

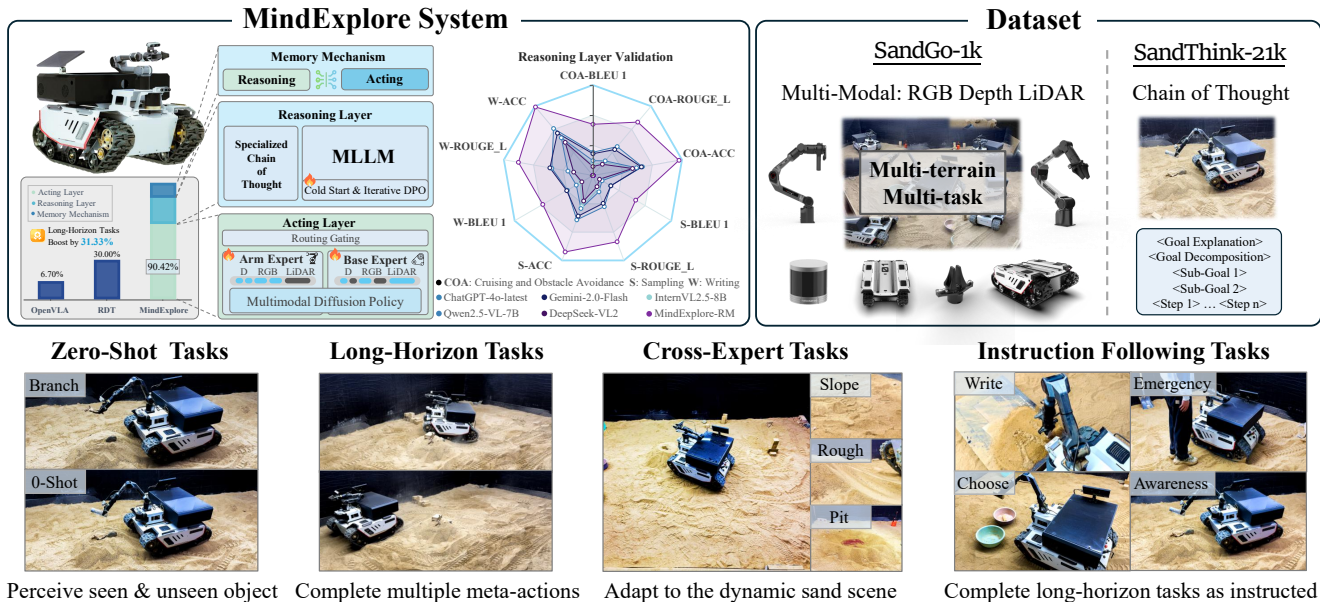
Towards Long-Horizon Vision-Language-Action System: Reasoning, Acting and Memory

Daixun Li^{1,*} Yusi Zhang^{1,*} Mingxiang Cao^{1,*} Donglai Liu^{1,*} Weiyang Xie^{1,†} Tianlin Hui¹

Lunkai Lin² Zhiqiang Xie² Yunsong Li¹

¹Xidian University ²AgileX Robotics

<https://xdu-imagelab.github.io/MindExplore/>



Perceive seen & unseen object Complete multiple meta-actions Adapt to the dynamic sand scene Complete long-horizon tasks as instructed

Figure 1. **Overview of MindExplore System:** A hierarchical embodied intelligence system including reasoning, acting, and memory, offering state-of-the-art generalizability in highly dynamic scenes. Unlike conventional planar embodied intelligence models that consolidate all capabilities within a single layer, MindExplore introduces a novel expert-level hierarchical embodied system architecture.

Abstract

Vision-Language-Action (VLA) is crucial for autonomous decision-making in embodied systems. While current methods have advanced single-skill abilities, their short-horizon capability limits applicability in real-world scenarios. To address this challenge, we innovatively propose **MindExplore**, a general hierarchical VLA system with cross-skill for long-horizon tasks in highly dynamic sand. The key insight is to iteratively align the knowledge domain of task planning and action execution. Thus, this task-oriented action enables outstanding generalization across a wide range of real-world scenarios. In the reasoning layer, task-specific chains of thought (CoT) are designed for planning long-horizon task sequences and providing meta-action signals. In the acting layer, a simple but powerful Mixture of Pol-

icy Experts strategy is built inspired by signals and multi-modal inputs for adaptively selecting skill experts and generating closed-loop action sequences. Also, it integrates a lightweight Multimodal Diffusion Policy (MMDP) to enhance spatial perception by fusing multi-visual modality features. Besides, the pioneering memory mechanism establishes feedback between the reasoning and acting layers, facilitating adaptive execution of long-horizon tasks and real-time replanning. Notably, we create **SandGo-1k** and **SandThink-21k**, the first expert-level multimodal embodied dataset and CoT dataset tailored for sandy environments. At a high execution frequency of 30 FPS, MindExplore is $3.01 \times$ more successful than existing methods in unstructured and dynamic environments.

* Equal contribution. † Corresponding authors.

1. Introduction

Current Visual-Language-Action (VLA) data and methods mainly focus on short sequence tasks of single ability [6, 20], that is, single-skill sequences limited to either arm operations or base movements, making them effective for simple tasks requiring only “muscle memory”. [13, 20, 24]. However, these approaches fail to generalize for long-horizon tasks in real-world scenarios, particularly in complex terrain action [14, 33]. This requires necessitates the agent to handle long-horizon and multi-skill instructions in highly dynamic environments [37]. Thus, successful execution in such settings needs deep consideration—Reasoning, precise control—Acting, and adaptive feedback—Memory [5].

To achieve this, three key challenges must be addressed: (1) Task-oriented Chain of Thought (CoT) reasoning requires the hierarchical decomposition of coarse-grained human instructions into executable sequences of meta-actions, ensuring long-horizon alignment in dynamic environments [10, 20, 23]; (2) High-Precision Multimodal Control through a lightweight spatial perception framework fuses RGB, depth, and LiDAR inputs for robust motion calibration and manipulation accuracy [17, 35]; (3) Adaptive State Tracking that dynamically memorizes system states and task progress, enabling real-time closed-loop feedback between planning and acting to mitigate error propagation from perceptual drift or control misalignment [1, 43]. Consequently, this remains an open and worthwhile challenge so far.

In this paper, we propose MindExplore for long-horizon tasks in complex and dynamic environments with a novel expert-level hierarchical embodied system architecture (see Figure 1). Our MindExplore is inspired by the discovery that continuous iterative alignment of the knowledge domain between task planning and action execution enables strong generalization across diverse real-world scenarios. On one hand we pursue task-oriented CoT in the Reasoning layer for effectively capturing meta-action sequences, while on the other hand we design a straightforward Mixture of Policy Experts (MoPE) strategy in the Acting Layer for flexibly generating closed-loop action sequences. Moreover, Memory Mechanism is designed to establish feedback between the reasoning and acting layers. Thus, the proposed MindExplore formulates long-horizon task execution into trainable skill expert policies, providing innovative guidance for hierarchical embodied intelligence systems without failures in highly dynamic environments. We also introduce a lightweight Multimodal Diffusion Policy (MMDP) that is dedicated to integrating RGB, depth and LiDAR data into a feature embedding space, thus enhancing spatial perception and motion control capabilities. To accurately evaluate complex tasks, we construct SandGo-1k and SandThink-21k as the first expert-level multimodal

embodied dataset and CoT dataset, respectively. We also carry out experiments to show that, the system maintains high accuracy and efficient execution even in highly dynamic environments such as mobile and unstructured sand, which validates our architecture. Our key contributions are summarized as follows:

- We propose **MindExplore**, a novel expert-level hierarchical embodied system to adapt long-horizon tasks in unstructured and dynamic environments. To our best knowledge, MindExplore is the first work to formulate Reasoning, Acting and Memory into a unified architecture.
- We create **SandGo-1k**, the first embodied intelligence dataset for mobile operations in complex environments. It includes text-vision (RGB, depth, and LiDAR) data covering complex movement episodes. Building on this, we further construct a multimodal CoT dataset for highly dynamic terrains, **SandThink-21k**.
- We empirically evaluate the accuracy on real-world experiments across diverse complex terrains with 24 distinct tasks, such as steep slopes, pits, and dunes. These experiments highlight MindExplore exceptional performance for long-horizon tasks in real-world scenarios.

2. Related Work

2.1. Visual Chain-of-Thought

CoT reasoning [32] has become fundamental in MLLMs [40, 41], with applications in science [21], mathematics [22], and advanced question answering [36]. Recently, visual CoT has been applied to robotics [19, 29, 37] to tackle task-specific challenges. However, existing methods focus on single-instruction, single-step tasks, failing in complex scenarios that demand long-horizon task decomposition. We argue that orchestrating concise instructions into structured long-horizon tasks is a critical step toward autonomous robot decision-making.

2.2. Embodied System

The long-horizon tasks in open-world environments has been a major challenge in robotics research [9]. Numerous studies leverage MLLMs with extensive world knowledge to decompose instructions into subtasks, which are executed by lightweight VLA models [4, 16, 24, 28]. The key challenges can be summarized as follows:

Challenge 1: How can instructions be expanded into long-horizon action sequences and executed flawlessly?

Challenge 2: How to effectively accomplish movement and manipulation in complex and dynamic terrain?

Challenge 3: How to overcome the data missing caused by the difficulty of simulation of highly dynamic scenes?

3. The System Dataset

To address Challenge 3, we collect a multi-task embodied operation dataset **SandGo-1k**, using a tracked mobile robot with a single-arm manipulator (as shown in Figure 2) in a real sand. Based on SandGo-1k, we generate a reasoning-acting embodied instruction dataset **SandThink-21k**.

3.1. Embodied Dataset: SandGo-1k

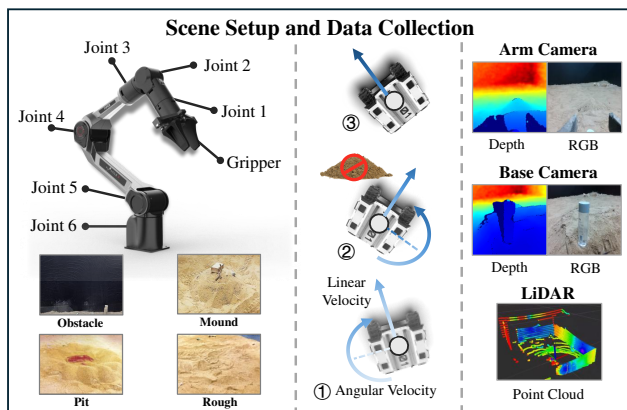


Figure 2. **Left:** AgileX Robotic arm degrees of freedom and environmental layout (rugged terrain, obstacles, pits, dunes). **Middle:** Base movement states. **Right:** Collected data content.

Environmental Layout. In real-world tasks [26, 30, 31, 42], uneven sandy terrains dominate outdoor applications, making them essential for studying mobile manipulation systems. The high dynamic fluidity of sand demands adaptive adjustments, while its visual noise complicates perception and calibration. Enhancing generalization in unstructured environments remains a key challenge, and our experiments show that a system robust in sandy conditions can more easily generalize to simpler terrains like factory floors, asphalt roads, or regular soil. Therefore, to investigate the capabilities of mobile manipulation systems in highly dynamic and complex environments, we select sand as the primary experimental setting, including potholes, steep slopes, mounds, and rough roads, and construct a 6×8 meters real-world sandbox for data collection and evaluation, as depicted in Figure 2.

Task Setup and Collection Process. We design three long-horizon meta-tasks: (1) Cruising and Obstacle Avoidance, training the model for environmental perception and motion planning in complex terrains; (2) Sampling, enabling the model to perform instruction-following, target localization, and coordinated base-arm movement; (3) Writing, enhancing the model’s fine manipulation capabilities. Based on these, we can decompose and reassemble them into five meta-actions: (1) Moving to a Target Location, (2) Crossing Obstacles, (3) Grasping a Specified Object, (4) Placing the Object into a designated Container, and (5)

Writing. During data collection, task instructions are manually written and decomposed into sequences of meta-actions with individual instructions. Each meta-action is saved as an episode, serving as the smallest unit of data collection.

Data Content. The collected data includes: (1) Sensor information, consisting of RGB-Depth images from the two cameras and LiDAR point clouds; (2) Robot state and manipulation commands, including the arm’s pose and velocity, as well as the base’s angular and linear velocities; (3) Text instructions for long-horizon tasks and meta-actions.

3.2. CoT Dataset: SandThink-21k

Effective robotic task execution requires goal comprehension, subgoal decomposition, and precise action sequencing. However, existing methods often lack structured reasoning mechanisms for long-horizon task decisions. To address this, we introduce a CoT reasoning dataset that enables deep reasoning before action generation, transforming concise instructions into structured long-horizon tasks for optimal execution. Our dataset follows a four-stage structure: (1) Goal Explanation; (2) Subgoal Decomposition; (3) Action Sequence Generation; (4) Action Sequence Modeling. By explicitly structuring these stages, our dataset mitigates fragmented planning and enhances decision-making. To improve model comprehension, each stage is marked with dedicated tags (e.g., `<Goal Explanation>`, `...</Goal Explanation>`). Except for the final action sequence, intermediate responses are generated using GPT-4o, with task instructions rewritten for dataset expansion. Additionally, we incorporate 5,700 public complex surface images to leverage their similarity to sand terrain. The final dataset, SandThink-21k, consists of 21k CoT samples for model training.

4. The MindExplore System

4.1. Reasoning Layer

To enhance reasoning abilities for long-horizon tasks, we employ a two-stage training strategy consisting of cold start followed by iterative direct preference optimization [7].

4.1.1. Cold Start with Curriculum Learning

To promote structured reasoning in complex robotic environments, we employ SigLIP to match text instructions with collected images of Sandgo-21k in feature space, filtering out blurred image data caused by jitter, and finally cold-start training all linear layers of the Qwen2-VL-7B-Instruct model with LoRA. The training goal at this stage is to maximize the probability of generating an reasoning result given the image and instruction input. Let X_v denotes the image input and X_q denote the

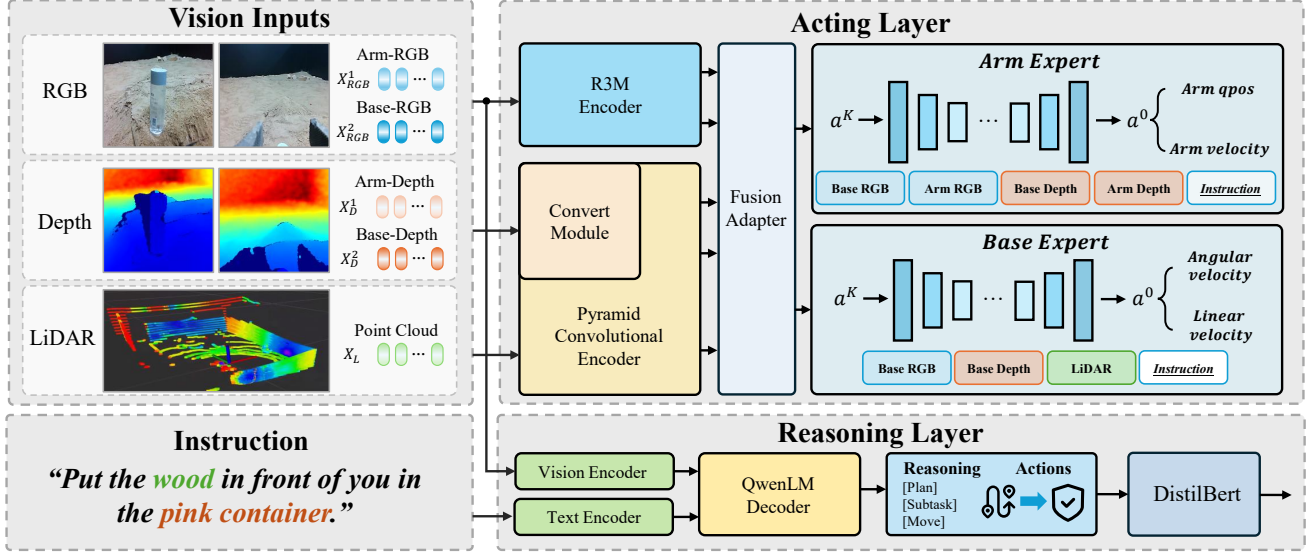


Figure 3. **MindExplore System Overview.** MindExplore comprises a reasoning layer for task decomposition and an acting layer for execution. The **Reasoning Layer**, built on Qwen2-VL, applying Chain-of-Thought (CoT) reasoning to generate structured meta-action instructions. The acting Layer, a Mixture of Policy Experts (MoPE), includes two Multimodal Diffusion Policies (MMDPs): a base expert using RGB-Depth and LiDAR for movement, and an arm expert using RGB-Depth for manipulation. The Memory Mechanism ensures real-time feedback, enabling the reasoning layer to dynamically adjust the actions of the acting layer for optimal execution.

instruction input. The objective formula is as follows:

$$L_{cold-start} = \sum_{i=1}^L \log P(X_{a,i} | X_v, X_q, X_{a,<i}), \quad (1)$$

where X_r denotes the reasoning output and L is its length. To effectively guide training, we employ a curriculum learning approach [3], where data is arranged in order from easy to difficult. Initially, the model is trained on structured complex surface data, followed by real-world sand scenes. This progression ensures that the model first learns basic reasoning skills before transitioning to more complex tasks. By gradually increasing the task complexity, the model acquires structured reasoning capabilities and improves its ability to handle real-world uncertainty.

4.1.2. Enhancing Reasoning with RL

While cold start establishes structured reasoning, reinforcement learning is used to further refine the model’s reasoning quality. We employ Iterative DPO to optimize the model’s decisions in multimodal scenarios. Given an image I and a task instruction Q , the model generates an inference chain R . DPO is based on the Bradley-Terry (BT) model:

$$P(R^+ > R^- | I, Q) = \frac{\exp(r_\varphi(I, Q, R^+))}{\exp(r_\varphi(I, Q, R^+)) + \exp(r_\varphi(I, Q, R^-))}. \quad (2)$$

r_φ is a parameterized reward model, R^+ and R^- denote preferred and non-preferred reasoning sequences.

The policy model π_θ is trained to maximize the preferred alignment inference, while reference model π_{ref} is used to stabilize training process. The initial objective function is:

$$\max_{\pi_\theta} \mathbb{E}_{(I, Q, R^+, R^-)} [r_\varphi(I, Q, R)] - \beta D_{KL}(\pi_\theta(R|I, Q) \| \pi_{ref}(R|I, Q)). \quad (3)$$

To stabilize training, we also incorporate a standard supervised fine-tuning loss, ensuring that the model retains the structured reasoning of supervised fine-tuning while benefiting from preference-based optimization. The loss function is given by the following formula:

$$L_{DPO}(\theta) = -\mathbb{E}_{(I, Q, R^+, R^-)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(R^+ | I, Q)}{\pi_{ref}(R^+ | I, Q)} - \beta \log \frac{\pi_\theta(R^- | I, Q)}{\pi_{ref}(R^- | I, Q)} \right) + \alpha \frac{\log \pi_\theta(R^+ | I, Q)}{|R^+|} \right]. \quad (4)$$

where θ denotes the trainable parameters of the model, $|R^+|$ denotes preferred reasoning length, the hyperparameter α balances the two loss terms and β is a scaling factor.

To build high-quality preference data, we refine the selection process in 3 rounds of training. In each iteration, we sample 8 inference outputs for each task and rank them using GPT-4o with preference scores ranging from 0 to 100. The selection follows a progressive rejection strategy:

- **Iteration 1:** The highest scoring response is selected as R^+ , while the lowest scoring response is taken as R^- .
- **Iteration 2:** R^- is selected as the highest ranked response below 60, increasing the task complexity.
- **Iteration 3:** R^- is the lowest ranked response above 60, making model learn fine-grained reasoning distinctions.

This approach enables models to excel in structured multimodal reasoning and long-horizon robotic planning.

4.2. Acting Layer

The Acting Layer generates a set of manipulation parameters and focuses on optimizing the joint probability distribution of all outputs. Therefore, generative diffusion models excelling in expressive capability and sampling quality [12] can be applied. Since the data dimensions of manipulation parameters is much smaller than that of image, the acting layer could get fast generation speed to provide real-time feedback to dynamic changes [18, 38, 39].

4.2.1. Multimodal Diffusion Policy

Most VLAs use only single-modal sensor data, which struggles to provide precise manipulation for instruction following in complex environments. In challenging visual settings, RGB images often contain significant noise, such as the similarity in color between the cube in Figure 2 and the sandy terrain. On the other hand, models based solely on depth or point cloud data lack the ability to perceive scene details and cannot align with the semantic features of instructions. To address these, we propose a multimodal diffusion Policy model (MMDP) that jointly exploits information of multi-sensor data and textual instructions.

We employ the robot manipulation visual encoder R3M, proposed in [25], to encode the RGB images. Instead of concatenating the depth and RGB images along the channel dimension, we transform them into point cloud data using a perspective projection model:

$$X(u, v) = \frac{(u - c_x) \cdot D(u, v)}{f_x}, \quad (5)$$

$$Y(u, v) = \frac{(v - c_y) \cdot D(u, v)}{f_y}, \quad (6)$$

$$Z(u, v) = D(u, v), \quad (7)$$

where u and v are the pixel coordinates in the depth image, $D(u, v)$ is the depth value of the pixel, f_x and f_y are the camera’s focal lengths, c_x and c_y are the image’s principal point coordinates, and (X, Y, Z) represents the 3D coordinates of the point cloud data. We use the pyramid convolutional encoder from iDP3 [38] to encode point cloud data, while DistilBERT [27] is employed to encode textual instructions into feature vectors. Additionally, we propose a fusion adapters for both the sensor data and text, consisting of two linear layers, to map encoder outputs into a unified feature space.

The manipulation parameter generation process of the MMDP is based on the conditional denoising diffusion generation of a lightweight 1D convolutional diffusion model. Specifically, we sample a random Gaussian noise a^K and perform k iterations of denoising using a conditional denoising network ϵ_θ , conditioned on RGB data r , depth information d , LiDAR data p , instruction data i , and the robot’s

state q , to recover a^K into the manipulation parameters a^0 :

$$a^{k-1} = \alpha_k (a^k - \gamma_k \epsilon_\theta (a^k, k, r, d, p, i, q)) + \sigma_k \mathcal{N}(0, I), \quad (8)$$

where $\mathcal{N}(0, I)$ is Gaussian noise, α_k , γ_k , and σ_k are functions of k and depend on the noise scheduler [12]. During training, we iteratively add random noise ϵ^k for k steps to the manipulation parameters, resulting in the noisy action $\bar{\alpha}_k a^0 + \bar{\beta}_k \epsilon^k$, where $\bar{\alpha}_k$ and $\bar{\beta}_k$ are the noise schedule parameters that performs one step noise adding. The denoising network ϵ_θ is then used to predict the random noise ϵ^k , and the training loss function is as follows:

$$\mathcal{L} = \text{MSE} (\epsilon^k, \epsilon_\theta(\bar{\alpha}_k a^0 + \bar{\beta}_k \epsilon^k, k, r, d, p, i, q)). \quad (9)$$

4.2.2. Mixture of Policy Experts

The acting layer needs to provide precise feedback to dynamic environmental changes with relatively low model complexity, we argue that a unified manipulation model and unfiltered sensor inputs may interfere with the acting layer’s accurate manipulation of specific meta-actions. Therefore, we propose a Mixture of Policy Experts Model that decouples manipulation of the robot’s base and arm. The base expert is primarily responsible for executing meta-actions related to movement, outputting only the linear and angular velocities of base. Multimodal inputs are filtered to include RGB images from the base camera, base point cloud data with depth values less than 1.5m, and LiDAR point clouds within a horizontal radius of 1.2m. The arm expert performs meta-actions for grasping and placing, outputting only the pose and velocity parameters for arm manipulation. Its sensor inputs are RGB images from both the base and arm cameras, base point cloud data with depth values less than 0.8m, and arm point clouds with depth values less than 0.5m.

4.3. Memory Mechanism

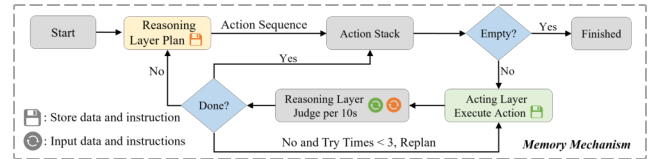


Figure 4. The details of Memory Mechanism

To enhance the coordination between the Reasoning Layer and the Acting Layer and achieve global control over both software and hardware, we design a memory mechanism with real-time feedback and gated routing. As shown in Figure 4, upon receiving a human instruction, the memory mechanism first processes multi-view RGB images and task instructions, which are fed into reasoning layer to generate a sequence of meta-action instructions. The gated routing network then sequentially assigns each meta-action to the corresponding VLA action expert, while real-time

Table 1. The success rates of different VLA models across 24 tasks in four long-horizon tasks.

Coarse-Grained Long-Horizon Tasks								Intensive Instruction-Following Tasks						
Method	Task1	Task2	Task3	Task4	Task5	Task6	Avg. (%)	Task7	Task8	Task9	Task10	Task11	Task12	Avg. (%)
OpenVLA	0/10	0/10	3/10	0/10	2/10	1/10	10.00%	1/10	3/10	0/10	0/10	1/10	2/10	11.67
RDT	5/10	3/10	0/10	3/10	6/10	5/10	36.67%	5/10	3/10	2/10	0/10	5/10	5/10	33.33
MindExplore	10/10	8/10	8/10	7/10	10/10	10/10	88.33	10/10	9/10	10/10	8/10	9/10	10/10	93.33
Cross-Expert Tasks								Zero-Shot Generalization Tasks						
Method	Task13	Task14	Task15	Task16	Task17	Task18	Avg. (%)	Task19	Task20	Task21	Task22	Task23	Task24	Avg. (%)
OpenVLA	0/10	0/10	1/10	0/10	0/10	0/10	1.67%	0/10	0/10	0/10	2/10	0/10	0/10	3.33
RDT	4/10	3/10	5/10	1/10	2/10	6/10	35.00%	2/10	1/10	3/10	0/10	0/10	3/10	15.00
MindExplore	9/10	10/10	10/10	9/10	9/10	9/10	93.33	10/10	8/10	8/10	9/10	7/10	10/10	86.67

multimodal sensor data is provided to the corresponding acting layer’s action expert to generate control parameters, which are subsequently used to execute actions via the hardware. The gated routing network can be formatted as:

$$w_{AE}, w_{BE} = \text{softmax}(\text{FFN}(\text{DistillBert}(mi))), \quad (10)$$

$$MoPE(mi) = \begin{cases} f_{AE}(mi, I_{AE}), & \text{if } w_{AE} > w_{BE} \\ f_{BE}(mi, I_{BE}), & \text{if } w_{AE} \leq w_{BE} \end{cases}, \quad (11)$$

where mi represents a single meta-action instruction, while I_{AE} and I_{BE} denote the multimodal sensor inputs required by the arm expert and base expert, respectively. During this process, the memory mechanism stores the initial multi-view RGB images received by reasoning layer and continuously feeds the current state of the robot back to Acting Layer while it generates control parameters, enabling real-time adjustments. Additionally, the memory mechanism supplies reasoning layer with both current and historical RGB sensor images, equipping reasoning layer with multi-temporal environmental perception to determine the completion status of each meta-action. If a meta-action is successfully executed, the memory mechanism forwards the next action to acting layer. If an error occurs during execution, it triggers reasoning layer to replan the entire task, generating a new sequence of meta actions, thereby improving execution stability and robustness.

5. Experiments

5.1. Implementation Details

We train the Reasoning Layer with SandThink-21k and the acting layer using individual meta-actions from the SandGo-1k, with each meta-action representing an episode. Movement and crossing obstacle are used to train the base expert, while grasping, placing, and writing are used to train the arm expert. During training, we randomly apply GPT-4o for instruction augmentation to simulate task instructions with varying levels of granularity. The SAM model [15] is used for semantic segmentation of RGB images, replacing the black background with different colors for data aug-

mentation. We evaluate the MindExplore system directly in the same sandpit scenario, using success rates as the metric, which is the ratio of successful trials to the total number of experiments. The tests are conducted with varying scene configurations and ensures that all comparison models are tested under the same conditions.

5.2. System Performance

5.2.1. Performance on Real World Tasks

We evaluate 24 real-world robotic tasks categorized into four types, each containing six distinct tasks. All these evaluated tasks can be divided into four categories according to their ability to complete long-horizon tasks, and we conducted 615 trials on real robots, ensuring its accuracy on meta-action and long-horizon tasks.

Coarse-Grained Long-Horizon Tasks. These tasks involve high-level, ambiguous instructions (e.g., “grab the branch”), requiring the system to first interpret and structure precise task sequences before execution. The robot must break down vague commands into actionable steps, Such as object grasping after obstacle avoidance (Task 1, 2), going to position writing (Task 3, 4), and navigating challenging terrain to complete sampling (Task 5, 6).

Intensive Instruction Following Tasks. These tasks consist of sequential, interdependent sub-tasks, testing the system’s compositional reasoning and execution fluency. The robot must understand and follow multi-step instructions with precision, such as continuous pick-and-place operations are performed in a dynamically changing environment (Task 7-9), and sampling is performed after multiple obstacle avoidance (Task 10-12).

Cross-Expert Tasks. These tasks demand seamless coordination between base movement and arm manipulation, requiring the integration of multiple operational skills (e.g., moving, turning, grasping, and placing) through cross-modal reasoning. The robot must reason about motion and manipulation tasks in combination, such as grasping a specified object to place in a container (Task 13-15) and writing at a specified location (Task 16-18).

Zero-Shot Generalization Tasks. These tasks evaluate

Instructions Follow Long-horizon Tasks

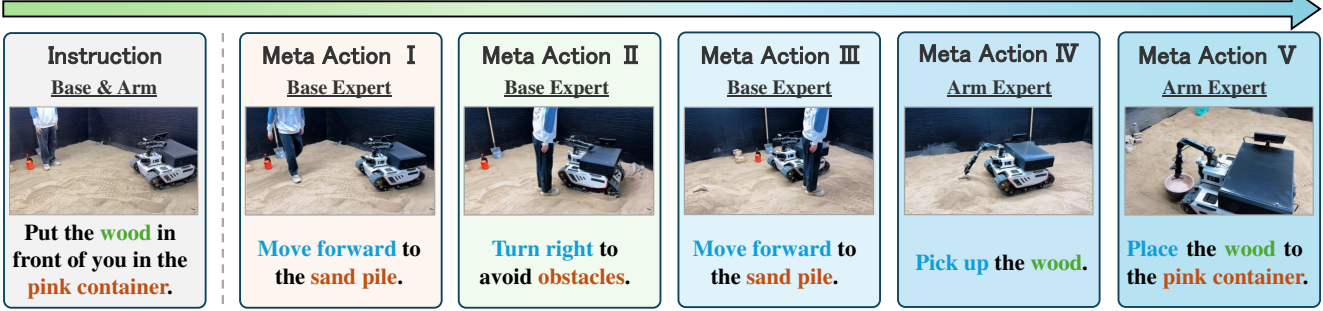


Figure 5. During long-horizon tasks, the reasoning layer decomposes the instruction into a structured sequence of executable steps. The acting layer then orchestrates the hierarchical execution by leveraging the **Base Expert** and **Arm Expert**.

Table 2. Comparison results with the most powerful closed-source and open-source MLLMs on three tasks.

Model	Coarse-Grained Long-Horizon Tasks		Intensive Instruction Following Tasks		Cross-Expert Multi-Tasks		The Zero-Shot Generalization Tasks	
	ROUGE.L (%)	Acc. (%)	ROUGE.L (%)	Acc. (%)	ROUGE.L (%)	Acc. (%)	ROUGE.L (%)	Acc. (%)
GPT-4o-latest	33.02	47.00	37.54	50.00	41.59	61.00	29.17	21.00
Gemini-2.0-Flash	29.52	43.00	33.69	49.00	43.53	55.00	24.36	18.00
InternVL2.5-8B	14.71	33.00	17.45	40.00	19.37	43.00	19.88	15.00
Qwen2.5-VL-7B	17.32	35.00	18.93	44.00	25.32	49.00	23.73	19.00
DeepSeek-VL2	11.84	27.00	14.61	40.00	19.49	43.00	7.18	5.00
Reasoning Layer	69.35	87.00	70.23	81.00	75.35	89.00	63.17	73.00

the ability of system to adapt previously unseen environmental terrain (Tasks 19 to 21) and object types (Tasks 22 to 24). The robot must generalize over different objects (branches, bottles, and stones), recognize and manipulate objects despite variations in texture, shape, and weight, and adjust to new ground conditions (rugged, steep slopes).

Table 3. Success Rates of the **Acting Layer** across 5 meta-actions: Moving to a Target Location (MTL), Crossing Obstacles (CO), Grasping a Specified Object (GSO), Placing the Object into a Designated Container (POC), Writing (Wr). Avg. - Average Success Rate, BE - Base Expert, AE - Arm Expert.

Method	MTL	CO	GSO	POC	Wr	Avg. (%)
ACT	13/100	3/65	37/100	55/60	19/50	36.86
iDP3	47/100	27/65	49/100	43/60	21/50	50.24
BE	85/100	58/65	-	-	-	87.12
AE	-	-	79/100	60/60	48/50	91.67

We select the OpenVLA [14] and the diffusion model RDT [18] as comparison methods. As shown in Table 1, in instruction-following, MindExplore’s performance improves as instruction granularity increases, whereas OpenVLA and RDT perform better with coarse-grained instructions than with intensive instructions. We attribute this to MindExplore’s strong reasoning capability, where more informative instructions guide the model to execute more precise actions, whereas OpenVLA and RDT tend to favor simple instruction descriptions and lack deep thoughts for long-horizon tasks. In multi-skill coordination, MindExplore outperforms single-expert models with a higher success rate due to the MoPE. In terms of generalization, compared to end-to-end MLLM and VLAs without world knowledge,

MindExplore demonstrates strong generalization capabilities in unseen scenarios. Notably, the average sequence length for the cruising and obstacle avoidance task is 3.12, for the sampling task is 4.3, and for the writing task is 2. We find that the success rates of OpenVLA and RDT decrease progressively with the increasing meta-action length across three tasks. In Figure 5, MindExplore maintains a consistent success rates across sequences of varying lengths, benefiting from its superior task planning, decomposition capabilities, and the memory mechanism.

5.2.2. Performance of Different Components

We also assess the performance of the two key components, reasoning and acting layer. For reasoning layer, we collect 480 RGB-instruction pairs across four test tasks as the evaluation dataset, and select the most powerful closed-source and open-source MLLMs for comparative experiments, including GPT-4o-latest, Gemini-2.0-Flash, InternVL2.5-8B [8], Qwen2.5-VL-7B [2], and DeepSeek-VL2 [34]. We use prompt to force all models to output a reasoning process similar to reasoning layer, and only focus on the output content of action sequence modeling, using ROUGE.L and meta-action planning accuracy to evaluate model performance. The experimental results are shown in Table 2. Reasoning layer performs best on all indicators and tasks.

For acting layer, we compare it with two typical manipulation models, the RGB-based ACT [11] and the point-cloud-based iDP3 [38]. Since both methods are not instruction-following models, a separate manipulation model must be trained for each meta-action. In contrast, the acting layer only requires training two expert models to

Table 4. Ablation results with training strategy on three tasks. “+” stands for accumulation.

Model	Coarse-grained Long-horizon tasks		Intensive instruction following tasks		Cross-Expert multi-tasks		The zero-shot Generalization task	
	ROUGE.L (%)	Acc. (%)	ROUGE.L (%)	Acc. (%)	ROUGE.L (%)	Acc. (%)	ROUGE.L (%)	Acc. (%)
Qwen2-VL-7B	13.54	29.00	15.22	35.00	23.17	37.00	7.31	19.00
+ Cold Start	47.35	62.00	51.42	68.00	64.57	75.00	33.16	43.00
+ iteration 1	54.18	69.00	54.54	70.00	67.15	79.00	35.34	55.00
+ iteration 2	61.58	75.00	63.29	79.00	70.23	81.00	46.57	64.00
+ iteration 3	69.35	87.00	70.23	81.00	75.35	89.00	63.17	73.00

Table 5. Ablation study of the MindExplore system on the cruising and obstacle avoidance task.

Method	Task1	Task2	Task7	Task8	Task13	Task14	Task19	Task20	Avg. (%)
w/o Reasoning & Memory	5/10	5/10	8/10	6/10	4/10	6/10	9/10	5/10	60.00
w/o Memory	8/10	7/10	10/10	7/10	5/10	8/10	9/10	7/10	76.25
MindExplore	10/10	8/10	10/10	9/10	9/10	10/10	10/10	8/10	92.50

Table 6. Ablation Study on Input Modalities for Meta-Actions. “PC” - Point Cloud, “2D” - 2D Images.

Method	Input Modality	Depth Type	MTL	GSO
AE	RGB	-	-	23/50
AE	Depth+RGB	PC	-	40/50
BE	RGB	-	5/50	-
BE	LiDAR+Depth	PC	37/50	-
BE	LiDAR+Depth+RGB	2D	32/50	-
BE	LiDAR+Depth+RGB	PC	42/50	-

control different hardware structures, which is parameter efficient. As shown in Table 3, the ACT performs worst on actions requiring long-distance movement, such as MTL and CO, indicating that point cloud with spatial information is crucial for movement manipulation. By combining spatial information with color and texture information from RGB images, the BE and AE achieve better alignment between object semantics and text instructions in real-world scenes, achieving the highest success rates across all meta-actions.

5.3. Ablation Study

Ablation results of Reasoning Layer. Compared to DeepSeek-VL2 and InternVL2.5, although they are on par with Qwen2-VL in some benchmark indicators, they lack scalability for real-world tasks in embodied systems, such as spatial calibration, semantic perception, and fine-grained object recognition in noisy backgrounds (e.g., inability to distinguish between sand and yellow branches). Therefore, we chose Qwen2-VL as the base model of the reasoning layer. To prove the effectiveness of the Reasoning Layer training method, we performed detailed ablation experiments on three tasks, which involved adding our training methods one by one to the basic model. As shown in Table 4, cold start enables the model to obtain structured reasoning capabilities. The three rounds of iterative DPO training further improve the model’s reasoning quality and achieve self-improvement by continuously building high-quality preference datasets.

Ablation of Acting Layer. We explore the impact of different input modalities on acting layer. As shown in Table 6, the performance of the AE using pure vision declined due to the lack of spatial information, and the RGB-based BE achieves just a 10% success rate in MTL. Similarly, the point-cloud-based BE also shows lower success rates, as point cloud data failed to align effectively with the semantic content of instructions. Additionally, we test concatenating Depth images with RGB images along the channel dimension. The results show that point cloud data in acting layer better leverages the spatial information from Depth images.

Ablation of MindExplore System. We conduct an ablation study on the core components of the MindExplore system. In Table 5, in the reasoning layer and acting layer system without the memory mechanism, Reasoning layer performs a single task decomposition and sequentially sends meta-action instructions to acting layer. However, without timely state feedback and error correction from acting layer, the system is highly susceptible to error accumulation, resulting in a decrease in success rate. Further, when the Reasoning layer component is removed and only the acting layer is used, the success rate on the cruising and obstacle avoidance task significantly drops due to the lack of deep reasoning and acting of the task instructions. Nonetheless, it still outperforms other VLAs, demonstrating that the MoPE structure can ensure accurate completion of base tasks.

6. Conclusion

We propose MindExplore, a hierarchical embodied system designed for long-horizon tasks, including Reasoning, Acting, and Memory Mechanism. By aligning planning with action, MindExplore achieves strong generalization in dynamic environments. Besides, we introduce SandGo-1k and SandThink-21k, the first expert-level multimodal embodied and CoT datasets, bridging the data gap in challenging terrains. Real-world evaluations demonstrate MindExplore achieves better performance, ensuring high reliability in complex environments.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant 62322117, 62371365, and U24B20136), the Fundamental Research Funds for the Central Universities, and the Innovation Fund of Xidian University (Grant No. YJSJ25007).

References

- [1] Muneeb Ahmed, Brejesh Lall, Rajesh Kumar, and Arzad A Kherani. Towards estimation of human intent in assistive robotic teleoperation using kinaesthetic and visual feedback. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1928–1934, 2023. 2
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 7
- [3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 41–48, 2009. 4
- [4] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π 0: A vision-language-action flow model for general robot control, 2024. *arXiv preprint arXiv:2410.24164*. 2
- [5] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π 0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 2
- [6] Shizhe Chen, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Sugar: Pre-training 3d visual representations for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18049–18060, 2024. 2
- [7] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024. 3
- [8] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 7
- [9] David B D’Ambrosio, Saminda Abeyruwan, Laura Graesser, Atıl İscen, Heni Ben Amor, Alex Bewley, Barney J Reed, Krista Reymann, Leila Takayama, Yuval Tassa, et al. Achieving human level competitive robot table tennis. *arXiv preprint arXiv:2408.03906*, 2024. 2
- [10] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3164–3174, 2020. 2
- [11] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024. 7
- [12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 5
- [13] Stephen James, Kentaro Wada, Tristan Laidlow, and Andrew J Davison. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13739–13748, 2022. 2
- [14] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 2, 7
- [15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 6
- [16] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18061–18070, 2024. 2
- [17] Haitao Lin, Yanwei Fu, and Xiangyang Xue. Pourit!: Weakly-supervised liquid perception from a single image for visual closed-loop robotic pouring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 241–251, 2023. 2
- [18] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024. 5, 7
- [19] Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, et al. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *arXiv preprint arXiv:2501.10074*, 2025. 2
- [20] Jingpei Lu, Florian Richter, and Michael C Yip. Markerless camera-to-robot pose estimation via self-supervised sim-to-real transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21296–21306, 2023. 2
- [21] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 2507–2521, 2022. 2

- [22] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 2
- [23] Xiao Ma, Sumit Patidar, Iain Haughton, and Stephen James. Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18081–18090, 2024. 2
- [24] Weixin Mao, Weiheng Zhong, Zhou Jiang, Dong Fang, Zhongyue Zhang, Zihan Lan, Fan Jia, Tiancai Wang, Haoqiang Fan, and Osamu Yoshie. Robomatrix: A skill-centric hierarchical framework for scalable robot task planning and execution in open-world. *arXiv preprint arXiv:2412.00171*, 2024. 2
- [25] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022. 5
- [26] Max Polzin, Qinghua Guan, and Josie Hughes. Robotic locomotion through active and passive morphological adaptation in extreme outdoor environments. *Science Robotics*, 10(99): eadp6419, 2025. 3
- [27] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 5
- [28] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2998–3009, 2023. 2
- [29] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospacial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. *arXiv preprint arXiv:2411.16537*, 2024. 2
- [30] Enrica Tricomi, Francesco Missiroli, Michele Xiloyannis, Nicola Lotti, Xiaohui Zhang, Marios Stefanakis, Maximilian Theisen, Jürgen Bauer, Clemens Becker, and Lorenzo Masia. Soft robotic shorts improve outdoor walking efficiency in older adults. *Nature Machine Intelligence*, 6(10):1145–1155, 2024. 3
- [31] Kasun Weerakoon, Adarsh Jagan Sathyamoorthy, Mohamed Elnoor, and Dinesh Manocha. Vapor: Legged robot navigation in unstructured outdoor environments using offline reinforcement learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10344–10350. IEEE, 2024. 3
- [32] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:24824–24837, 2022. 2
- [33] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2409.12514*, 2024. 2
- [34] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 7
- [35] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4737–4746, 2023. 2
- [36] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, 2024. 2
- [37] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024. 2
- [38] Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, Xue Bin Peng, and Jiajun Wu. Generalizable humanoid manipulation with improved 3d diffusion policies. *arXiv preprint arXiv:2410.10803*, 2024. 5, 7
- [39] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024. 5
- [40] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 2
- [41] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 2
- [42] Da Zhao, Haobo Luo, Yuxiao Tu, Chongxi Meng, and Tin Lun Lam. Snail-inspired robotic swarms: a hybrid connector drives collective adaptation in unstructured outdoor environments. *Nature Communications*, 15(1):3647, 2024. 3
- [43] Allan Zhou, Moo Jin Kim, Lirui Wang, Pete Florence, and Chelsea Finn. Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17907–17917, 2023. 2