

AnyBimanual: Transferring Unimanual Policy for General Bimanual Manipulation

Guanxing Lu^{1,2*}, Tengbo Yu^{1,2*}, Haoyuan Deng³, Season Si Chen^{1,2}, Yansong Tang^{1,2†}, Ziwei Wang³

*Equal contributor †Corresponding author

¹Tsinghua Shenzhen International Graduate School, ²Tsinghua University

³School of Electrical and Electronic Engineering, Nanyang Technological University

{lgx23@mails., ytb23@mails., season.chen@, tang.yansong@}sz.tsinghua.edu.cn

{E230112@e., ziwei.wang@}ntu.edu.sg

<https://anybimanual.github.io/>

Abstract

General-purpose bimanual manipulation is challenging due to high-dimensional action spaces and expensive data collection. In contrast, unimanual policy has recently demonstrated impressive generalizability across a wide range of tasks because of scaled model parameters and training data, which can provide sharable manipulation knowledge for bimanual systems. We propose a plug-and-play method named **AnyBimanual**, which transfers pre-trained unimanual policy to general bimanual manipulation policy with few bimanual demonstrations. Specifically, we first introduce a skill manager to dynamically schedule the skill representations discovered from pre-trained unimanual policy for bimanual manipulation tasks, which linearly combines skill primitives with task-oriented compensation to represent the bimanual manipulation instruction. To mitigate the observation discrepancy between unimanual and bimanual systems, we present a visual aligner to generate soft masks for visual embedding, which aims to align visual input of unimanual policy model for each arm with those during pretraining stage. **AnyBimanual** shows superiority on 12 simulated tasks from *RLBench2* with a sizable 17.33% improvement in success rate over previous methods. Experiments on 9 real-world tasks further verify its practicality with an average success rate of 84.62%.

1. Introduction

Bimanual systems play an important role in robotic manipulation due to the high capacity of completing diverse tasks in household service [69], robotic surgery [32], and component assembly in factories [9]. Compared to unimanual systems, bimanual systems enlarge the workspace and are able to handle more complex manipulation tasks by stabi-

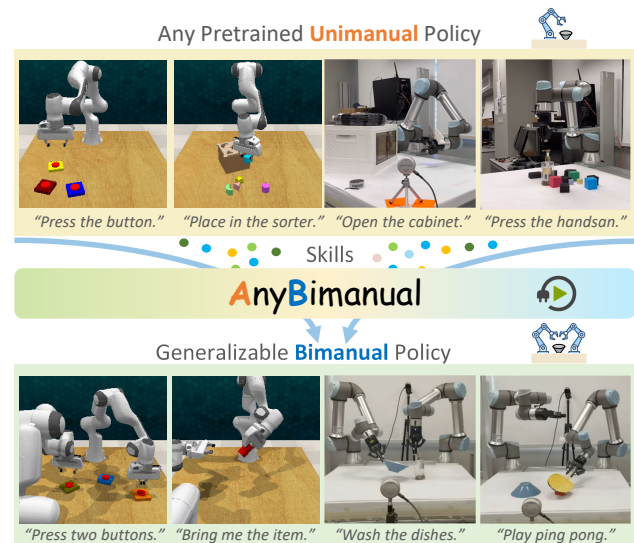


Figure 1. **AnyBimanual** enables plug-and-play transferring from pre-trained unimanual policies to bimanual manipulation policy, which preserves the generalizability with the proposed skill scheduling framework.

lizing the target with one arm and interacting with that using another arm [29, 44]. Even for the tasks that unimanual policies can handle, bimanual systems are often more efficient because multiple action steps can be simultaneously accomplished [30]. Since modern robotic applications require the robot to interact with different tasks and objects, it is desirable to design a generalizable policy model for bimanual manipulation.

To enhance the generalization ability of the manipulation agent, prior works present to leverage the high-level reasoning and semantic understanding capabilities of foundation models like Large Language Models (LLMs) and Vi-

sion Language Models (VLMs) to breakdown tasks into executable sub-tasks that can be solved by external low-level controllers [25, 26, 33, 36, 43], which thus struggles with contact-rich tasks that require complex and precise low-level motion. To generalize to contact-rich tasks, recent methods [1, 40, 50] tend to learn robotic foundation models directly from large-scale teleoperation data [17, 39, 54], which has shown impressive generalizability across various unimanual tasks. However, bimanual demonstrations are extremely expensive to acquire in real-world, which often need specialized teleoperation systems with additional sensors and fine-grained calibration with high human laborer cost [12, 18, 22, 55, 59, 70]. To address this challenge, recent methods aim to simplify the learning budget by exploiting human inductive bias like parameterized atomic movements that detail the position and rotation [5, 15] or assigning stabilizing and acting roles for each arm [24, 44, 53], thereby reducing the need for extensive expert data. Nevertheless, shareable atomic movements and fixed cooperation patterns struggle to generalize across different bimanual manipulation tasks, which limits the deployment scenario of these classes of methods.

In this paper, we propose a plug-and-play module named AnyBimanual that transfers any pre-trained unimanual policy to bimanual manipulation policy with limited demonstrations. Since unimanual policy model [38, 49] has demonstrated impressive generalization ability across tasks due to the large model sizes and numerous training demonstrations, we realize high generalizability across diverse language-conditioned bimanual manipulation tasks by mining and transferring the commonsense knowledge in pre-trained unimanual policies. More specifically, we first introduce a skill manager that dynamically schedules discovered skill representations. Skill representations are formed by skill primitives that store shareable manipulation knowledge across embodiments, with task-oriented importance weights and compensation. To enhance the transferability of the pretrained unimanual policy in bimanual manipulation tasks, the observation discrepancy between unimanual and bimanual systems should be minimized. We propose a voxel aligner to generate spatial soft masks to highlight relevant visual clues for different arms, whose goal is to align the visual input of the unimanual policy model for each arm with those during the pretraining stage. We evaluate AnyBimanual on a comprehensive task suite composed of 12 simulated tasks from RLBench2 [30] and 9 real-world tasks, where our method surpasses the previous state-of-the-art method by a large margin. The contributions are summarized as follows:

- We propose a model-agnostic plug-and-play framework named AnyBimanual that transfers an arbitrary pretrained unimanual policy to generalizable bimanual manipulation policy with limited bimanual demonstrations.

- We introduce a skill manager to dynamically schedule skill representations for unimanual policy transferring and a visual aligner to mitigate the observation discrepancy between unimanual and bimanual systems for transferability enhancement.
- We conduct extensive experiments of 12 tasks from RLBench2 and 9 tasks from the real world. The results demonstrate that our method achieves a higher success rate than the state-of-the-art methods.

2. Related Work

Generalizable Bimanual Manipulation. Bimanual manipulation agents [6, 11, 22, 24, 26, 30, 37, 44, 65, 68] are able to handle a large variety of tasks by predicting a trajectory of bimanual operation, which is of great significance in complex applications from household service [69], robotic surgery [32], to component assembly in factories [9]. To achieve multi-task learning for generalizable Bimanual manipulation, earlier studies attempted to leverage the emerged general understanding and reasoning capacities of pretrained foundation models like LLMs [52] and VLMs [10], where the foundation model was prompted to generate a high-level plan for low-level executors. However, the performance of directly leveraging foundation models in a training-free manner is bottlenecked by the predefined low-level executor, which struggles to generalize to more contact-rich tasks like straightening a rope. To overcome this challenge, robotic foundation models [7, 8, 17, 40, 45, 50] that pretrained on large-scale real-world demonstrations were proposed under the unimanual setting, which has shown high generalizability across everyday manipulation tasks. However, bimanual tasks demand precise coordination of two high degree-of-freedom arms, making the teleoperation of demonstrations for training generalizable policies also costly. Although some recent approaches [13, 16, 18, 22, 62, 70] have developed more specialized teleoperation systems to reduce these costs, scaling up demonstrations for high generalization ability remains a challenge. To address the limited availability of demonstrations, alternative methods [29, 44, 57] proposed to simplify the learning of bimanual policies by decoupling the bimanual system into a stabilizing arm and an acting arm. Nevertheless, these methods often rely on predefined roles for each arm, which precludes their applicability to tasks requiring more flexible collaboration patterns. In contrast to these approaches, our work presents a novel method that transfers generalizable unimanual policies to bimanual tasks, which eliminates the necessity for explicit inductive bias like role specification.

Skill-based Methods. Skill learning [67] is the process where intelligent agents acquire new abilities that are transferable across different tasks, which is of great significance for cross-task generalization. Thus, skill learning is being

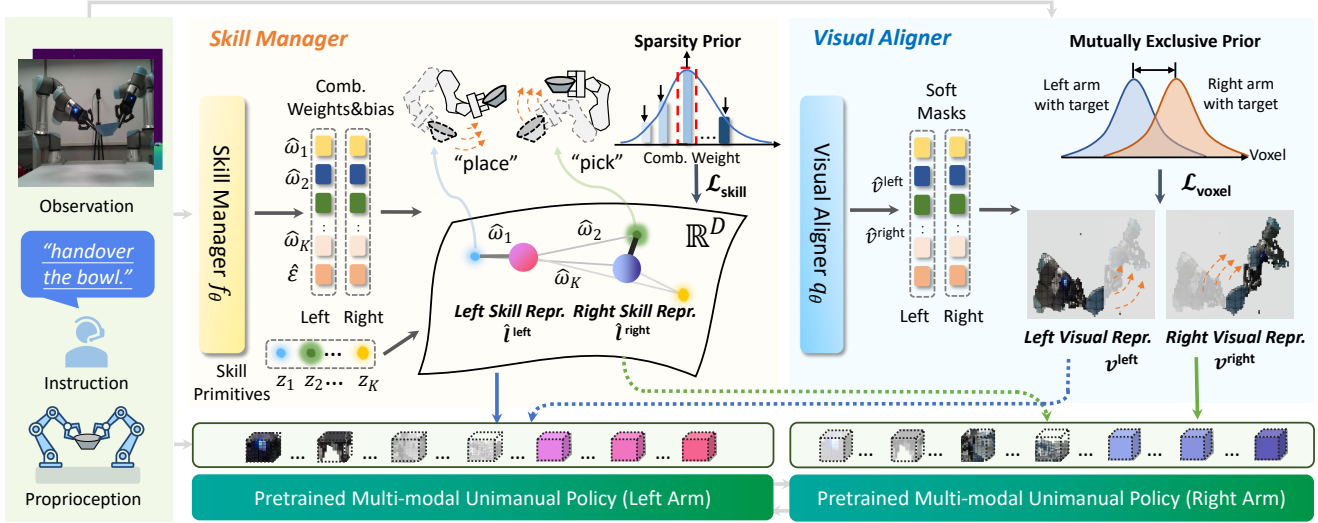


Figure 2. **The overall pipeline of AnyBimanual**, which primarily consists of a skill manager and a visual aligner. The skill manager adaptively coordinates primitive skills for each robot arm, while the visual aligner mitigates the distributional shift from unimanual to bimanual by decomposing the 3D voxel observation for each arm.

attractive in enhancing the generalizability of different models, such as game agents [56], robotic manipulation [42], and autonomous driving [20]. The initial attempt to utilize skill learning was orchestrating a set of predefined skill [47], which hindered their scalability to unseen tasks. To overcome this limitation, [15, 42] proposed to learn shareable skill primitives from data. For example, skill diffuser [42] introduced a hierarchical planning framework that integrates learnable skill embedding into conditional trajectory generation, which realized accurate execution of diverse compositional tasks. In the field of bimanual manipulation, skill learning was dominated by handcrafted primitives. For example, [2–4, 14, 19, 21, 23, 28, 31, 61, 63] propose to utilize parameterized atomic movements to shrink the high dimensionality of the bimanual action space, which has shown impressive performance on templated bimanual manipulation tasks. While the predefined atomic movements did boost the success rate on specific tasks, they are often difficult for even human users to specify, which restricts the deployment scenarios of these methods. In this paper, we propose leveraging learnable skill primitives to represent the learned commonsense of the pretrained unimanual policy, so that the knowledge can be transferred across different levels of tasks.

3. Approach

In this section, we first introduce required preliminaries (Section 3.1), and then we present a pipeline overview (Section 3.2). Then, we introduce a skill manager (Section 3.3) that schedules skills to each arm to form general bimanual manipulation policies. We build a visual aligner (Sec-

tion 3.4) to mitigate the observation discrepancy between bimanual and unimanual systems for better generalization. Finally, we outline the training objectives (Section 3.5).

3.1. Problem Formulation

The task of policy learning for general bimanual manipulation can be defined as follows. To complete a wide range of manipulation tasks specified in natural language, the bimanual agent is required to interactively predict the actions of both end-effectors based on the visual observation and robot states, where the motion is acquired by a low-level planner (*e.g.*, RRT-Connect). The observation o_t at the t_{th} time step includes the voxel v_t converted from RGB and depth images [30, 49] and the robot proprioception p_t . The action a_t for each end-effector at the t_{th} time step contains the location a_{trans} , orientation a_{rot} , gripper open state a_{open} and usage of collision avoidance in the motion planner a_{col} . For the training data, human demonstrators produce a limited set of M offline expert trajectories $\mathcal{D} = \{(o_1, a_1^{left}, a_1^{right}), \dots, (o_M, a_M^{left}, a_M^{right})\}$ for each task instruction l , where a_t^{arm} , $arm \in \mathcal{A} = \{left, right\}$ demonstrates the actions for left and right grippers. Given an off-the-shelf general unimanual manipulation model, our goal is to finetune a generalizable bimanual model using limited M demonstrations per task.

3.2. Overall Pipeline

Overall pipeline of our AnyBimanual method is shown in Figure 2. For the language branch, we employed a pretrained text encoder [48] to parse the bimanual instruction to language embeddings with high-level semantics, where the

skill manager schedules the skill primitives with composition and compensation to enhance the language embeddings that instruct relevant subtasks for different arms. Therefore, the pre-trained unimanual policy model can be prompted to generate feasible manipulation policy for each arm with high generalization ability across tasks with the sharable manipulation knowledge. For the visual branch, the visual aligner generates a soft spatial mask to align the visual representation of unimanual policy model with its representation during pretraining, so that the observation discrepancy between unimanual and bimanual systems can be minimized for policy transferability enhancement. We employ two pretrained unimanual models to predict the left and right robot actions based on the text embeddings and visual representations, where the pretrained unimanual policy can be multi-modal transformer-based policy [7, 8, 17, 40, 49] or diffusion-based policy [38, 50]. Besides, the unimanual policy can incorporate any explicit coupling techniques like shared latent or leader-follower architecture.

3.3. Scheduling Unimanual Skill Primitives

In order to transfer unimanual manipulation policy to bimanual settings without generalizability drops, we propose a skill manager to decompose the action policies from unimanual foundation models into skill primitives and integrate primitives for bimanual systems. However, the given offline expert demonstrations \mathcal{D} do not contain any explicit intermediate skill primitives or sub-task boundaries, but only low-level end-effector poses and high-level natural language instruction are provided. Therefore, we design an automatic skill discovery method in an unsupervised manner to learn skill representations and their schema from offline bimanual manipulation datasets during training. In the test phase, the skill manager predicts different weighted combinations of primitive skills to orchestrate each arm given high-level language instruction, which enables effective transfer of pre-trained unimanual policy to diverse bimanual manipulation tasks.

Skill Manager. We start with a discrete primitive skill set $\mathcal{Z} = \{z_1, z_2, \dots, z_K\}$, where K is a hyper-parameter that indicates the number of skill primitives. To realize end-to-end skill discovery and scheduling, each potential skill is an implicit embedding $z_k \in \mathbb{R}^D$, which can be initialized with the corresponding language template tokens of the pre-trained unimanual policy to mitigate the domain gap. By combining the primitives from the skill set, the language embedding for the unimanual policy model can be represented as a linear combination of these primitives. Hence, the reconstructed language embedding as skill representation can be expressed as:

$$\hat{l}_t^{\text{left}} = \sum_{k=1}^K \hat{w}_{k,t}^{\text{left}} z_k + \epsilon_t^{\text{left}}, \quad \hat{l}_t^{\text{right}} = \sum_{k=1}^K \hat{w}_{k,t}^{\text{right}} z_k + \epsilon_t^{\text{right}} \quad (1)$$

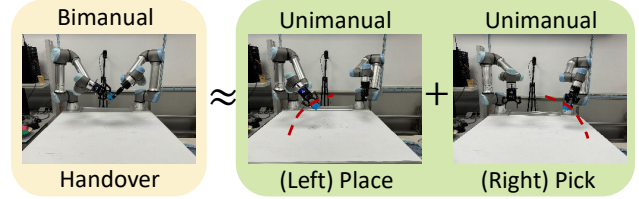


Figure 3. **Shareable skills across unimanual and bimanual settings.** Our key assumption is that bimanual tasks are often originated from the combination of unimanual sub-tasks, which thus can be solved by effectively coordinating unimanual skills synchronously or asynchronously.

where \hat{l}_t^{arm} is the decomposed unimanual language embedding for one arm of the bimanual system, and $\hat{w}_t^{\text{arm}} \in \mathbb{R}^K$ denotes the importance weight for linear combination for both values of arm $\in \mathcal{A}$. We also introduce a task-oriented compensation $\epsilon_t^{\text{arm}} \in \mathbb{R}^D$ to introduce the embodiment-specific knowledge for the policy transfer. The upscript arm of the variables can be selected from left or right to indicate the embodiment in the bimanual systems. As depicted in Figure 3, considering a bimanual task `Handover`, it can be explicitly solved by scheduling two unimanual primitive skills, i.e., the left arm `Place` the block to the right gripper while the right arm `Pick` it from the left gripper.

Though every language embedding that passed through the pretrained single policy can be represented as a linear combination of the skill set, the combination weights that specify the importance of each skill are dynamic across the task completion process. We propose to parametrize a multi-model transformer named skill manager to dynamically predict the combination weight for each arm at each time step. Therefore, our skill manager f can be formulated as $(\hat{w}_t^{\text{left}}, \epsilon_t^{\text{left}}, \hat{w}_t^{\text{right}}, \epsilon_t^{\text{right}}) = f_\theta(v_t, l, p_t)$, which takes the overall bimanual visual and language embeddings, proprioception as input, and assigns the reconstructed unimanual language embedding for each arm as output to schedule the skill primitive of each arm dynamically. θ represents the learnable parameters. Finally, the combined skill primitives are concatenated with the initial bimanual language embedding to enhance the global context, which is then forwarded to the corresponding unimanual policy.

Learning Generalizable Skill Representations. To update the skill library, we expect the discovered skill representations are informative to encode fundamental robot motions that are sharable across a variety of tasks, thereby enhancing the generalizability of our framework. To realize this, the learning objective of the skill manager is designed as a sparse representation problem [58]:

$$\mathcal{L}_{\text{skill}} = \|\hat{w}^{\text{left}}\|_1 + \|\hat{w}^{\text{right}}\|_1 + \lambda_\epsilon (\|\epsilon^{\text{left}}\|_{2,1} + \|\epsilon^{\text{right}}\|_{2,1}) \quad (2)$$

where $\|\cdot\|_1$, and $\|\cdot\|_{2,1}$ denote the ℓ_1 and $\ell_{2,1}$ norm, respectively. λ_ϵ is a hyper-parameter that balances the denoising

term. By minimizing this sparse regularization term, the skill manager is encouraged to reconstruct the skill representation by selecting fewer skill primitives, which is further surrogated by minimizing a differentiable ℓ_1 regularization [51]. Therefore, the skill subspaces are required to be orthogonal and disjoint with each other to reconstruct the language embedding with sparse combination and compensation, which implicitly enforces each skill representation to capture an independent fundamental motion.

3.4. Aligning Unimanual Visual Representations

Visual Aligner. Despite the skill manager enabling generalization in the language modality, the distributional shift from the unimanual to bimanual workspace in terms of the visual context still may harm the model performance. To mitigate the observation discrepancy, we present a visual aligner q that predict two spatial soft masks at each step t to edit the voxel space so that the decomposed subspace of unimanual policy model for each arm aligns with those during pretraining stage: $(\hat{v}_t^{\text{left}}, \hat{v}_t^{\text{right}}) = q_\theta(v_t, l, p_t)$. The decomposed observation represents the locality of the workspace, which is then augmented by the bimanual observation to form the final visual embedding:

$$v_t^{\text{left}} = (\hat{v}_t^{\text{left}} \odot v_t) \oplus v_t, \quad v_t^{\text{right}} = (\hat{v}_t^{\text{right}} \odot v_t) \oplus v_t, \quad (3)$$

where \odot is the element-wise multiplication, and \oplus refers to the concatenation operator. As a result, the augmented visual representations for each arm contains both embodiment-specific information and the global context, which is then passed through the two unimanual policy models to decode the optimal bimanual action.

Learning Aligned Visual Representations. Our goal is to mitigate the visual domain gap between the unimanual and bimanual setting, so that the pretrained commonsense knowledge in unimanual policy can be transferred with high adaptation ability. Since we can not access the unimanual pretraining data in common usages, we instead impose a mutually exclusive prior to the visual aligner. This prior is regularized by optimizing a Jensen-Shannon (JS) divergence regularization term:

$$\mathcal{L}_{\text{voxel}} = -D_{KL}(\hat{v}_t^{\text{left}} \parallel \hat{v}_t^{\text{right}})/2 - D_{KL}(\hat{v}_t^{\text{right}} \parallel \hat{v}_t^{\text{left}})/2 \quad (4)$$

where D_{KL} means the Kullback-Leibler (KL) divergence operator. To provide further explanation, bimanual manipulation tasks often involve asynchronous collaboration that requires the left and right arm to attend to different areas of the whole workspace to act as different roles, such as stabilizing and acting. As a result, the mutually exclusive division of the entire bimanual workspace will naturally separate one arm and its target from the other, which highly resembles the unimanual configuration. Hence by maximizing the divergence between the two soft masks, the voxel

input of the bimanual manipulation agent can be disentangled into unimanual visual representations that align with those in the pretraining phase effectively.

3.5. Learning Objectives

The decomposed skill and visual representation are used to pass through the two pretrained unimanual policies to predict the optimal actions of the two end-effectors. We assume access to a pretrained unimanual policy p , which is fundamentally a multi-model multi-task neural network that takes visual and language embedding as inputs and outputs end-effector actions. Our AnyBimanual is a model-agnostic plug-and-play method, which indicates that the architecture of pretrained unimanual policy p is flexible in different architectures such as multi-modal transformer-based policies [40, 49] and diffusion policies [38, 50]. To supervise the model with the provided expert demonstrations for behavior cloning, we leverage the cross-entropy loss to learn accurate action prediction for each arm:

$$\mathcal{L}_{\text{BC}} = \sum_{\text{arm} \in \mathcal{A}} CE(p_{\text{trans}}^{\text{arm}}, p_{\text{rot}}^{\text{arm}}, p_{\text{open}}^{\text{arm}}, p_{\text{col}}^{\text{arm}}) \quad (5)$$

where $p_{\text{trans}}^{\text{arm}}, p_{\text{rot}}^{\text{arm}}, p_{\text{open}}^{\text{arm}}, p_{\text{col}}^{\text{arm}}$ represents the distribution of the ground-truth actions for translation, rotation, gripper openness, and collision avoidance for the corresponding robot arm, respectively. The behavior cloning loss is then combined with the two regularization terms described above to learn informative skill manager and visual aligner. To sum up, the training objective of AnyBimanual is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BC}} + \lambda_{\text{skill}} \mathcal{L}_{\text{skill}} + \lambda_{\text{voxel}} \mathcal{L}_{\text{voxel}}, \quad (6)$$

where λ_{skill} and λ_{voxel} refer to hyper-parameters that balance the importance of the regularizations.

4. Experiments

In this section, we first introduce the experimental setups (Section 4.1). Then, we transfer various unimanual methods to bimanual manipulation via AnyBimanual, and compare them with the state-of-the-art to show superiority (Section 4.2). We conduct an ablation study to evaluate validity of the proposed components (Section 4.3). We further interpret learned skill representation and decomposed volumetric representation by visualization (Section 4.4). Finally, we report real-robot results to demonstrate effectiveness of AnyBimanual in real-world applications (Section 4.5).

4.1. Experiment Setup

Simulation. For benchmarking, our simulation experiments are conducted on RLBench2 [30], a bimanual version extended from the widely-used RLBench [35] benchmark in prior works [27, 34, 38, 46, 60, 66]. Following the

Table 1. **Multi-Task Test Results in Simulator.** Average success rates (%) of general bimanual manipulation agents trained with 20 or 100 demonstrations per task and evaluated over 100 episodes. ‘LF’ means the leader-follower architecture that transfers unimanual manipulation policy for bimanual manipulation. AnyBimanual enables plug-and-play transfers for multiple unimanual manipulation policies.

Method	pick laptop		pick plate		straighten rope		lift ball		lift tray		push box	
	20	100	20	100	20	100	20	100	20	100	20	100
	RVT [27]-LF	1	2	1	1	1	2	3	3	4	6	7
RVT-LF + AnyBimanual	1	3	2	4	2	5	3	4	4	6	11	21
PerAct [49]-LF	1	2	3	4	5	11	4	7	6	12	7	17
PerAct-LF + AnyBimanual	2	2	3	5	12	14	8	8	15	17	23	29
PerAct ² [30]	3	4	2	4	6	8	4	4	4	3	5	6
PerAct ² + Pretraining	3	5	5	7	13	17	9	14	2	5	23	31
PerAct + AnyBimanual	4	7	6	8	17	24	22	36	9	14	31	46

Method	put in fridge		press buttons		handover item		sweep to dustpan		take out tray		handover easy	
	20	100	20	100	20	100	20	100	20	100	20	100
	RVT [27]-LF	0	0	10	22	0	0	0	0	2	2	3
RVT-LF + AnyBimanual	1	3	17	32	1	2	4	13	1	2	2	7
PerAct [49]-LF	2	7	2	9	0	3	22	47	0	3	2	5
PerAct-LF + AnyBimanual	4	9	12	14	1	5	34	57	1	4	3	13
PerAct ² [30]	7	16	23	41	3	6	25	52	1	3	22	29
PerAct ² + Pretraining	10	22	27	40	3	7	29	55	3	8	24	33
PerAct + AnyBimanual	13	26	39	73	7	15	43	67	9	24	31	44

Table 2. **Comparison of AnyBimanual with Different Techniques.** We categorize the long-term tasks that require more than 6.5 keyframes to Long, tasks that involve multiple variations to Generalized and tasks that involve synchronization of both arms to Sync for further interpretability.

Row ID	Skill Manager	Visual Aligner	Long	Generalized	Sync	Average
1	-	-	16.29	23.50	3.50	14.67
2	✗	✗	19.57	25.50	9.50	16.75
3	✗	✓	21.57	44.00	15.50	19.75
4	✓	✗	23.71	42.00	17.00	25.67
5	✓	✓	27.29	44.00	25.00	32.00

setup in [30], we utilize 12 language-conditioned bimanual manipulation tasks varying from different challenge levels. The diverse task suite requires the agent to acquire and correctly schedule shareable skills to achieve high success rates, rather than merely imitating limited expert demonstrations. For observation, we employ six cameras with a resolution of 256×256 to cover the entire workspace. During the training phase, we provide 20 or 100 demonstrations for each task, and we evaluate 100 episodes per task in the testing set to mitigate the randomness.

Real Robot. The real-world setup for our experiments involves two Universal Robots UR5e manipulators equipped with Robotiq 2F-85 grippers, controlled by two Xbox joysticks for collecting demonstrations. A calibrated front RGB-D Realsense camera provides 640×480 resolution images at 30 Hz for observation. We collect 30 real-world human demonstrations per task for training, while the evaluations are conducted using a Nvidia RTX 4080 GPU.

Baselines. We compare our AnyBimanual with the state-of-

the-art general bimanual manipulation agents, including the voxel-based method PerAct² [30] and its leader-follower version PerAct-LF, both are modified from the well-known unimanual policy PerAct [49], as well as the multi-view image-based method RVT [27]-LF. To exclude the influence of model parameters, we also implement a counterpart that directly combines two pre-trained PerAct [49] policies. Note that the proposed method is model-agnostic, which supports different communication mechanisms between single-arm policies, and thus we transfer all 3 baselines to validate the versatility of AnyBimanual. The evaluation metric is the task success rate, which is defined as the percentage of episodes where the agent successfully completes the instructed goal within 25 steps.

Implementation Details. For fair comparisons, all compared methods are trained for 100k iterations on two NVIDIA RTX 3090 GPUs with a total batch size of 4. We use the LAMB optimizer [64] with a constant learning rate of 5×10^{-4} to update model parameters, in line with the previous arts [27, 30, 49].

4.2. Comparison with the State-of-the-Art Methods

In this section, we compare our AnyBimanual with previous state-of-the-art approaches on RL Bench tasksuite. Table 1 presents a comparison of the average success rates for each task and the average performance is shown in Figure 1. Our method achieves the highest overall performance, with an average success rate of 32.00%, setting a new state-of-the-art in general bimanual manipulation. AnyBimanual leverages the knowledge distilled from unimanual models and successfully transfers it to guide general bimanual manip-

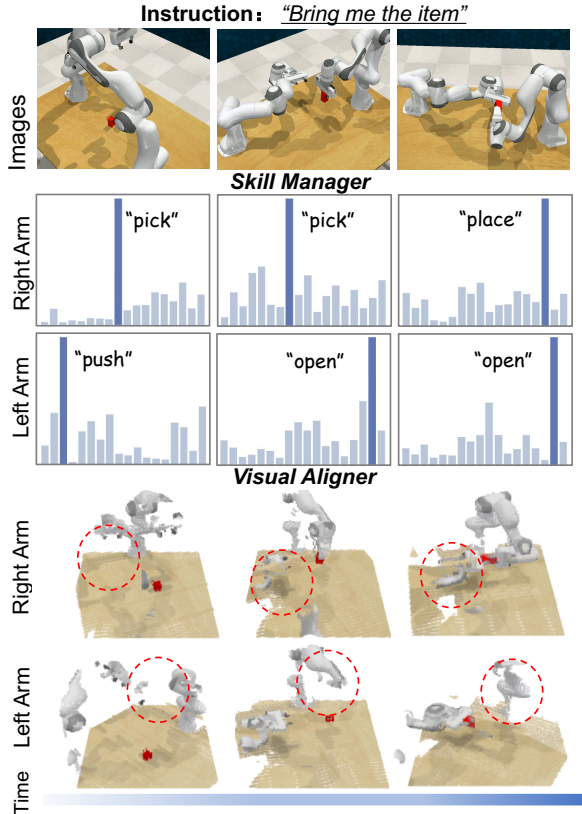


Figure 4. **Visualization of AnyBimanual.** This figure shows in different key timesteps, how the skill manager dynamically schedules skill weights and how the visual aligner decomposes volumetric observation. We use a logarithmic scale for visualization.

ulation. As a result, our method outperforms the cutting-edge bimanual method PerAct² by a significant improvement of 17.33% on average. With a limited set of 20 demonstrations, our method still defeats the baseline with a sizable margin of 10.50%. Note that original PerAct² results were reported in single-task settings, while we focus on multi-task evaluation to emphasize generalizability. Especially, we observe that AnyBimanual shows greater improvement in long-horizon tasks like `put in fridge`, multi-variations tasks like `press buttons`, and tasks demand synchronous coordinations like `straighten rope`. Moreover, AnyBimanual enables plug-and-play transferring of PerAct-LF and RVT, which also leads to relative boosts of 72.76% (4.92% to 8.50%) and 39.41% (10.58% to 14.75%) on average. However, incorporating AnyBimanual slightly affects the performance on the simple, short-horizon `Lift ball`, due to the additional complexity in input processing.

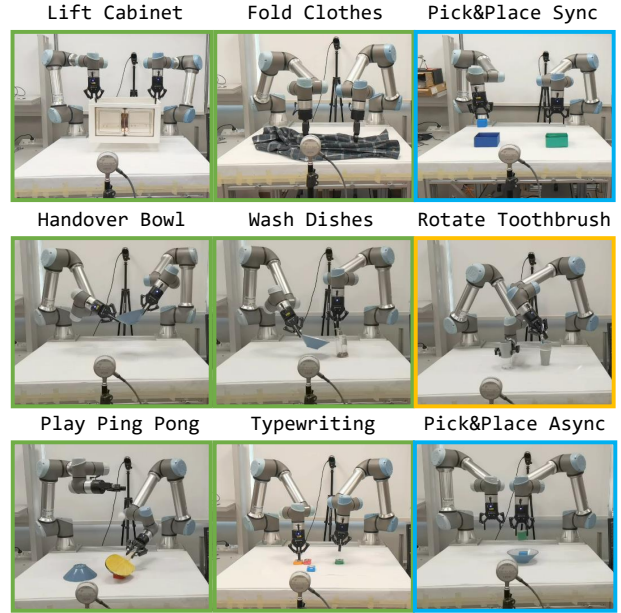


Figure 5. **Real-World Tasks.** The real-world experiments are performed in a tabletop setup with objects randomized in location every episode. AnyBimanual can simultaneously conduct 9 complex real-world bimanual manipulation tasks with one model. Different colors mean different success rates.

4.3. Ablation Study

Our AnyBimanual framework leverages a skill manager to dynamically coordinate unimanual skills, while mitigating the distributional shift between unimanual and bimanual visual inputs with a visual aligner. We conduct an ablation study in Table 2 by transferring the powerful unimanual baseline PerAct [49] to bimanual tasks. We first implement a vanilla baseline without any of the proposed techniques (Row 1), where we load the pre-trained PerAct model and directly finetune the baseline to predict bimanual action, which shows enhanced long-horizon task execution and generalizability with an un neglected performance drop in tasks that require proper coordination.

Skill Manager. By employing the skill manager in the experiment, we observe that the average success rate increases by 8.92% (Row 4 vs. Row 2). Notably, in tasks requiring long-term manipulation within the `Long` category, it significantly outperforms the vanilla version, demonstrating the skill manager’s effectiveness in long-horizon tasks.

Visual Aligner. Additionally, we incorporate the spatial soft masking from the visual aligner, resulting in a substantial performance improvement of 3.00% (Row 3 vs. Row 2). Although the inclusion of spatial soft masking mechanism may slightly affect performance in simpler tasks due to increased input processing complexity, it leads to significant gains in overall task success. The gains are even more no-

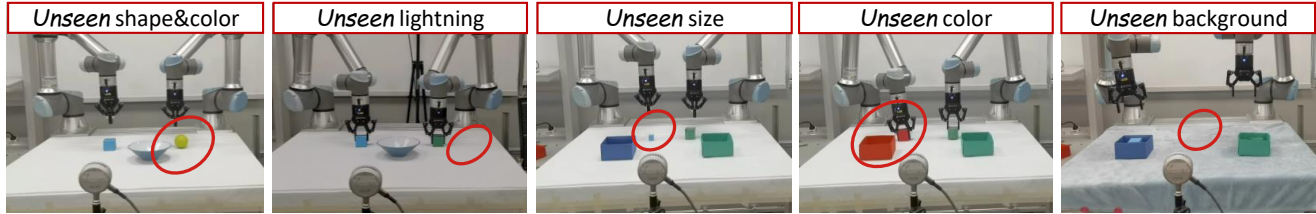


Figure 6. **Unseen Data Generalization.** We include multiple distractors in real-world experiments, and find AnyBimanual generalizes to these settings successfully by unlocking the commonsense held by unimanual base models.

table in Generalized and Sync tasks, which demand a high degree of adaptability to scene variations and synchronization for cooperative actions. The results demonstrate the visual aligner preserves the generalizability stored in the unimanual policy perfectly, facilitating seamless coordination between the arms. By integrating all techniques in our AnyBimanual, the success rate improves from 14.67% to 32.00% (Row 5), validating the importance of properly mining unimanual knowledge for bimanual manipulation.

4.4. Qualitative Analysis

In Figure 4, we visualize linear combination of the skill set and the decomposition of volumetric observation in one complete task to further interpret AnyBimanual. For illustration, we use 18 task embeddings from PerAct [49] as the initial skill set and explicit soft mask. At timestep 1, the right arm primarily follows the `pick blocks in sorter` task (*skill 7*) from the skill set, as the right arm needs to pick up a block from the table at first. The left arm is guided by the `push button` task (*skill 2*), as its motion only involves a downward push without interacting with any objects. At timestep 2, the right arm holds the red block stationary, and the left arm approaches and grasps it, similar to the motion in `open drawer` task (*skill 16*). Additionally, the visual aligner effectively decomposes the voxel space, allowing each arm to focus on relevant information. For instance, at timestep 2, details of the right arm are soft-masked in the left arm’s voxel (as shown in the circled area), which allows the left arm only focus on the red cube without interference from the other arm, enabling the left arm to grasp the red cube more effectively.

4.5. Real-Robot Results

We further validate our approach through real-robot experiments. We train one multi-task AnyBimanual agent that transfers the unimanual policy PerAct [49] to 9 real-world tasks and report the average success rates on Table 3. Guided by [41], these tasks are designed to cover different challenge levels, such as synchronous and asynchronous coordination, short and long-horizon execution, etc. In the real-robot experiments, our keyframes are manually extracted, providing more stable task guid-

Table 3. **Real-world Results.** Average success rates (%) of our multi-task model trained and evaluated on 9 real-world tasks.

Task	Sync.	Object	Variation	Episode	Keyframe	Average
Lift Cabinet	✓	1	1	5	3.0	100.0
Fold Clothes	✓	1	1	5	3.0	100.0
Pick&Place Sync	✓	7	3	15	3.0	80.0
Handover Bowl	✗	1	1	5	4.0	100.0
Wash Dishes	✗	2	1	5	4.0	100.0
Play Ping Pong	✗	3	1	5	3.0	100.0
Rotate Toothbrush	✗	3	1	5	4.0	20.0
Typewriting	✗	4	1	5	6.0	100.0
Pick&Place Async	✗	5	3	15	4.0	80.0

ance compared to heuristic extraction methods in simulation. The overall success rate across 65 test episodes in total is 84.62%, which illustrates a significant practicality of AnyBimanual in real-world settings. Especially, in multi-variant tasks that require high generalizability like `Pick&Place Sync` and `Pick&Place Async` shown in Figure 6, our AnyBimanual completes 80% tasks of different initial positions, object colors, object size, lighting and backgrounds successfully. The failure cases occur on tasks that demand precise rotation, e.g., `Rotate Toothbrush`, which could be mitigated by leveraging high-capacity unimanual policy [40] or balancing the weight of the rotation term in behavior cloning.

5. Conclusion

In this paper, we have introduced AnyBimanual, a framework designed to transfer pretrained unimanual manipulation policies to multi-task bimanual manipulation with few bimanual demonstrations. We develop a skill manager to dynamically schedule skill primitives discovered from unimanual policies, enabling their effective adaptation for bimanual tasks. To address the observation discrepancies between unimanual and bimanual systems, we propose a visual aligner that generates spatial soft masks, aligning the visual embeddings of each arm with those used during the pretraining stage of the unimanual policy model. Extensive experiments across 12 simulated and 9 real-world tasks demonstrate the effectiveness of AnyBimanual. The limitations are discussed in the supplementary file.

Acknowledgements

This work was supported by Shenzhen Science and Technology Program (JCYJ20240813111903006), Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2025B1515020012), MoE AcRF Tier 1 Seed RS17/24 and NTU Ignition Research Grant 024920-00001.

References

- [1] Michael Ahn, Debidatta Dwibedi, Chelsea Finn, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Karol Hausman, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, et al. Autort: Embodied foundation models for large scale orchestration of robotic agents. *arXiv preprint arXiv:2401.12963*, 2024. 2
- [2] Fabio Amadio, Adrià Colomé, and Carme Torras. Exploiting symmetries in reinforcement learning of bimanual robotic tasks. *IEEE Robotics and Automation Letters (RAL)*, 4(2): 1838–1845, 2019. 3
- [3] Yahav Avigal, Lars Berscheid, Tamim Asfour, Torsten Kröger, and Ken Goldberg. Speedfolding: Learning efficient bimanual folding of garments. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2022.
- [4] Aleksandar Batinica, Bojan Nemeč, Aleš Ude, Mirko Raković, and Andrej Gams. Compliant movement primitives in a bimanual setting. In *IEEE-RAS International Conference on Humanoid Robotics (Humanoids)*, pages 365–371, 2017. 3
- [5] Peter Baumgartner, Alexander Fuchs, and Cesare Tinelli. Lemma learning in the model evolution calculus. In *International Conference on Logic for Programming Artificial Intelligence and Reasoning (LPAR)*, pages 572–586, 2006. 2
- [6] Christian Bersch, Benjamin Pitzer, and Sören Kammel. Bimanual robotic cloth manipulation for laundry folding. In *Proceedings of Robotics: Science and Systems (RSS)*, pages 1413–1419, 2011. 2
- [7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 2, 4
- [8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, pages 2165–2183, 2023. 2, 4
- [9] Jens F Buhl, Rune Grønhøj, Jan K Jørgensen, Guilherme Mateus, Daniela Pinto, Jacob K Sørensen, Simon Bøgh, and Dimitrios Chrysostomou. A dual-arm collaborative robot system for the smart factories of the future. *Procedia manufacturing*, 38:333–340, 2019. 1, 2
- [10] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. 2
- [11] Yuanpei Chen, Yiran Geng, Fangwei Zhong, Jiaming Ji, Jiechuang Jiang, Zongqing Lu, Hao Dong, and Yaodong Yang. Bi-dexhands: Towards human-level bimanual dexterous manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 46(5):2804–2818, 2024. 2
- [12] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024. 2
- [13] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024. 2
- [14] Rohan Chitnis, Shubham Tulsiani, Saurabh Gupta, and Abhinav Gupta. Intrinsic motivation for encouraging synergistic behavior. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 3
- [15] Rohan Chitnis, Shubham Tulsiani, Saurabh Gupta, and Abhinav Gupta. Efficient bimanual manipulation using learned task schemas. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1149–1155, 2020. 2, 3
- [16] Ian Chuang, Andrew Lee, Dechen Gao, and Iman Soltani. Active vision might be all you need: Exploring active vision in bimanual robotic manipulation. *arXiv preprint arXiv:2409.17435*, 2024. 2
- [17] Embodiment Collaboration and Abby O’Neill et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903, 2024. 2, 4
- [18] Runyu Ding, Yuzhe Qin, Jiyue Zhu, Chengzhe Jia, Shiqi Yang, Ruihan Yang, Xiaojuan Qi, and Xiaolong Wang. Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning. *arXiv preprint arXiv:2407.03162*, 2024. 2
- [19] Giovanni Franzese, Leandro de Souza Rosa, Tim Verburg, Luka Peternel, and Jens Kober. Interactive imitation learning of bimanual movement primitives. *IEEE/ASME Transactions on Mechatronics (T-Mech)*, 2023. 3
- [20] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 910–919, 2024. 3
- [21] Letian Fu, Huang Huang, Lars Berscheid, Hui Li, Ken Goldberg, and Sachin Chitta. Safe self-supervised learning in real of visuo-tactile feedback policies for industrial insertion. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 10380–10386, 2023. 3
- [22] 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. 2
- [23] Aditya Ganapathi, Priya Sundaresan, Brijen Thananjeyan, Ashwin Balakrishna, Daniel Seita, Jennifer Grannen, Minho

- Hwang, Ryan Hoque, Joseph E Gonzalez, Nawid Jamali, et al. Learning dense visual correspondences in simulation to smooth and fold real fabrics. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 11515–11522, 2021. 3
- [24] Jianfeng Gao, Xiaoshu Jin, Franziska Krebs, Noémie Jaquier, and Tamim Asfour. Bi-kvil: Keypoints-based visual imitation learning of bimanual manipulation tasks. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 16850–16857, 2024. 2
- [25] Zeyu Gao, Yao Mu, Jinye Qu, Mengkang Hu, Lingyue Guo, Ping Luo, and Yanfeng Lu. Dag-plan: Generating directed acyclic dependency graphs for dual-arm cooperative planning. *arXiv preprint arXiv:2406.09953*, 2024. 2
- [26] Koffivi Fidèle Gbagbe, Miguel Altamirano Cabrera, Ali Alabbas, Oussama Alyunes, Artem Lykov, and Dzmityr Tsetserukou. Bi-vla: Vision-language-action model-based system for bimanual robotic dexterous manipulations. *arXiv preprint arXiv:2405.06039*, 2024. 2
- [27] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning (CoRL)*, pages 694–710, 2023. 5, 6
- [28] Jennifer Grannen, Priya Sundaresan, Brijen Thananjeyan, Jeffrey Ichnowski, Ashwin Balakrishna, Vainavi Viswanath, Michael Laskey, Joseph Gonzalez, and Ken Goldberg. Untangling dense knots by learning task-relevant keypoints. In *Conference on Robot Learning (CoRL)*, pages 782–800, 2020. 3
- [29] Jennifer Grannen, Yilin Wu, Brandon Vu, and Dorsa Sadigh. Stabilize to act: Learning to coordinate for bimanual manipulation. In *Conference on Robot Learning (CoRL)*, pages 563–576, 2023. 1, 2
- [30] Markus Grotz, Mohit Shridhar, Yu-Wei Chao, Tamim Asfour, and Dieter Fox. Peract2: Benchmarking and learning for robotic bimanual manipulation tasks. In *Conference on Robot Learning (CoRL)*, 2024. 1, 2, 3, 5, 6
- [31] Huy Ha and Shuran Song. Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding. In *Conference on Robot Learning (CoRL)*, pages 24–33, 2021. 3
- [32] Zhaoyang Jacopo Hu, Ziwei Wang, Yanpei Huang, Aran Sena, Ferdinando Rodriguez y Baena, and Etienne Burdet. Towards human-robot collaborative surgery: Trajectory and strategy learning in bimanual peg transfer. *IEEE Robotics and Automation Letters (RAL)*, 8(8):4553–4560, 2023. 1, 2
- [33] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *Conference on Robot Learning (CoRL)*, pages 540–562, 2023. 2
- [34] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4651–4664, 2021. 5
- [35] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters (RAL)*, 5(2):3019–3026, 2020. 5
- [36] Frank Joublin, Antonello Ceravola, Pavel Smirnov, Felix Ocker, Joerg Deigmoeller, Anna Belardinelli, Chao Wang, Stephan Hasler, Daniel Tanneberg, and Michael Gienger. Copal: corrective planning of robot actions with large language models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 8664–8670, 2024. 2
- [37] Satoshi Kataoka, Seyed Kamyar Seyed Ghasemipour, Daniel Freeman, and Igor Mordatch. Bi-manual manipulation and attachment via sim-to-real reinforcement learning. *arXiv preprint arXiv:2203.08277*, 2022. 2
- [38] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*, 2024. 2, 4, 5
- [39] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024. 2
- [40] Moo Jin Kim and Karl Pertsch et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 2, 4, 5, 8
- [41] Franziska Krebs and Tamim Asfour. A bimanual manipulation taxonomy. *IEEE Robotics and Automation Letters (RAL)*, 7(4):11031–11038, 2022. 8
- [42] Zhixuan Liang, Yao Mu, Hengbo Ma, Masayoshi Tomizuka, Mingyu Ding, and Ping Luo. Skilldiffuser: Interpretable hierarchical planning via skill abstractions in diffusion-based task execution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16467–16476, 2024. 3
- [43] Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*, 2023. 2
- [44] I Liu, Chun Arthur, Sicheng He, Daniel Seita, and Gaurav Sukhatme. Voxact-b: Voxel-based acting and stabilizing policy for bimanual manipulation. *arXiv preprint arXiv:2407.04152*, 2024. 1, 2
- [45] Songming Liu and Lingxuan Wu et al. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024. 2
- [46] Guanxing Lu, Shiyi Zhang, Ziwei Wang, Changliu Liu, Jiwen Lu, and Yansong Tang. Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation. In *European Conference on Computer Vision (ECCV)*, pages 349–366, 2025. 5
- [47] Asim Munawar, Giovanni De Magistris, Tu-Hoa Pham, Daiki Kimura, Michiaki Tatsubori, Takao Moriyama, Ryuki Tachibana, and Grady Booch. Maestrob: A robotics framework for integrated orchestration of low-level control and high-level reasoning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 527–534, 2018. 3

- [48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 3
- [49] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning (CoRL)*, 2023. 2, 3, 4, 5, 6, 7, 8
- [50] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. 2, 4, 5
- [51] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996. 5
- [52] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2
- [53] Jake Varley, Sumeet Singh, Deepali Jain, Krzysztof Choromanski, Andy Zeng, Somnath Basu Roy Chowdhury, Avinava Dubey, and Vikas Sindhwani. Embodied ai with two arms: Zero-shot learning, safety and modularity. *arXiv preprint arXiv:2404.03570*, 2024. 2
- [54] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, pages 1723–1736, 2023. 2
- [55] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024. 2
- [56] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research (TMLR)*, 2024, 2024. 3
- [57] Woodrow Zhouyuan Wang, Andy Shih, Annie Xie, and Dorsa Sadigh. Influencing towards stable multi-agent interactions. In *Conference on Robot Learning (CoRL)*, pages 1132–1143, 2022. 2
- [58] John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S Huang, and Shuicheng Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010. 4
- [59] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. *arXiv preprint arXiv:2309.13037*, 2023. 2
- [60] Zhou Xian, Nikolaos Gkanatsios, Theophile Gervet, Tsung-Wei Ke, and Katerina Fragkiadaki. Chaineddiffuser: Unifying trajectory diffusion and keypose prediction for robotic manipulation. In *Conference on Robot Learning (CoRL)*, pages 2323–2339, 2023. 5
- [61] Fan Xie, Alexander Chowdhury, M. Clara De Paolis Kaluza, Linfeng Zhao, Lawson L. S. Wong, and Rose Yu. Deep imitation learning for bimanual robotic manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3
- [62] Shiqi Yang, Minghuan Liu, Yuzhe Qin, Runyu Ding, Jialong Li, Xuxin Cheng, Ruihan Yang, Sha Yi, and Xiaolong Wang. Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation. *arXiv preprint arXiv:2408.11805*, 2024. 2
- [63] Xiaochuan Yin and Qijun Chen. Learning nonlinear dynamical system for movement primitives. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3761–3766, 2014. 3
- [64] Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training BERT in 76 minutes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 6
- [65] Dongjie Yu, Hang Xu, Yizhou Chen, Yi Ren, and Jia Pan. Bicc: Keypose-conditioned consistency policy for bimanual robotic manipulation. *arXiv preprint arXiv:2406.10093*, 2024. 2
- [66] Yanjie Ze, Ge Yan, Yueh-Hua Wu, Annabella Macaluso, Yuying Ge, Jianglong Ye, Nicklas Hansen, Li Erran Li, and Xiaolong Wang. Gnfactor: Multi-task real robot learning with generalizable neural feature fields. In *Conference on Robot Learning (CoRL)*, pages 284–301, 2023. 5
- [67] Jesse Zhang, Jiahui Zhang, Karl Pertsch, Ziyi Liu, Xiang Ren, Minsuk Chang, Shao-Hua Sun, and Joseph J. Lim. Bootstrap your own skills: Learning to solve new tasks with large language model guidance. In *Conference on Robot Learning (CoRL)*, pages 302–325. PMLR, 2023. 2
- [68] Minghao Zhang, Pingcheng Jian, Yi Wu, Huazhe Xu, and Xiaolong Wang. Dair: Disentangled attention intrinsic regularization for safe and efficient bimanual manipulation. *arXiv preprint arXiv:2106.05907*, 2021. 2
- [69] Tianle Zhang, Dongjiang Li, Yihang Li, Zecui Zeng, Lin Zhao, Lei Sun, Yue Chen, Xuelong Wei, Yibing Zhan, Lu-song Li, et al. Empowering embodied manipulation: A bimanual-mobile robot manipulation dataset for household tasks. *arXiv preprint arXiv:2405.18860*, 2024. 1, 2
- [70] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 2