Universal Actions for Enhanced Embodied Foundation Models

Abstract

Training on diverse, internet-scale data is a key factor in the success of recent large foundation models. Yet, using the same recipe for building embodied agents has faced noticeable difficulties. Despite the availability of many crowd-sourced embodied datasets, their action spaces often exhibit significant heterogeneity due to distinct physical embodiment and control interfaces for different robots, causing substantial challenges in developing embodied foundation models using cross-domain data. In this paper, we introduce UniAct, a new embodied foundation modeling framework operating in a ***tokenized Universal Action Space****. Our learned universal actions capture the generic atomic behaviors across diverse robots by exploiting their shared structural features, and enable enhanced cross-domain data utilization and cross-embodiment generalizations by eliminating the notorious heterogeneity. The universal actions can be efficiently translated back to heterogeneous actionable commands by simply adding embodiment-specific details, from which fast adaptation to new robots becomes simple and straightforward. Our 0.5B instantiation of UniAct outperforms 14X larger SOTA embodied foundation models in extensive evaluations on various real-world and simulation robots, showcasing exceptional cross-embodiment control and adaptation capability, highlighting the crucial benefit of adopting universal actions.*

1. Introduction

In fields like natural language processing and computer vision, foundation models trained on vast and diverse data sources have demonstrated remarkable success and strong generalization ability, highlighting the benefits of learning general-purpose models over task-specific counterparts [1, 5, 40, 62]. Inspired by these successes, developing versatile embodied foundation models that are capable of handling cross-task, cross-environment, and cross-embodiment generalization, offers a promising pathway towards building general-purpose embodied agent [8, 17, 18, 28, 31, 61, 68].

However, significant challenges arise from the substantial heterogeneity of embodied data [20, 49, 64]. Such heterogeneity is evident not only in visual discrepancies caused by variations in camera placements (*e.g.*, wrist or third-person view) and environmental conditions (*e.g.*, lighting or background variations), but more critically, in *action heterogeneity* [17, 48, 65]. *1)* Robots with different embodiment (*e.g.*, different degrees of freedom or distinctions across robotic arms, quadrupeds, and cars) possess entirely distinct action spaces [68]. *2)* Furthermore, the diversity in control interfaces (*e.g.*, end effector (EEF) position or velocity controller for robotics arms) leads to fundamentally different physical meanings for action commands [49]. *3)* Even when actions are collected from the same robotic platform but by different human manipulators, the multi-modality in human behaviors also exacerbates such heterogeneity [14, 34, 41, 55]. Consequently, embodied action data collected across different robots and institutions tend to reside on largely disjoint manifolds within the original physical spaces (e.g., position and rotation of end-effectors) [17, 65], significantly complicating data sharing across different data sources.

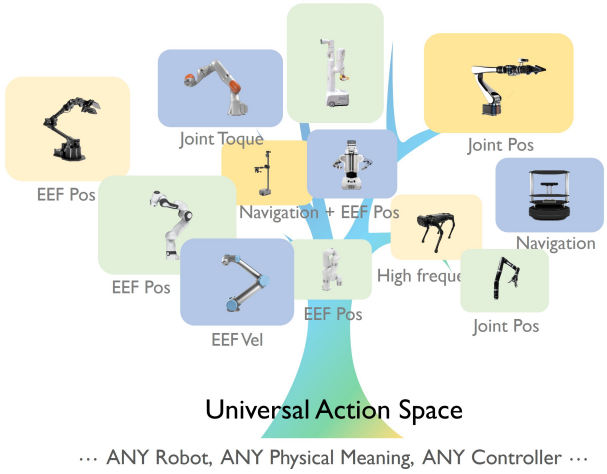

Figure 1. Illustration of the *Universal Action Space*. Universal actions should be versatile to ANY domain-specific actions.

Currently, no existing solution can adequately address the issue of action heterogeneity. Most prior studies forcibly treat different action spaces as equivalent and apply the same discretization or normalization techniques, leading to a potentially conflicted action space where similar action

encodings could represent entirely different physical meanings [31, 49, 61]. While some efforts attempt to design a physically interpretable action space that is applicable across various robotic systems by naïvely aggregating all individual action spaces [17, 41]. This requires extensive human engineering efforts and fails to uncover and leverage the inherent connections across different embodied action spaces, impeding the effective development of general-purpose embodied foundation models.

In this paper, we introduce *UniAct* (Embodied foundation models with Universal Actions), a novel embodied modeling framework that is constructed in the *Universal Action Space* rather than the troublesome heterogeneous action spaces. The universal actions learned in *UniAct* encode the generic atomic behaviors that are generalizable across diverse robotics platforms without being constrained to specific embodiment mechanics and control interfaces. For instance, different robots should perform similar behaviors of "moving forward" when facing a target directly ahead, despite exhibiting totally different control signals due to their embodiment gaps. This abstraction transcends specific embodiment and control constraints, allowing it to be universally applicable across diverse robotics platforms, providing significant potential to enhance cross-embodiment data utilization and model generalization. In this sense, minimal parameters and data are sufficient to decode the universal action to an embodiment-specific one, since the general motion behaviors have been captured in the universal space and the decoder simply needs to add some embodiment details for each robotic platform, therefore enabling efficient adaptation across new robotic systems during deployment.

In details, UniAct employs a shared Vision Language Model (VLM) [6, 32, 35, 39, 40, 72] to construct the universal action space as a vector-quantized codebook [63]. Akin to a learnable skill library [34, 45, 55], each code encapsulates an atomic behavior versatile enough to be executed by diverse robots. This setup acts as a crucial information bottleneck, driving the VLM to identify and exploit shared primitive behaviors across diverse action spaces. This extraction scheme enables effective generalization of behaviors for cross-embodiment control, making our 0.5B instantiation, UniAct-0.5B, surpasses models 14x larger, such as OpenVLA [31] with 7B parameters, in a wide range of tasks. The derived universal actions can be translated into precise, actionable commands for various embodiments through streamlined heterogeneous decoders. These decoders take the universal action as conditional input and augment it with embodiment-specific features from their unique observational data. This allows for flexible customization according to specific requirements, such as the inclusion or exclusion of proprioceptive features or variations in the number of camera views. Fast adaptation to new domains or robotic platforms can be achieved by simply adding new heads as lightweight decoders for new tasks. Our comprehensive evaluations on challenging task settings, including large view changes and an unseen robot not present in the training data, confirm UniAct's remarkable transferability, demonstrating great advantages of developing embodied foundation models within a universal action space over the conventional heterogeneous spaces.

2. Related Work

Multimodal Foundation Models. Large Language Models (LLMs) [1, 5, 40, 62] have exhibited remarkable capabilities across a variety of tasks, showcasing impressive zero-shot and in-context learning capabilities [16]. Building on this, large Vision Language Models (VLMs) have been developed by integrating vision and language into a unified tokenized space, demonstrating outstanding multimodal instruction-following abilities [6, 21, 32, 35, 58, 59, 72]. Their success is primarily attributed to extensive internet-scale pretraining, which leverages vast and diverse high-quality data corpora from the Internet.

Embodied Foundation Models. When developing embodied foundation models, an additional crucial modality—action (the deployable control signals that robots can interpret and execute, *e.g.*, EEF position/velocity)—is incorporated during training. State-of-the-art models are often constructed as Vision Language Action models (VLA) [9, 10, 18, 31, 49], integrating visual and linguistic inputs with actionable outputs. However, the action labels collected from different robotics platforms and labs exhibit significant heterogeneity [30, 49, 64], impeding effective data sharing across different sources. To sidestep this challenge, many works employ large-scale action-free vision language data, such as out-of-domain human activities [15, 24, 25], to firstly obtain a good embodied VLMs, then finetune it to a specialized VLA given a small set of action labels from a specific robot platform [3, 4, 7, 10, 12, 19, 22, 29, 36, 37, 42, 47, 66, 67, 69, 72]. While these methods can enhance sample efficiency for specific robots on a narrow set of tasks, they suffer from serious performance bottlenecks towards building a generalist embodied agent [7, 19, 56], as the action data gathered from any single robot platform is far less comprehensive than crowed-sourced data collected globally [20, 30, 49].

Some recent works leverage the abundant heterogeneous action labels to develop generalist robot policies for cross-embodiment control. RT-X series [49], Octo [61] and OpenVLA [31] leverage data from different 7-DoF robots to enhance generalization over the one trained on single robot source. Step further, CrossFormer [17], RDT [41], π_0 [8] and Yang et al. [68] explore the potential of using data from robots with totally distinct mechanical structures, such as those in manipulation and navigation, and from single-arm versus bi-manual systems. However, existing works

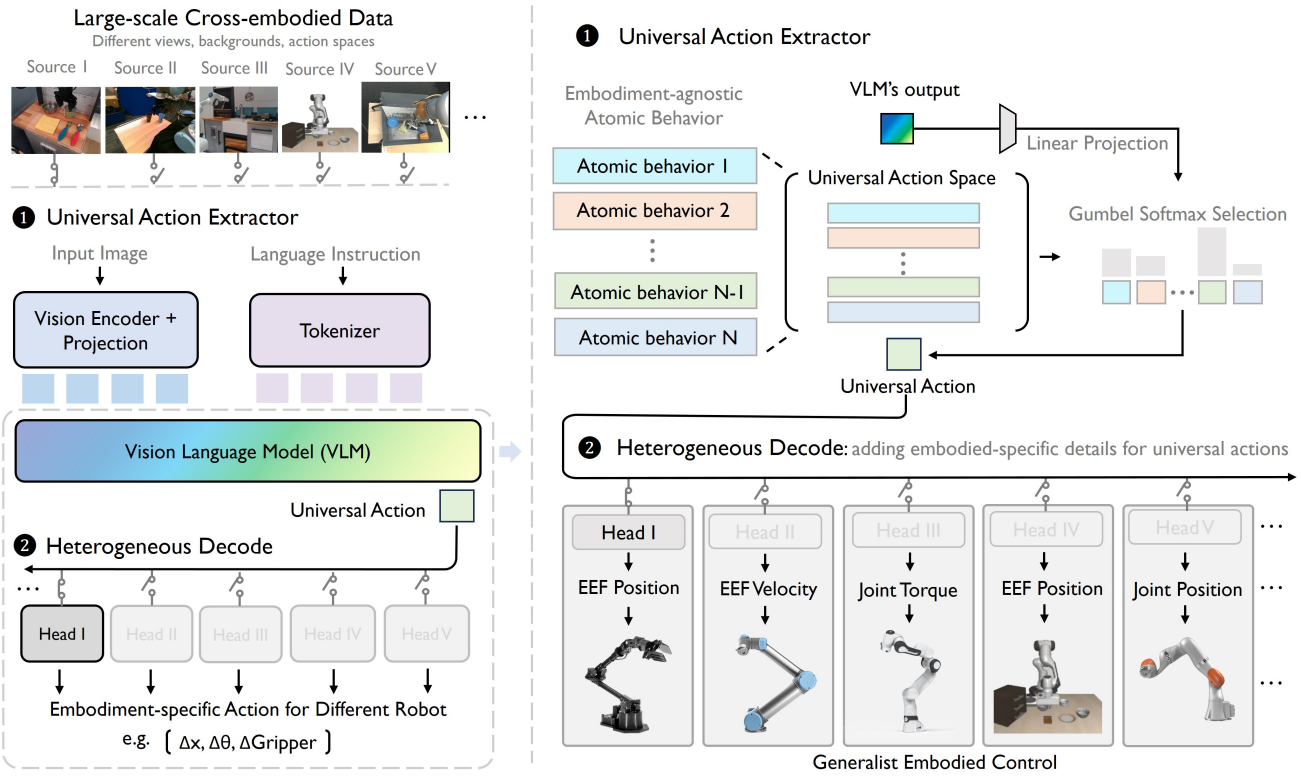

Figure 2. UniAct leverages diverse data sources for generalist embodied control. A shared Vision Language Model (VLM) extracts transferable features across various data sources; The output tokens are converted into universal actions, represented as vector-quantized codes where each code captures common atomic behaviors across different robots; The Gumbel Softmax-selected universal action is then translated back into specific commands through different heads, each encoding embodiment-specific features for individual robots.

either ignore the heterogeneous properties of action spaces of different sources, crudely treating them as equal without considering their inherent conflicts [31, 49, 61], or naïvely aggregate all action spaces together, failing to exploit the underlying shared commonalities across different robots [8, 17, 41, 68].

Embodied Models with Latent Action Spaces. Our work aims to extract a versatile universal action space, akin to a latent space but encodes common atomic control behaviors and patterns across various robotic platforms. Some works develop embodied models in latent spaces [4, 13, 34, 45, 54, 55, 69, 70]. Among them, LAPA [69], IGOR[4], and LAPO [54] develop a latent action space through joint self-supervised training of inverse and forward dynamics models on action-free videos [11]. However, the latent actions extracted in this way primarily focus on explaining the changes between video frames, lacking embodiment considerations or direct causal connections to actual control signals. To see why this is problematic, assuming we add a new object in front of the robot, the visual inputs will change but this has nothing to do with the control behavior, an ideal encoded action should not capture such distracted information. BeT [55], VQ-BeT [34] and QueST [45] also build a discrete codebook of actions via K-means clustering [43] or Vector Quantization [33, 44, 63], where each code in the codebook encodes a different clustering center for action labels. These works primarily focus on simpler domains with a single embodiment type, which enhances the ability to model complex human demonstrations with multiple modes, but struggles to address action heterogeneity across different embodiments. In contrast, our universal actions integrate goal information from the embodiment-agnostic language modality with supervision on the actual action signal, providing a versatile and abstracted skill library to facilitate cross-embodiment sharing. Moreover, our research delves into a more complex heterogeneous setting and develops a large embodied foundation model, moving beyond the limited scopes considered in previous studies.

3. The UniAct Framework

We introduce UniAct, an embodied foundation modeling framework designed to operate in a *Universal Action Space*, adept at bridging domain gaps and facilitating training on large-scale heterogeneous data. We first discuss the desir-

able properties of universal actions, and then provide a detailed discussion about the model architecture and learning scheme for extracting and decoding universal actions from heterogeneous cross-embodiment data.

3.1. Universal Action Space

The desired universal action space is that all movements, driven by heterogeneous control signals from various embodiments, can be distilled into shared latent atomic behaviors, despite their distinct physical meanings. We refer to these abstract behavior representations as *universal actions*, which are shared across all physical embodiments.

We are particularly interested in exploring a discrete universal action space. This is motivated by the robust capabilities of discrete representations in complex reasoning, planning, and predictive learning, as demonstrated by the success of LLMs and VLMs [1, 5, 6, 60] and Vector Quantized Variational Autoencoders [52, 63]. In this paper, we model the universal action space as $\mathcal{U} \in \mathbb{R}^{N \times D}$ and implement it with a vector quantized codebook [63], represented as

$$\mathcal{U} = (u_1, u_2, u_3 ... u_N), \tag{1}$$

where N is the space size and each u_i is a D-dimensional vector embedding that represents a generic atomic behavior.

Several prior studies [4, 11, 69] pursued a similar concept of constructing generic, latent actions by inferring them as the dynamic changes observed between two visual states. However, this scheme suffers two key limitations, leading to suboptimal and noisy universal action spaces:

- The observational changes encompass not only the control outcomes of robots, but also external factors (e.g., environment variation, the appearance of new objects, human intervention, etc.) that have no causal connection to the actual control.
- The interval between two observations critically influences the semantic interpretation of the extracted atomic behaviors, making standardizing behavioral interpretation complicated across different data sources.

3.2. Universal Action Extraction

To derive the desirable universal action space, we propose a new method for extracting universal actions, pivoting away from solely focusing on explaining observational changes, but more on understanding task progression. Specifically, we fine-tune a large vision language model as the universal action extractor, which outputs the likelihood $p(u|o, g)$ of selecting universal action u given observation o and task goal g (e.g., language instruction). We want the corresponding adopted universal action u^* that matches the atomic behavior encoded in the embodied data to satisfy:

$$u^* = \arg\max_{u \in \mathcal{U}} p(u|o, g) \tag{2}$$

Akin to planning in the latent space [13, 50, 51], the extractor aims to infer the most relevant universal action for solving a given task g under the observation o, thereby crafting universal actions directly related to task progression rather than merely identifying the noisy observational changes. We employ a VLM to achieve this purpose due to its strong visual-language reasoning capability. Moreover, fine-tuning a pre-trained VLM also greatly improves the sample efficiency when learning the universal actions. This extractor creates a crucial information bottleneck for cross-domain generalization, as different robots are forced to use the same discrete codebook $\mathcal{U}$ to capture the generic and shared atomic behaviors across all domains.

To implement this, however, the non-differentiable $\arg\max$ impedes the gradient propagation. So, we use categorical reparametrization during the training, utilizing the Gumbel-Softmax technique to facilitate gradient estimation [27]. Specifically, the forward procedure is as follows:

$$u^* = \sum_{i=1}^{n} w_i u_i, \tag{3}$$

where w_i is the weight for each universal action u_i, computed using the Gumbel Softmax function:

$$w_i = \frac{\exp((\log p(u_i|o, g) + \epsilon_i)/\tau}{\sum_{k=1}^{n} \exp((\log p(u_k|o, g) + \epsilon_k)/\tau} \tag{4}$$

Here, ϵ_i is a Gumbel noise sampled from the Gumbel distribution, and τ is the temperature to smooth the probability distribution. To promote parameter space exploration in the early training stage and the stability of model convergence, we gradually decay temperature τ during the training process. Our proposed universal action extractor is illustrated in Figure 2, please refer to the Appendix A for more details.

3.3. Heterogeneous Decoding

To efficiently translate the highly abstract behaviors in the universal action space into precise, embodiment-specific control signals, it is crucial to integrate more embodiment detail, such as control type, proprioception, and distinct observations. To address this, we introduce a series of lightweight decoder heads to adapt for each type of embodiment, denoted as $\mathcal{H} = (h_1 ... h_k ... h_K)$, K is number of training domains. Each head h_k is specifically designed to learn the mapping from universal action u^* and visual observation o to heterogeneous control signals for the embodiment in domain k. The operation of each decoder head h_k can be formulated as:

$$\hat{a}^{(k)} = h_k(u^*, o), \tag{5}$$

where $\hat{a}^{(k)}$ is the predicted control signals. As overly complex decoding heads with excessive parameters could overfit to the data distribution of the target domain, all heterogeneous heads are implemented as simple MLP networks

that take the u^* and visual features o extracted by a shared vision backbone as inputs. By keeping the decoding heads lightweight, we ensure that the majority of learning is conducted for the universal actions, thereby maximally improving generalization across different embodiments.

3.4. Training Procedure

The primary learning objective of UniAct is to distill a universal action space that is shared across diverse embodiments, with the critical feature that these universal actions can be precisely translated back into domain-specific control signals. To facilitate this, the model is trained using a comprehensive collection of K heterogeneous datasets $\mathcal{D} = (\mathcal{D}_1 ... \mathcal{D}_k ... \mathcal{D}_K)$. Each $\mathcal{D}_k$ comprises a set of robotic control trajectories, represented as $\tau_i^{(k)} = \{o_{i,t}^{(k)}, a_{i,t}^{(k)}, g_i\}_{1 \leq t \leq T}$, where $\tau_i^{(k)}$ is the i-th trajectory of maximum length T, containing observations, actions, and goals. Concretely, UniAct ingests o and g as inputs to predict the universal action u^* using the universal action extractor, which is then mapped along with o to $\hat{a}^{(k)}$ using the heterogeneous decoding heads. The overall training objective is as follows:

$$\min_{\mathcal{U},\theta} \sum_{k=1}^{K} \mathbb{E}_{a_i \in \tau_i, \tau_i \in \mathcal{D}_k} \mathcal{L}_k(\hat{a}^{(k)}, a_i^{(k)}) \tag{6}$$

Here, $\mathcal{L}_k$ is the behavior cloning loss, which is customizable based on the nature of the action labels in each dataset, for example, employing Cross-Entropy for discrete actions and MSE, Huber loss, or diffusion loss [14, 26] for continuous actions. We optimize the above objective to learn both the universal action codebook $\mathcal{U}$ as well as the parameters θ of the universal action extractor and all heterogeneous decoding heads. Importantly, while $\mathcal{U}$ and the universal action extractor are concurrently updated throughout each training iteration, the heterogeneous heads h_k are updated based on the domain-specific sampled training batches. This training strategy mirrors the philosophy in many meta-learning methods [23], which learns both globally shared parameters that allow adaptation to related tasks, as well as task-specific components that guarantee downstream task performance. Through this approach, UniAct strives to refine a robust, adaptive universal action space as well as a decoding strategy that can be seamlessly integrated with diverse embodiments and their specific operational contexts.

4. Experiments

In this section, we first describe the detailed implementation of the UniAct framework and then present the evaluation experiments conducted to answer the following questions:

- Can universal actions enhance execution performance across various embodiments with large domain gaps?
- Can universal actions be seamlessly transferred to new, unseen embodiments?
- Dose UniAct learn a meaningful universal action space?

4.1. Experiments Setup

Implementation Details. In this paper, we build a 0.5B instantiation of UniAct on heterogeneous embodied data sources to explore a universal action space ($\mathcal{U} \in \mathbb{R}^{256 \times 128}$). Specifically, UniAct-0.5B is built upon *LLaVA-OneVion-0.5B* [35], a well-trained VLM which can provide comprehensive multi-modal representations. The training of UniAct-0.5B is carried out on 64 A100 GPUs with DeepSpeed [53] over a span of 10 days, utilizing 1 million demonstrations gathered from 28 distinct embodiments. The training data combines several open-sourced robot collections, including Open-XEmbodiment [49], Libero [38], and Droid [30], standardized to include third-person visual observations and language instructions while preserving action heterogeneity. For more details about training and data constructions, please refer to the Appendix A.

Baseline Setup. We select two state-of-the-art open-source vision-language-action models as baselines: Octo [61] and OpenVLA [31]. Octo is a 0.1B diffusion-based policy, and OpenVLA employs a 7B-parameter auto-regressive architecture with discrete actions. Both models are trained on about 1 million carefully curated robot demonstrations without action heterogeneity, such as pre-processing all absolute EEF positions to relative EEF positions and removing joint position actions. In contrast, UniAct-0.5B is trained on a similar scale of data from the same data sources but does not employ such tedious data cleaning. We compare UniAct with the baseline models to demonstrate its effectiveness in extracting universal actions from heterogeneous data.

4.2. Main Results

To assess the cross-embodiment generalization capabilities of UniAct-0.5B, we conduct "out-of-the-box" evaluations on both a real-world WidowX robot [64] and a simulation Franka robot from Liu et al. [38]. Both platforms are commonly used in previous works to test the effectiveness of generalist robot policies [49, 61, 64], and possess substantial domain gaps. Given that our training dataset includes data from these two embodiments, we can leverage the pre-trained heterogeneous heads to translate the universal actions seamlessly back into deployable control signals.

Real-World Robot Evaluation. Following Kim et al. [31], we define a comprehensive set of evaluation tasks for real-world robots, covering several dimensions of generalization: **visual** (unseen backgrounds/distractions/object appearances); **motion** (unseen object positions/orientations); **physical** (unseen object sizes/shapes); **semantic** (unseen target objects/instructions/concepts from the Internet); and **language grounding** (manipulate the object specified in language). Overall, each model is evaluated across 190 rollouts, distributed over 19 tasks with 10 trials each. See

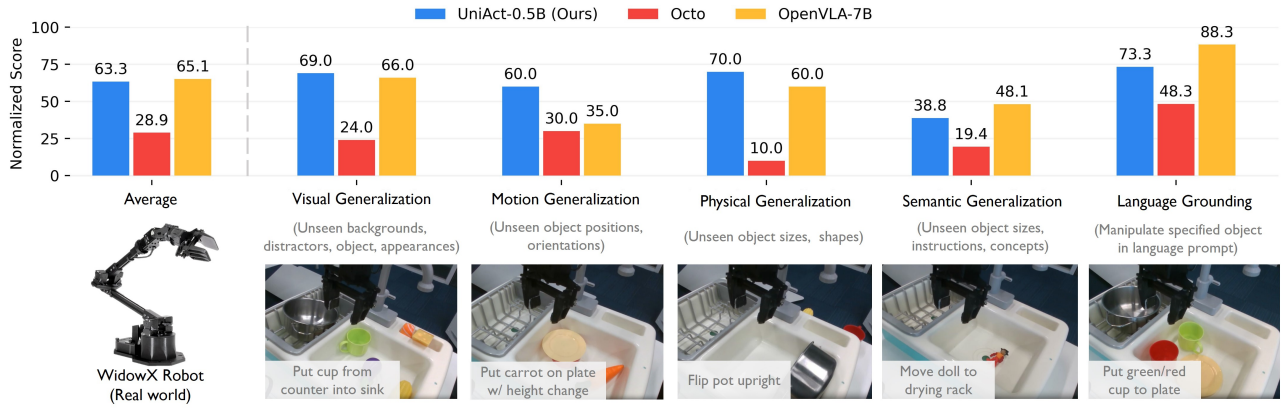

Figure 3. Evaluation on the WidowX robot. We reproduce the BridgeV2 real-world platform [64] and follow the evaluation protocol of OpenVLA [31] to ensure fair comparisons. We evaluate different axes of generalization ability, covering a total of 19 tasks. Representative tasks are presented in the figure. For each task, we evaluate 10 trials for each model and report the average success rate for each task suites.

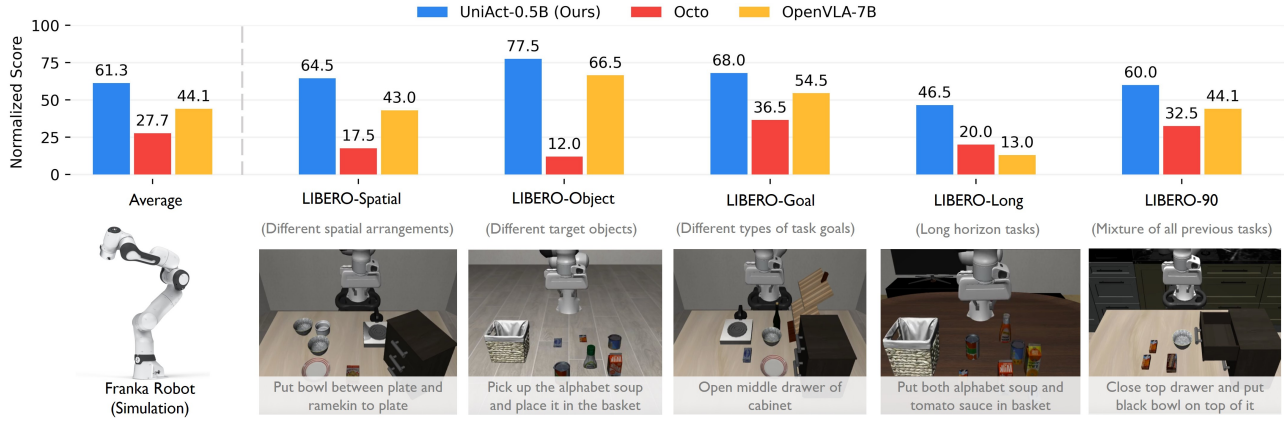

Figure 4. For benchmarking purpose, we evaluate on the LIBERO benchmark [38] in simulations. LIBERO contains several task suites to examine different axes of policy abilities. LIBERO-90 contains 90 tasks and others each contain 10 tasks, covering 130 tasks. For each task, we evaluate 20 trials for each model and report the average success rate for each task suites.

Appendix B for more details. Representative tasks and results are illustrated in Figure 3. UniAct-0.5B outperforms the 14X larger OpenVLA-7B in visual, motion, and physical generalization tasks. This demonstrates the substantial benefits of extracting universal actions from heterogeneous data in enhancing robustness to visual distractions and low-level control generalization. While OpenVLA leverages a 7B vision-language backbone for superior semantic understanding and language grounding capability, UniAct-0.5B achieves comparable performance in semantic generalization and language grounding tasks, underscoring its efficiency and effectiveness.

Simulation Evaluation. We utilize the LIBERO Benchmarks [38] for evaluation. Notably, the baseline models are not initially trained on the simulation data, thus we utilize their official codebase and training guidelines to fine-tune them on the LIBERO platform. The simulation data used to train both UniAct and the baseline models are fully aligned in terms of task type, number of expert trajectories, and image quality. For more details about the fine-tuning procedure, please refer to the Appendix A. The benchmarks comprise 130 robotic simulation tasks across five distinct suites: LIBERO-Spatial, -Object, -Goal, -Long, and -90. The LIBERO-90 suite includes 90 tasks, while each of the other four suites contains 10 tasks. The deployment examples and performance for UniAct-0.5B can be found in Fig 4. UniAct-0.5B surpasses the baseline models in all task suites, demonstrating a significant improvement with an overall average accuracy that is **17.2%** higher than the 7B OpenVLA and **33.6%** higher than Octo. This superior performance can be attributed to the ability of UniAct to bridge domain gaps and extract generalizable atomic behaviors. By leveraging demonstrations from various domains to learn the universal actions, UniAct significantly enhances task performance on the LIBERO benchmarks.

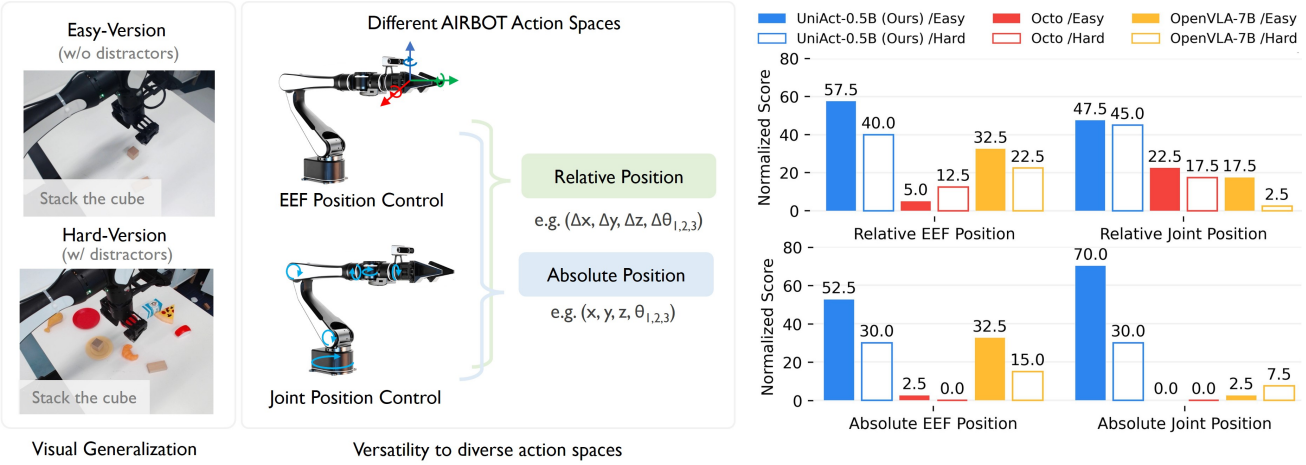

Figure 5. Fast adaptation to new robots. We fine-tune models on an unseen AIRBOT robot with different controller interfaces including relative/absolute joint/EEF positions. The task is to stack the small cube on the small cuboid, requiring precise action execution. The fine-tuning data is collected w/o any distractors, but we also evaluate a hard version, where the policy should be robust to diverse clutters.

4.3. Fast Adaptation to New Embodiment

Experiment Setup. To assess the fast adaptation ability, we evaluate on a new real-world robot, AIRBOT, with four drastically different controller interfaces: relative/absolute EEF position and relative/absolute joint position. Neither UniAct nor the baselines were pre-trained on AIRBOT data. We collect 100 demonstrations on this new robot platform with the four different types of control interfaces. Considering the significant heterogeneity among these control interfaces, we put a lot of effort into fine-tuning baselines and make sure the model convergence meets the official requirements (e.g., 95% prediction accuracy for OpenVLA). Please refer to Appendix B for details.

Fast Adaptation with UniAct. Unlike baseline models that require extensive training to bridge the adaptation gap across different action types, UniAct can rapidly adapt to new embodiments and control interfaces. Having already learned cross-embodiment behaviors, we facilitate fast adaptation by freezing the codebook and the universal action extractor. Concurrently, we train four heterogeneous decoding heads from scratch for each type of actions with the collected demonstrations. Each newly introduced head is implemented with a simple MLP that takes the u^* and vision features o from a shared vision backbone as inputs.

Evaluation. We conducted evaluations using both an easy and a hard version of the task "stack the cube on another cube". The results can be found in Figure 5. UniAct-0.5B demonstrates consistently strong generalization across all types of control signals, surpassing the two baseline models. Notably, the number of parameters UniAct-0.5B used for fine-tuning versus the total model size is the smallest (4M / 500M: 0.8%). In comparison, OpenVLA and Octo utilize 1.4% (97M / 7000M) and 2% (2M / 100M) of their total model sizes, respectively. This efficient parameter utilization highlights UniAct's effectiveness and adaptability, showcasing its superior performance in applying learned universal actions to new tasks and embodiments with minimal parameter space expansion.

4.4. In-depth Analysis of Universal Actions

In this section, we demonstrate that UniAct constructs a meaningful universal action space from two perspectives: 1) Consistent semantical behaviors are encoded as the same universal actions across diverse embodiments; 2) Universal action extractor can efficiently exploit this shared structure in the universal action space across different robots.

Interpretation of Universal Actions. We manually inspect the decoded behaviors for all 256 universal actions across different robots and observe at least 40% of them exhibit exact consistency. Fig 6 shows that the same universal action can be decoded back to consistent behaviors for different robots even with huge gaps. For instance, different robots with different viewpoints, and even those undergoing huge sim-to-real gaps, can execute similar semantical meaningful behaviors given the same universal action.

Control with the Universal Action. Therefore, we can directly interact with the robot to perform desired behaviors by choosing a sequence of universal actions. Fig 7 clearly demonstrates that we can control the robot using universal actions without any robotic knowledge, such as learning complex forward/inverse kinematics transformations. This also underscores the potential of utilizing the universal action extractor as an *action tokenizer* to facilitate future deployments of more advanced embodied foundation models by planning in this discrete universal action space.

Statistical Analysis For Universal Action Utilization.

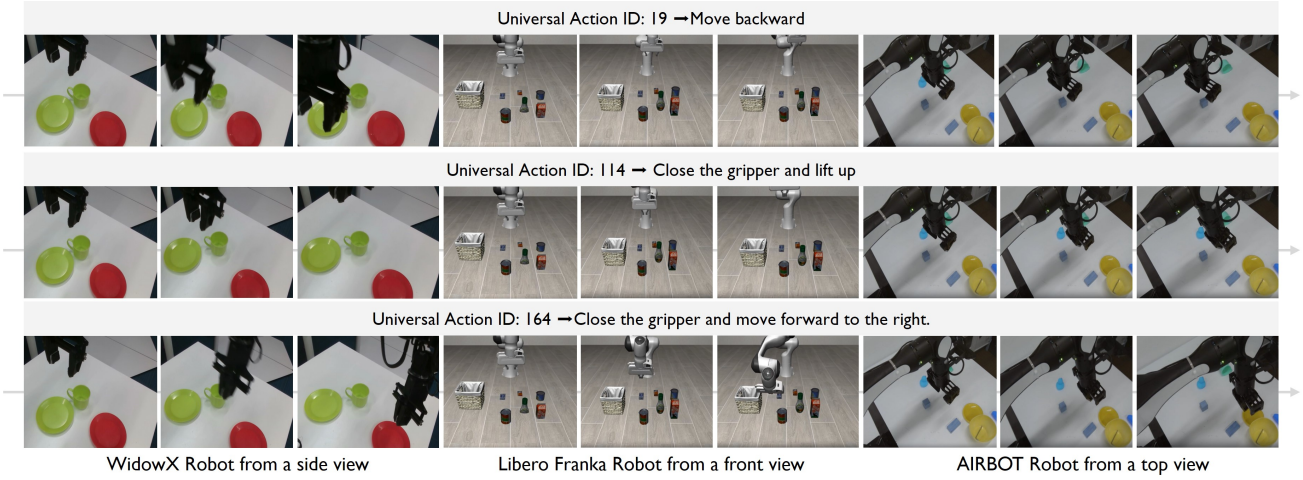

Figure 6. We manually check the decoded behavior for all 256 universal actions across different robots and observe at least 40% show exact consistency, even with varying viewpoints and significant domain gaps between the real world and simulation. Visualizations of repeated executions of the same universal action across different robots reveal clear, semantically meaningful, and consistent behaviors.

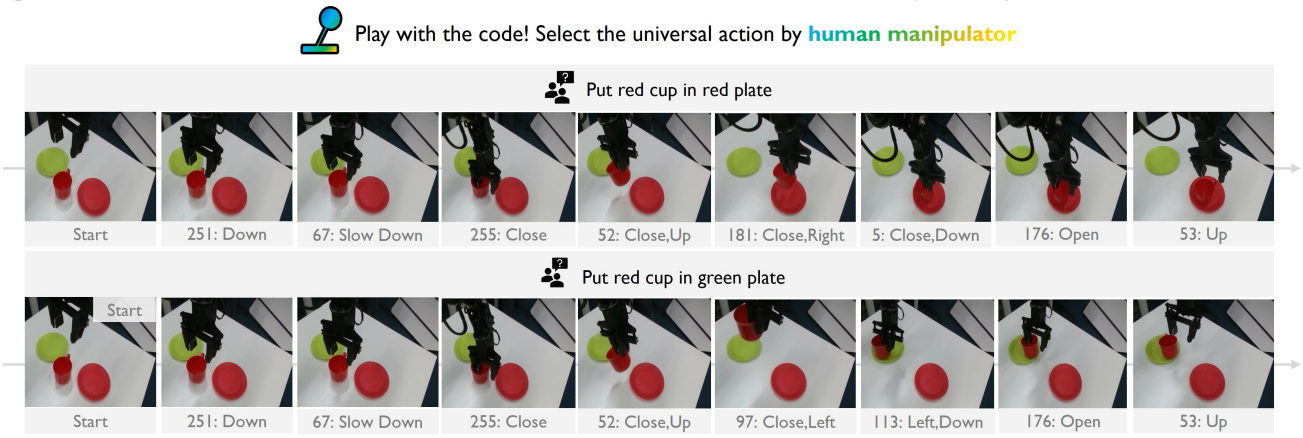

Figure 7. Since all universal actions encode semantic meaningful behaviors, we can directly play with it! By selecting the desired behavior using its corresponding universal action ID, we can manually control the robot to perform complex tasks such as `Pick & Place`.

	P_{WidowX}	P_{Franka}	O_{WidowX}	O_{Franka}
P_{WidowX}	0.	0.34	0.45	0.51
P_{Franka}	0.34	0.	0.60	0.58
O_{WidowX}	0.45	0.60	0.	0.44
O_{Franka}	0.51	0.58	0.44	0.

Table 1. We calculate the JS divergence of the universal action utilization distributions for two tasks in two distinct domains (lower means more consistent). P and O denote "pick up the bowl" and "open the drawer", respectively. WidowX and Franka denote the robot in the real world and simulation in Figure 3-4, respectively.

Here, we summarize the universal action utilization distributions for different tasks across different robots. Table 1 clearly shows that the utilization distributions for the same tasks and different robots are similar, but for different tasks and same robots are different, demonstrating that the universal action extractor indeed correctly exploits these embodiment-agnostic atomic behaviors by focusing more on task progressions over embodiment details.

5. Conclusion

We introduce UniAct, an innovative embodied foundation modeling framework that operates in a *Universal Action Space* to address the challenge of action heterogeneity. This universal action space encodes shareable atomic behaviors across diverse embodied action spaces to significantly enhance cross-domain data utilization and facilitate cross-embodiment generalization, enabling our 0.5B parameter model to outperform SOTA models that are 14 times larger. Also, the learned universal actions can be precisely translated to any embodiment-specific actions with minimal parameters through heterogeneous decoding, thus allowing for

fast adaptation to new robots possessing distinct control interfaces and physical properties. Moreover, our learned universal action extractor can also be used as a universal action tokenizer to power the construction of future large-scale embodied foundation models. Currently, UniAct is trained with a 0.5B parameter instantiation and evaluated mostly on single-arm robotics platforms due to resource constraints. Future work will focus on scaling up UniAct to larger models and extending its application to a broader range of embodiments, including bi-manual robots and even autonomous driving, further leveraging its versatile capability and effectiveness in more robotic applications.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2, 4

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 15

[3] Anurag Ajay, Seungwook Han, Yilun Du, Shuang Li, Abhi Gupta, Tommi Jaakkola, Josh Tenenbaum, Leslie Kaelbling, Akash Srivastava, and Pulkit Agrawal. Compositional foundation models for hierarchical planning. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[4] Anonymous. IGOR: Image-GOal representations are the atomic building blocks for next-level generalization in embodied AI. In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. under review. 2, 3, 4

[5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1, 2, 4

[6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 4

[7] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Rich Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In *The Twelfth International Conference on Learning Representations*. 2, 15

[8] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. 1, 2, 3

[9] 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. *arXiv preprint arXiv:2212.06817*, 2022. 2

[10] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 2

[11] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 3, 4

[12] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024. 2

[13] Xi Chen, Ali Ghadirzadeh, Tianhe Yu, Jianhao Wang, Alex Yuan Gao, Wenzhe Li, Liang Bin, Chelsea Finn, and Chongjie Zhang. Lapo: Latent-variable advantage-weighted policy optimization for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:36902–36913, 2022. 3, 4

[14] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668. 1, 5

[15] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 2

[16] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 2

[17] Ria Doshi, Homer Rich Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. In *8th Annual Conference on Robot Learning*. 1, 2, 3

[18] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8469–8488, 2023. 1, 2

[19] Yilun Du, Sherry Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, Leslie Pack Kaelbling, et al. Video language planning. In *The Twelfth International Conference on Learning Representations*. 2, 15

[20] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 653–660. IEEE, 2024. 1, 2

[21] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1): 3094, 2022. 2

[22] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12462–12469. IEEE, 2024. 2

[23] Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard Turner. Meta-learning probabilistic inference for prediction. In *International Conference on Learning Representations*, 2019. 5

[24] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 2

[25] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[27] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4

[28] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022. 1

[29] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023. 2

[30] 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, 5

[31] 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. 1, 2, 3, 5, 6, 12, 13, 15

[32] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2

[33] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022. 3

[34] Seungjae Lee, Yibin Wang, Haritheja Etukuru, H. Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Behavior generation with latent actions. In *Forty-first International Conference on Machine Learning*, 2024. 1, 2, 3

[35] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 5, 12

[36] Jianxiong Li, Jinliang Zheng, Yinan Zheng, Liyuan Mao, Xiao Hu, Sijie Cheng, Haoyi Niu, Jihao Liu, Yu Liu, Jingjing Liu, Yaqin Zhang, and Xianyuan Zhan. Decisionnce: Embodied multimodal representations via implicit preference learning. In *Forty-first International Conference on Machine Learning*. 2

[37] Jianxiong Li, Zhihao Wang, Jinliang Zheng, Xiaoai Zhou, Guanming Wang, Guanglu Song, Yu Liu, Jingjing Liu, Ya-Qin Zhang, Junzhi Yu, et al. Robo-mutual: Robotic multimodal task specification via unimodal learning. *arXiv preprint arXiv:2410.01529*, 2024. 2

[38] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36, 2024. 5, 6, 14

[39] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2

[41] 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. 1, 2, 3

[42] Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. Liv: Language-image representations and rewards for robotic control. In *International Conference on Machine Learning*, pages 23301–23320. PMLR, 2023. 2

[43] J. MacQueen. Some methods for classification and analysis of multivariate observations. 1967. 3

[44] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *The Twelfth International Conference on Learning Representations*. 3

[45] Atharva Mete, Haotian Xue, Albert Wilcox, Yongxin Chen, and Animesh Garg. Quest: Self-supervised skill abstractions for learning continuous control. *arXiv preprint arXiv:2407.15840*, 2024. 2, 3

[46] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. EmbodiedGPT: Vision-language pre-training via embodied chain of thought. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 15

[47] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *6th Annual Conference on Robot Learning*. 2

[48] Haoyi Niu, Jianming Hu, Guyue Zhou, and Xianyuan Zhan. A comprehensive survey of cross-domain policy transfer for embodied agents. In *The 33rd International Joint Conference on Artificial Intelligence*, 2024. 1

[49] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 1, 2, 3, 5

[50] Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with hilbert representations. In *Forty-first International Conference on Machine Learning*. 4

[51] Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. Hiql: Offline goal-conditioned rl with latent states as actions. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[52] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 4

[53] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020. 5

[54] Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *The Twelfth International Conference on Learning Representations*. 3

[55] Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Arty Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone. *Advances in neural information processing systems*, 35:22955–22968, 2022. 1, 2, 3

[56] Lucy Xiaoyang Shi, Zheyuan Hu, Tony Z Zhao, Archit Sharma, Karl Pertsch, Jianlan Luo, Sergey Levine, and Chelsea Finn. Yell at your robot: Improving on-the-fly from language corrections. *arXiv preprint arXiv:2403.12910*, 2024. 2, 15

[57] 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. 15

[58] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[59] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 2

[60] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 4

[61] 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. 1, 2, 3, 5, 12

[62] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2

[63] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 3, 4, 13

[64] 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*, pages 1723–1736. PMLR, 2023. 1, 2, 5, 6, 12

[65] Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *arXiv preprint arXiv:2409.20537*, 2024. 1

[66] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023. 2, 15

[67] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *The Twelfth International Conference on Learning Representations*. 2

[68] Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, Quan Vuong, Chelsea Finn, Dorsa Sadigh, and Sergey Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. *arXiv preprint arXiv:2402.19432*, 2024. 1, 2, 3

[69] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024. 2, 3, 4

[70] Haoqi Yuan, Zhancun Mu, Feiyang Xie, and Zongqing Lu. Pre-training goal-based models for sample-efficient reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[71] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. In *8th Annual Conference on Robot Learning*, 2024. 15

[72] Jinliang Zheng, Jianxiong Li, Sijie Cheng, Yinan Zheng, Jiaming Li, Jihao Liu, Yu Liu, Jingjing Liu, and Xianyuan Zhan. Instruction-guided visual masking. *Advances in neural information processing systems*, 2024. 2

A. Training Details

Training hyper-parameters UniAct-0.5B utilizes the pre-trained parameters from LLava-One-Vision-0.5B [35] to initialize the VLM. For visual feature extraction in heterogeneous decoding heads, we deploy an ImageNet pretrained ResNet18, which is commonly employed in vision-based policy learning [64]. This model was jointly trained during the large-scale pre-training process to enhance perceptual capability in manipulation task scenarios. We adopt resolutions of 384×384 for the VLM and 224×224 for the ResNet18, consistent with their original configurations. Image augmentation settings and more hyper-parameters can be found in Table 2 and Table 3, respectively.

Augmentation	value
RandomResize	ratio=(0.75, 1.3333) scale=(0.5, 1.0) interpolation=BICUBIC
RandomHorizontalFlip	p=0.5
ColorJitter	contrast=(0.6, 1.4) brightness=(0.6, 1.4) saturation=(0.6, 1.4)

Table 2. Image augmentation settings during training

config	value
optimizer	AdamW
batch size	1024
learning rate	2×10^{-5}
weight decay	0.
optimizer momentum	β_1, β_2=0.9,0.95
iters	500K
model precision	BFloat16

Table 3. Training hyper-parameters

Data construction As illustrated in Table 4, we detail the composition of our training data, which includes the number of trajectories, samples, and the control interfaces for the 28 distinct embodiments. Following previous works [31, 61], we assign different sampling rates to each dataset during training to ensure a balanced mix of embodiments, tasks, and scenes. These sampling rates are specified in Table 4. It is important to note that many of the datasets may contain distinctly action spaces such as EEF position and Joint position. In addition, images from different datasets may contain multiple view points. By default, we use only the first third-person perspective following [31, 61] to maintain consistency.

Dataset	Trajectory	Samples	Sample rate(%)	Control Interface
Utaustin Mutex	1500	361871	1.0	EEF Position
Berkeley Cable Routing	1557	38789	0.2	EEF velocity
NYU Franka Play	454	44412	1.0	EEF velocity
Kuka	557893	7130157	12.7	EEF Position
Austin Sailor	240	352110	2.2	EEF velocity
Fmb	8611	1136907	1.0	EEF velocity
Berkeley Autolab	1000	95310	1.2	EEF Position
Viola	150	75614	0.9	EEF Position
Dobbe	5208	1139874	1.4	EEF Position
Iamlab CMU	631	146241	0.9	EEF Position
Austin Buds	50	34110	0.2	EEF Position
Language Table	441911	6602077	4.4	EEF Position
Stanford Hydra	569	357137	4.4	EEF Position
Robo Set	18246	1419838	5.0	Joint position
Austin Sirius	559	279724	1.7	EEF velocity
Dlr Edan	104	8824	0.1	EEF Position
Fractal	86599	3607028	12.7	EEF Position
TOTO	1003	324669	2.0	Joint position
Berkeley Fanuc	415	58660	0.7	EEF Position
CMU Stretch	135	25012	0.2	EEF Position
Roboturk	1934	186910	2.3	EEF Position
Jaco Play	976	66094	0.4	EEF Position
Taco Play	3603	230966	3.0	EEF Position
BC-Z	42811	5957097	7.5	EEF Position
Droid	92115	27043929	10.0	EEF Position
Furniture Bench	5100	3905717	2.4	EEF velocity
Bridge	28933	899685	13.3	EEF Position
Libero	6500	1007204	5.0	EEF Position

Table 4. Details of the training data composition, including the number of trajectories, number of samples, and control interfaces of the 28 different data sources. Following [31], we set different sampling probabilities for different data, which we also report in this table.

Categorical Reparameterization. Except for the Gumbel-Softmax utilized for training UniAct, we also explored another commonly used reparameterization technique: the Straight-Through Estimator (STE) [63]. However, empirical findings indicate that using STE can lead to severe collapses in the universal action codebook. An intuitive explanation for this is that each universal action requires numerous optimization steps to learn the highly abstracted behaviors. Since STE involves hard sampling, it results in only one universal action being selected and optimized for each training sample. This lack of gradient distribution among all potential actions can stifle the learning process and reduce the diversity of learned actions.

B. Evaluation Setups

In this section, we delve deeper into the evaluation experiments for the three robotic embodiments used in our study: WidowX Robot, Franka Robot, and AIRBOT. We provide detailed scores for each embodiment to offer a clearer, more intuitive comparison of their performance.

B.1. WidowX Robot in Real World

The experiments on WidowX aim to assess the models' generalization capabilities across five distinct dimensions, as illustrated in Figure 8. Instead of merely reporting task success rates, we calculate scores based on the progress made towards completing the task for some complex tasks. This scoring method can more intuitively reflect the model's understanding and generalization ability [31]. Detailed scores are available in Table 5. In the subsequent sections, we will provide a comprehensive description of the task settings.

Visual Generalization: This dimension evaluates the model's ability to adapt to different visual environments characterized by variations in lighting, background, and object textures. Following the setups outlined in [31], we design three specific tasks, detailed in Figure 8. Each task is set against a distinct background environment, features dif-

ferent lighting brightness levels, and involves target objects in two colors: red and green. Additionally, we introduce various visual distractors unrelated to the task to assess the model's ability to generalize visually and maintain robustness against such distractions. To ensure a fair comparison, all variables related to the model's testing conditions are kept consistent across all models. All tasks in this suite are assigned binary score: 0 for failure and 1 for success.

Motion Generalization: Tasks in this dimension aim to evaluate the robot's capability to perform appropriate motions while recognizing the target object's position and orientation, which may not have been encountered during training. Specifically, we have designed two tasks: *Lift eggplant* and *Put carrot on plate*, as detailed in Figure 8. The target objects are placed in several predetermined positions and oriented in pre-designed directions. This helps in assessing the robot's adaptability to changes in physical task parameters. All tasks in this suite are assigned binary score: 0 for failure and 1 for success.

Physical Generalization: In this dimension, we examine the model's capability to manage physical variations in objects, such as size, weight, and material properties. The tasks are crafted to evaluate the robot's adaptability to these fluctuations, which significantly influence manipulation strategies. Detailed task setups are listed in Figure 8. The objects involved in these tasks, such as carrots and AAA batteries, often have irregular shapes or are placed in unusual positions (e.g., an overturned pot). Successfully handling and completing tasks with these objects requires a generalized policy that can accurately recognize the physical attributes of the target objects and execute appropriate strategies. All tasks in this suite are assigned binary score: 0 for failure and 1 for success.

Semantic Generalization: In this dimension, we assess the model's ability to understand and generalize across various semantic contexts. Specifically, the tasks especially the target objects included in this dimension have never been encountered during the training procedures for both UniAct and the baseline models. This approach tests the model's capability to interpret and adapt to new instructions and environments, emphasizing its flexibility and learning efficiency. The purple grape is highly smooth, so we assign scores as follows: 0.25 for touching the grape, 0.5 for grasping it, 0.75 for moving it toward the pot, and 1 for successfully completing a pick-and-place. Also, for the stack green cup on red cup task, we assign scores as follows: 0.25 for touching the green cup, 0.5 for grasping it, 0.75 for moving it toward the red cup, and 1 for successfully stacking. Other tasks in this suite are assigned binary scores.

Language Grounding: This dimension evaluates the model's ability to comprehend and execute commands that are grounded in natural language. The focus is on assessing the model's proficiency in following novel instructions to manipulate specific objects as described verbally. While there are similarities with Semantic Generalization in terms of handling unseen scenarios, Language Grounding task distinctly tests the model's capacity to accurately understand and act on language-based directives within potentially misleading environments. An illustrative example is placing two cups of different colors on a table and instructing the model to manipulate one of them using language. The model must accurately ground the language instruction to the correct object in a complex real-world setting. This tests the model's ability to connect linguistic descriptions directly with physical actions in dynamic and visually diverse environments. For all tasks in this suite, we assign scores as follows: 0.5 for correct grounding, 1 for a successful task completion.

B.2. Franka Robot in Simulation

We include 6500 expert demonstrations of 130 different tasks collected with Franka Robot in LIBERO [38] simulation to train our UniAct-0.5B and then follow the LIBERO Benchmark [38] to evaluate our models. The input images are rendered by the emulator and we use the default 128×128 resolution. As the two open-source baseline models, OpenVLA and Octo, were not initially trained with the simulation data, substantial effort was put into fine-tuning them to facilitate fair comparisons with our UniAct framework. The fine-tuning process for OpenVLA and Octo was conducted on 8 A6000 GPUs and 2 4090 GPUs, lasting 7 hours and 4 hours, respectively. Details of the training hyperparameters are provided in Table 6. Notably, we manually cleaned the "no-op'" data before fine-tuning OpenVLA following its official guidance, a step that proved crucial for achieving convergence. However, as shown in Figure 9, even with an increased number of training steps, the model's action accuracy remained low. In contrast, our UniAct framework demonstrated robustness to noisy data.

B.3. Fast adaptation to AIRBOT

To assess the fast adaptation capability of UniAct-0.5B and baselines, we fine-tune them using newly collected demonstrations on an unseen embodiment during the pretraining phase, AIRBOT. In this section, we provide detailed information about the fine-tuning processes.

Fine-tuning Settings For UniAct: We utilized 4 A100 GPUs to fine-tune UniAct-0.5B with DeepSpeed. Notably, we train a new MLP network from scratch as the heterogeneous head for AIRBOT while keeping the other modules frozen. The fine-tuning was conducted over a span of 1 hours. The training hyper-parameters employed during this process are detailed in Table 7.

Fine-tuning Settings For Baseline Models. The fine-tuning for the baseline models, OpenVLA and Octo, was executed on 8 A6000 GPUs and 2 4090 GPUs, lasting 1.5

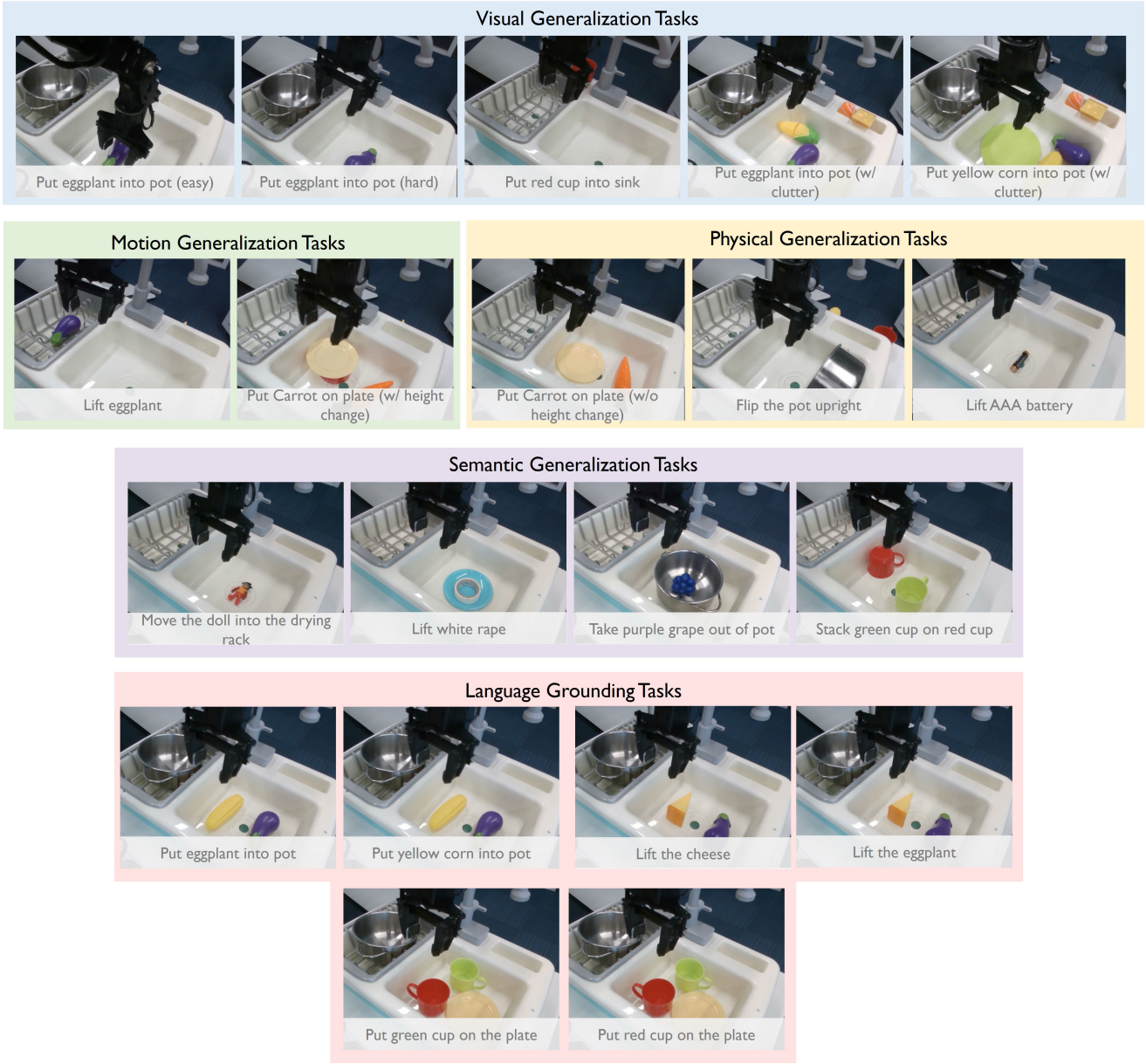

Figure 8. Illustration for WidowX evaluation tasks.

hours and 0.5 hours, respectively. To ensure robust performance, the training of OpenVLA continued until it reached an accuracy rate of 95% on the training dataset, aligning with the official requirements in [31]. Training hyperparameters for this process are illustrated in Tab 10.

C. More Related Works

Embodied Models with Hierarchical Structures. UniAct resembles a hierarchical-style structure, which firstly infers an universal action and then translate it into actionable actions. Some works also explore hierarchical structures, utilizing the planning capabilities of LLMs or VLMs to decompose original instructions into linguistic [2, 46, 56, 57, 71] or visual [7, 19, 66] sequences of subgoals. However, transferring these language or visual subgoals into precise actions still requires extensive action labels and struggles to leverage the shared structures across diverse action spaces. The universal action space in UniAct, however, serves as a more fine-grained skill library that can be efficiently adapted to physically grounded actions. This space is also end-to-end trained, maximizing the VLMs' capabilities to develop a flexible and comprehensive structure.

Task Category	Task Name	Octo	OpenVLA	UniAct-0.5B(ours)
Visual Generalization	*put eggplant into pot(easy)*	3.0	**10.0**	**10.0**
	put eggplant into pot	6.0	6.0	**10.0**
	put cup from counter into sink	0.0	**2.0**	1.0
	put eggplant into pot	3.0	**8.0**	5.0
	put corn on plate	0.0	7.0	**8.0**
	Average Score	2.4	6.6	**6.8**
Motion Generalization	*lift eggplant*	4.0	3.0	**6.0**
	put carrot on plate	2.0	4.0	**6.0**
	Average Score	3.0	3.5	**6.0**
Physical Generalization	*put carrot on plate*	2.0	**9.0**	**9.0**
	flip pot upright	1.0	2.0	**7.0**
	lift AAA battery	0.0	**7.0**	5,0
	Average Score	1.0	6.0	**7.0**
Sematic Generalization	*move doll into drying rack*	0.0	5.0	**6.0**
	lift white rape	0.0	1.0	**2.0**
	take purple grapes out of pot	4.5	**5.0**	**5.0**
	stack green cup on red cup	3.25	**8.25**	2.5
	Average Score	1.9	**4.8**	3.9
Language Grounding	*put eggplant into pot*	6.0	7.0	5.0
	put yellow corn into pot	7.0	10.0	10.0
	lift cheese	1.0	**8.0**	6.0
	lift eggplant	6.0	8.0	**9.0**
	put green cup on plate	3.0	**10.0**	**6.0**
	put red cup on plate	6.0	**10.0**	**8.0**
	Average Score	4.8	**8.8**	7.3

Table 5. Comparison between UniAct-0.5B and baseline models on detailed scores in WidowX evaluations.

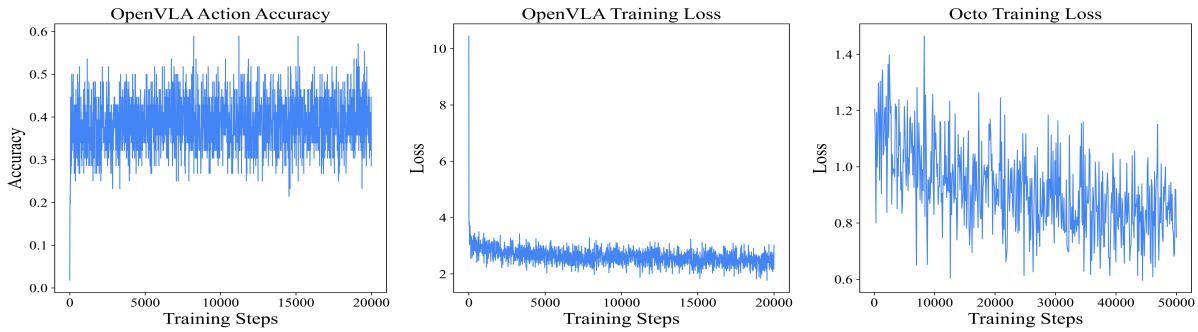

Figure 9. The learning curves for OpenVLA and Octo in LIBERO fintuning.

D. Limitations and Future Works

In this section, we discuss limitations in this work and the corresponding solutions. We hope this will inspire more interesting works.

More Embodiments: In this study, we explore the concept of a universal action space that can be shared across different action spaces. In this current version, UniAct is mostly evaluated with varied control interfaces on single robotic arms. The underlying motivation stems from an intuitive observation: despite differences in control interfaces, these robotic arms inherently exhibit common physical movements. However, this raises a crucial question: Is the commonality of physical movement exclusive to robotic arms or similar embodiments?

We argue that the answer is no. Future work will explore

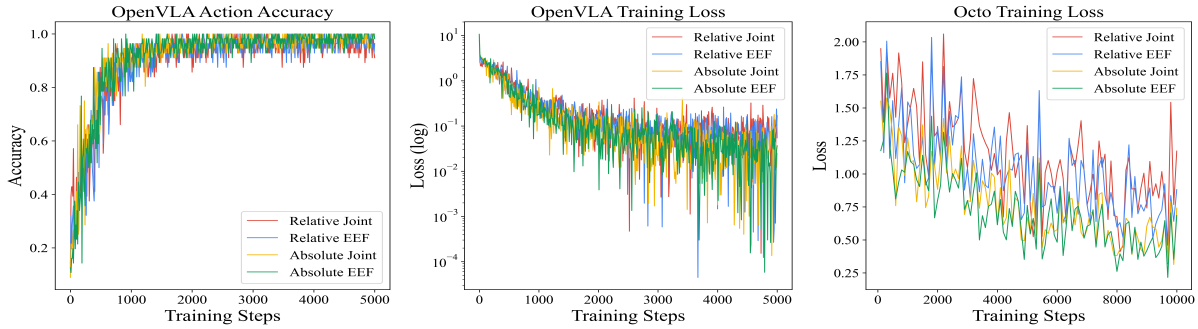

Figure 10. The learning curves for OpenVLA and Octo in AIRBOT finetuning.

Model-Task	Config	Value
OpenVLA-LIBERO	Optimizer	AdamW
	Batch size	64
	Learning rate	3×10^{-4}
	Weight decay	1×10^{-2}
	Optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$
Octo-LIBERO	Optimizer	AdamW
	Batch size	64
	Learning rate	3×10^{-4}
	Weight decay	1×10^{-2}
	Optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$
	Warmup iters	2000
OpenVLA-AIRBOT	Optimizer	AdamW
	Batch size	64
	Learning rate	3×10^{-4}
	Weight decay	1×10^{-2}
	Optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$
Octo-AIRBOT	Optimizer	AdamW
	Batch size	64
	Learning rate	3×10^{-4}
	Weight decay	1×10^{-2}
	Optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$
	Warmup iters	2000

Table 6. Fine-tuning hyper-parameters for baseline models in simulation.

config	value
optimizer	AdamW
batch size	128
learning rate	3×10^{-4}
weight decay	0.
optimizer momentum	β_1, β_2=0.9,0.95
iters	10K
model precision	BFloat16

Table 7. Hyper-parameters for fine-tuning UniAct-0.5B on AIRBOT

the universal actions for more complex embodiments, such as dual robotic arms, dexterous robotic hands, quadrupeds and even autonomous driving cars, which differ in degrees of freedom and mechanical structures. Despite their strong heterogeneity, these systems may also share some fundamental movements with simpler robotic arms, holding the potential to be incorporated into the same universal action space. Therefore, we hope to develop a "truly" universal action space that is capable of encoding movements across ANY physical embodiment while recognizing unique characteristics and prioritizing shared commonalities. This offers a promising direction for future research, with strong potential for enabling cross-embodiment control across diverse systems.

More Flexible Network Design: In our current implementation, UniAct utilizes identical decoding heads—a simple MLP network—for each embodiment to minimize the risk of over-fitting and facilitating training for the universal action extractor. However, it seems more reasonable that the complexity and number of parameters in the decoding heads should be tailored according to the control complexity of each embodiment and the diversity of its training data. For example, a bi-manual robot, with its increased degrees of freedom, would logically require a more complex decoding head than a single-arm robot to effectively learn and decode universal actions.

Looking forward, as the dataset expands to include more demonstrations from more diverse embodiments, it will become crucial to develop specialized decoding heads that consider the specific control complexities of each embodiment. Additionally, incorporating embodiment-specific information, such as proprioceptive data or different views, into these decoding heads could significantly enhance their performance.

Scaling Law For Universal Action Training: While our efforts to train UniAct with a vast array of open-source data have yielded commendable results, an intriguing question has arisen: Does more data unequivocally improve the universal action space? This question calls for a thorough examination from multiple perspectives. Firstly, does the incorporation of more embodiments inherently enhance the action space, or is there greater value in accumulating diverse task demonstrations for the same embodiment? More-

over, it's critical to consider whether the more data always bring better result.

Studying the relationship between the amount of data and how well our universal action space performs could be very useful. By understanding these relationships, we can better plan our data collection to improve model training effectively and efficiently.

More Utilization of Universal Action. Our work has highlighted the robust capabilities of universal actions in deploying cross-embodiment robot policies. However, we believe that the potential applications of universal actions extend far beyond what has been explored so far.

One promising direction is in the development of world models, where universal actions serve as a form of 'tokenizer'. This role involves breaking down complex actions into standardized, understandable components. By employing a universal action space for planning, these world models can more accurately predict and simulate outcomes across various scenarios and environments. This uniform approach to understanding and interacting with the world is invaluable for advanced planning and decision-making in robotics.