# Universal Actions for Enhanced Embodied Foundation Models

增强实体基础模型的通用动作

Jinliang Zheng\* , Jianxiong Li\* , Dongxiu Liu\* , Yinan Zheng ,

郑金亮\* ,李建雄\* ,刘东旭\* ,郑一楠 ,

Zhihao Wang , Zhonghong Ou , Yu Liu , Jingjing Liu , Ya-Qin Zhang , Xianyuan Zhan

王志浩 ,欧中宏 ,刘宇 ,刘晶晶 ,张亚勤 ,詹先元

AIR, Tsinghua University, Sensetime Research, Peking University

清华大学AIR, 商汤研究院, 北京大学

Beijing University of Posts and Telecommunications, Shanghai AI Lab

北京邮电大学, 上海人工智能实验室

{zhengjl23, li-jx21}@mails.tsinghua.edu.cn

{zhengjl23, li-jx21}@mails.tsinghua.edu.cn

zhanxianyuan@mail.tsinghua.edu.cn

# Abstract

# 摘要

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. Project page: https://github.com/2toinf/UniAct

在多样化的互联网规模数据上进行训练是近期大型基础模型成功的关键因素。然而，使用相同的方法构建实体代理却面临显著困难。尽管有许多众包的实体数据集可用，但由于不同机器人的物理实体和控制接口的差异，它们的动作空间往往表现出显著的异质性，这给使用跨领域数据开发实体基础模型带来了巨大挑战。在本文中，我们介绍了UniAct，一种在标记化通用动作空间中运行的实体基础建模框架。我们学习的通用动作通过利用不同机器人的共享结构特征，捕捉了跨机器人的通用原子行为，并通过消除臭名昭著的异质性，增强了跨领域数据利用和跨实体泛化能力。通过简单地添加实体特定的细节，通用动作可以高效地转换回异构的可执行命令，从而快速适应新机器人变得简单直接。我们的0.5B实例化UniAct在各种现实世界和仿真机器人的广泛评估中优于14倍大的SOTA实体基础模型，展示了卓越的跨实体控制和适应能力，凸显了采用通用动作的关键优势。项目页面:https://github.com/2toinf/UniAct

# 1. Introduction

# 1. 引言

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 . 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].

在自然语言处理和计算机视觉等领域，基于大量多样化数据源训练的基础模型展示了显著的成功和强大的泛化能力，凸显了学习通用模型相对于任务特定模型的优势 。受这些成功的启发，开发能够处理跨任务、跨环境和跨实体泛化的多功能实体基础模型，为构建通用实体代理提供了一条有前景的途径 [8, 17, 18, 28, 31, 61, 68]。

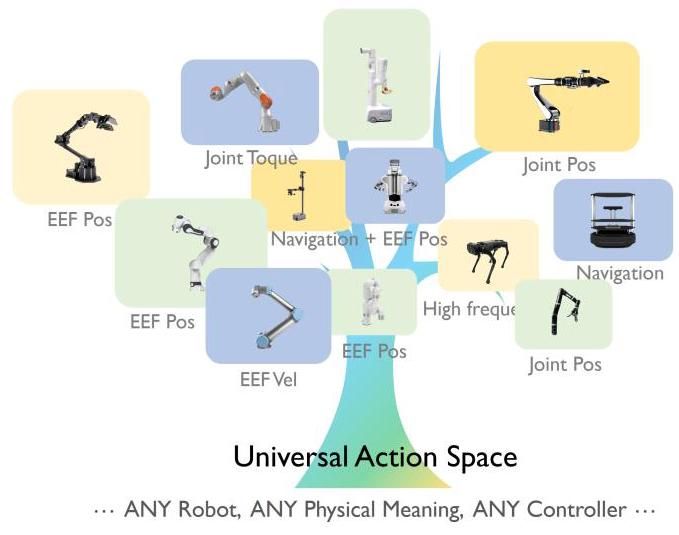


Figure 1. Illustration of the Universal Action Space. Universal actions should be versatile to ANY domain-specific actions.

图1. 通用动作空间的示意图。通用动作应适用于任何特定领域的动作。

However, significant challenges arise from the substantial heterogeneity of embodied data . 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 multimodality 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) , significantly complicating data sharing across different data sources.

然而，实体数据的显著异质性带来了重大挑战 。这种异质性不仅体现在由相机位置(例如，手腕或第三人称视角)和环境条件(例如，光照或背景变化)引起的视觉差异上，更关键的是体现在动作异质性上 [17, 48, 65]。1) 具有不同实体(例如，不同自由度或机器人手臂、四足动物和汽车之间的区别)的机器人拥有完全不同的动作空间 [68]。2) 此外，控制接口的多样性(例如，机器人手臂的末端执行器(EEF)位置或速度控制器)导致动作命令的物理意义根本不同 [49]。3) 即使动作是从同一机器人平台但由不同人类操作员收集的，人类行为的多模态性也加剧了这种异质性 [14, 34, 41, 55]。因此，跨不同机器人和机构收集的实体动作数据往往位于原始物理空间(例如，末端执行器的位置和旋转)中很大程度上不相交的流形上 ，这显著复杂化了不同数据源之间的数据共享。

[[1]](#footnote-28)

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 . 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.

目前，尚无现有解决方案能够充分解决动作异质性问题。大多数先前的研究强行将不同的动作空间视为等效，并应用相同的离散化或归一化技术，导致潜在冲突的动作空间，其中相似的动作编码可能代表完全不同的物理意义 。尽管一些努力尝试通过简单聚合所有个体动作空间来设计一个适用于各种机器人系统的物理可解释动作空间[17, 41]。这需要大量的人工工程努力，并且未能揭示和利用不同具身动作空间之间的内在联系，阻碍了通用具身基础模型的有效开发。

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.

在本文中，我们介绍了UniAct(具身基础模型与通用动作)，这是一种新颖的具身建模框架，构建在通用动作空间而非麻烦的异质动作空间中。在UniAct中学习的通用动作编码了可跨多种机器人平台推广的通用原子行为，而不受特定具身机制和控制接口的限制。例如，不同的机器人在面对正前方的目标时，应执行相似的“向前移动”行为，尽管由于它们的具身差异而表现出完全不同的控制信号。这种抽象超越了特定的具身和控制约束，使其能够普遍适用于各种机器人平台，为跨具身数据利用和模型泛化提供了显著潜力。从这个意义上说，最少的参数和数据足以将通用动作解码为特定具身的动作，因为通用运动行为已在通用空间中被捕获，解码器只需为每个机器人平台添加一些具身细节，从而在部署期间实现跨新机器人系统的高效适应。

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 , 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.

具体而言，UniAct采用共享的视觉语言模型(VLM)[6, 32, 35, 39, 40, 72]将通用动作空间构建为向量量化码本[63]。类似于可学习技能库 ，每个码封装了一个原子行为，足够通用，可由多种机器人执行。这种设置充当了关键的信息瓶颈，推动VLM识别和利用跨不同动作空间的共享原始行为。这种提取方案实现了跨具身控制行为的有效泛化，使我们的0.5B实例UniAct-0.5B在广泛任务中超越了参数规模14倍的模型，如具有7B参数的OpenVLA [31]。通过简化的异质解码器，可以将导出的通用动作转换为各种具身的精确可操作命令。这些解码器将通用动作作为条件输入，并通过其独特的观测数据增强具身特定特征。这允许根据特定需求进行灵活定制，例如包含或排除本体感觉特征或相机视图数量的变化。通过简单地添加新头作为新任务的轻量级解码器，可以快速适应新领域或机器人平台。我们在具有挑战性的任务设置上的全面评估，包括大视角变化和训练数据中未见过的机器人，证实了UniAct的显著可迁移性，展示了在通用动作空间中开发具身基础模型相对于传统异质空间的巨大优势。

# 2. Related Work

# 2. 相关工作

Multimodal Foundation Models. Large Language Models (LLMs) 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.

多模态基础模型。大型语言模型(LLMs) 在各种任务中展示了显著的能力，展示了令人印象深刻的零样本和上下文学习能力[16]。在此基础上，通过将视觉和语言整合到统一的标记化空间中，开发了大型视觉语言模型(VLMs)，展示了出色的多模态指令跟随能力[6, 21, 32, 35, 58, 59, 72]。它们的成功主要归功于广泛的互联网规模预训练，利用了来自互联网的广泛且多样化的高质量数据语料库。

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 , 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 , . 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 , as the action data gathered from any single robot platform is far less comprehensive than crowed-sourced data collected globally [20, 30, 49].

具身基础模型。在开发具身基础模型时，训练过程中会加入一个额外的关键模态——动作(机器人可以解释和执行的可部署控制信号，例如末端执行器位置/速度)。最先进的模型通常构建为视觉语言动作模型(VLA)[9, 10, 18, 31, 49]，将视觉和语言输入与可操作输出相结合。然而，从不同机器人平台和实验室收集的动作标签表现出显著的异质性 ，阻碍了不同来源之间的有效数据共享。为了规避这一挑战，许多工作采用大规模无动作的视觉语言数据，例如领域外的人类活动[15, 24, 25]，首先获得一个良好的具身VLM，然后根据特定机器人平台的一小部分动作标签对其进行微调 ， 。虽然这些方法可以提高特定机器人在狭窄任务集上的样本效率，但它们在构建通用具身代理方面存在严重的性能瓶颈 ，因为从任何单一机器人平台收集的动作数据远不如全球收集的众包数据全面[20, 30, 49]。

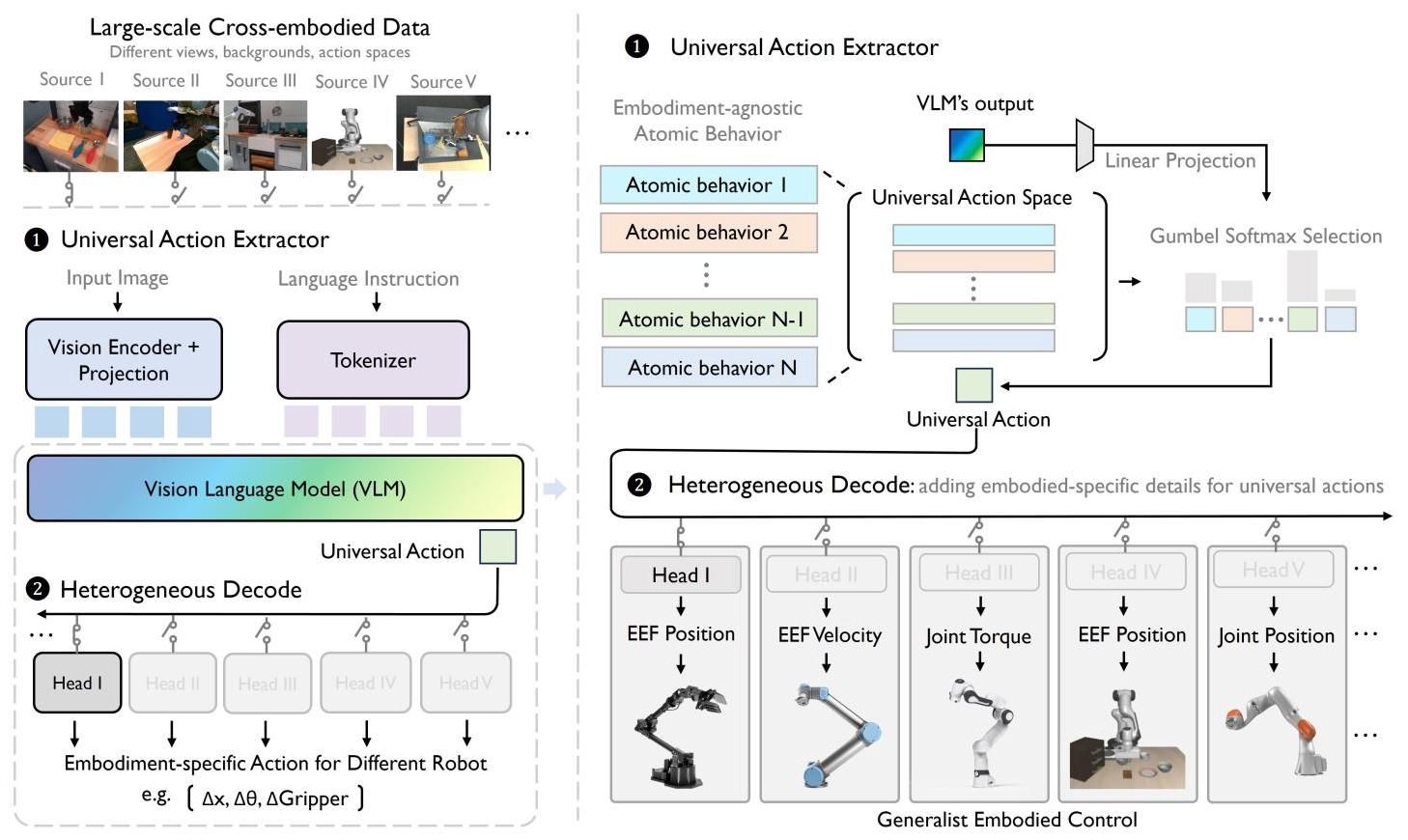


Figure 2. UniAct leverages diverse data sources for generalist embodied control. A shared Vision Language Model (VLM) extracts transferable features across various data sources; The output tokens are converted into universal actions, represented as vector-quantized codes where each code captures common 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.

图2. UniAct利用多样化的数据源进行通用具身控制。共享的视觉语言模型(VLM)提取跨各种数据源的可转移特征；输出标记被转换为通用动作，表示为向量量化代码，其中每个代码捕捉不同机器人之间的共同原子行为；通过Gumbel Softmax选择的通用动作随后通过不同的头转换回特定命令，每个头编码个体机器人的具身特定特征。

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 Open-VLA [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], [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 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 .

最近的一些工作利用丰富的异质动作标签来开发用于跨具身控制的通用机器人策略。RT-X系列[49]、Octo[61]和Open-VLA[31]利用来自不同7自由度机器人的数据来增强在单一机器人源上训练的模型的泛化能力。更进一步，CrossFormer[17]、RDT[41]、 [8]和Yang等人[68]探索了使用具有完全不同机械结构的机器人数据的潜力，例如操作和导航中的机器人，以及单臂与双臂系统的数据。然而，现有工作要么忽略了不同来源动作空间的异质特性，粗暴地将它们视为等同而不考虑其内在冲突[31, 49, 61]，要么天真地将所有动作空间聚合在一起，未能利用不同机器人之间的潜在共享共性 。

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 , . 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.

具有潜在动作空间的具身模型。我们的工作旨在提取一个多功能的通用动作空间，类似于潜在空间，但编码了跨各种机器人平台的共同原子控制行为和模式。一些工作在潜在空间中开发具身模型 ， 。其中，LAPA[69]、IGOR[4]和LAPO[54]通过在无动作视频上联合自监督训练逆动力学模型和前向动力学模型来开发潜在动作空间[11]。然而，以这种方式提取的潜在动作主要集中在解释视频帧之间的变化，缺乏具身考虑或与实际控制信号的直接因果关系。为了理解为什么这有问题，假设我们在机器人面前添加一个新物体，视觉输入会发生变化，但这与控制行为无关，理想的编码动作不应捕捉这种分散注意力的信息。BeT[55]、VQ-BeT[34]和QueST[45]也通过K-means聚类[43]或向量量化[33, 44, 63]构建了一个离散的动作代码本，其中代码本中的每个代码编码了动作标签的不同聚类中心。这些工作主要集中在单一具身类型的简单领域，这增强了建模具有多种模式的复杂人类演示的能力，但难以解决跨不同具身的动作异质性问题。相比之下，我们的通用动作集成了来自与具身无关的语言模态的目标信息，并对实际动作信号进行监督，提供了一个多功能和抽象的技能库，以促进跨具身共享。此外，我们的研究深入探讨了更复杂的异质环境，并开发了一个大型具身基础模型，超越了先前研究中考虑的有限范围。

# 3.The UniAct Framework

# 3.UniAct框架

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 desirable 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.

我们介绍了UniAct，一个旨在通用动作空间中运行的具身基础建模框架，擅长弥合领域差距并促进大规模异构数据的训练。我们首先讨论了通用动作的理想特性，然后详细讨论了从异构跨具身数据中提取和解码通用动作的模型架构和学习方案。

# 3.1. Universal Action Space

# 3.1. 通用动作空间

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.

理想的通用动作空间是，尽管来自各种具身的异构控制信号驱动的所有运动具有不同的物理意义，但它们可以被提炼为共享的潜在原子行为。我们将这些抽象的行为表示称为通用动作，它们在所有物理具身中共享。

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 and Vector Quantized Variational Autoencoders [52, 63]. In this paper, we model the universal action space as and implement it with a vector quantized codebook [63], represented as

我们特别感兴趣的是探索离散的通用动作空间。这是由离散表示在复杂推理、规划和预测学习中的强大能力所驱动的，正如LLMs和VLMs 以及向量量化变分自编码器 [52, 63] 的成功所证明的。在本文中，我们将通用动作空间建模为 并使用向量量化码本 [63] 实现，表示为

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

其中 是空间大小，每个 是一个 维向量嵌入，表示一个通用的原子行为。

Several prior studies 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:

一些先前的研究 追求了通过推断两个视觉状态之间观察到的动态变化来构建通用的潜在动作的类似概念。然而，这种方案存在两个关键限制，导致次优且嘈杂的通用动作空间:

* 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.
* 观察到的变化不仅包括机器人的控制结果，还包括与实际控制没有因果联系的外部因素(例如环境变化、新物体的出现、人为干预等)。
* The interval between two observations critically influences the semantic interpretation of the extracted atomic behaviors, making standardizing behavioral interpretation complicated across different data sources.
* 两次观察之间的间隔对提取的原子行为的语义解释有重要影响，使得在不同数据源之间标准化行为解释变得复杂。

# 3.2. Universal Action Extraction

# 3.2. 通用动作提取

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 of selecting universal action given observation and task goal (e.g., language instruction). We want the corresponding adopted universal action that matches the atomic behavior encoded in the embodied data to satisfy:

为了获得理想的通用动作空间，我们提出了一种新的提取通用动作的方法，不再仅仅关注解释观察到的变化，而是更多地理解任务进展。具体来说，我们微调了一个大型视觉语言模型作为通用动作提取器，它输出在给定观察 和任务目标 (例如语言指令)下选择通用动作 的可能性 。我们希望采用的相应通用动作 与具身数据中编码的原子行为相匹配，以满足:

Akin to planning in the latent space , the extractor aims to infer the most relevant universal action for solving a given task under the observation , 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 to capture the generic and shared atomic behaviors across all domains.

类似于在潜在空间 中进行规划，提取器旨在推断在观察 下解决给定任务 的最相关通用动作，从而创建与任务进展直接相关的通用动作，而不仅仅是识别嘈杂的观察变化。我们使用VLM来实现这一目的，因为它具有强大的视觉语言推理能力。此外，微调预训练的VLM也大大提高了学习通用动作时的样本效率。该提取器为跨领域泛化创建了一个关键的信息瓶颈，因为不同的机器人被迫使用相同的离散码本 来捕捉所有领域中的通用和共享的原子行为。

To implement this, however, the non-differentiable 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:

然而，为了实现这一点，不可微的 阻碍了梯度传播。因此，我们在训练期间使用分类重参数化，利用Gumbel-Softmax技术来促进梯度估计 [27]。具体来说，前向过程如下:

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

其中 是每个通用动作 的权重，使用Gumbel Softmax函数计算:

Here, is a Gumbel noise sampled from the Gumbel distribution, and 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 during the training process. Our proposed universal action extractor is illustrated in Figure 2, please refer to the Appendix A for more details.

这里， 是从Gumbel分布中采样的Gumbel噪声， 是用于平滑概率分布的温度。为了在早期训练阶段促进参数空间探索和模型收敛的稳定性，我们在训练过程中逐渐降低温度 。我们提出的通用动作提取器如图2所示，更多细节请参见附录A。

# 3.3. Heterogeneous Decoding

# 3.3. 异构解码

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 is number of training domains. Each head is specifically designed to learn the mapping from universal action and visual observation to heterogeneous control signals for the embodiment in domain . The operation of each decoder head can be formulated as:

为了有效地将通用动作空间中的高度抽象行为转化为精确的、特定于具体实现的控制信号，整合更多实现细节(如控制类型、本体感觉和独特观察)至关重要。为此，我们引入了一系列轻量级解码器头，以适应每种类型的实现，记为 是训练域的数量。每个头 专门设计用于学习从通用动作 和视觉观察 到域 中实现的异构控制信号的映射。每个解码器头 的操作可以表示为:

where 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 and visual features 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.

其中 是预测的控制信号。由于参数过多的复杂解码头可能会过拟合目标域的数据分布，所有异构头都实现为简单的MLP网络，它们以 和共享视觉骨干提取的视觉特征 作为输入。通过保持解码头轻量级，我们确保大部分学习是针对通用动作进行的，从而最大限度地提高不同实现之间的泛化能力。

# 3.4. Training Procedure

# 3.4. 训练过程

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 heterogeneous datasets . Each comprises a set of robotic control trajectories, represented as , where is the -th trajectory of maximum length , containing observations, actions, and goals. Concretely, Uni-Act ingests and as inputs to predict the universal action using the universal action extractor, which is then mapped along with to using the heterogeneous decoding heads. The overall training objective is as follows:

UniAct的主要学习目标是提炼一个跨多种实现共享的通用动作空间，其关键特征是这些通用动作可以精确地转换回特定域的控制信号。为了实现这一点，模型使用 个异构数据集 的综合集合进行训练。每个 包含一组机器人控制轨迹，表示为 ，其中 是最大长度为 的第 条轨迹，包含观察、动作和目标。具体来说，Uni-Act将 和 作为输入，使用通用动作提取器预测通用动作 ，然后将其与 一起映射到 ，使用异构解码头。整体训练目标如下:

Here, 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 as well as the parameters of the universal action extractor and all heterogeneous decoding heads. Importantly, while and the universal action extractor are concurrently updated throughout each training iteration, the heterogeneous heads 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.

这里， 是行为克隆损失，可以根据每个数据集中动作标签的性质进行自定义，例如，对离散动作使用交叉熵，对连续动作使用MSE、Huber损失或扩散损失[14, 26]。我们优化上述目标以学习通用动作代码本 以及通用动作提取器和所有异构解码头的参数 。重要的是，虽然 和通用动作提取器在每个训练迭代中同时更新，但异构头 是基于特定域采样的训练批次进行更新的。这种训练策略反映了许多元学习方法[23]中的哲学，即学习允许适应相关任务的全局共享参数，以及保证下游任务性能的任务特定组件。通过这种方法，UniAct努力提炼一个稳健、自适应的通用动作空间以及可以与多种实现及其特定操作环境无缝集成的解码策略。

# 4. Experiments

# 4. 实验

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:

在本节中，我们首先描述UniAct框架的详细实现，然后介绍为回答以下问题而进行的评估实验:

* Can universal actions enhance execution performance across various embodiments with large domain gaps?
* 通用动作能否在具有大域差距的各种实现中提高执行性能？
* Can universal actions be seamlessly transferred to new, unseen embodiments?
* 通用动作能否无缝转移到新的、未见过的实现？
* Dose UniAct learn a meaningful universal action space?
* UniAct是否学习到了一个有意义的通用动作空间？

# 4.1. Experiments Setup

# 4.1. 实验设置

Implementation Details. In this paper, we build a 0.5B instantiation of UniAct on heterogeneous embodied data sources to explore a universal action space . 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 Deep-Speed [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.

实现细节。在本文中，我们在异构的具身数据源上构建了一个0.5B的UniAct实例，以探索通用动作空间 。具体来说，UniAct-0.5B基于LLaVA-OneVion-0.5B [35]构建，这是一个训练有素的视觉语言模型(VLM)，能够提供全面的多模态表示。UniAct-0.5B的训练在64个A100 GPU上使用Deep-Speed [53]进行，历时10天，利用了从28个不同具身实例中收集的100万次演示。训练数据结合了多个开源的机器人数据集，包括Open-XEmbodiment [49]、Libero [38]和Droid [30]，标准化为包含第三人称视觉观察和语言指令，同时保持动作的异质性。有关训练和数据构建的更多细节，请参阅附录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.

基线设置。我们选择了两个最先进的开源视觉-语言-动作模型作为基线:Octo [61]和OpenVLA [31]。Octo是一个0.1B的基于扩散的策略，而OpenVLA则采用了一个7B参数的自回归架构，具有离散动作。这两个模型都在大约100万次精心挑选的机器人演示上进行训练，没有动作异质性，例如将所有绝对末端执行器(EEF)位置预处理为相对EEF位置，并移除关节位置动作。相比之下，UniAct-0.5B在相同数据源的类似规模数据上进行训练，但没有采用这种繁琐的数据清理。我们将UniAct与基线模型进行比较，以展示其在从异构数据中提取通用动作方面的有效性。

# 4.2. Main Results

# 4.2. 主要结果

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.

为了评估UniAct-0.5B的跨具身泛化能力，我们在现实世界的WidowX机器人[64]和Liu等人[38]的仿真Franka机器人上进行了“开箱即用”的评估。这两个平台在之前的工作中常用于测试通用机器人策略的有效性[49, 61, 64]，并且存在显著的领域差距。鉴于我们的训练数据集包含来自这两个具身实例的数据，我们可以利用预训练的异构头将通用动作无缝转换回可部署的控制信号。



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.

图3. 在WidowX机器人上的评估。我们复现了BridgeV2现实世界平台[64]，并遵循OpenVLA [31]的评估协议以确保公平比较。我们评估了不同维度的泛化能力，涵盖了总共19个任务。图中展示了代表性任务。对于每个任务，我们对每个模型进行10次试验，并报告每个任务套件的平均成功率。

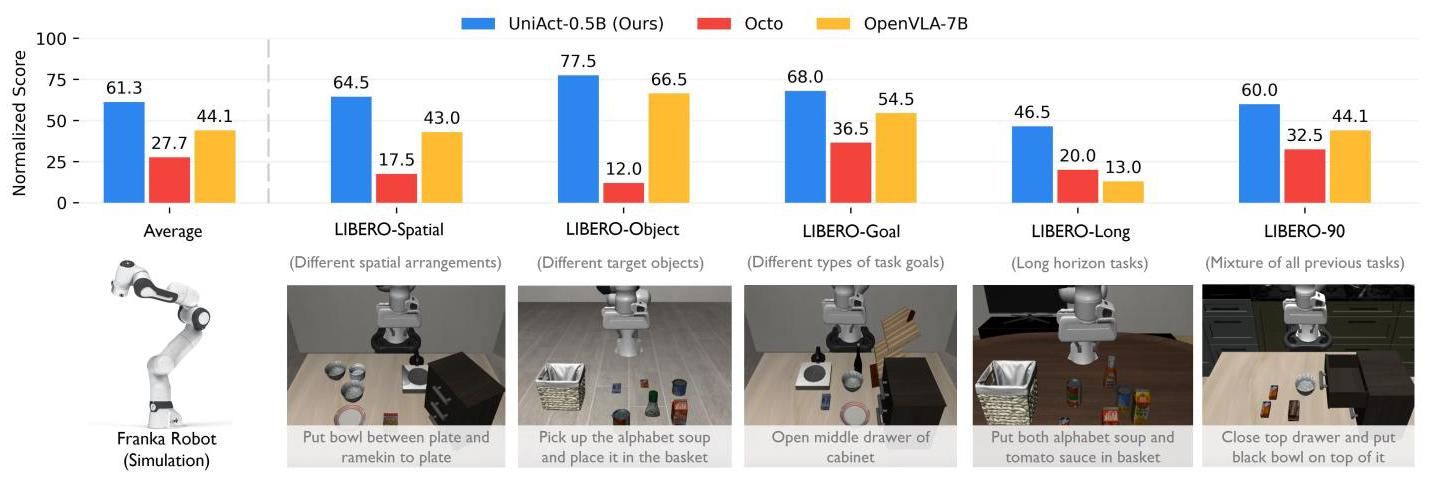


Figure 4. For benchmarking purpose, we evaluate on the LIBERO benchmark [38] in simulations. LIBERO contains several task suites to examine different axes of policy abilities. LIBERO-90 contains 90 tasks and others each contain 10 tasks, covering 130 tasks. For each task, we evaluate 20 trials for each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

图4. 出于基准测试的目的，我们在仿真中对LIBERO基准[38]进行了评估。LIBERO包含多个任务套件，用于检查策略能力的不同维度。LIBERO-90包含90个任务，其他每个套件包含10个任务，总共涵盖130个任务。对于每个任务，我们对每个模型进行20次试验，并报告每个任务套件的平均成功率。

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 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.

现实世界机器人评估。遵循Kim等人[31]的方法，我们为现实世界机器人定义了一套全面的评估任务，涵盖了多个维度的泛化:视觉(未见过的背景/干扰/物体外观)；运动(未见过的物体位置/方向)；物理(未见过的物体大小/形状)；语义(未见过的目标物体/指令/来自互联网的概念)；以及语言基础(操作语言中指定的物体)。总体而言，每个模型在190次试验中进行评估，分布在19个任务上，每个任务进行10次试验。更多细节请参见附录B。代表性任务和结果如图3所示。UniAct-0.5B在视觉、运动和物理泛化任务中优于14倍大的OpenVLA-7B。这展示了从异构数据中提取通用动作在增强对视觉干扰的鲁棒性和低级控制泛化方面的显著优势。虽然OpenVLA利用7B的视觉语言骨干网络在语义理解和语言基础能力上表现出色，但UniAct-0.5B在语义泛化和语言基础任务中实现了可比的性能，突显了其效率和有效性。

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 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.

仿真评估。我们使用LIBERO基准[38]进行评估。值得注意的是，基线模型最初并未在仿真数据上进行训练，因此我们利用其官方代码库和训练指南在LIBERO平台上对其进行微调。用于训练UniAct和基线模型的仿真数据在任务类型、专家轨迹数量和图像质量方面完全一致。有关微调过程的更多详细信息，请参阅附录A。基准包括五个不同套件中的130个机器人仿真任务:LIBERO-Spatial、-Object、-Goal、-Long和-90。LIBERO-90套件包含90个任务，而其他四个套件各包含10个任务。UniAct-0.5B的部署示例和性能见图4。UniAct-0.5B在所有任务套件中均优于基线模型，显示出显著改进，其总体平均准确率比7B OpenVLA高 ，比Octo高33.6%。这一卓越性能归因于UniAct能够弥合领域差距并提取可推广的原子行为。通过利用来自不同领域的示范来学习通用动作，UniAct显著提高了在LIBERO基准上的任务表现。

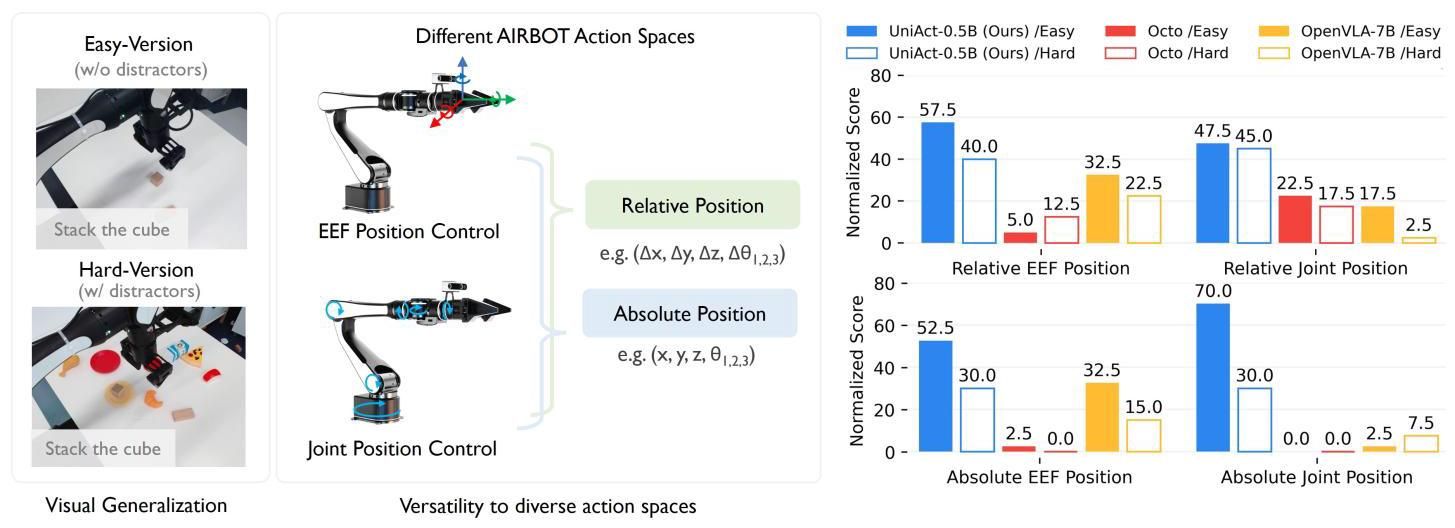


Figure 5. Fast adaptation to new robots. We fine-tune models on an unseen AIRBOT robot with different controller interfaces including relative/absolute joint/EEF positions. The task is to stack the small cube on the small cuboid, requiring precise action execution. The fine-tuning data is collected w/o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

图5. 快速适应新机器人。我们在一个未见过的AIRBOT机器人上对模型进行微调，该机器人具有不同的控制器接口，包括相对/绝对关节/末端执行器位置。任务是将小立方体堆叠在小长方体上，需要精确的动作执行。微调数据是在没有任何干扰物的情况下收集的，但我们也评估了一个困难版本，其中策略应对各种杂物具有鲁棒性。

# 4.3. Fast Adaptation to New Embodiment

# 4.3. 快速适应新实体

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.

实验设置。为了评估快速适应能力，我们在一个新的真实世界机器人AIRBOT上进行评估，该机器人具有四种截然不同的控制器接口:相对/绝对末端执行器位置和相对/绝对关节位置。UniAct和基线模型均未在AIRBOT数据上进行预训练。我们在这四种不同类型的控制接口上收集了100个示范。考虑到这些控制接口之间的显著异质性，我们投入了大量精力对基线模型进行微调，并确保模型收敛符合官方要求(例如，OpenVLA的预测准确率为95%)。详情请参阅附录B。

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 and vision features from a shared vision backbone as inputs.

使用UniAct快速适应。与需要大量训练以弥合不同动作类型之间适应差距的基线模型不同，UniAct能够快速适应新实体和控制接口。由于已经学习了跨实体的行为，我们通过冻结代码本和通用动作提取器来促进快速适应。同时，我们从头开始为每种动作类型训练四个异构解码头，使用收集的示范进行训练。每个新引入的头由一个简单的MLP实现，该MLP以共享视觉骨干的 和视觉特征 作为输入。

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 and 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.

评估。我们使用“将立方体堆叠在另一个立方体上”任务的简单和困难版本进行了评估。结果见图5。UniAct-0.5B在所有类型的控制信号上均表现出一致的强泛化能力，优于两个基线模型。值得注意的是，UniAct-0.5B用于微调的参数量与总模型大小之比最小(4M / 500M:0.8%)。相比之下，OpenVLA和Octo分别使用了其总模型大小的 和 。这种高效的参数利用突显了UniAct的有效性和适应性，展示了其在将学习到的通用动作应用于新任务和实体时，以最小的参数空间扩展实现卓越性能的能力。

# 4.4. In-depth Analysis of Universal Actions

# 4.4. 通用动作的深入分析

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.

在本节中，我们从两个角度展示UniAct构建了一个有意义的通用动作空间:1)一致的语义行为被编码为跨不同实体的相同通用动作；2)通用动作提取器能够有效利用不同机器人之间通用动作空间中的共享结构。

Interpretation of Universal Actions. We manually inspect the decoded behaviors for all 256 universal actions across different robots and observe at least 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.

通用动作的解释。我们手动检查了所有256个通用动作在不同机器人上的解码行为，并观察到至少 的动作表现出完全一致性。图6显示，相同的通用动作可以解码回不同机器人上的一致行为，即使存在巨大差距。例如，具有不同视角的不同机器人，甚至那些经历巨大仿真到现实差距的机器人，在给定相同通用动作的情况下，可以执行类似的语义上有意义的行为。

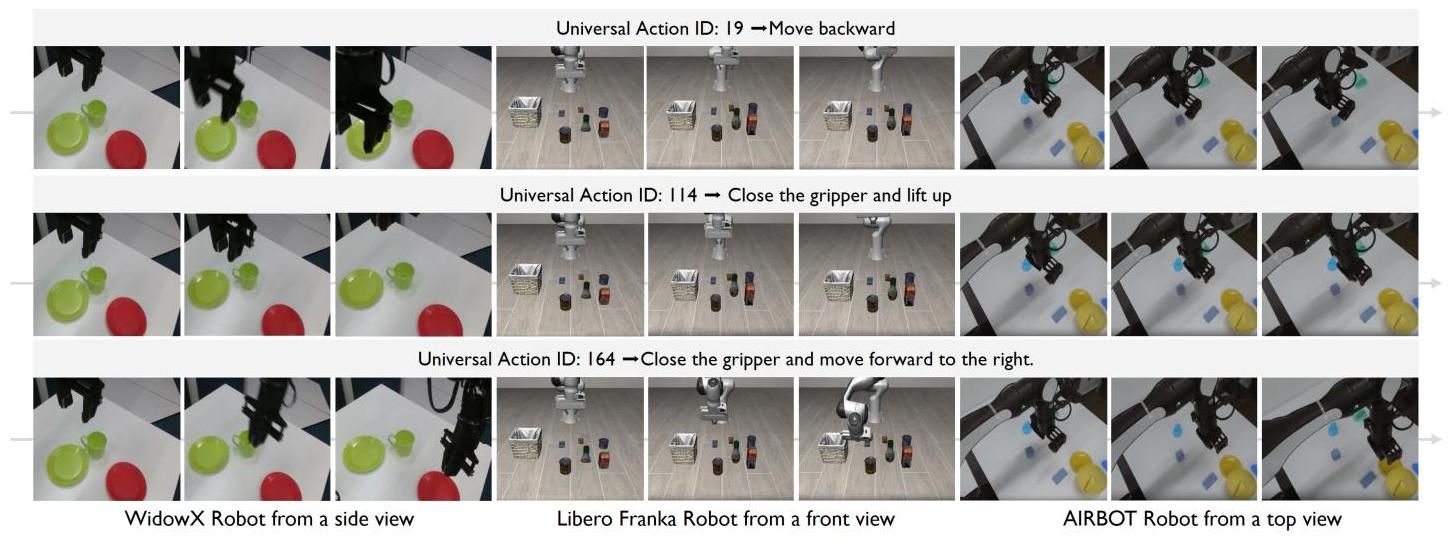


Figure 6. We manually check the decoded behavior for all 256 universal actions across different robots and observe at least show exact consistency, even with varying viewpoints and significant domain gaps between the real world and simulation. Visualizations of repeated executions of the same universal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

图6. 我们手动检查了不同机器人上所有256个通用动作的解码行为，观察到至少有 显示出完全一致性，即使视角不同且现实世界与模拟之间存在显著的领域差距。在不同机器人上重复执行同一通用动作的可视化结果揭示了清晰、语义明确且一致的行为。

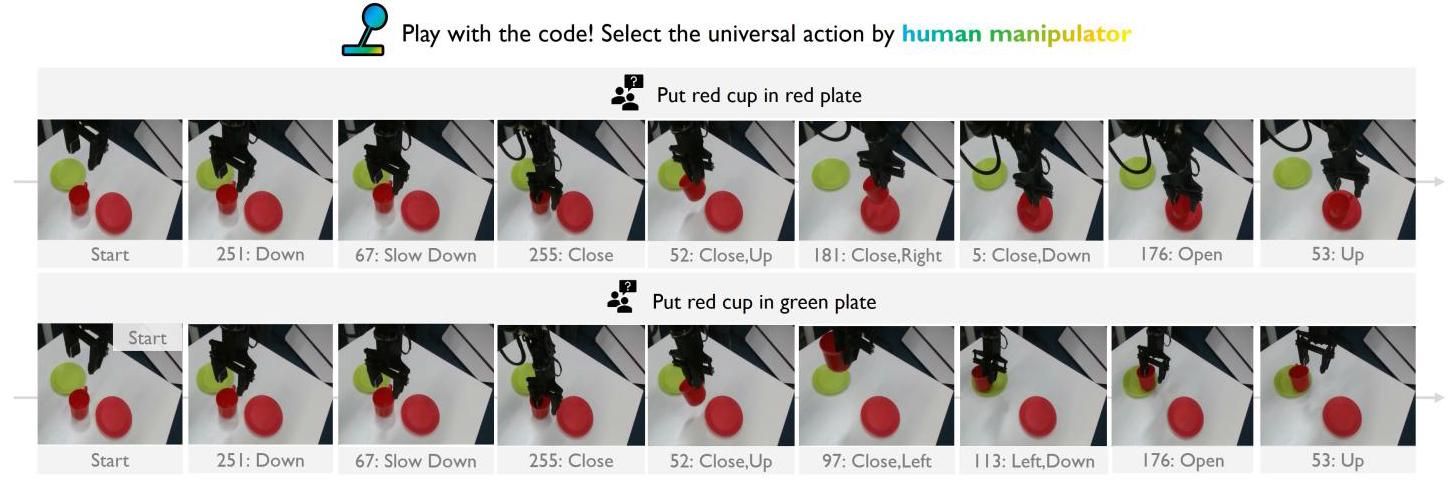


Figure 7. Since all universal actions encode semantic meaningful behaviors, we can directly play with it! By selecting the desired behavior using its corresponding universal action ID, we can manually control the robot to perform complex tasks such as Pick & Place.

图7. 由于所有通用动作都编码了具有语义意义的行为，我们可以直接与之互动！通过选择相应通用动作ID的期望行为，我们可以手动控制机器人执行诸如拾取与放置等复杂任务。

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.

使用通用动作进行控制。因此，我们可以通过选择一系列通用动作直接与机器人互动以执行期望的行为。图7清楚地展示了我们可以在没有任何机器人知识的情况下使用通用动作控制机器人，例如学习复杂的前向/逆向运动学变换。这也强调了利用通用动作提取器作为动作标记器的潜力，通过在这个离散的通用动作空间中进行规划，促进未来更先进的具身基础模型的部署。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | 0. | 0.34 | 0.45 | 0.51 |
|  | 0.34 | 0. | 0.60 | 0.58 |
|  | 0.45 | 0.60 | 0. | 0.44 |
|  | 0.51 | 0.58 | 0.44 | 0. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | 0. | 0.34 | 0.45 | 0.51 |
|  | 0.34 | 0. | 0.60 | 0.58 |
|  | 0.45 | 0.60 | 0. | 0.44 |
|  | 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). and 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.

表1. 我们计算了两个不同领域中两个任务的通用动作利用分布的JS散度(值越低表示一致性越高)。 和 分别表示“拿起碗”和“打开抽屉”。WidowX和Franka分别表示图3-4中的现实世界机器人和模拟机器人。

Statistical Analysis For Universal Action Utilization. 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.

通用动作利用的统计分析。在此，我们总结了不同机器人在不同任务中的通用动作利用分布。表1清楚地表明，相同任务和不同机器人的利用分布相似，但不同任务和相同机器人的利用分布不同，这表明通用动作提取器确实通过更多地关注任务进展而非具体实现细节，正确利用了这些与具体实现无关的原子行为。

# 5. Conclusion

# 5. 结论

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 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 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.

我们介绍了UniAct，一个创新的具身基础建模框架，它在通用动作空间中运行，以解决动作异质性的挑战。这个通用动作空间编码了跨不同具身动作空间的可共享原子行为，显著提高了跨领域数据利用率，并促进了跨具身泛化，使我们的 参数模型能够超越比其大14倍的SOTA模型。此外，学习到的通用动作可以通过异构解码以最少的参数精确转换为任何特定具身的动作，从而快速适应具有不同控制接口和物理特性的新机器人。此外，我们学习到的通用动作提取器还可以用作通用动作分词器，为未来大规模具身基础模型的构建提供支持。目前，由于资源限制，UniAct以 参数实例进行训练，并主要在单臂机器人平台上进行评估。未来的工作将集中在将UniAct扩展到更大的模型，并将其应用扩展到更广泛的具身领域，包括双手机器人甚至自动驾驶，进一步发挥其在更多机器人应用中的多功能性和有效性。

# References

# 参考文献

[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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, 等. Gpt-4技术报告. arXiv预印本 arXiv:2303.08774, 2023. 1, 2, 4

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Cheb-otar, 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

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Cheb-otar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, 等. 按我能做的做，而不是按我说的做:将语言扎根于机器人能力中. arXiv预印本 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

Anurag Ajay, Seungwook Han, Yilun Du, Shuang Li, Abhi Gupta, Tommi Jaakkola, Josh Tenenbaum, Leslie Kaelbling, Akash Srivastava, 和 Pulkit Agrawal. 用于分层规划的组合基础模型. 神经信息处理系统进展, 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

匿名. IGOR:图像-目标表示是具身AI中下一级泛化的原子构建块. 提交至第十三届国际学习表示会议, 2024. 审稿中.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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, 等. Qwen技术报告. arXiv预印本 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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, 和 Jingren Zhou. Qwen-vl:具有多功能能力的前沿大型视觉语言模型. arXiv预印本 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.2,3,15

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Rich Walke, Chelsea Finn, Aviral Kumar, 和 Sergey Levine. 使用预训练图像编辑扩散模型进行零样本机器人操作. 第十二届国际学习表示会议.2,3,15

[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. : A vision-language-action flow model for general robot control, 2024. 1, 3

凯文·布莱克、诺亚·布朗、丹尼·德里斯、阿德南·埃斯梅尔、迈克尔·埃奎、切尔西·芬恩、尼科洛·富赛、拉奇·格鲁姆、卡罗尔·豪斯曼、布莱恩·伊克特、雅各布扎克·雅库布扎克、蒂姆·琼斯、柯立明、谢尔盖·莱文、阿德里安·李-贝尔、莫希特·莫图库里、苏拉杰·奈尔、卡尔·佩奇、施晓阳·露西、詹姆斯·坦纳、权·武昂、安娜·沃林、王浩轩和尤里·日林斯基。 :一种用于通用机器人控制的视觉-语言-动作流模型，2024年。1, 3

[9] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakr-ishnan, 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

安东尼·布罗汉、诺亚·布朗、贾斯蒂斯·卡巴哈尔、叶夫根·切博塔尔、约瑟夫·达比斯、切尔西·芬恩、基尔塔纳·戈帕拉克里希南、卡罗尔·豪斯曼、亚历克斯·赫尔佐格、贾斯敏·许等。Rt-1:用于大规模现实世界控制的机器人变压器。arXiv预印本 arXiv:2212.06817，2022年。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

安东尼·布罗汉、诺亚·布朗、贾斯蒂斯·卡巴哈尔、叶夫根·切博塔尔、陈曦、克日什托夫·霍罗曼斯基、丁天立、丹尼·德里斯、阿维纳瓦·杜贝、切尔西·芬恩等。Rt-2:将网络知识迁移到机器人控制的视觉-语言-动作模型。arXiv预印本 arXiv:2307.15818，2023年。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

杰克·布鲁斯、迈克尔·丹尼斯、阿什利·爱德华兹、杰克·帕克-霍尔德、施雨歌、爱德华·休斯、马修·赖、阿迪蒂·马瓦兰卡、里奇·斯泰格沃尔德、克里斯·阿普斯等。Genie:生成式交互环境。在第四十一届国际机器学习会议上，2024年。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

郑志林、陈广增、景雅、孔涛、李航、李逸峰、刘宇霄、吴洪涛、徐嘉峰、杨一初等。Gr-2:一种具有网络规模知识的生成式视频-语言-动作模型，用于机器人操作。arXiv预印本 arXiv:2410.06158，2024年。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

陈曦、阿里·加迪尔扎德、余天和、王建浩、高元、李文哲、梁斌、切尔西·芬恩和张崇杰。Lapo:用于离线强化学习的潜在变量优势加权策略优化。神经信息处理系统进展，35:36902-36913，2022年。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 02783649241273668.2,5

程驰、徐振佳、冯思远、埃里克·库西诺、杜一伦、本杰明·伯奇菲尔、拉斯·泰德拉克和宋书然。扩散策略:通过动作扩散进行视觉运动策略学习。国际机器人研究杂志，页码:02783649241273668。2,5

[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

迪马·达门、黑兹尔·多尔蒂、乔瓦尼·玛丽亚·法里内拉、桑贾·菲德勒、安东尼诺·弗纳里、埃万杰洛斯·卡扎科斯、达维德·莫尔蒂桑蒂、乔纳森·芒罗、托比·佩雷特、威尔·普莱斯等。扩展自我中心视觉:epic-kitchens数据集。在欧洲计算机视觉会议(ECCV)上，页码:720-736，2018年。2

[16] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu

董庆秀、李磊、戴大麦、郑策、马景元、李瑞、夏鹤鸣、徐晶晶、吴志勇、田宇

Liu, et al. A survey on in-context learning. arXiv preprint arXiv:2301.00234, 2022. 2

刘等。关于上下文学习的综述。arXiv预印本 arXiv:2301.00234，2022年。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

里亚·多希、霍默·里奇·沃克、奥伊尔·米斯、苏迪普·达萨里和谢尔盖·莱文。扩展跨体学习:一种用于操作、导航、移动和航空的策略。在第八届机器人学习年会上。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

丹尼·德里斯、夏飞、梅赫迪·SM·萨贾迪、科里·林奇、阿坎莎·乔杜里、布莱恩·伊克特、阿伊赞·瓦希德、乔纳森·汤普森、权·武昂、余天和等。Palm-e:一种具身多模态语言模型。在第四十届国际机器学习会议上，页码:8469-8488，2023年。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.2,3,15

杜一伦、杨雪莉、皮特·佛罗伦斯、夏飞、瓦希德·阿扎安、皮埃尔·塞尔曼内、余天和、彼得·阿比尔、约书亚·B·特南鲍姆、莱斯利·帕克·凯布林等。视频语言规划。在第十二届国际学习表征会议上。2,3,15

[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, 2024.1,3

方浩舒、方洪杰、唐振宇、刘继荣、王晨曦、王俊波、朱浩毅和卢策武。Rh20t:一个全面的机器人数据集，用于一次性学习多种技能。在2024年IEEE国际机器人与自动化会议(ICRA)上，第653-660页。IEEE，2024年。1,3

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

费南一、卢志武、高一钊、杨国兴、霍宇琪、温静远、卢浩宇、宋瑞华、高鑫、向涛等。通过多模态基础模型迈向人工通用智能。《自然通讯》，13(1): 3094, 2022年。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

高杰森、萨卡尔·比迪普塔、夏飞、肖特德、吴佳俊、布莱恩·伊克特、马朱姆达尔·阿尼鲁达和萨迪格·多萨。基于物理的视觉语言模型用于机器人操作。在2024年IEEE国际机器人与自动化会议(ICRA)上，第12462-12469页。IEEE，2024年。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

乔纳森·戈登、约翰·布朗斯基尔、马蒂亚斯·鲍尔、塞巴斯蒂安·诺沃辛和理查德·特纳。元学习概率推理用于预测。在国际学习表征会议上，2019年。5

[24] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michal-ski, 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

拉加夫·戈亚尔、萨米拉·易卜拉欣·卡胡、文森特·米哈尔斯基、乔安娜·马特尔津斯卡、苏珊娜·韦斯特法尔、赫娜·金、瓦伦丁·哈内尔、英戈·弗伦德、彼得·亚尼洛斯、莫里茨·穆勒-弗赖塔格等。“某物某物”视频数据库用于学习和评估视觉常识。在IEEE国际计算机视觉会议论文集上，第5842-5850页，2017年。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

克里斯滕·格劳曼、安德鲁·韦斯特伯里、尤金·伯恩、扎卡里·查维斯、安东尼诺·弗纳里、罗希特·吉尔达尔、杰克逊·汉堡、姜浩、刘苗、刘星宇等。Ego4d:在3000小时的第一人称视频中环游世界。在IEEE/CVF计算机视觉与模式识别会议论文集上，第18995-19012页，2022年。2

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

乔纳森·何、阿贾伊·贾恩和彼得·阿比尔。去噪扩散概率模型。《神经信息处理系统进展》，33:6840-6851, 2020年。5

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

埃里克·张、顾世翔和本·普尔。使用Gumbel-Softmax进行类别重参数化。arXiv预印本arXiv:1611.01144, 2016年。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

埃里克·张、亚历克斯·伊尔潘、莫希·坎萨里、丹尼尔·卡普勒、弗雷德里克·埃伯特、科里·林奇、谢尔盖·莱文和切尔西·芬恩。BC-z:通过机器人模仿学习实现零样本任务泛化。在机器人学习会议上，第991-1002页。PMLR，2022年。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

西达尔特·卡拉姆切蒂、苏拉杰·奈尔、安妮·S·陈、托马斯·科拉尔、切尔西·芬恩、多萨·萨迪格和珀西·梁。语言驱动的机器人表示学习。arXiv预印本arXiv:2302.12766, 2023年。2

[30] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yun-liang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. arXiv preprint arXiv:2403.12945, 2024. 2, 3, 5

亚历山大·哈扎茨基、卡尔·佩尔奇、苏拉杰·奈尔、阿什温·巴拉克里希纳、苏迪普·达萨里、西达尔特·卡拉姆切蒂、索鲁什·纳西里亚尼、莫汉·库马尔·斯里拉马、陈云亮、柯斯蒂·埃利斯等。Droid:一个大规模野外机器人操作数据集。arXiv预印本arXiv:2403.12945, 2024年。2, 3, 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

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, 等. Openvla: 一个开源的视觉-语言-动作模型. arXiv 预印本 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

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, 和 Jiaya Jia. Lisa: 通过大语言模型进行推理分割. 在 IEEE/CVF 计算机视觉与模式识别会议论文集, 页码 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

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, 和 Wook-Shin Han. 使用残差量化进行自回归图像生成. 在 IEEE/CVF 计算机视觉与模式识别会议论文集, 页码 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. 2, 3

Seungjae Lee, Yibin Wang, Haritheja Etukuru, H. Jin Kim, Nur Muhammad Mahi Shafiullah, 和 Lerrel Pinto. 使用潜在动作进行行为生成. 在第四十一届国际机器学习会议, 2024. 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

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, 和 Chunyuan Li. Llava-onevision: 简单的视觉任务迁移. arXiv 预印本 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

Jianxiong Li, Jinliang Zheng, Yinan Zheng, Liyuan Mao, Xiao Hu, Sijie Cheng, Haoyi Niu, Jihao Liu, Yu Liu, Jingjing Liu, Yaqin Zhang, 和 Xianyuan Zhan. Decisionnce: 通过隐式偏好学习实现的多模态表示. 在第四十一届国际机器学习会议. 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

Jianxiong Li, Zhihao Wang, Jinliang Zheng, Xiaoai Zhou, Guanming Wang, Guanglu Song, Yu Liu, Jingjing Liu, Ya-Qin Zhang, Junzhi Yu, 等. Robo-mutual: 通过单模态学习实现机器人多模态任务规范. arXiv 预印本 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

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, 和 Peter Stone. Libero: 终身机器人学习知识转移的基准测试. 神经信息处理系统进展, 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

Haotian Liu, Chunyuan Li, Yuheng Li, 和 Yong Jae Lee. 通过视觉指令调优改进基线. 在 IEEE/CVF 计算机视觉与模式识别会议论文集, 页码 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

Haotian Liu, Chunyuan Li, Qingyang Wu, 和 Yong Jae Lee. 视觉指令调优. 神经信息处理系统进展, 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. 2, 3

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, 和 Jun Zhu. Rdt-1b: 用于双手操作的扩散基础模型. arXiv 预印本 arXiv:2410.07864, 2024. 2, 3

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

Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bas-tani, 和 Dinesh Jayaraman. Liv: 用于机器人控制的语言-图像表示和奖励. 在国际机器学习会议, 页码 23301-23320. PMLR, 2023. 2

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

J. MacQueen. 一些用于分类和分析多变量观测的方法. 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

Fabian Mentzer, David Minnen, Eirikur Agustsson, 和 Michael Tschannen. 有限标量量化: 简化的 VQ-VAE. 在第十二届国际学习表示会议. 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

Atharva Mete, Haotian Xue, Albert Wilcox, Yongxin Chen, 和 Animesh Garg. Quest: 用于学习连续控制的自监督技能抽象. arXiv 预印本 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

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, 和 Ping Luo. EmbodiedGPT: 通过具身思维链进行视觉语言预训练. 在第三十七届神经信息处理系统会议上, 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

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, 和 Abhinav Gupta. R3m: 用于机器人操作的通用视觉表示. 在第六届机器人学习年会上. 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

Haoyi Niu, Jianming Hu, Guyue Zhou, 和 Xianyuan Zhan. 具身智能体跨域策略迁移的全面综述. 在第三十三届国际人工智能联合会议上, 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, 6

Abby O’Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, 等. Open x-embodiment: 机器人学习数据集和 rt-x 模型. arXiv 预印本 arXiv:2310.08864, 2023. 1, 2, 3, 5, 6

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

Seohong Park, Tobias Kreiman, 和 Sergey Levine. 基于希尔伯特表示的基础策略. 在第四十一届国际机器学习会议上. 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

Seohong Park, Dibya Ghosh, Benjamin Eysenbach, 和 Sergey Levine. Hiql: 以潜在状态为动作的离线目标条件强化学习. 神经信息处理系统进展, 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

William Peebles 和 Saining Xie. 基于变压器的可扩展扩散模型. 在 IEEE/CVF 国际计算机视觉会议论文集, 第 4195-4205 页, 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

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, 和 Yuxiong He. Deepspeed: 系统优化使得训练超过 1000 亿参数的深度学习模型成为可能. 在第 26 届 ACM SIGKDD 国际知识发现与数据挖掘会议论文集, 第 3505-3506 页, 2020. 5

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

Dominik Schmidt 和 Minqi Jiang. 学习无动作的行动. 在第十二届国际学习表示会议上. 3

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

Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Arty Al-tanzaya, 和 Lerrel Pinto. 行为变压器: 一石二鸟克隆 模式. 神经信息处理系统进展, 35:22955-22968, 2022. 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. 3, 15

Lucy Xiaoyang Shi, Zheyuan Hu, Tony Z Zhao, Archit Sharma, Karl Pertsch, Jianlan Luo, Sergey Levine, 和 Chelsea Finn. 对你的机器人喊叫: 通过语言即时纠正进行改进. arXiv 预印本 arXiv:2403.12910, 2024. 3, 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

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, 和 Yu Su. Llm-planner: 使用大语言模型进行具身智能体的少样本规划. 在 IEEE/CVF 国际计算机视觉会议 (ICCV) 论文集, 第 2998-3009 页, 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

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, 和 Xinlong Wang. Emu: 多模态生成预训练. 在第十二届国际学习表示会议上, 2023. 2

[59] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiy-ing 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

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiy-ing Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, 和 Xinlong Wang. 生成多模态模型是上下文学习者. 在 IEEE/CVF 计算机视觉与模式识别会议论文集, 第 14398-14409 页, 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

Gemma团队，Thomas Mesnard，Cassidy Hardin，Robert Dadashi，Surya Bhupatiraju，Shreya Pathak，Laurent Sifre，Morgane Rivière，Mihir Sanjay Kale，Juliette Love等。Gemma:基于Gemini研究和技术的开放模型。arXiv预印本 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:2405.12213,2024.1,2,3,5,6,12

Octo模型团队，Dibya Ghosh，Homer Walke，Karl Pertsch，Kevin Black，Oier Mees，Sudeep Dasari，Joey Hejna，Tobias Kreiman，Charles Xu等。Octo:一个开源的通用机器人策略。arXiv预印本 arXiv:2405.12213,2024.1,2,3,5,6,12

[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

Hugo Touvron，Thibaut Lavril，Gautier Izacard，Xavier Martinet，Marie-Anne Lachaux，Timothée Lacroix，Baptiste Rozière，Naman Goyal，Eric Hambro，Faisal Azhar等。Llama:开放且高效的基础语言模型。arXiv预印本 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

Aaron Van Den Oord，Oriol Vinyals等。神经离散表示学习。神经信息处理系统进展，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

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等。Bridgedata v2:一个用于大规模机器人学习的数据集。在机器人学习会议上，页码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, 2

Lirui Wang，Xinlei Chen，Jialiang Zhao，和Kaiming He。通过异构预训练变压器扩展本体感知视觉学习。arXiv预印本 arXiv:2409.20537, 2024. 1, 2

[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

Chuan Wen，Xingyu Lin，John So，Kai Chen，Qi Dou，Yang Gao，和Pieter Abbeel。任意点轨迹建模用于策略学习。arXiv预印本 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

Hongtao Wu，Ya Jing，Chilam Cheang，Guangzeng Chen，Jiafeng Xu，Xinghang Li，Minghuan Liu，Hang Li，和Tao Kong。释放大规模视频生成预训练用于视觉机器人操作。在第十二届国际学习表示会议上。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, 3

Jonathan Yang，Catherine Glossop，Arjun Bhorkar，Dhruv Shah，Quan Vuong，Chelsea Finn，Dorsa Sadigh，和Sergey Levine。推动跨体现学习的极限用于操作和导航。arXiv预印本 arXiv:2402.19432, 2024. 1, 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

Seonghyeon Ye，Joel Jang，Byeongguk Jeon，Sejune Joo，Jianwei Yang，Baolin Peng，Ajay Mandlekar，Reuben Tan，Yu-Wei Chao，Bill Yuchen Lin等。从视频中进行潜在动作预训练。arXiv预印本 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

Haoqi Yuan，Zhancun Mu，Feiyang Xie，和Zongqing Lu。预训练基于目标的模型用于样本高效的强化学习。在第十二届国际学习表示会议上，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

Michał Zawalski，William Chen，Karl Pertsch，Oier Mees，Chelsea Finn，和Sergey Levine。通过体现链式推理进行机器人控制。在第八届机器人学习年会上，2024. 15

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

Jinliang Zheng，Jianxiong Li，Sijie Cheng，Yinan Zheng，Ji-aming Li，Jihao Liu，Yu Liu，Jingjing Liu，和Xianyuan Zhan。指令引导的视觉掩码。神经信息处理系统进展，2024. 2

# A. Training Details

# A. 训练细节

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 for the VLM and 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.

训练超参数 UniAct-0.5B 使用来自 LLava-One-Vision-0.5B [35] 的预训练参数来初始化视觉语言模型(VLM)。对于异构解码头中的视觉特征提取，我们部署了一个在 ImageNet 上预训练的 ResNet18，该模型通常用于基于视觉的策略学习 [64]。该模型在大规模预训练过程中联合训练，以增强在操作任务场景中的感知能力。我们采用与原始配置一致的 分辨率用于 VLM， 分辨率用于 ResNet18。图像增强设置和更多超参数分别可以在表 2 和表 3 中找到。

|  |  |
| --- | --- |
| 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) |

|  |  |
| --- | --- |
| 增强 | 值 |
| 随机调整大小 | 比例=(0.75, 1.3333) 缩放=(0.5, 1.0) 插值=双三次 |
| 随机水平翻转 | p=0.5 |
| 颜色抖动 | 对比度=(0.6, 1.4) 亮度=(0.6, 1.4) 饱和度=(0.6,1.4) |

Table 2. Image augmentation settings during training

表2. 训练期间的图像增强设置

|  |  |
| --- | --- |
| config | value |
| optimizer | AdamW |
| batch size | 1024 |
| learning rate |  |
| weight decay | 0. |
| optimizer momentum |  |
| iters | 500K |
| model precision | BFloat16 |

|  |  |
| --- | --- |
| 配置 | 值 |
| 优化器 | AdamW |
| 批量大小 | 1024 |
| 学习率 |  |
| 权重衰减 | 0. |
| 优化器动量 |  |
| 迭代次数 | 500K |
| 模型精度 | BFloat16 |

Table 3. Training hyper-parameters

表3. 训练超参数

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 , 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 to maintain consistency.

数据构建 如表4所示，我们详细说明了训练数据的组成，其中包括轨迹数量、样本数量以及28个不同实体的控制接口。根据之前的工作 ，我们在训练期间为每个数据集分配不同的采样率，以确保实体、任务和场景的平衡混合。这些采样率在表4中指定。需要注意的是，许多数据集可能包含不同的动作空间，例如末端执行器位置(EEF position)和关节位置(Joint position)。此外，来自不同数据集的图像可能包含多个视角。默认情况下，我们仅使用第一个第三人称视角，以保持一致性。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | 轨迹 | 样本 | 采样率(%) | 控制界面 |
| UT奥斯汀互斥锁 | 1500 | 361871 | 1.0 | 末端执行器位置 |
| 伯克利电缆布线 | 1557 | 38789 | 0.2 | 末端执行器速度 |
| 纽约大学Franka Play | 454 | 44412 | 1.0 | 末端执行器速度 |
| 库卡 | 557893 | 7130157 | 12.7 | 末端执行器位置 |
| 奥斯汀水手 | 240 | 352110 | 2.2 | 末端执行器速度 |
| Fmb | 8611 | 1136907 | 1.0 | 末端执行器速度 |
| 伯克利自动化实验室 | 1000 | 95310 | 1.2 | 末端执行器位置 |
| 维奥拉 | 150 | 75614 | 0.9 | 末端执行器位置 |
| 多贝 | 5208 | 1139874 | 1.4 | 末端执行器位置 |
| 卡内基梅隆大学Iamlab | 631 | 146241 | 0.9 | 末端执行器位置 |
| 奥斯汀花蕾 | 50 | 34110 | 0.2 | 末端执行器位置 |
| 语言表 | 441911 | 6602077 | 4.4 | 末端执行器位置 |
| 斯坦福九头蛇 | 569 | 357137 | 4.4 | 末端执行器位置 |
| 机器人套装 | 18246 | 1419838 | 5.0 | 关节位置 |
| 奥斯汀天狼星 | 559 | 279724 | 1.7 | 末端执行器速度 |
| 德国航空航天中心Edan | 104 | 8824 | 0.1 | 末端执行器位置 |
| 分形 | 86599 | 3607028 | 12.7 | 末端执行器位置 |
| TOTO | 1003 | 324669 | 2.0 | 关节位置 |
| 伯克利发那科 | 415 | 58660 | 0.7 | 末端执行器位置 |
| 卡内基梅隆大学Stretch | 135 | 25012 | 0.2 | 末端执行器位置 |
| Roboturk | 1934 | 186910 | 2.3 | 末端执行器位置 |
| Jaco Play | 976 | 66094 | 0.4 | 末端执行器位置 |
| Taco Play | 3603 | 230966 | 3.0 | 末端执行器位置 |
| BC-Z | 42811 | 5957097 | 7.5 | 末端执行器位置 |
| 机器人 | 92115 | 27043929 | 10.0 | 末端执行器位置 |
| 家具工作台 | 5100 | 3905717 | 2.4 | 末端执行器速度 |
| 桥梁 | 28933 | 899685 | 13.3 | 末端执行器位置 |
| 利贝罗 | 6500 | 1007204 | 5.0 | 末端执行器位置 |

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.

表4. 训练数据组成的详细信息，包括28个不同数据源的轨迹数量、样本数量和控制接口。根据[31]，我们为不同数据设置了不同的采样概率，这些概率也在本表中报告。

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.

分类重参数化。除了用于训练UniAct的Gumbel-Softmax，我们还探索了另一种常用的重参数化技术:直通估计器(STE)[63]。然而，实证结果表明，使用STE可能导致通用动作代码书的严重崩溃。一个直观的解释是，每个通用动作需要大量的优化步骤来学习高度抽象的行为。由于STE涉及硬采样，它导致每个训练样本只选择一个通用动作进行优化。这种在所有潜在动作中缺乏梯度分布的情况可能会抑制学习过程，并减少学习到的动作的多样性。

# B. Evaluation Setups

# B. 评估设置

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.

在本节中，我们深入探讨了研究中使用的三个机器人实体的评估实验:WidowX机器人、Franka机器人和AIRBOT。我们为每个实体提供了详细的分数，以便更清晰、更直观地比较它们的性能。

# B.1. WidowX Robot in Real World

# B.1. 现实世界中的WidowX机器人

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.

WidowX的实验旨在评估模型在五个不同维度上的泛化能力，如图8所示。我们不仅仅报告任务成功率，还为一些复杂任务基于任务完成进度计算分数。这种评分方法可以更直观地反映模型的理解和泛化能力[31]。详细分数见表5。在后续章节中，我们将全面描述任务设置。

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 different 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.

视觉泛化:该维度评估模型适应不同视觉环境的能力，这些环境以光照、背景和物体纹理的变化为特征。根据[31]中的设置，我们设计了三个具体任务，详见图8。每个任务都设置在不同的背景环境中，具有不同的光照亮度，并涉及红色和绿色两种颜色的目标物体。此外，我们引入了与任务无关的各种视觉干扰物，以评估模型的视觉泛化能力及其在面对此类干扰时的鲁棒性。为了确保公平比较，所有与模型测试条件相关的变量在所有模型中保持一致。该套件中的所有任务都分配了二元分数:0表示失败，1表示成功。

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.

运动泛化:该维度的任务旨在评估机器人在识别目标物体位置和方向时执行适当运动的能力，这些位置和方向可能在训练期间未曾遇到过。具体来说，我们设计了两项任务:举起茄子和将胡萝卜放在盘子上，详见图8。目标物体被放置在几个预定位置，并朝向预先设计的方向。这有助于评估机器人对物理任务参数变化的适应性。该套件中的所有任务都分配了二元分数:0表示失败，1表示成功。

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.

物理泛化:在该维度中，我们检查模型管理物体物理变化的能力，如大小、重量和材料属性。这些任务旨在评估机器人对这些波动的适应性，这些波动显著影响操作策略。详细的任务设置列于图8中。这些任务中涉及的物体，如胡萝卜和AAA电池，通常具有不规则的形状或放置在异常位置(例如，翻倒的锅)。成功处理并完成这些任务需要一个能够准确识别目标物体物理属性并执行适当策略的通用策略。该套件中的所有任务都分配了二元分数:0表示失败，1表示成功。

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.

语义泛化:在该维度中，我们评估模型理解和泛化各种语义上下文的能力。具体来说，该维度中包含的任务尤其是目标物体在UniAct和基线模型的训练过程中从未遇到过。这种方法测试模型解释和适应新指令和环境的能力，强调其灵活性和学习效率。紫色葡萄非常光滑，因此我们分配分数如下:触摸葡萄为0.25，抓住葡萄为0.5，将其移向锅为0.75，成功完成拾取和放置为1。此外，对于将绿色杯子堆叠在红色杯子上的任务，我们分配分数如下:触摸绿色杯子为0.25，抓住绿色杯子为0.5，将其移向红色杯子为0.75，成功堆叠为1。该套件中的其他任务分配了二元分数。

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.

语言基础(Language Grounding):这一维度评估模型在理解和执行基于自然语言的命令方面的能力。重点是评估模型在遵循新颖指令以操作特定对象方面的熟练程度，这些指令是通过语言描述的。尽管在处理未见过的场景方面与语义泛化(Semantic Generalization)有相似之处，但语言基础任务明确测试模型在潜在误导性环境中准确理解并执行基于语言的指令的能力。一个示例是将两个不同颜色的杯子放在桌子上，并指示模型使用语言操作其中一个。模型必须准确地将语言指令与复杂现实环境中的正确对象对应起来。这测试了模型在动态和视觉多样化的环境中将语言描述直接与物理动作连接的能力。对于本套件中的所有任务，我们按以下方式分配分数:正确基础得0.5分，成功完成任务得1分。

# B.2. Franka Robot in Simulation

# B.2. 仿真中的Franka机器人

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 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 hy-perparameters are provided in Table 6. Notably, we manually cleaned the "no-op" data before fine-tuning Open-VLA 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 Uni-Act framework demonstrated robustness to noisy data.

我们包含了在LIBERO [38]仿真中使用Franka机器人收集的130个不同任务的6500个专家演示，用于训练我们的UniAct-0.5B，然后按照LIBERO基准[38]评估我们的模型。输入图像由仿真器渲染，我们使用默认的 分辨率。由于两个开源基线模型OpenVLA和Octo最初并未使用仿真数据进行训练，我们投入了大量精力对它们进行微调，以便与我们的UniAct框架进行公平比较。OpenVLA和Octo的微调过程分别在8个A6000 GPU和2个4090 GPU上进行，分别持续了7小时和4小时。训练超参数的详细信息见表6。值得注意的是，我们在微调OpenVLA之前按照其官方指南手动清理了“无操作”数据，这一步骤对于实现收敛至关重要。然而，如图9所示，即使增加了训练步骤，模型的动作准确率仍然较低。相比之下，我们的UniAct框架表现出对噪声数据的鲁棒性。

# B.3. Fast adaptation to AIRBOT

# B.3. 快速适应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.

为了评估UniAct-0.5B和基线模型的快速适应能力，我们使用在预训练阶段未见过的AIRBOT上收集的新演示对它们进行微调。在本节中，我们提供了微调过程的详细信息。

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.

UniAct的微调设置:我们使用了4个A100 GPU和DeepSpeed对UniAct-0.5B进行微调。值得注意的是，我们从头开始训练一个新的MLP网络作为AIRBOT的异构头，同时保持其他模块冻结。微调过程持续了1小时。在此过程中使用的训练超参数详见表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 hours and 0.5 hours, respectively. To ensure robust performance, the training of OpenVLA continued until it reached an accuracy rate of on the training dataset, aligning with the official requirements in [31]. Training hyper-parameters for this process are illustrated in Tab 10.

基线模型的微调设置。基线模型OpenVLA和Octo的微调分别在8个A6000 GPU和2个4090 GPU上进行，分别持续了1.5小时和0.5小时。为了确保鲁棒性能，OpenVLA的训练持续到其在训练数据集上达到 的准确率，符合[31]中的官方要求。此过程的训练超参数见表10。

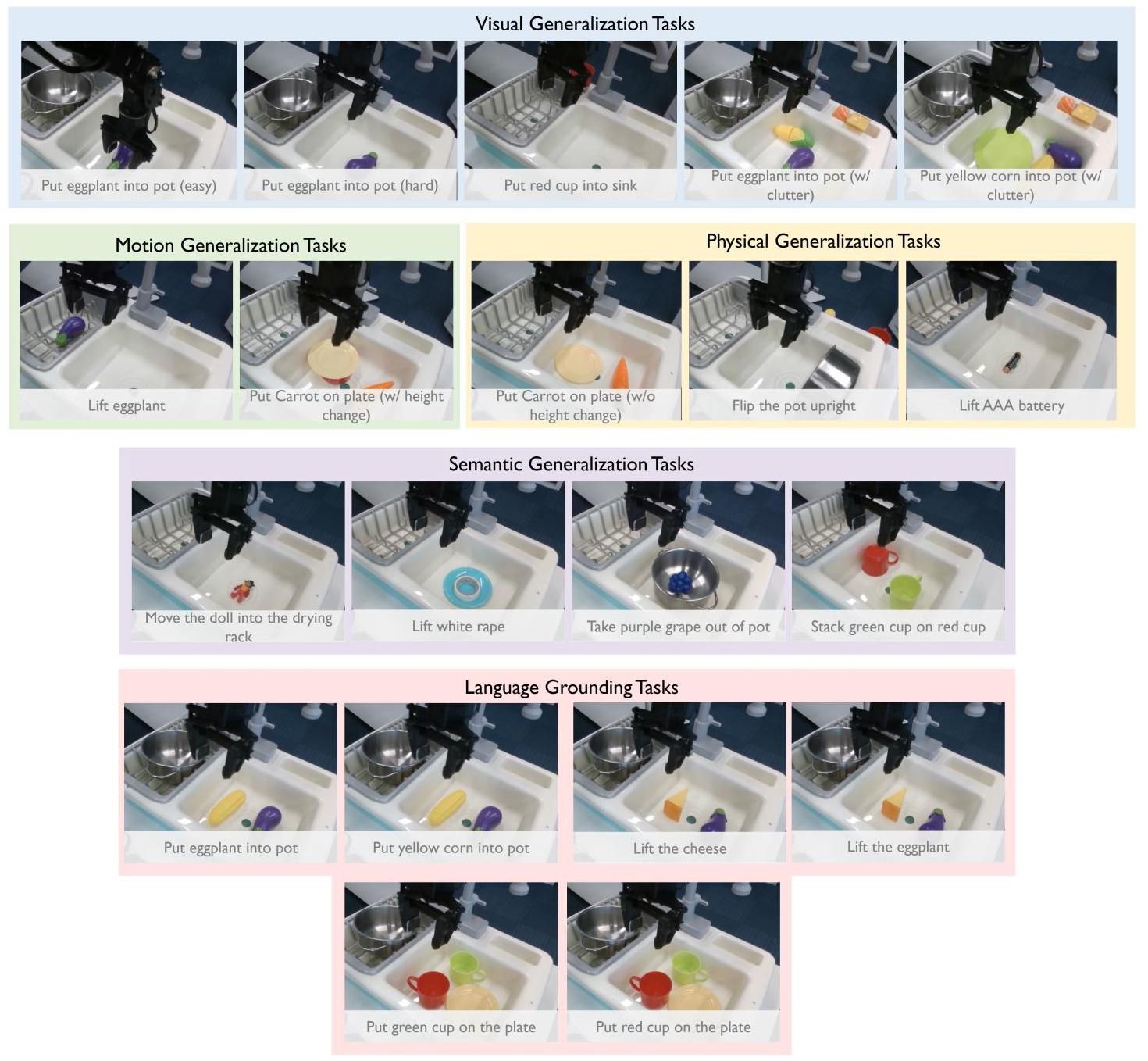


Figure 8. Illustration for WidowX evaluation tasks.

图8. WidowX评估任务示意图。

# C. More Related Works

# C. 更多相关工作

Embodied Models with Hierarchical Structures. Uni-Act 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 , 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.

具有分层结构的具身模型。UniAct类似于一种分层结构，首先推断出一个通用动作，然后将其转化为可执行的动作。一些工作也探索了分层结构，利用LLM或VLM的规划能力将原始指令分解为语言 , 57, 71]或视觉[7, 19, 66]子目标序列。然而，将这些语言或视觉子目标转化为精确动作仍然需要大量的动作标签，并且难以利用不同动作空间之间的共享结构。UniAct中的通用动作空间则作为一个更细粒度的技能库，可以有效地适应物理基础动作。该空间也是端到端训练的，最大化VLM的能力，以开发一个灵活且全面的结构。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 任务类别 | 任务名称 | Octo | OpenVLA | UniAct-0.5B(我们的) |
| 视觉泛化 | 将茄子放入锅中(简单) | 3.0 | 10.0 | 10.0 |
| 将茄子放入锅中 | 6.0 | 6.0 | 10.0 |
| 将杯子从台面放入水槽 | 0.0 | 2.0 | 1.0 |
| 将茄子放入锅中 | 3.0 | 8.0 | 5.0 |
| 将玉米放在盘子上 | 0.0 | 7.0 | 8.0 |
| 平均得分 | 2.4 | 6.6 | 6.8 |
| 运动泛化 | 拿起茄子 | 4.0 | 3.0 | 6.0 |
| 将胡萝卜放在盘子上 | 2.0 | 4.0 | 6.0 |
| 平均得分 | 3.0 | 3.5 | 6.0 |
| 物理泛化 | 将胡萝卜放在盘子上 | 2.0 | 9.0 | 9.0 |
| 将锅翻转直立 | 1.0 | 2.0 | 7.0 |
| 拿起AAA电池 | 0.0 | 7.0 | 5,0 |
| 平均得分 | 1.0 | 6.0 | 7.0 |
| 语义泛化 | 将玩偶移到晾衣架上 | 0.0 | 5.0 | 6.0 |
| 拿起白萝卜 | 0.0 | 1.0 | 2.0 |
| 将紫葡萄从锅中取出 | 4.5 | 5.0 | 5.0 |
| 将绿色杯子叠放在红色杯子上 | 3.25 | 8.25 | 2.5 |
| 平均得分 | 1.9 | 4.8 | 3.9 |
| 语言基础 | 将茄子放入锅中 | 6.0 | 7.0 | 5.0 |
| 将黄色玉米放入锅中 | 7.0 | 10.0 | 10.0 |
| 拿起奶酪 | 1.0 | 8.0 | 6.0 |
| 拿起茄子 | 6.0 | 8.0 | 9.0 |
| 将绿色杯子放在盘子上 | 3.0 | 10.0 | 6.0 |
| 将红色杯子放在盘子上 | 6.0 | 10.0 | 8.0 |
| 平均得分 | 4.8 | 8.8 | 7.3 |

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

表5. UniAct-0.5B与基线模型在WidowX评估中详细得分的比较。

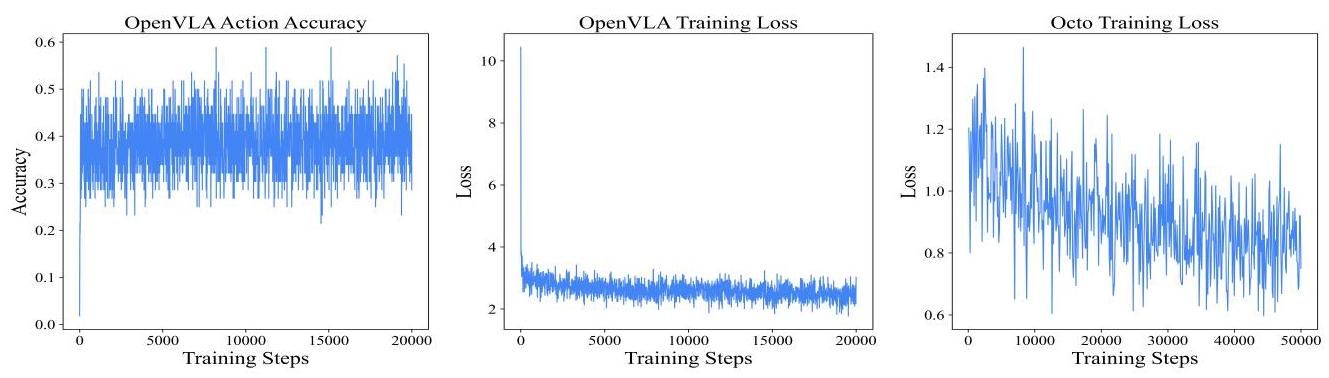


Figure 9. The learning curves for OpenVLA and Octo in LIBERO fintuning.

图9. OpenVLA和Octo在LIBERO微调中的学习曲线。

# D. Limitations and Future Works

# D. 局限性与未来工作

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

在本节中，我们讨论了本工作的局限性及相应的解决方案。我们希望这将激发更多有趣的研究。

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?

更多实现:在本研究中，我们探索了一个可以在不同动作空间之间共享的通用动作空间的概念。在当前版本中，UniAct主要在单机械臂上使用不同的控制接口进行评估。其背后的动机源于一个直观的观察:尽管控制接口存在差异，这些机械臂本质上表现出共同的物理运动。然而，这引发了一个关键问题:物理运动的共同性是否仅限于机械臂或类似的实现？

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

我们认为答案是否定的。未来的工作将探索更复杂实现的通用动作，例如双机械臂、灵巧机械手、四足机器人甚至自动驾驶汽车，这些系统在自由度和机械结构上存在差异。尽管它们具有很强的异质性，这些系统也可能与更简单的机械臂共享一些基本运动，有可能被纳入同一个通用动作空间。因此，我们希望开发一个“真正”通用的动作空间，能够编码任何物理实现的运动，同时识别独特特征并优先考虑共享的共性。这为未来研究提供了一个有前景的方向，具有跨系统跨实现控制的强大潜力。

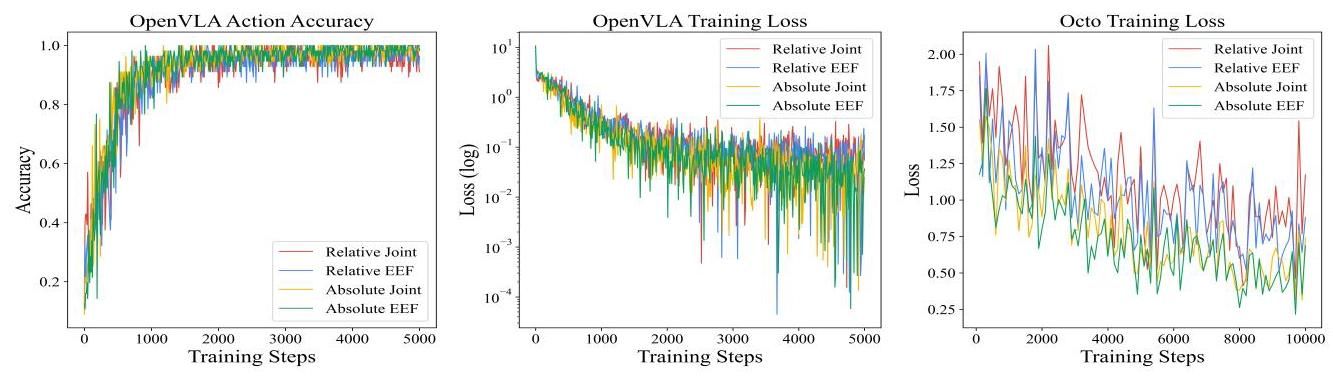


Figure 10. The learning curves for OpenVLA and Octo in AIRBOT finetuning.

图10. OpenVLA和Octo在AIRBOT微调中的学习曲线。

|  |  |  |
| --- | --- | --- |
| Model-Task | Config | Value |
| OpenVLA-LIBERO | Optimizer | AdamW |
| Batch size | 64 |
| Learning rate |  |
| Weight decay |  |
| Optimizer momentum |  |
| Octo-LIBERO | Optimizer | AdamW |
| Batch size | 64 |
| Learning rate |  |
| Weight decay |  |
| Optimizer momentum |  |
| Warmup iters | 2000 |
| OpenVLA-AIRBOT | Optimizer | AdamW |
| Batch size | 64 |
| Learning rate |  |
| Weight decay |  |
| Optimizer momentum |  |
| Octo-AIRBOT | Optimizer | AdamW |
| Batch size | 64 |
| Learning rate |  |
| Weight decay |  |
| Optimizer momentum |  |
| Warmup iters | 2000 |

|  |  |  |
| --- | --- | --- |
| 模型任务 | 配置 | 值 |
| OpenVLA-LIBERO | 优化器 | AdamW |
| 批量大小 | 64 |
| 学习率 |  |
| 权重衰减 |  |
| 优化器动量 |  |
| Octo-LIBERO | 优化器 | AdamW |
| 批量大小 | 64 |
| 学习率 |  |
| 权重衰减 |  |
| 优化器动量 |  |
| 预热迭代次数 | 2000 |
| OpenVLA-AIRBOT | 优化器 | AdamW |
| 批量大小 | 64 |
| 学习率 |  |
| 权重衰减 |  |
| 优化器动量 |  |
| Octo-AIRBOT | 优化器 | AdamW |
| 批量大小 | 64 |
| 学习率 |  |
| 权重衰减 |  |
| 优化器动量 |  |
| 预热迭代次数 | 2000 |

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

表6. 模拟中基线模型的微调超参数。

|  |  |
| --- | --- |
| config | value |
| optimizer | AdamW |
| batch size | 128 |
| learning rate |  |
| weight decay | 0. |
| optimizer momentum |  |
| iters | 10K |
| model precision | BFloat16 |

|  |  |
| --- | --- |
| 配置 | 值 |
| 优化器 | AdamW |
| 批量大小 | 128 |
| 学习率 |  |
| 权重衰减 | 0. |
| 优化器动量 |  |
| 迭代次数 | 10K |
| 模型精度 | BFloat16 |

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

表7. 在AIR-BOT上微调UniAct-0.5B的超参数

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.

更灵活的网络设计:在我们当前的实现中，UniAct为每个具体化使用相同的解码头——一个简单的MLP网络——以最小化过拟合风险并促进通用动作提取器的训练。然而，解码头的复杂性和参数数量应根据每个具体化的控制复杂性和其训练数据的多样性进行调整，这似乎更为合理。例如，一个双手机器人由于其增加的自由度，逻辑上需要一个比单臂机器人更复杂的解码头，以有效地学习和解码通用动作。

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.

展望未来，随着数据集扩展到包括更多来自不同具体化的演示，开发考虑每个具体化特定控制复杂性的专用解码头将变得至关重要。此外，将具体化特定信息(如本体感受数据或不同视图)纳入这些解码头可能会显著提高其性能。

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? Moreover, it’s critical to consider whether the more data always bring better result.

通用动作训练的扩展定律:虽然我们使用大量开源数据训练UniAct的努力取得了值得称赞的结果，但一个有趣的问题出现了:更多的数据是否明确地改善了通用动作空间？这个问题需要从多个角度进行彻底检查。首先，纳入更多具体化是否本质上增强了动作空间，还是为同一具体化积累多样化的任务演示更有价值？此外，考虑更多数据是否总是带来更好的结果也至关重要。

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.

研究数据量与我们的通用动作空间性能之间的关系可能非常有用。通过理解这些关系，我们可以更好地规划数据收集，以有效和高效地改进模型训练。

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.

更多利用通用动作。我们的工作突出了通用动作在部署跨具体化机器人策略中的强大能力。然而，我们相信通用动作的潜在应用远远超出了迄今为止的探索范围。

One promising direction is in the development of world models, where universal actions serve as a form of ’tok-enizer’. 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.

一个有前景的方向是世界模型的开发，其中通用动作作为一种“标记器”。这一角色涉及将复杂动作分解为标准化的、可理解的组成部分。通过使用通用动作空间进行规划，这些世界模型可以更准确地预测和模拟各种场景和环境中的结果。这种统一的理解和与世界互动的方法对于机器人中的高级规划和决策具有无可估量的价值。

1. \*Equal contribution

   \*同等贡献

   Corresponding author

   通讯作者 [↑](#footnote-ref-28)