous. The observation space $\mathcal{O}$ includes a robot-centric local occupancy map $\boldsymbol{m}^t$ and a state vector $\boldsymbol{v}^t$ consisting of the previous actions $\boldsymbol{a}^{t-1}$, the joint states of robot, the EE's pose and velocity, and the desired pose of EE. Such an EE-centric MM framework inherits decades of experience in the field of robotics and incorporates the advantages of robot learning while ensuring robustness and practicality.

4 Methodology

In the following section, we propose leveraging sample-efficient episodic recall to reinforce and spatially generalize RL-based MM skills. We first present how to employ AES-enhanced RSSM to memorize informative experience fragments for dynamic imagination (§4.1). Then, we present how to achieve MPFP based on RSSM to boost RL-based MM skill learning and promote spatial generalization (§4.2).

4.1 AES-Enhanced RSSM for Dynamic Imagination

Recurrent State-Space Model. RSSM is responsible for modeling state transitions in a latent space to capture the coupled dynamics between mobile base and manipulator. We adopt a structured latent representation consisting of two components: the stochastic representation $\boldsymbol{z}^t$ and the recurrent state $\boldsymbol{h}^t$, maintaining consistent MM progress while incorporating new observations. An encoder e_ϕ is used to encode the observation $\boldsymbol{o}^{t-1} = \{\boldsymbol{m}^{t-1}, \boldsymbol{v}^{t-1}\}$ as a stochastic representation $\boldsymbol{z}^{t-1}$, which is integrated into the recurrent state $\boldsymbol{h}^{t-1}$ to form a comprehensive state representation for MM decision-making. A recurrent module f_ϕ is utilized as the fundamental component of RSSM for latent state modeling, ensuring consistent temporal dynamics across state inference and future prediction. A decoder p_ϕ is used to predict observations and rewards from the recurrent state. As shown in Fig. 2 (a), the components of our dynamic imagination

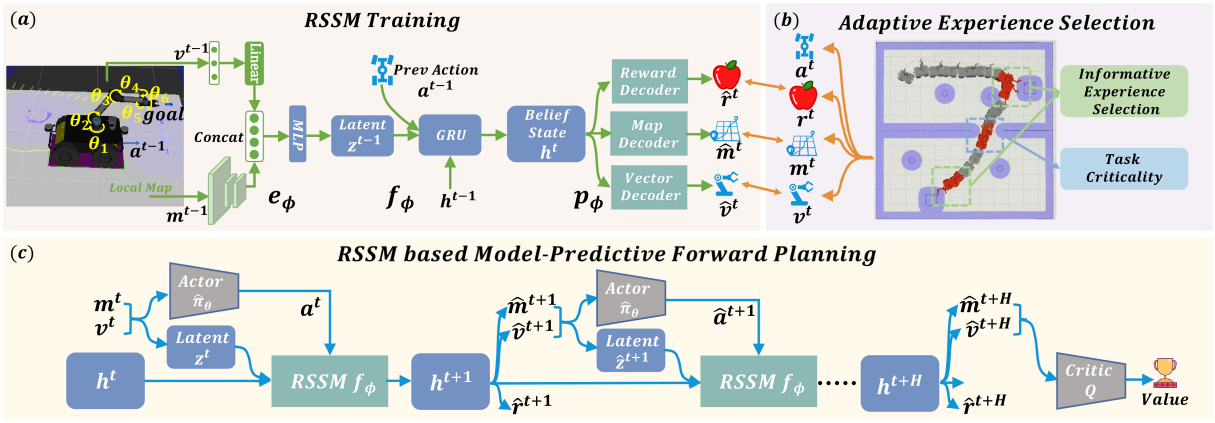

Figure 2: (a) An illustration of the structure, unfolding, and training process of RSSM. (b) An illustration of which key experience fragments the AES focuses on, including the experience of interacting with objects, opening doors, and passing through narrow areas. These experience segments are likely to have collisions and IK solving failures that affect task success. (c) An illustration of RSSM-based MPFP.

model are outlined below:

$$\begin{aligned} \text{Observation Encoder:} \boldsymbol{z}^{t-1} &= e_\phi(\boldsymbol{o}^{t-1}), \\ \text{Dynamic Model: } \boldsymbol{h}^{t} &= f_\phi(\boldsymbol{h}^{t-1}, \boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1}), \\ \text{Observation Predictor: } \hat{\boldsymbol{o}}^{\boldsymbol{t}} &\sim p_\phi(\hat{\boldsymbol{o}}^{\boldsymbol{t}} | \boldsymbol{h}^{\boldsymbol{t}}), \\ \text{Reward Predictor: } \hat{\boldsymbol{r}}^{t} &\sim p_\phi(\hat{\boldsymbol{r}}^{t} | \boldsymbol{h}^{t}). \end{aligned} \tag{2}$$

According to the problem definition in Sec. 3, the dynamic model f_ϕ approximates $\mathcal{P}$ by modeling state transitions. The training of e_ϕ, f_ϕ, and q_ϕ is parallel to RL-based MM skill learning, and they share the same experience buffer. The difference is that RSSM's training data is well-selected by AES as follows. The loss functions used to train RSSM are consistent with Dreamer V3 [Hafner *et al.*, 2025].

Adaptive Experience Selection. MM skills learned in sample-inefficient manner are difficult to generalize to new spatial layouts. On the one hand, standard uniform sampling is insufficient, as it tends to overfit the abundant navigation data while neglecting the critical manipulation experience strongly associated with task success. On the other hand, focusing solely on hard failures risks catastrophic forgetting of foundational navigation capabilities. To address this challenge, we introduce an AES mechanism that pays more attention to critical state transitions in MM trajectories, as shown in Fig. 2 (b). We aim to maintain a balance between common navigation and critical failure experience, thus fostering a dynamic model that is robust across all skills and diverse spatial layouts to enhance generalization and prevent skill forgetting. Technically, the RSSM is trained on fixed-length experience fragments that are extracted from episodes collected by the RL-based MM agent. For each MM episode that is put into the experience buffer, the experience segments are sliding experience windows that contain consecutive state transitions that have the highest total prioritization. Prioritization is achieved by assigning the i-th state transition a composite priority score P^i_{total}, which integrates three complementary signals:

(1) Informative Experience Selection: From a data distribution perspective, informative state transitions that exhibit significant changes usually account for a small percentage and can thus be considered outliers. These state transitions usually correspond to moments such as opening a door, picking up or placing the target object, and avoiding collisions, as shown in Fig. 2 (b). Therefore, we propose to transform the task of selecting critical experience into the detection of outliers in the feature space of observations. Specifically, given a sequence of observations $\{\boldsymbol{o}^1, \boldsymbol{o}^2, ..., \boldsymbol{o}^T\}$, where $\boldsymbol{o}^t = \{\boldsymbol{m}^t, \boldsymbol{v}^t\}$, we use $d(\boldsymbol{o}^i, \boldsymbol{o}^j)$ to denote the distance in feature space:

$$d(\boldsymbol{o}^i, \boldsymbol{o}^j) = ||\boldsymbol{m}^j - \boldsymbol{m}^i||_1 + \lambda \cdot (1 - \frac{\boldsymbol{v}^{i\top} \boldsymbol{v}^j}{||\boldsymbol{v}^i||_2 \cdot ||\boldsymbol{v}^j||_2}), \tag{3}$$

where λ is a weight used to balance magnitude. Based on Eq. (3), we define the k-distance neighborhood: $N_k(\boldsymbol{o}^i) = \{\boldsymbol{o}^j | d(\boldsymbol{o}^i, \boldsymbol{o}^j) \leq d_k(\boldsymbol{o}^i)\}$, where $d_k(\boldsymbol{o}^i)$ is the distance to the k-th nearest neighbor. For each neighbor $\boldsymbol{o}^j \in N_k(\boldsymbol{o}^i)$, the reachability distance is defined as reach-$d_k(\boldsymbol{o}^i, \boldsymbol{o}^j) = max\{d_k(\boldsymbol{o}^j), d(\boldsymbol{o}^i, \boldsymbol{o}^j)\}$. Using this, the **L**ocal **O**utlier **F**actor (LOF) [Breunig *et al.*, 2000] of $\boldsymbol{o}^i$ is the inverse of its mean reachability distance:

$$\text{LRD}_k(\boldsymbol{o}^i) = \left(\frac{1}{|N_k(\boldsymbol{o}^i)|} \sum_{\boldsymbol{o}^j \in N_k(\boldsymbol{o}^i)} \text{reach-}d_k(\boldsymbol{o}^i, \boldsymbol{o}^j) \right)^{-1}. \tag{4}$$

Then, the LOF is calculated as the relative density of $\boldsymbol{o}^i$ compared to its neighbors:

$$P^i_{ice} = \text{LOF}_k(\boldsymbol{o}^i) = \frac{1}{|N_k(\boldsymbol{o}^i)|} \sum_{\boldsymbol{o}^j \in N_k(\boldsymbol{o}^i)} \frac{\text{LRD}_k(\boldsymbol{o}^j)}{\text{LRD}_k(\boldsymbol{o}^i)} \tag{5}$$

Based on this formulation, the observation is selected when $\text{LOF}_k(\boldsymbol{o}^i) > 1$. A high $\text{LOF}_k(\boldsymbol{o}^i)$ indicates a novel and unusual state, receiving proportional priority.

(2) Task Criticality: This signal marks state transitions containing IK solving failures of the manipulator. These informative failure modes receive significant prioritization P_{tc}, ensuring that RSSM consistently focuses on learning to navigate to coordinate with successful IK solving, thereby preventing future failures.

(3) Prediction Error: We follow the idea of prioritized experience replay [Schaul *et al.*, 2015], employing the RSSM's

overall training loss as a dynamic indicator of learning difficulty. The state transitions that are modeled incorrectly by RSSM are given higher priorities P_{per}, which the RSSM should focus on in the following training.

Overall, the final priority P^i_{total} for i-th state transition is computed as a weighted sum of three signals:

$$P^i_{total} = w_1 \max\big(0,\, P^i_{ice} - 1\big) + w_2 P^i_{tc} \cdot \mathbb{I} + w_3 P^i_{per}, \quad (6)$$

where w_1, w_2, w_3 are respective weights and $\mathbb{I}$ indicates whether an IK solving failure occurs. In practice, all high-priority experience fragments are structured as a SumTree [Wang *et al.*, 2024] to efficiently manage and sample them based on normalized prioritization.

4.2 Dynamic Imagination-based Foresighted MM

Since RSSM and MM skills are trained together, AES-enhanced RSSM can provide foresighted guidance for RL-based MM skill learning through dynamic imagination. In addition, we emphasize using the AES-enhanced RSSM to generalize fully trained MM agents to new spatial layouts without additional learning. Following common practice [Jauhri *et al.*, 2022; Honerkamp *et al.*, 2023b], we model a maximum-entropy MM policy $\hat{\pi}_\theta$ to explore the state space of a given task using the **S**oft **A**ctor-**C**ritic (SAC) algorithm [Haarnoja *et al.*, 2018]. The following RSSM-based MPFP is introduced to reinforce and spatially generalize MM skills during training and evaluation, respectively.

Model-Predictive Forward Planning. During MM decision-making in both training and evaluation, the agent does not directly execute the action predicted by the policy $\hat{\pi}_\theta$ learned by RL. Instead, it leverages RSSM as an internal simulator for dynamic imagination. As shown in Fig. 2 (c), this process integrates the concurrently trained actor-critic networks $(\hat{\pi}_\theta, Q)$, which are used for the action rollout in dynamic imagination and the value estimation of the last imagined step, respectively. At each timestep, the MM agent samples candidate action sequences from a Gaussian distribution $\mathcal{N}(\mu, \sigma)$, where the mean μ is initialized by the actor policy $\hat{\pi}_\theta(\boldsymbol{o}^t; \boldsymbol{g})$ based on the current observation $\boldsymbol{o}^t$ and goal $\boldsymbol{g}$. The sampled actions are then iteratively refined through CEM for model-predictive planning. During each iteration of **C**ross-**E**ntropy **M**ethod (CEM) [Rubinstein, 1999], L candidate action sequences of horizon H are drawn and evaluated in parallel by unrolling the RSSM:

$$\begin{aligned}
\text{Dynamic Imagination: } & \boldsymbol{h}_i^{k+1} = f_\phi(\boldsymbol{h}_i^k, \boldsymbol{z}_i^k, \boldsymbol{a}_i^k),\\
\text{Observation Rollout: } & \hat{\boldsymbol{o}}_i^{k+1} \sim p_\phi(\hat{\boldsymbol{o}}_i^k | \hat{\boldsymbol{h}}_i^k),\\
\text{Action Rollout: } & \hat{\boldsymbol{a}}_i^{k+1} \sim \mathcal{N}(\mu_i^{k+1}, \sigma_i^{k+1}),\\
\text{Observation Encoding: } & \hat{\boldsymbol{z}}_i^{k+1} = e_\phi(\hat{\boldsymbol{o}}_i^k),\\
\text{Reward Prediction: } & \hat{\boldsymbol{r}}_i^{k+1} \sim p_\phi(\hat{\boldsymbol{r}}_i^{k+1} | \boldsymbol{h}_i^{k+1}),
\end{aligned} \quad (7)$$

where $i = 0, ..., L-1$ and $k \in [t, t+H-1]$. As shown in Fig. 2 (c), each H-horizon dynamic imagination yields a cumulative return:

$$\hat{\mathcal{R}}^t = \sum_{k=t}^{t+H-1} \gamma^{k-t} \hat{\boldsymbol{r}}_i^k + \gamma^{t+H} Q(\hat{\boldsymbol{o}}_i^{t+H}, \hat{\pi}_\theta(\hat{\boldsymbol{o}}_i^{t+H}; \boldsymbol{g})), \quad (8)$$

which is an estimate of

$$\gamma \int_{\boldsymbol{s}^{t+1}} \mathcal{P}(\boldsymbol{s}^{t+1} \mid \hat{\boldsymbol{o}}^t, \boldsymbol{a}^t) V^*(\boldsymbol{s}^{t+1}; g)\, d\boldsymbol{s}^{t+1}$$

in Eq. (1). Notably, the critic Q provides a terminal value estimate based on the last imagined observation $\hat{\boldsymbol{o}}_i^{t+H}$, which is decoded from the belief state $\boldsymbol{h}_i^{t+H}$. The top-K sequences with the highest returns are selected as elites, and the distribution parameters (μ, σ) of $\hat{\pi}_\theta$ are refitted to their empirical statistics. This sampling-evaluation-refitting loop repeats for a fixed number of iterations. After convergence, only the first action of the final mean sequence is executed by the MM agent, and the overall process repeats at the next timestep, forming a standard model-predictive cycle. The combination of RSSM rollouts, MM decision-making, and CEM optimization enables robust and foresighted planning under complex dynamics and uncertainty.

5 Experiments

Based on the EE-centric MM framework introduced in Sec. 3, we answer the following research questions through comparative and ablative studies: **(1)** How does our method perform compared to strong baselines? **(2)** Does RSSM-based MPFP help to generalize the model-free RL-based MM skills to new spatial layouts? **(3)** Does our proposed AES mechanism help improve sample efficiency? **(4)** Does RSSM-based MPFP help to reinforce the MM policy? **(5)** How well does our MM policy perform on the real-world robot?

5.1 Experimental Setup

Experimental Configurations. As shown in Fig. 3, we follow the experimental configurations in N^{2}M^2 [Honerkamp *et al.*, 2023b] and consider four different experimental configurations: cross-room, warehouse-oriented, home-scene-oriented, and dynamic-scene-oriented mobile manipulations. We first train and evaluate the proposed method and baselines individually in each of the four scenes. Since cross-room MM involves more diverse skills, including navigation, collision avoidance, picking up, opening doors, and placing, we evaluate the performance of generalizing the MM agent fully trained in such a configuration to the other three configurations. This refers to generalizing MM agents to new spatial layouts without additional training. Please see the supplementary material for more details.

Baselines and Evaluation Metrics. We compare our proposed method with several model-free RL-based MM policies, including the SoTA EE-centric MM method N^{2}M^2 [Honerkamp *et al.*, 2023b], an end-to-end whole-body control method adapted from N^{2}M^2, the award-winning hierarchical method BHyRL [Jauhri *et al.*, 2022], and the SoTA approach based on auxiliary task distillation [Harish *et al.*, 2024]. In addition, we compare our method with two world model-based MM policies: **(1) Dreamer V3-based MM** [Hafner *et al.*, 2025]: This baseline adapts the Dreamer V3 architecture to the MM setting by learning a latent world model and performing latent imagination rollout for policy optimization. We modify the open-source implementation of Dreamer V3 to make it suitable for EE-centric MM to obtain experimental results. **(2) TD-MPC2-based MM** [Hansen *et al.*, 2024]:

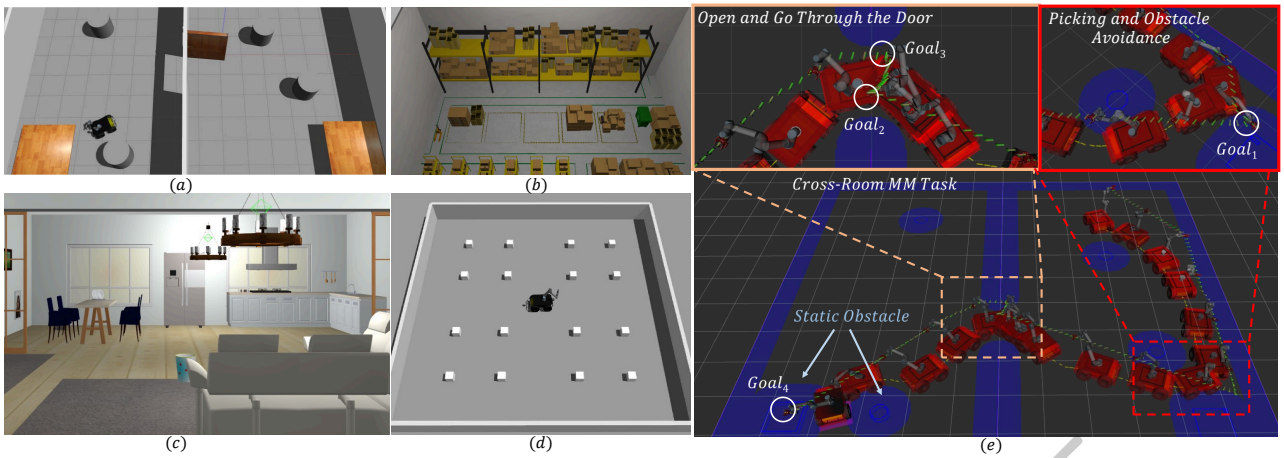

Figure 3: We consider four different experimental configurations: (a) cross-room, (b) warehouse-oriented, (c) home-scene-oriented, and (d) dynamic-scene-oriented (white dynamic obstacles) mobile manipulations. (e) An illustration of the cross-scene MM task. In each episode, the positions of the robot, the obstacles, the desks, the picking point (Goal_1), the start point of opening the door (Goal_2), the end point of opening the door (Goal_3), and the placing point (Goal_4) are randomly initialized. This task requires the robot to pick up an object and then open a door to another room to place the object in a specified position in the room. More details can be found in the supplementary material.

Method	Cross-room			
	AIKF↓	ABC↓	TCR↑	PSR↑
End-to-End MM	-	-	$0_{\pm 0}$	$0_{\pm 0}$
BHyRL	$12.1_{\pm 1.2}$	$2.8_{\pm 0.5}$	$21_{\pm 3}$	$19_{\pm 4}$
N^2M^2	$11.3_{\pm 1.3}$	$2.2_{\pm 0.6}$	$29_{\pm 4}$	$17_{\pm 2}$
Dreamer V3	$8.1_{\pm 1.3}$	$2.0_{\pm 0.5}$	$54_{\pm 4}$	$40_{\pm 1}$
TD-MPC2	$9.9_{\pm 1.6}$	$\mathbf{1.2}_{\pm 0.4}$	$46_{\pm 3}$	$41_{\pm 1}$
AuxDistill	$8.8_{\pm 1.5}$	$1.3_{\pm 0.3}$	$57_{\pm 6}$	$42_{\pm 7}$
Ours	$\mathbf{4.1}_{\pm 1.1}$	$2.0_{\pm 0.2}$	$\mathbf{81}_{\pm 3}$	$\mathbf{54}_{\pm 2}$

Table 1: Comparative studies are conducted in the cross-room experimental configuration to demonstrate the superiority of our method. Each method was evaluated over 100 episodes across 3 random seeds. We report the mean and standard deviation of each metric.

This baseline applies TD-MPC2 to the EE-centric MM setting as a model-based MPC-style planner that uses online trajectory optimization in the latent state space. Similar to our approach, only the first action from the optimized sequence is executed in the environment, resulting in a receding-horizon control loop suitable for complex MM dynamics.

We report the **A**verage **IK** solution **F**ailures (AIKF), **A**verage **B**ase **C**ollisions (ABC), **T**ask **C**ompletion **R**ate (TCR), and **P**erfect **S**uccess **R**ate (PSR) as evaluation metrics. A lower AIKF means a better cooperation between the mobile base and the manipulator. TCR implies that the MM robotic system completes the task while allowing for IK solving failures and base collisions. We set the threshold for allowing both kinds of failures to 20. PSR means completing the MM tasks with zero failures. For implementation details, please see Section 1 of the supplementary material.

5.2 Comparative Studies

To answer question (1), we first conduct comparative studies of the proposed method with all baselines in the challenging cross-room experimental configuration, and the experimental results are shown in Tab. 1. Our approach significantly outperforms the SoTA EE-centric MM policy (N^2M^2), the RL-based maximum entropy policy (BHyRL), the world model-based methods (Dreamer V3 and TD-MPC2), and the auxiliary task distillation-based method (AuxDistill) in terms of reducing IK solving failures and increasing task success. As stated in Section 2 of the supplementary material, our method converges faster and better compared with strong baselines during training, reflecting that our method is more sample-efficient. The results of End-to-End MM indicate that it is difficult to learn MM policies based on whole-body control in an end-to-end manner using model-free RL. The results in Tab. 1 reflect that our method sometimes achieved sub-optimal ABC metrics. This is mainly because our approach reduces early task termination due to IK solving failures, and thus the extended episodes increase the likelihood of collisions. Furthermore, we train and evaluate our method and strong baselines in three other experimental configurations, with results shown in Tab. 2. Either in static or dynamic scenes, our approach significantly reduces the number of IK solving failures (AIKF). In addition, our approach achieves competitive ABC metrics while significantly improving MM success rates (i.e., TCR and PSR). These results reflect the significant superiority of our method over the strong baselines. Training curves reflecting the sample efficiency can be found in the supplementary material. These advances are due to the fact that (1) AES enables our method paying more attention to those key state transitions that determine the success of the MM task; (2) RSSM-based MPFP helps to generate foresighted and collision-free mobile base motions to cooperate with the IK solving of the manipulator.

To answer question (2), both our method and three strong baselines trained in the cross-room experimental configuration are directly evaluated in the other three configurations without additional training. The results in Tab. 3 reflect that our approach can better generalize the RL-based MM agent to new spatial layouts, while N^2M^2 generalizes poorly. Com-

Method	Home-scene				Warehouse				Dynamic-scene			
	AIKF↓	ABC↓	TCR↑	PSR↑	AIKF↓	ABC↓	TCR↑	PSR↑	AIKF↓	ABC↓	TCR↑	PSR↑
N^2M^2	$11.0_{\pm 0.9}$	$2.3_{\pm 0.5}$	$32_{\pm 4}$	$18_{\pm 2}$	$7.1_{\pm 1.5}$	$0.0_{\pm 0.0}$	$68_{\pm 2}$	$56_{\pm 3}$	$16.5_{\pm 1.1}$	$6.4_{\pm 0.8}$	$12_{\pm 2}$	$3_{\pm 1}$
Dreamer V3	$9.3_{\pm 0.7}$	$1.6_{\pm 0.5}$	$64_{\pm 2}$	$45_{\pm 2}$	$5.3_{\pm 1.2}$	$0.4_{\pm 0.2}$	$80_{\pm 3}$	$61_{\pm 1}$	$9.0_{\pm 0.9}$	$\mathbf{3.1}_{\pm 0.4}$	$44_{\pm 3}$	$\mathbf{40}_{\pm 3}$
TD-MPC2	$10.7_{\pm 1.1}$	$1.5_{\pm 0.4}$	$56_{\pm 3}$	$32_{\pm 2}$	$6.0_{\pm 1.0}$	$0.5_{\pm 0.5}$	$74_{\pm 2}$	$60_{\pm 1}$	$8.9_{\pm 1.0}$	$5.0_{\pm 0.6}$	$40_{\pm 2}$	$35_{\pm 1}$
Ours	$\mathbf{3.6}_{\pm 1.1}$	$\mathbf{0.1}_{\pm 0.1}$	$\mathbf{90}_{\pm 2}$	$\mathbf{66}_{\pm 4}$	$\mathbf{3.0}_{\pm 1.0}$	$\mathbf{0.0}_{\pm 0.0}$	$\mathbf{96}_{\pm 4}$	$\mathbf{76}_{\pm 3}$	$\mathbf{5.8}_{\pm 1.6}$	$5.6_{\pm 0.4}$	$\mathbf{58}_{\pm 3}$	$39_{\pm 1}$

Table 2: Comparative studies are conducted in three other experimental configurations to demonstrate the superiority of our method.

Method	Home-scene				Warehouse				Dynamic-scene			
	AIKF↓	ABC↓	TCR↑	PSR↑	AIKF↓	ABC↓	TCR↑	PSR↑	AIKF↓	ABC↓	TCR↑	PSR↑
N^2M^2	$15.1_{\pm 1.3}$	$1.4_{\pm 0.3}$	$29_{\pm 1}$	$15_{\pm 1}$	$14.3_{\pm 1.2}$	$0.0_{\pm 0.0}$	$19_{\pm 1}$	$11_{\pm 1}$	$18.7_{\pm 1.6}$	$4.1_{\pm 0.5}$	$10_{\pm 1}$	$5_{\pm 0}$
Dreamer V3	$9.7_{\pm 0.9}$	$1.9_{\pm 0.5}$	$44_{\pm 2}$	$40_{\pm 1}$	$7.2_{\pm 0.8}$	$0.0_{\pm 0.0}$	$45_{\pm 2}$	$39_{\pm 2}$	$9.9_{\pm 0.7}$	$\mathbf{3.7}_{\pm 0.3}$	$39_{\pm 3}$	$\mathbf{33}_{\pm 2}$
TD-MPC2	$13.8_{\pm 1.4}$	$0.7_{\pm 0.2}$	$32_{\pm 1}$	$16_{\pm 1}$	$9.5_{\pm 1.0}$	$2.2_{\pm 0.5}$	$40_{\pm 4}$	$32_{\pm 3}$	$9.2_{\pm 1.3}$	$5.5_{\pm 0.8}$	$35_{\pm 2}$	$25_{\pm 1}$
Ours w/o MPFP	$12.3_{\pm 1.2}$	$0.6_{\pm 0.2}$	$46_{\pm 3}$	$22_{\pm 3}$	$8.6_{\pm 1.1}$	$0.0_{\pm 0.0}$	$45_{\pm 2}$	$41_{\pm 1}$	$10.5_{\pm 1.6}$	$4.4_{\pm 0.7}$	$35_{\pm 3}$	$29_{\pm 2}$
Ours	$\mathbf{8.7}_{\pm 1.8}$	$\mathbf{0.5}_{\pm 0.1}$	$\mathbf{61}_{\pm 2}$	$\mathbf{42}_{\pm 1}$	$\mathbf{4.3}_{\pm 1.2}$	$\mathbf{0.0}_{\pm 0.0}$	$\mathbf{70}_{\pm 1}$	$\mathbf{60}_{\pm 1}$	$\mathbf{6.3}_{\pm 1.2}$	$6.9_{\pm 1.0}$	$\mathbf{45}_{\pm 3}$	$32_{\pm 3}$

Table 3: The models trained in the cross-room experimental configuration are transferred to three new experimental configurations for evaluation without additional training.

Ablation	AIKF↓	ABC↓	TCR↑	PSR↑
w/o All AES	$16.7_{\pm 1.3}$	$1.3_{\pm 0.5}$	$10_{\pm 2}$	$5_{\pm 1}$
w/o AES-p_{tc}	$10.1_{\pm 1.0}$	$0.9_{\pm 0.4}$	$51_{\pm 3}$	$39_{\pm 1}$
w/o AES-p_{ies}	$12.7_{\pm 1.1}$	$1.5_{\pm 0.5}$	$34_{\pm 4}$	$21_{\pm 2}$
w/o MPFP in train	$12.8_{\pm 1.4}$	$2.6_{\pm 0.5}$	$31_{\pm 1}$	$20_{\pm 1}$
w/o MPFP in eval	$7.13_{\pm 1.4}$	$\mathbf{0.8}_{\pm 0.7}$	$68_{\pm 4}$	$46_{\pm 3}$
Our Full Method	$\mathbf{4.19}_{\pm 1.2}$	$2.0_{\pm 0.2}$	$\mathbf{81}_{\pm 3}$	$\mathbf{54}_{\pm 2}$

Table 4: Ablation studies on components of AES and MPFP in the challenging cross-room experimental configuration.

pared to world model-based baselines, our approach is effective in reducing IK solving failures and collisions while improving task success. In addition, we find that RSSM-based MPFP is necessary because the spatial generalization of our method is significantly weaker when it is absent. The results in Tab. 2 and Tab. 3 also reflect that our method sometimes achieved sub-optimal ABC metrics. Such a phenomenon is especially prominent in dynamic scenes. Nevertheless, significant gains in the TCR metrics reflect the robustness and spatial generalization of our approach.

5.3 Ablation Studies

We answer questions (3) and (4) by respectively ablating different signals in AES and MPFP in the challenging cross-room experimental configuration. Based on the experimental results in Tab. 4, we find that the absence of all AES signals significantly degrades the MM performance, especially by increasing IK solving failures and decreasing the TCR and PSR metrics. In addition, the absence of both informative experience selection (AES-p_{ies}) and task criticality (AES-p_{tc}) impairs MM performance, with AES-p_{ies} having a greater impact. These results demonstrate the importance of AES for memorizing key experience fragments for robust MM. In addition, training curves in the supplementary material demonstrate that our AES-based MM policy is sample-efficient. By ablating the RSSM-based MPFP in training and evaluation, respectively, we find that it always helps to facilitate the collaboration between the mobile base and the manipulator through farsighted dynamic imagination, which improves the MM performance. More studies on parameter H and computational efficiency can be found in Section 3 of the supplementary material.

5.4 Real-World Experiments

We deploy our method to a real robotic platform consisting of a Husky A200 mobile base and a UR5 manipulator to verify its practicality. The robot's upper computer is a laptop with an Intel Core i9-13900HX CPU and GeForce RTX 4060 GPU. The robot uses a RealSense D435 RGB-D camera and YOLO v7 [Wang *et al.*, 2023] to identify target objects. We use ROS to organize the hardware and software resources of the robotic system. Considering the security, we set the maximum speed and decision frequency of the mobile base to 0.3 m/s and 1 Hz, respectively. Please see Section 4 of the supplementary material and videos for sim-to-real experiments.

6 Conclusion and Limitations

This paper investigates the problems of low sample efficiency and poor spatial generalization faced by MM in the context of embodied AI. Correspondingly, we propose the AES mechanism and RSSM-based MPFP to address these challenges. We find that AES, especially informative experience selection, can help RSSM memorize key experience fragments to significantly reinforce MM agents. RSSM-based prospective dynamic imagination can better generalize MM policies to new spatial layouts without additional training. Our problem modeling and experiments are carried out in an EE-centric MM framework, which fully combines the advantages of traditional robotic motion planning and robot learning. The practicality of our approach is verified by migrating it to a real-world robotic system without additional training.

Limitations. To focus on the contribution of the AES-enhanced RSSM to MM, we assumed an ideal observation space, especially using ground truth local occupancy maps directly. In practice, the construction of low-noise local occupancy maps from sensors is essential.

Acknowledgments

This work was supported in part by the Key Project of Xiangjiang Laboratory under 25XJ02003 and in part by the National Natural Science Foundation of China under 62272489, 62332020, and 62350004. This work was carried out in part using computing resources at the High-Performance Computing Center of Central South University.

References

[Arduengo *et al.*, 2021] Miguel Arduengo, Carme Torras, and Luis Sentis. Robust and adaptive door operation with a mobile robot. *Intelligent Service Robotics*, 14(3):409–425, 2021.

[Breunig *et al.*, 2000] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[Cen *et al.*, 2025] Jun Cen, Siteng Huang, Yuqian Yuan, Kehan Li, Hangjie Yuan, Chaohui Yu, Yuming Jiang, Jiayan Guo, Xin Li, Hao Luo, et al. Rynnvla-002: A unified vision-language-action and world model. *arXiv preprint arXiv:2511.17502*, 2025.

[Chen *et al.*, 2023] Bolei Chen, Jiaxu Kang, Ping Zhong, Yongzheng Cui, Siyi Lu, Yixiong Liang, and Jianxin Wang. Think holistically, act down-to-earth: A semantic navigation strategy with continuous environmental representation and multi-step forward planning. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5):3860–3875, 2023.

[Chen *et al.*, 2024] Bolei Chen, Jiaxu Kang, Ping Zhong, Yixiong Liang, Yu Sheng, and Jianxin Wang. Embodied contrastive learning with geometric consistency and behavioral awareness for object navigation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4776–4785, 2024.

[Chen *et al.*, 2025] Sixiang Chen, Jiaming Liu, Siyuan Qian, Han Jiang, Lily Li, Renrui Zhang, Zhuoyang Liu, Chenyang Gu, Chengkai Hou, Pengwei Wang, et al. Ac-dit: Adaptive coordination diffusion transformer for mobile manipulation. *arXiv preprint arXiv:2507.01961*, 2025.

[Fang *et al.*, 2022] Kuan Fang, Toki Migimatsu, Ajay Mandlekar, Li Fei-Fei, and Jeannette Bohg. Active task randomization: Learning visuomotor skills for sequential manipulation by proposing feasible and novel tasks. *CoRR*, 2022.

[Feng *et al.*, 2025] Xiaoxu Feng, Takato Horii, and Takayuki Nagai. Predictive reachability for embodiment selection in mobile manipulation behaviors. *IEEE Robotics and Automation Letters*, 2025.

[Gu *et al.*, 2022] Jiayuan Gu, Devendra Singh Chaplot, Hao Su, and Jitendra Malik. Multi-skill mobile manipulation for object rearrangement. *arXiv preprint arXiv:2209.02778*, 2022.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[Hafner *et al.*, 2025] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pages 1–7, 2025.

[Hansen *et al.*, 2024] Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations (ICLR)*, 2024.

[Harish *et al.*, 2024] Abhinav Narayan Harish, Larry Heck, Josiah P Hanna, Zsolt Kira, and Andrew Szot. Reinforcement learning via auxiliary task distillation. In *European Conference on Computer Vision*, pages 214–230. Springer, 2024.

[He *et al.*, 2025] Guanqi He, Xiaofeng Guo, Luyi Tang, Yuanhang Zhang, Mohammadreza Mousaei, Jiahe Xu, Junyi Geng, Sebastian Scherer, and Guanya Shi. Flying hand: End-effector-centric framework for versatile aerial manipulation teleoperation and policy learning. *arXiv preprint arXiv:2504.10334*, 2025.

[Honerkamp *et al.*, 2023a] Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. N2m2: Learning navigation for arbitrary mobile manipulation motions in unseen and dynamic environments. *IEEE Transactions on Robotics*, 39(5):3601–3619, 2023.

[Honerkamp *et al.*, 2023b] Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. N^2m^2: Learning navigation for arbitrary mobile manipulation motions in unseen and dynamic environments. *IEEE Transactions on Robotics*, 2023.

[Huang *et al.*, 2023] Xiaoyu Huang, Dhruv Batra, Akshara Rai, and Andrew Szot. Skill transformer: A monolithic policy for mobile manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10852–10862, 2023.

[Jauhri *et al.*, 2022] Snehal Jauhri, Jan Peters, and Georgia Chalvatzaki. Robot learning of mobile manipulation with reachability behavior priors. *IEEE Robotics and Automation Letters*, 7(3):8399–8406, 2022.

[Jauhri *et al.*, 2024] Snehal Jauhri, Sophie Lueth, and Georgia Chalvatzaki. Active-perceptive motion generation for mobile manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1413–1419. IEEE, 2024.

[Kang *et al.*, 2024] Jiaxu Kang, Bolei Chen, Ping Zhong, Haonan Yang, Yu Sheng, and Jianxin Wang. Hspnav: Hierarchical scene prior learning for visual semantic navigation towards real settings. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9434–9440, 2024.

[Lin *et al.*, 2025a] Min Lin, Xiwen Liang, Bingqian Lin, Liu Jingzhi, Zijian Jiao, Kehan Li, Yuhan Ma, Yuecheng Liu, Shen Zhao, Yuzheng Zhuang, et al. Echovla: Robotic vision-language-action model with synergistic declarative memory for mobile manipulation. *arXiv preprint arXiv:2511.18112*, 2025.

[Lin *et al.*, 2025b] Zhenyang Lin, Yurou Chen, and Zhiyong Liu. Autoskill: Hierarchical open-ended skill acquisition for long-horizon manipulation tasks via language-modulated rewards. *IEEE Transactions on Cognitive and Developmental Systems*, 2025.

[Liu *et al.*, 2025] Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025.

[Nguyen and La, 2019] Hai Nguyen and Hung La. Review of deep reinforcement learning for robot manipulation. In *2019 Third IEEE international conference on robotic computing (IRC)*, pages 590–595. IEEE, 2019.

[Rubinstein, 1999] Reuven Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and computing in applied probability*, 1(2):127–190, 1999.

[Schaul *et al.*, 2015] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

[Sun *et al.*, 2023] Weinan Sun, Madhu Advani, Nelson Spruston, Andrew Saxe, and James E Fitzgerald. Organizing memories for generalization in complementary learning systems. *Nature neuroscience*, 26(8):1438–1448, 2023.

[Szot *et al.*, 2022] Andrew Szot, Karmesh Yadav, Alex Clegg, Vincent-Pierre Berges, Aaron Gokaslan, Angel Chang, Manolis Savva, Zsolt Kira, and Dhruv Batra. Habitat rearrangement challenge 2022, 2022.

[Tafazoli *et al.*, 2025] Sina Tafazoli, Flora M Bouchacourt, Adel Ardalan, Nikola T Markov, Motoaki Uchimura, Marcelo G Mattar, Nathaniel D Daw, and Timothy J Buschman. Building compositional tasks with shared neural subspaces. *Nature*, pages 1–9, 2025.

[Vahrenkamp *et al.*, 2013] Nikolaus Vahrenkamp, Tamim Asfour, and Rüdiger Dillmann. Robot placement based on reachability inversion. In *2013 IEEE International Conference on Robotics and Automation*, pages 1970–1975. IEEE, 2013.

[Wang *et al.*, 2023] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023.

[Wang *et al.*, 2024] Can Wang, Jiaheng Zhang, Aoqi Wang, Zhen Wang, Nan Yang, Zhuoli Zhao, Chun Sing Lai, and Loi Lei Lai. Prioritized sum-tree experience replay td3 drl-based online energy management of a residential microgrid. *Applied Energy*, 368:123471, 2024.

[Wen *et al.*, 2025] Yuqing Wen, Hebei Li, Kefan Gu, Yucheng Zhao, Tiancai Wang, and Xiaoyan Sun. Lladavla: Vision language diffusion action models. *arXiv preprint arXiv:2509.06932*, 2025.

[Zhang *et al.*, 2025] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, Fan Lu, He Wang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *arXiv preprint arXiv:2507.04447*, 2025.