Robotics: Science and Systems 2025
Los Angeles, CA, USA, June 21–25, 2025

SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Model

Delin Qu*[1,2], Haoming Song*[1,3], Qizhi Chen*[1,4], Dong Wang†[1], Yuanqi Yao[1], Xinyi Ye[1], Yan Ding[1], Zhigang Wang[1] Jiayuan Gu[5], Bin Zhao†[1,6], Xuelong Li[1,6]
[1]Shanghai AI Laboratory, [2]Fudan University, [3]Shanghai Jiao Tong University, [4]Zhejiang University, [5]ShanghaiTech University, [6]Northwestern Polytechnical University
https://spatialvla.github.io

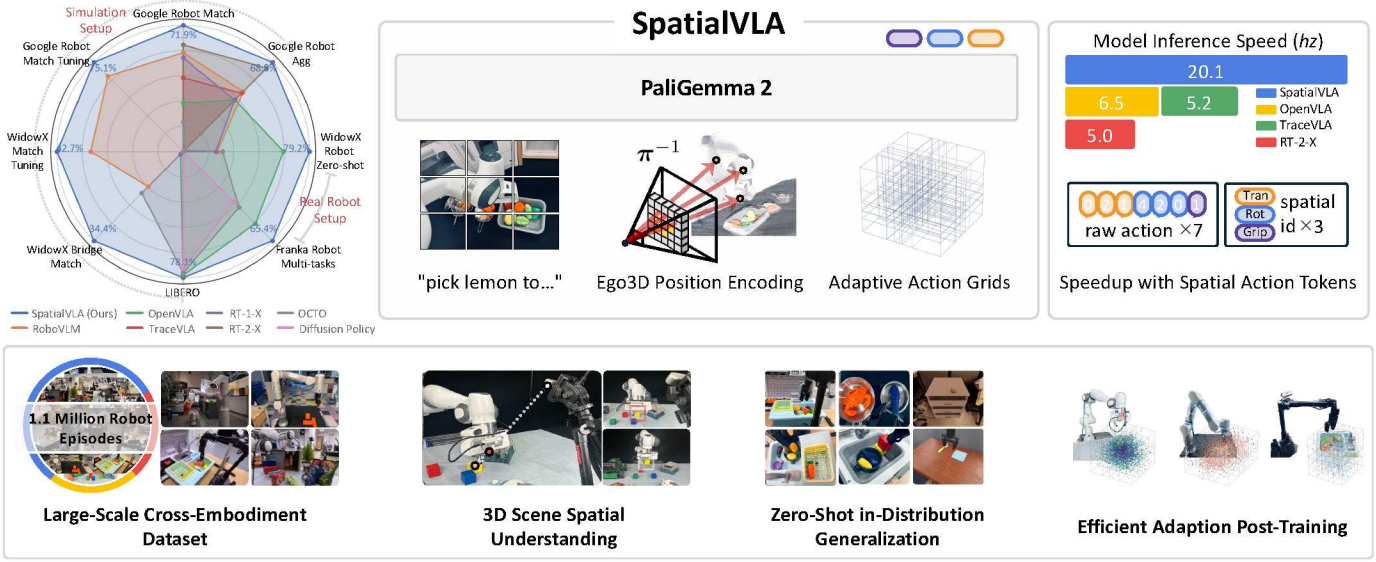


Fig. 1: We present SpatialVLA, a spatial-enhanced vision-language-action model that is trained on 1.1 Million real robot episodes. The model is equipped with Ego3D Position Encoding and Adaptive Action Grids to explore spatial representations for generalist robot policies, achieving superior 3D scene spatial understanding, zero-shot in-distribution generalization, and efficient adaption to new robot setups. The model achieves state-of-the-art performance across a diverse range of evaluations and shows significantly faster inference speed with fewer tokens per action.

***Abstract*—In this paper, we claim that spatial understanding is the keypoint in robot manipulation, and propose SpatialVLA to explore effective spatial representations for the robot foundation model. Specifically, we introduce *Ego3D Position Encoding* to inject 3D information into the input observations of the visual-language-action model, and propose *Adaptive Action Grids* to represent spatial robot movement actions with adaptive discretized action grids, facilitating learning generalizable and transferrable spatial action knowledge for cross-robot control. SpatialVLA is first pre-trained on top of a vision-language model with 1.1 Million real-world robot episodes, to learn a generalist manipulation policy across multiple robot environments and tasks. After pre-training, SpatialVLA is directly applied to perform numerous tasks in a zero-shot manner. The superior results in both simulation and real-world robots demonstrate its advantage of inferring complex robot motion trajectories and its strong in-domain multi-task generalization ability. We further show the proposed *Adaptive Action Grids* offer a new and effective way to fine-tune the pre-trained SpatialVLA model for new simulation and real-world setups, where the pre-learned action grids are re-discretized to capture robot-specific spatial action movements of new setups. The superior results from extensive evaluations demonstrate the exceptional in-distribution generalization and out-of-distribution adaptation capability, highlighting the crucial benefit of the proposed spatial-aware representations for generalist robot policy learning. All the details and codes are open-sourced.**

∗ Authors contributed equally: dlqu22@m.fudan.edu.cn. † Corresponding authors: dongwang.dw93@gmail.com.

I. Introduction

Generalist robot policies that are capable of interacting with the physical environment, adapting to various embodiments, and performing complex tasks have been a long-standing pursuit in robotics [6, 3, 16, 8, 65]. Recent advances in Vision-Language-Action (VLA) models [7, 30, 5, 33] show a promising paradigm in building such generalist policy by fine-tuning the pre-trained Vision-Language Models (VLMs) [1, 55, 50, 37] on diverse robot data [13, 29, 18]. The key to the success of this paradigm lies in adapting the generalization power of VLMs to numerous robot manipulation tasks, as well

as specific architectural designs that synergize the VLM backbone and robot action output head. Nonetheless, existing VLA models are primarily confined to 2D observation inputs and lack precise perception and comprehension of the 3D physical world — where humans instinctively construct rich, structured mental representations of space, effortlessly aligning objects within a canonical, intuitive, and even personally tailored workspace for manipulation [20, 40, 52, 63, 67]. Therefore, an essential question for the field now is ***how to effectively equip the VLA models with a profound spatial understanding of the 3D physical world?***

However, developing such generalist robot policies with 3D spatial intelligence encounters two primary challenges in the aspects of robot observation and action. Firstly, the observations from different robot embodiments are not 3D-aligned, because the camera sensors of different robots are various and mounted at different places (*e.g.* wrist and/or third-person), resulting in non-calibrated 3D observation spaces. Secondly, different robots have different action movement characteristics to accomplish diverse tasks, due to different degrees of freedom, motion controllers, workspace configurations, and task complexity, leading to significant difficulty in learning generalizable spatial actions. Despite some attempts in generalist policy learning across heterogeneous robots [48, 13, 30, 65], advancement in 3D spatial understanding abilities of generalist policy has significantly lagged behind. This is largely attributed to the heterogeneity in robot observation and action information. The solutions to the above challenges require spatial-aligned robot observation and action representations for cross-embodiment control and adaptation in the universal 3D physical world.

In this work, as illustrated in Fig. 1, we propose a generalist robot policy SpatialVLA, which equips the VLA model with 3D spatial intelligence by exploring aligned spatial representations of robot observation and action signals. SpatialVLA perceives 3D world through ***Egocentric 3D (Ego3D) Position Encoding*** to integrate 3D spatial context with semantic features. This position encoding is derived in the egocentric camera frame that eliminates the need for specific robot-camera calibration, which is universally applicable to various robot embodiments. As for robot actions, SpatialVLA unifies the action space of various robots via ***Adaptive Action Grids***, which discretizes the continuous robot actions into adaptive spatial grids according to statistical action distributions on the whole robot episodes and learns spatial action tokens on these grids to align cross-robot actions with the 3D spatial structure of the physical world. Crucially, after pre-training, the learned spatial action grids demonstrate a superior capability in adapting to new robot environments via adaptively grid re-discretization, providing a flexible and effective approach to robot-specific post-training. We find that the proposed model SpatialVLA bridges observation inputs and action outputs in a universal robot-agnostic manner, which explores powerful 3D spatial-aware representations to enhance the VLA model.

We extensively evaluate and ablate SpatialVLA on diverse robot manipulation tasks and different robot embodiments in both simulation and real-world, including 24 real-robot tasks and 3 simulation environments. To broadly test SpatialVLA as a generalist robot policy, we examine the model's abilities in zero-shot in-distribution robot control and new robot setup adaption abilities with instruction following, 3D scene structure understanding, and fine-tuning to new robot environments. The evaluation setups include view/texture/lighting change, unseen objects, unseen robot environment, and challenging spatial layout changes in robot setups and environments, demonstrating remarkable generalizability and transferability of SpatialVLA with spatial-aware representations. In summary, the contributions of this work consist of a novel generalist robot policy that explores spatial representations for robot foundation models, sophisticated designs on Ego3D Position Encoding and Adaptive Action Grids for effective 3D-awareness injection, and superior evaluation results across various robot setups and tasks.

II. Related Work

Generalist Robot Polices. Recent advances in robotics have witnessed a trend towards developing multi-task "generalist" robot policies to perform diverse tasks, rather than one specific task. Some early works [49, 57, 21, 6, 61, 76, 22] achieve great success in learning a language-conditioned visual multi-task policy on a single embodiment with pre-trained visual/text encoder, thereby lacking the ability to adapt new robot embodiment. More recent efforts [48, 39, 65] explore to use large-scale, cross-embodiment robot datasets [13] for generalist polices pre-training, supporting effective fine-tuning to new robot setups. Notably, Octo [48] proposes a flexible transformer-based architecture to unify different configurations in Open X-Embodiment (OXE) dataset [13], and the trained policy can solve a variety of in-domain tasks in zero-shot and achieves strong performance in the new embodiment after fine-tuning. With the same cross-embodiment robot datasets, RDT [39] pre-trains a 1.2B-parameter diffusion-based generalist model and fine-tunes it for complex bimanual manipulation. Moreover, HPT [65] proposes a modular architecture to align data across heterogeneous embodiments into a shared representation via embodiment-specific stem module, embracing the heterogeneity in data through pre-training.

Vision-Language-Action Models. Recently, several studies [34, 7, 30, 31, 71, 33, 70, 51] propose to build generalist robot policies by extending pre-trained VLMs with ability to robot action generation. As a pioneer, RT-2 [7] fine-tune VLM PaLI-X [11] on both large-scale vision-language data and robot demonstration data via autoregressive next token prediction, where robot actions are discretized into 256 bins and represented as separate tokens analogous to text tokens. OpenVLA [30] adopts a similar action discretization approach and fine-tune Prismatic VLM [28] only on the OXE dataset [13], which consists of robot data from 22 different robot embodiments across 21 institutions. CogACT [31] and TraceVLA [71] continue to fine-tune the trained OpenVLA model with the new attached diffusion action module and visual trace prompting separately. Moreover, π_0 [5] adapts

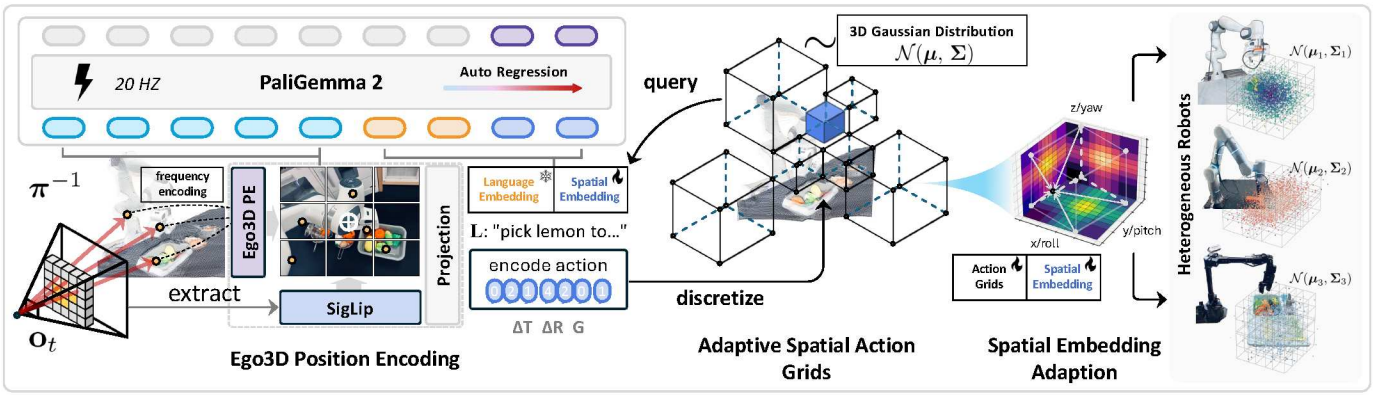


Fig. 2: **Overview of SpatialVLA**. Given an image observation $\mathbf{o}_t$ and a task instruction $\mathbf{L}$, the model processes the image using Ego3D Position Encoding and auto-regressively predicts spatial action tokens, which are then de-tokenized to generate continuous actions $\mathbf{A}_t$ for robot control. The model comprises three key components: (1) SigLIP vision encoder extracts 2D semantic features, which are then infused with 3D spatial context via Ego3D Position Encoding; (2) continuous 7D actions $\Delta\mathbf{T}, \Delta\mathbf{R}, \mathbf{G}$ are translated to 3 spatial action tokens by querying Adaptive Action Grids and auto-regressively predicted and de-tokenized for robot control; (3) in post-training, action grids and spatial embeddings are adapted from new Gaussian distributions to facilitate effective transfer to new robot setups.

PaliGemma VLM to robot control by adding a separate action expert module that produces continuous actions via flow matching, and the model can then be prompted for zero-shot control or fine-tuned on high-quality data to enable complex dexterous manipulation tasks. Notably, while these models benefit from VLMs' capabilities and show some zero-shot capabilities, a sophisticated fine-tuning step with new data is essential and required for complex tasks or new robot setups.

3D Foundation Models for Robotics. Some researches [73, 10, 19, 24, 53, 25, 69] have focused on extending the generalist ability of LLMs and VLMs from language-vision towards the 3D world. 3D-LLM [24] integrates a 3D feature extractor with 2D VLMs backbone and train 3D-LLMs on a wide variety of tasks, including dense captioning, 3D question answering, task decomposition, 3D grounding, 3D-assisted dialog, navigation, and so on. LLaVA-3D [10] extends the 2D LLaVA's capabilities with the proposed 3D patches to bridge 2D features within a 3D space for 3D spatial understanding. Similarly, LEO [25] trains an embodied multi-modal generalist agent that can take egocentric 2D images, 3D point clouds, and texts as task input and handle comprehensive tasks within the 3D environment. Moreover, 3D-VLA [69] builds a generative world model on top of 3D-based LLM to perform 3D reasoning and localization, multimodal goal generation, and embodied action planning. LEO and 3D-VLA are closely related to our work, but their attention is on 3D world understanding and prediction, ignoring the 3D spatial characteristics in the robot action space.

III. Methodology

In this section, we describe SpatialVLA model and its training framework in detail. Our model with the proposed Ego3D position encoding and adaptive action grids to capture and learn 3D spatial knowledge for generalizable robot control, which we describe in Sec. III-A. Next, we detail the training procedure of SpatialVLA that consists of a pre-training stage and a post-training stage in Sec. III-B. The pre-training aims to learn generalizable knowledge with large-scale cross-robot data and the goal of post-training is to adapt pre-trained model to specific downstream robot embodiments and tasks.

A. The SpatialVLA Model Architecture

As illustrated in Fig. 2, SpatialVLA is developed based on a vision-language model to inherit the general world knowledge. Formally, SpatialVLA takes image observations $\mathbf{o}_t = \{\mathbf{I}_t^1, .., \mathbf{I}_t^n\}$ and a natural language task instruction $\mathbf{L}$ as inputs, and then learns a mapping function $\tau(\cdot)$ to generate a sequence of robot actions $\mathbf{A}_t = [\mathbf{a}_t, \mathbf{a}_{t+1}, ..., \mathbf{a}_{t+H-1}]$, *i.e.*, $\mathbf{A}_t = \mathcal{F}(\mathbf{o}_t, L)$. To empower SpatialVLA with 3D spatial intelligence, we augment the VLM backbone with robotics-specific 3D-aware inputs and outputs, namely, *Ego3D Position Encoding* and *Adaptive Action Grids*. The ego3D position encoding representation $\mathbf{O_{3d}}$ aims to capture 3D scene structure via integrating 3D spatial information with 2D semantic features. The adaptive action grids are designed to represent the continuous distribution of robot actions $\mathbf{a}$ with a set of discrete spatial action tokens $\mathfrak{a} = \{\mathfrak{a}^1, ..., \mathfrak{a}^V\}$. During training, SpatialVLA model is trained to take the ego3D position encoding representation $\mathbf{O_{3d}}$ and natural language task instruction $\mathbf{L}$ as inputs, and autoregressively generate spatial action tokens $\tilde{\mathfrak{a}}_t$ using the cross-entropy objective $\mathcal{L}$,

$$\mathfrak{L}(\theta) = \mathbb{E}_{p(\mathbf{A}_t|\mathbf{o}_t)}\mathcal{L}(\mathfrak{a}_t, \tilde{\mathfrak{a}}_t)), \tag{1}$$

where the predicted action tokens $\tilde{\mathfrak{a}}_t = \tau(\mathbf{O_{3d}}, \mathbf{L})$ is the de-tokenized into continuous action signals $\mathbf{a}_t$ for robot control. More details of the model architecture and action encoding can be found in Appendix. B.

Ego3D Position Encoding. The proposed Ego3D position encoding integrates depth information from the camera frame and image pixels to construct an egocentric 3D coordinate system, which eliminates the need for robot-camera extrinsic calibration and is agnostic to specific robot setups. Specifically, we use ZoeDepth [4] to estimate depth map D and obtain

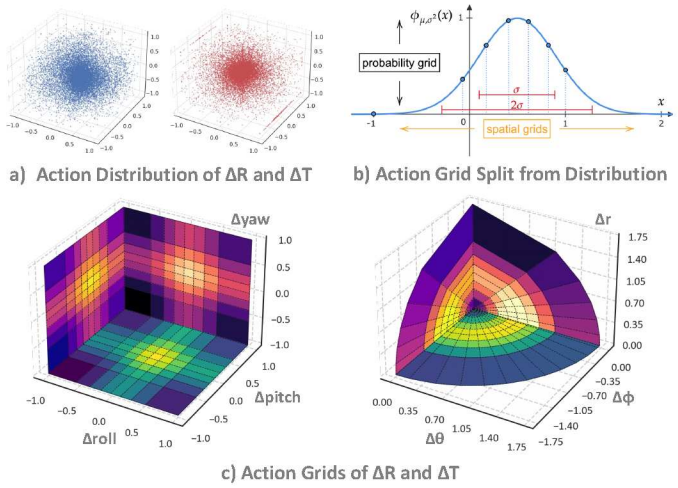


Fig. 3: Illustration of adaptive action grids. (a) Statistics of translation and rotation action movements on the whole pre-training mixture, (b) grids are split on each action variable according to the probability density function of fitted Gaussian distribution, and (c) the obtained adaptive action grids in translation and rotation action spaces.

pixel's 3D position $\mathbf{p} = \{x, y, z\}$ in the egocentric 3D coordinate system via back-projection π^{-1} with camera intrinsic parameters. Then, as illustrated in Fig. 2, we first employ SigLIP [68] visual encoder to extract 2D semantic visual features $\mathbf{X} \in \mathbb{R}^{d \times h \times w}$ to inherit the alignment between vision and language, and calculate their corresponding 3D positions $\mathbf{P} \in \mathbb{R}^{3 \times h \times w}$ in the egocentric 3D coordinate system. The egocentric 3D positions $\mathbf{P}$ are then encoded into 3D position embeddings $\mathbf{P}^{'} \in \mathbb{R}^{d \times h \times w}$ through a sinusoidal function $\boldsymbol{\gamma}(\cdot)$ following by a learnable MLP. The egocentric 3D spatial representations $\mathbf{O_{3d}} \in \mathbb{R}^{d \times h \times w}$ are obtained by adding 3D position embedding $\mathbf{P}^{'}$ and 2D path visual tokens $\mathbf{X}$, depicted as follows,

$$\mathbf{O_{3d}} = \mathbf{X} + \mathbf{P}^{'} = \mathbf{X} + \text{MLP}(\boldsymbol{\gamma}(\mathbf{P})). \tag{2}$$

Adaptive Action Grids. In order to auto-regressively generate continuous robot actions with pre-trained VLM backbone, we design Adaptive Action Grids to translate continuous robot actions to discrete grids that are represented as tokenized classes for prediction. Specifically, for a single-arm robot, its actions consist of seven dimensions for movement $\mathbf{a} = \{\text{x, y, z, roll, pitch, yaw, grip}\}$, and are split into three parts as follows,

$$\mathbf{a} = \{\mathbf{a}_{\text{trans}},\ \mathbf{a}_{\text{rot}},\ \mathbf{a}_{\text{grip}}\}, \tag{3}$$

where $\mathbf{a}_{\text{trans}} = \{\text{x, y, z}\}$ represents translation movements $\Delta\mathbf{T}$, $\mathbf{a}_{\text{rot}} = \{\text{roll, pitch, yaw}\}$ denotes rotation movements $\Delta\mathbf{R}$, and $\mathbf{a}_{\text{grip}} = \{\text{grip}\}$ consists of two discrete tokens that represent opening and closing gripper actions. Moreover, we transform the translation movements (x, y, z) into polar coordinates (ϕ, θ, r) to disentangle movement direction (ϕ, θ) and distance r.

As illustrated in Fig. 3, for tokenizing continuous translation and rotation movements, we first normalize each action variable into $[-1, 1]$ for each robot environment and statistic the translation and rotation movements $\Delta\mathbf{R} = \{\text{roll, pitch, yaw}\}$, $\Delta\mathbf{T} = \{\phi, \theta, r\}$ on the whole dataset mixture (see Appendix. G), following with a parameterized Gaussian distribution fitting $\mathcal{N}(\mu^a, \Sigma^a)$. Then, the continuous actions are split into M intervals $\mathbf{G}_{i=1,..,\text{M}} = \{[a_1 = -1, a_2), ..., [a_{\text{M-1}}, a_{\text{M}} = 1]\}$ with equal probability $1/\text{M}$ on each normalized action variable, *i.e.*,

$$a_2, ..., a_{\text{M}} = \underset{a_2,...,a_{\text{M}}}{\arg\min} \int_{a_i}^{a_{i+1}} f(x)dx - 1/\text{M},\ i = 1, ..., \text{M} \tag{4}$$

where $f(x)$ is the probability density function of Gaussian distribution $\mathcal{N}(\mu^a, \Sigma^a)$. Note that we split more grids on $\{\phi, \theta\}$ to capture fine-grained movement direction other than movement distance r. Suppose M_ϕ, M_θ, M_r are the numbers of the discrete bins on variable (ϕ, θ, r), then the translation space consists of $\text{M}_{\text{trans}} = \text{M}_\phi \cdot \text{M}_\theta \cdot \text{M}_r$ discrete spatial grids $\mathfrak{a}_{\text{trans}} = \{\mathfrak{a}^1, ..., \mathfrak{a}^{\text{M}_{\text{trans}}}\}$. Similarly, there are $\text{M}_{\text{rot}} = \text{M}_{\text{roll}} \cdot \text{M}_{\text{pitch}} \cdot \text{M}_{\text{yaw}}$ 3D discrete grids $\mathfrak{a}_{\text{rot}} = \{\mathfrak{a}^1, ..., \mathfrak{a}^{\text{M}_{\text{rot}}}\}$ in rotation 3D spatial space. Then, the associated learnable spatial action token embeddings are defined as follows,

$$\mathbf{E}_{\mathfrak{a}} = \{\mathbf{E}_{\text{trans}},\ \mathbf{E}_{\text{rot}},\ \mathbf{E}_{\text{grip}}\}, \tag{5}$$

where $\mathbf{E}_{\text{trans}} \in \mathbb{R}^{d \times \text{M}_{\text{trans}}}$, $\mathbf{E}_{\text{rot}} \in \mathbb{R}^{d \times \text{M}_{\text{rot}}}$, $\mathbf{E}_{\text{grip}} \in \mathbb{R}^{d \times 2}$ denote the translation, rotation, and gripper actions, and the total number of action tokens is $V = \text{M}_{\text{trans}} + \text{M}_{\text{rot}} + 2$. After training, these learned spatial action tokens capture general robot action knowledge and show a surprising ability in new robot embodiment adaption, as discussed in Sec. III-B. Moreover, it is worth noting that the model only needs to generate 3 tokens for one-step robot actions rather than 7 tokens as in RT-1 [6], RT-2 [7] and OpenVLA [30], achieving in fast model inference speed.

B. The Pre-training and Post-training Scheme

To obtain a generalist robot policy model, the training procedure of SpatialVLA consists of pre-training stage and post-training stage. Pre-training stage aims to learn generalizable knowledge across diverse tasks and robots from a large-scale dataset mixture, while the post-training stage adapts the pre-trained model into new robot embodiments or new tasks. In the following, we discuss the dataset mixture and key designs for implementing this two-stage training procedure.

Pre-training Procedure. We train SpatialVLA from Paligemma2 backbone [62] on a cross-robot dataset mixture with 1.1 Million real robot demonstrations $\{\zeta_1, ..., \zeta_n\}$, covering a diverse range of robot embodiments, scenes, and tasks. This pre-training dataset mixture consists of a subset of OXE [13] and the RH20T dataset [18] and we modify the mixture weights from OpenVLA [30] according to the real-word testing performance of individual dataset, which are exhibited in Appendix. A. At the beginning of pre-training, the embeddings $\mathbf{E}_{\mathfrak{a}}$ of spatial action tokens and parameters of MLP in egocentric 3D spatial representation are randomly

initialized, and then they are optimized during training, as well as the parameters of vision encoder and LLM backbone. At each training step, a batch of data pairs is extracted at random timesteps $t_1, ..., t_B$ from shuffled demonstrations $\{\zeta_i, ..., \zeta_j\}$, *i.e.*, a batch of tuple $[(\mathbf{o}_{t_1}, \mathbf{A}_{t_1}, \mathbf{L}_i), ..., (\mathbf{o}_{t_B}, \mathbf{A}_{t_B}, \mathbf{L}_j)]$, and then SpatialVLA is trained with a standard auto-regressive next-token prediction objective in eq. (1). Importantly, the embeddings of text tokens $\mathbf{E}_{\text{text}}$ are frozen to maintain the general world knowledge in pre-trained VLM, and the experimental results show this frozen operation is beneficial for the instruction following ability. Moreover, as discussed in OpenVLA [30], DROID dataset [29] are removed from the data mixture for the final third of pre-training to improve the quality of the pre-trained SpatialVLA model.

Post-training Designs. In the post-training stage, we fine-tune our model with robot-specific demonstrations to adapt it to new tasks and robot setups. Prior works have mainly focused on fine-tuning pre-trained VLA models using full-parameter or LoRA fine-tuning, with little attention to effective techniques for the post-training stage. In this paper, we investigate the potentials of the proposed spatial action tokenizer for quick adaption to new robot setups, namely ***Spatial Embedding Adaption***, providing a new and effective way for useful post-training. In detail, we fit a new Gaussian distribution $\mathcal{N}(\mu_{\text{new}}, \Sigma_{\text{new}})$ for each action variable on post-training datasets and create discrete spatial action grids $\mathbf{G}_{\text{new}}$ in translation and rotation movement to construct action grids $\mathbf{G}_{\text{new}}$ and tokens $\mathfrak{a}_{\text{new}}$, where the embeddings of new spatial action tokens $\mathbf{E}_{\mathfrak{a}_{\text{new}}}$ are initialized by trilinear interpolation with pre-trained action tokens $\mathbf{E}_{\mathfrak{a}}$. These action token embeddings $\mathbf{E}_{\mathfrak{a}_{\text{new}}}$ and model parameters are then optimized with the next-token prediction objective.

Formally, for new spatial action grids $\mathbf{G}_{\text{new}}$, suppose i-th 3D grid $\mathbf{G}^i_{\text{new}}$ in translation space $\mathbf{a}^{\text{new}}_{\text{trans}}$ with centroid $(\phi^i_{\text{new}},\ \theta^i_{\text{new}},\ r^i_{\text{new}})$ and its adjacent 3D grids from the pre-trained action grids are $\mathbf{G}^{\text{adj}} = \{\mathbf{G}^1, ..., \mathbf{G}^K\}$. The embedding of new i-th action token $\mathbf{e}^i_{\mathfrak{a}_{\text{new}}}$ are initialized by trilinear interpolation with $\mathbf{G}^{\text{adj}}$, as follows,

$$\mathbf{e}^i_{\mathfrak{a}_{\text{new}}} = \sum_{j=1}^{K} w_j \mathbf{e}^j, \tag{6}$$

where $\mathbf{e}^j_{\mathfrak{a}} \in \mathbb{R}^d$ are the embeddings of the pre-trained action grids, w_j is the weights calculated by the normalized distances between centroid $(\phi^i_{\text{new}},\ \theta^i_{\text{new}},\ r^i_{\text{new}})$ and adjacent centroids. Note that the new action tokens of rotation $\mathbf{E}^{\text{new}}_{\mathfrak{a}_{\text{rot}}}$ are initialized in the same way. With this embedding initialization, the new action tokenizer is capable of effectively transferring pre-trained spatial action knowledge to new robot setups.

IV. Experiment

The goal of our experimental evaluations is to test SpatialVLA's ability to serve as a generalist robot control policy out of the box, as well as be a good initialization for fine-tuning to new robot tasks. Our extensive experiments consist of zero-shot evaluations and adaption to downstream tasks in both simulation and real-world. SpatialVLA is compared to previous state-of-the-art robot foundation models and alternative designs in spatial representations. Concretely, experiments seek to answer the following research questions:

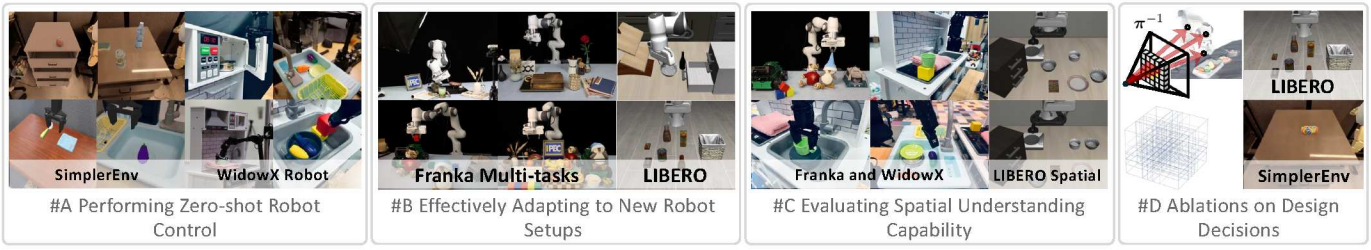
1) How well does SpatialVLA directly perform on a variety of in-distribution tasks after pre-training on large-scale robotic data mixture?
2) Can SpatialVLA be effectively fine-tuned on new robot setup and task?
3) How well does SpatialVLA perform in scenarios that require spatial understanding?
4) To what extent do Egocentric 3D Spatial Representations and Adaptive Spatial Action Grids improve the performance of SpatialVLA?

To answer these questions, as shown in Fig. 4, we evaluate SpatialVLA's capabilities across a representative spectrum of 7 different robot learning scenarios with 24 real-robot tasks and 3 simulation environments. Firstly, we evaluate SpatialVLA in both SimplerEnv [35] simulation and the real-world WidowX robot platform (BridgeV2 [64] [64] setups), testing its out-of-the-box control capabilities on different robots with setups matching the pre-training dataset. Second, we assess the fine-tuning efficacy of our method in both simulation and real-world settings, including LIBERO [36] and new Franka robot setups, to adapt to new robot environments and tasks. Then, we design 4 special tasks that require precise spatial understanding in 2 different real-world robot environments to test the effectiveness of spatial representations of SpatialVLA. Finally, we conduct comprehensive ablation studies on a mixture of Fractal [6] and BridgeV2 [64] datasets to verify our design decisions in SpatialVLA. For more details on evaluation setups, see Appendix. E.

Implementation Details. The SpatialVLA model is pre-trained with 1.1 Million real-robot demonstrations from the OXE [13] and RH20T dataset [18] on a cluster of 64 A100 GPUs for 10 days, using a batch size of 2048. For input robot observation, the SpatialVLA policy is only conditioned on one third-person camera and takes one image for constructing egocentric 3D spatial representations. For output robot actions, the SpatialVLA policy predicts a chunk of $T = 4$ future actions (12 spatial action tokens from total $V = 8194$ tokens) and executes the ensemble actions before predicting the next chunk. During inference, SpatialVLA requires 8.5GB of GPU memory and runs at approximately 20Hz on one NVIDIA RTX 4090 GPU to run evaluations in both simulation and real-world. For more details about model training and deployment, please refer to the Appendix. D.

A. *Performing Zero-shot Robot Control*

Evaluation Setups and Baselines. To assess the robustness of SpatialVLA in diverse environmental variations, we employ the SimplerEnv simulation benchmark [35] to evaluate visual matching and variant aggregation metrics. SimplerEnv features WidowX and Google Robot setups, providing diverse manipulation scenarios with varied lighting, color, textures, and robot camera pose conditions, bridging the visual appearance



#A Performing Zero-shot Robot Control | #B Effectively Adapting to New Robot Setups | #C Evaluating Spatial Understanding Capability | #D Ablations on Design Decisions

Fig. 4: **Experiment Setup.** We evaluate SpatialVLA across 7 robot learning scenarios, 16 real-robot tasks, and 48 simulation setups, focusing on three key aspects: zero-shot control, adaptability to new setups, and spatial understanding. We also conduct a thorough ablation study on a mixed Fractal and Bridge dataset to verify our design decisions.

TABLE I: **SimplerEnv evaluation across different policies on Google Robot tasks**. The zero-shot and fine-tuning results denote performance of OXE dataset [13] pre-trained models and Fractal dataset [6] fine-tuned models, respectively.

Model	Visual Matching				Variant Aggregation			
	Pick Coke Can	Move Near	Open/Close Drawer	#Average	Pick Coke Can	Move Near	Open/Close Drawer	#Average
RT-1 [6] (Begin)	2.7%	5.0%	13.9%	6.8%	2.2%	4.0%	6.9%	4.2%
RT-1 [6] (15%)	71.0%	35.4%	56.5%	60.2%	81.3%	44.6%	26.7%	56.2%
RT-1 [6] (Converged)	85.7%	44.2%	73.0%	74.6%	89.8%	50.0%	32.3%	63.3%
HPT [65]	56.0%	60.0%	24.0%	46.0%	———	———	———	———
TraceVLA [71]	28.0%	53.7%	57.0%	42.0%	60.0%	56.4%	31.0%	45.0%
RT-1-X [13]	56.7%	31.7%	59.7%	53.4%	49.0%	32.3%	29.4%	39.6%
RT-2-X [13]	78.7%	77.9%	25.0%	60.7%	82.3%	79.2%	35.3%	64.3%
Octo-Base [48]	17.0%	4.2%	22.7%	16.8%	0.6%	3.1%	1.1%	1.1%
OpenVLA [30]	16.3%	46.2%	35.6%	27.7%	54.5%	47.7%	17.7%	39.8%
RoboVLM (zero-shot) [32]	72.7%	66.3%	26.8%	56.3%	68.3%	56.0%	8.5%	46.3%
RoboVLM (fine-tuning) [32]	77.3%	61.7%	43.5%	63.4%	75.6%	60.0%	10.6%	51.3%
π_0^* (BF16 uniform) [5]	88.0%	80.3%	56.0%	70.1%				
SpatialVLA (zero-shot)	81.0%	69.6%	59.3%	**71.9%**	89.5%	71.7%	36.2%	**68.8%**
SpatialVLA (fine-tuning)	86.0%	77.9%	57.4%	**75.1%**	88.0%	72.7%	41.8%	**70.7%**

π_0^*: The results are referred from open-pi-zero.

TABLE II: **SimplerEnv evaluation across different policies on WidowX Robot tasks**. The zero-shot and fine-tuning results denote the performance of OXE dataset [13] pre-trained models and BridgeData V2 [64] fine-tuned models, respectively.

Model	Put Spoon on Towel		Put Carrot on Plate		Stack Green Block on Yellow Block		Put Eggplant in Yellow Basket		#Overall
	Grasp Spoon	Success	Grasp Carrot	Success	Grasp Green Block	Success	Grasp Eggplant	Success	Average
RT-1-X [13]	16.7%	0%	20.8%	4.2%	8.3%	0%	0.0%	0%	1.1%
Octo-Base [48]	34.7%	12.5%	52.8%	8.3%	31.9%	0%	66.7%	43.1%	16.0%
Octo-Small [48]	77.8%	47.2%	27.8%	9.7%	40.3%	4.2%	87.5%	56.9%	30.0%
OpenVLA [30]	4.1%	0%	33.3%	0%	12.5%	0%	8.3%	4.1%	1.0%
RoboVLM (zero-shot) [32]	37.5%	20.8%	33.3%	25.0%	8.3%	8.3%	0.0%	0%	13.5%
RoboVLM (fine-tuning) [32]	54.2%	29.2%	25.0%	25.0%	45.8%	12.5%	58.3%	58.3%	31.3%
SpatialVLA (zero-shot)	25.0%	20.8%	41.7%	20.8%	58.3%	25.0%	79.2%	70.8%	**34.4%**
SpatialVLA (fine-tuning)	20.8%	16.7%	29.2%	25.0%	62.5%	29.2%	100.0%	100.0%	**42.7%**

gap between real and simulated environments. We compare our model with the latest state-of-the-art generalist manipulation policies, including RT-1 [6], RT-1-X [13], RT-2-X [13], Octo [48], OpenVLA [30], HPT [65], TraceVLA [71], and RoboVLM [32]. Where RT-1-X, RT-2-X, Octo, OpenVLA, HPT, TraceVLA, and RoboVLM are trained with mixtures of OXE dataset [13]. Since RT-1 is trained with the Google Fractal Dataset [6], we also compare RT-1 with our method fine-tuned on the Google Fractal and BridgeData V2 [64].

For a more comprehensive evaluation, we conduct experiments on a real-world WidowX robot platform from the BridgeData V2 evaluation [64]. As shown in Fig. 5, we design seven task suites for the WidowX robot, encompassing **language grounding**, **semantic understanding** (unseen background and poses), and **motion distractors** (manually move the object). All generalist manipulation policies, including Octo, RT-1-X, OpenVLA, and RoboVLM, are evaluated across 7 task suites with 11 trials each, resulting in a total of 77 rollouts. A more detailed breakdown of all tasks and policy settings can be found in the Appendix. E.

SimplerEnv Evaluation of Google Robot and WidowX. Tab. I summarizes the zero-shot and fine-tuning results across different manipulation policies on the Google Robot setup. On average, SpatialVLA achieves the highest overall visual matching and variant aggregation performance with a significant margin. Our SpatialVLA model yields 71.9% and 75.1% Visual Matching scores in zero-shot and fine-tuning settings, surpassing the second-best policy, RoboVLM, by +15.6% and +11.7% margins. Notably, our model trained from scratch on OXE mixture with RH20T surpasses the state-of-the-art closed-source model RT-2-X [13], achieving superior performance in Visual Matching (71.9% vs 60.7%) and Variant Aggregation (68.8% vs 64.3%), while using significantly fewer model parameters (3.5B vs 55B). Qualitatively, we find that SpatialVLA exhibits greater generalizability and robustness across diverse robotic manipulation tasks and environmental

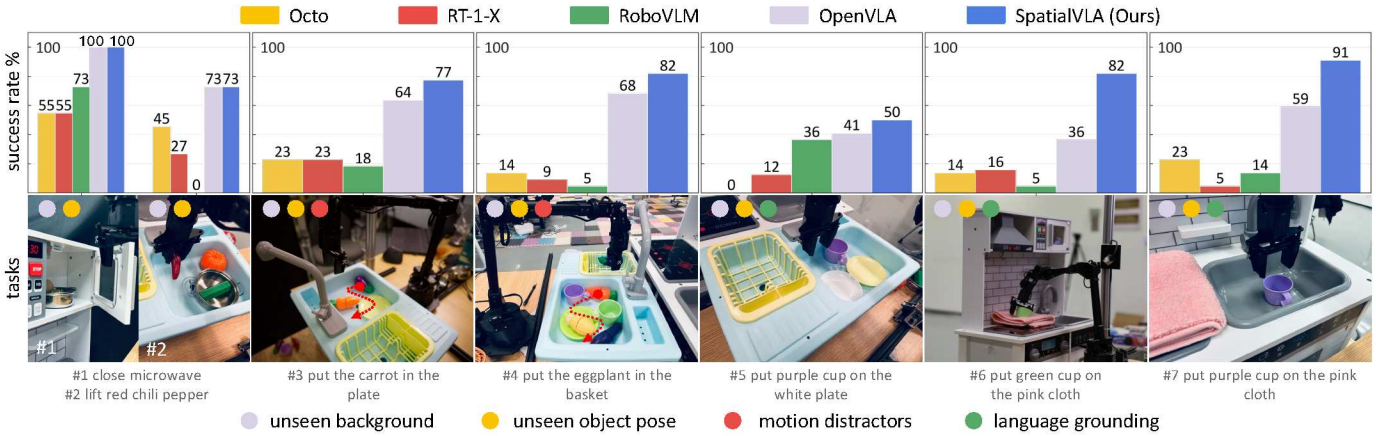


Fig. 5: **Zero-shot Robot Control Evaluation on WidowX Robot**. We evaluate SpatialVLA across 7 task suites to explore the language grounding, semantic understanding, and motion sensing capabilities, with varying backgrounds, poses, and motion distractors. SpatialVLA achieves the highest average success rate, outperforming all generalist manipulation policies.

conditions, characterized by varying visual appearances, which is further supported by its superior performance in variant aggregation. In particular, SpatialVLA also matches or outperforms the latest SOTA model π_0. Tab. II summarizes the results across different manipulation policies on the WidowX setup. Our model surpasses the state-of-the-art RoboVLM policy, achieving overall success rates of 34.4% and 42.7%. Fine-tuning on the BridgeV2 yields a remarkable 100% success rate in the "Put Eggplant in Yellow Basket" task, demonstrating the model's exceptional zero-shot manipulation capability.

Real-world WidowX Evaluation. Fig. 5 presents the results of the real-world "out-of-the-box" evaluation in WidowX robot platform. We observe that, in simple single-task scenarios (#1 close microwave), all the policies exhibit some generalizability, successfully completing tasks in unseen environments. However, in moderately complex tasks (#3-7), most policies, such as RT-1-X, Octo, and RoboVLM struggle with manipulation, frequently encountering issues like object misidentification and grasp failures. Compared to OpenVLA, our method demonstrates superior robustness in handling motion disturbances (human-induced dynamic object movement in tasks #3 and #4), successfully tracking and grasping carrot and eggplant. Furthermore, in the instruction-following tasks (#5-7), our method demonstrates strong instruction-following ability, accurately executing tasks like picking up a green cup and placing it on a white plate, not a pink one, based on color descriptions in the prompts, outperforming OpenVLA and other generalist policies. Overall, SpatialVLA achieves a higher average success rate, showcasing robust and generalizable operation capabilities in unseen scenarios, objects, language grounding, and dynamic motions.

B. Adapting to New Robot Setups

Evaluation Setup and Comparisons. We present the evaluation of SpatialVLA on the LIBERO simulation benchmark [36], which consists of a set of diverse robotic manipulation tasks in simulated environments. Following OpenVLA [30], we conduct experiments on four task suites, each comprising 10 tasks with 50 human-teleoperated demonstrations. These suites evaluate the model's understanding of spatial relationships (**LIBERO-Spatial**), object types (**LIBERO-Object**), task-oriented behaviors (**LIBERO-Goal**), and its ability to generalize to long-horizon tasks with diverse objects, layouts, and goals (**LIBERO-Long**). We compare our approach to several generalist manipulation policy methods, including Diffusion Policy [12], Octo [48], OpenVLA [30], and TraceVLA [71]. SpatialVLA is fine-tuned on the corresponding dataset for 200 epochs using LoRA ($r = 32$, $\alpha = 32$), which incorporates spatial embedding adaption in Sec. III-B from new Gaussian distribution.

To facilitate a more comprehensive evaluation, 13 Franka tasks are established to validate the model's manipulation performance, as shown in Fig. 6. The evaluation consists of three setups: **Single Task**, which includes four basic tasks: pick, place, push, and close; **Instruction Following**, which involves manipulating different objects in the same scene based on language instructions; and **Multi-tasks**, which involves training on a mixture of all four single-task data and tested on these tasks. We compare SpatialVLA with mainstream policies, including Diffusion Policy, Octo, and OpenVLA. More details can be found in the Appendix. E.

Evaluation Results. Tab. III present the LIBERO [36] experimental results. Notably, we observe that SpatialVLA can be effectively adapted to tasks in the LIBERO environments, as it obtains the highest average success rate of 78.1% and the first rank across all the policies. In particular, SpatialVLA achieves a remarkable 88.2% success rate on the LIBERO-Spatial task, which consists of different object layouts, demonstrating the model's strong understanding of spatial relationships. In most tasks, SpatialVLA outperforms the state-of-the-art generalist manipulation policies but struggles with long-horizon tasks in **LIBERO-Long**, due to the lack of architecture design for long-horizon observation.

Fig. 6 summarizes the results of the Franka robot fine-tuning evaluation. In single-task tests, SpatialVLA and Diffusion Policy show similar accuracy (82% vs 81%), outperforming

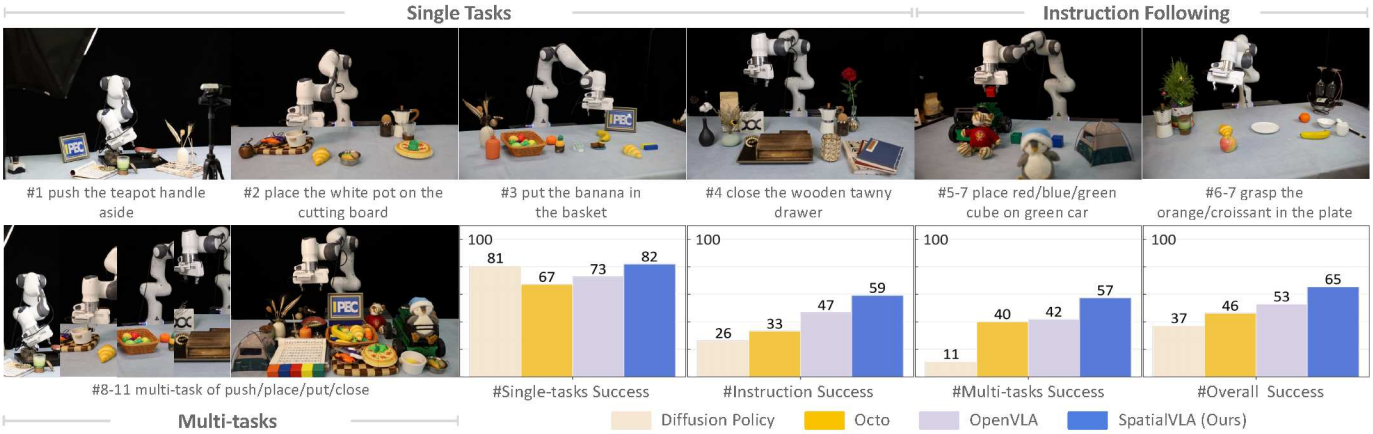


Fig. 6: **Adapting to New Robot Setups on Franka Robot**. SpatialVLA serves as a generalist robot control policy, achieving better performance across multiple setups, and can be effectively used as an initialization for fine-tuning to new robot tasks.

TABLE III: **LIBERO Simulation Benchmark Results.** We present the success rate (SR) and standard error for each method across four task suites, which are averaged over three random seeds with 500 trials. Fine-tuned SpatialVLA models achieve the highest average success rate and ranking, followed by fine-tuned OpenVLA [30] and Octo [48].

	LIBERO-Spatial		LIBERO-Object		LIBERO-Goal		LIBERO-Long		Average	
	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)
Diffusion Policy from scratch	78.3 ± 1.1%	5	**92.5 ± 0.7%**	1	68.3 ± 1.2%	5	50.5 ± 1.3%	5	72.4 ± 0.7%	5
Octo fine-tuned	78.9 ± 1.0%	4	85.7 ± 0.9%	4	**84.6 ± 0.9%**	1	51.1 ± 1.3%	4	75.1 ± 0.6%	3
OpenVLA fine-tuned	84.7 ± 0.9%	2	88.4 ± 0.8%	3	79.2 ± 1.0%	2	53.7 ± 1.3%	3	76.5 ± 0.6%	2
TraceVLA fine-tuned	84.6 ± 0.2%	3	85.2 ± 0.4%	5	75.1 ± 0.3%	4	54.1 ± 1.0%	2	74.8 ± 0.5%	4
SpatialVLA fine-tuned	**88.2 ± 0.5%**	1	**89.9 ± 0.7%**	2	78.6 ± 0.6%	3	**55.5 ± 1.0%**	1	**78.1 ± 0.7%**	1

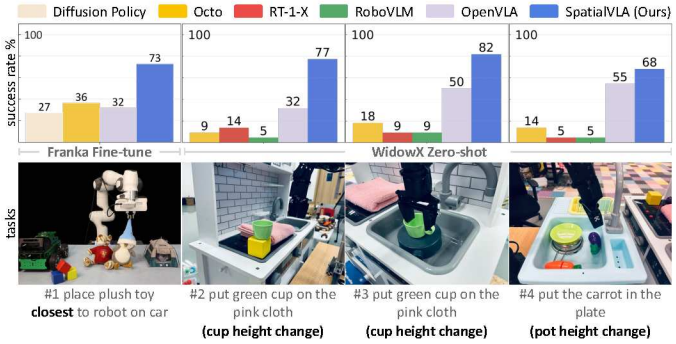


Fig. 7: **Spatial Understanding Capability Evaluation**. Benefiting from the proposed Ego3D Position Encoding, SpatialVLA exhibits superior performance in understanding spatial prompts and complex spatial layout tasks.

OpenVLA and Octo. However, in the instruction following tasks, SpatialVLA improves by +12% over OpenVLA, while Diffusion Policy struggles with a 26% success rate. In multi-tasks, SpatialVLA leverages its pre-training on OXE and 3D perception capabilities to achieve a 57% accuracy rate, surpassing other generalist policies. In summary, SpatialVLA demonstrates its versatility as a generalist robot control policy, achieving better performance across various tasks, and can be effectively used as an initialization for new robot tuning.

C. *Evaluating Spatial Understanding Capability*

Evaluation Setup and Comparisons. As shown in Fig. 7 and Tab. III, we evaluate the spatial understanding capabilities of SpatialVLA through on three robot setups: **Franka Robot** fine-tuning, **WidowX Robot** zero-shot, and **Libero-Spatial** fine-tuning. The tasks exhibit varying spatial complexities, with the Franka task involving prompt understanding (e.g., #1 place plush toy **closest to** robot on car), the WidowX task featuring explicit **height changes** (e.g., #2 put green cup on the pink cloth), and the LIBERO-Spatial task involving object layout variations. Seven mainstream policies, namely Diffusion Policy, Octo, RT-1-X, OpenVLA, TraceVLA, and RoboVLM, are employed for comparison.

Evaluation Results. Compared to existing policies, SpatialVLA shows superior spatial understanding, achieving 73% accuracy in Franka task #1, which involves spatial prompts, and significantly improving manipulation capabilities for complex positional changes in the out-of-distribution WidowX Zero-shot tasks #2-4. Similar results are observed in the LIBERO-Spatial task suite (88.2% success rate). Policies like Octo, Diffusion Policy, and OpenVLA, which lack integrated depth information, face significant challenges in adapting to spatial layout changes, yielding a success rate consistently lower than 50%. Consequently, we suggest integrating 3D information (Sec. III), including depth or point cloud, into the VLA framework to improve the model's adaptability and robustness in spatial layout variations.

D. *Ablations on Design Decisions*

In this section, we conduct ablation studies to investigate the effectiveness of the proposed 3D Spatial Presentation in both *pre-training* and *post-training* stages.

TABLE IV: **Pre-training Ablations on the Mixture Dataset of Google Fractal and BridgeData V2**. Initializing a high-resolution action grid from the data distribution and 3D position encoding enhances the model's generalization capability.

#setting		Pick Coke Can		Move Near		Put Carrot on Plate		Put Eggplant in Yellow Basket	
		variant aggregation	visual matching	variant aggregation	visual matching	grasp carrot	success	grasp eggplant	success
#All	[1]. SpatialVLA	81.6%	70.7%	79.2%	85.4%	41.7%	33.3%	91.7%	87.5%
Linear Discretization Distribution	[2]. ∼ linear 256 bins	40.7%	19.0%	47.1%	52.9%	41.7%	33.3%	87.5%	70.8%
	[3]. ∼ uniform distribution	77.9%	28.0%	64.2%	55.0%	45.8%	12.5%	79.2%	54.2%
Action Grids Resolution	[4]. resolution 1026	74.4%	67.3%	59.1%	54.2%	45.8%	25.0%	66.7%	54.2%
	[5]. resolution 4610	76.7%	68.0%	69.8%	79.2%	41.7%	33.3%	83.3%	75.0%
	[6]. resolution 6166	80.9%	74.0%	74.0%	79.2%	41.7%	33.3%	95.8%	87.5%
	[7]. resolution 8194	81.6%	70.7%	79.2%	85.4%	41.7%	33.3%	91.7%	87.5%
Encoding	[8]. − ego3d encoding	68.9%	70.3%	66.7%	62.0%	54.2%	12.5%	75.0%	37.5%
	[9]. − freeze llm embedding	70.2%	50.7%	63.1%	62.5%	33.3%	20.8%	95.8%	79.2%

TABLE V: **Fine-tuning Ablations in Domain Datasets**. Pretrained models are full parameter fine-tuned in individual Google Fractual and Bridge V2 Dataset. In LIBERO tasks, both full-tuning and LoRA-tuning are applied. fine-tuned with Gaussian adaptation from new dataset distribution helps align spatial grid features and improve initialization and accelerating convergence.

#setting	Pick Coke Can		Move Near		Put Carrot on Plate		Put Eggplant in Yellow Basket	
	variant aggregation	visual matching	variant aggregation	visual matching	grasp carrot	success	grasp eggplant	success
[1]. full params tuning	88.0%	77.0%	72.7%	75.0%	29.2%	20.8%	100%	91.7%
[2]. + Gaussian adaption	90.1%	86.0%	74.6%	77.9%	29.2%	25.0%	100%	100%

#setting	LIBERO-Spatial	LIBERO-Object	LIBERO-Goal	LIBERO-Long
[3]. Full params tuning	77.7 ± 0.4%	73.3 ± 0.4%	78.5 ± 0.5%	43.9 ± 0.8%
[4]. LoRA tuning	83.6 ± 0.7%	84.8 ± 0.9%	76.4 ± 0.2%	50.1 ± 0.3%
[5]. + Spatial embedding adaption	**88.2 ± 0.5%**	**89.9 ± 0.7%**	**78.6 ± 0.6%**	**55.5 ± 1.0%**

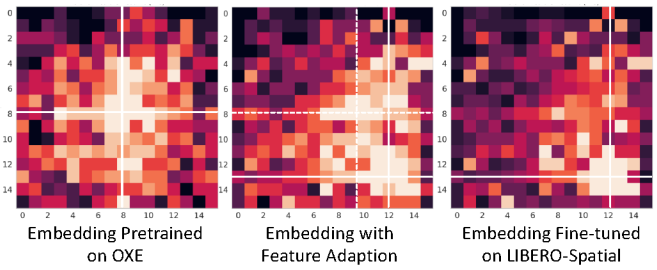


Fig. 8: Cross-sectional features visualization in spatial grids. The proposed spatial embedding adaptation aligns the pre-trained spatial grid features with those of the target fine-tuned model, improving initialization and accelerating convergence.

Pre-training in Mixture Dataset. The pre-training ablations in Tab. IV are conducted on a mixture dataset that combines Google Fractal [6] and BridgeData V2 [64]. All the models are trained from scratch on 8 A100 GPUs with 128 batch size for 120k steps. We select four tasks from the SimplerEnv benchmark [35], namely "Pick Coke Can" and "Move Near" on the Google Robot, as well as "Put Carrot on Plate" and "Put Eggplant in Yellow Basket" on the WidowX Robot, to dissect the model's component-wise performance.

In contrast to the conventional linear 256-bin action space discretization [6, 13, 30] (#1*v.s.*#2), the proposed *adaptive spatial action grids exhibits significant advantages*, particularly in the Google Robot task, with the promotion of +36.5% and +42.1% in variant aggregation and visual matching success rates, respectively. During model training, we also observe that models using linear 256-bin discretization converge slower, despite achieving lower L1 Loss. Another suggestion is to *initialize the grid partitioning based on the dataset distribution*, rather than using a uniform grid (#1*v.s.*#3), which enables the model to focus on high-frequency action spaces adaptively and further improves its generalization capabilities.

Compared to 1026-resolution action grids (#1*v.s.*#4), where $M_{trans} = M_{trans} = 512, M_{grip} = 2$, SpatialVLA with 8194-resolution action grids ($M_{trans} = M_{trans} = 4096, M_{grip} = 2$) achieves significant performance boosts, particularly in "move near" and "put eggplant in yellow basket" tasks, with success rate increments of +31.2% and +33.3%. Additionally, we find that *lower-resolution models tend to learn smaller actions, causing slow motion issues, and high-resolution models exhibit improved transfer performance in the fine-tuning stage*.

According to the ablation results (#1*v.s.*#8), the proposed egocentric 3D position encoding (ego3d), *incorporating 3D point cloud features, helps the model overcome varied lighting, color, textures, and camera poses*, yielding stronger generalizability in diverse manipulation scenarios. Models w/o ego3d suffer a significant performance drop in variant aggregation, from 81.6% and 79.2% to 68.9% and 66.7%, due to their inability to adapt to scene changes. During pre-training, we also observe from (#1*v.s.*#9) that *freezing the language embedding and sharing a trainable spatial embedding helps to improve the model's manipulation capabilities*, which is also beneficial for faster training and instruction following.

Post-training in Domain Dataset. We conduct post-training ablations in Tab. V, separately fine-tuning on large-scale datasets Google Fractal and BridgeData V2 and BridgeData V2, and comparing full fine-tuning and LoRA-tuning on the small-scale LIBERO datasets [36]. The spatial embedding adaption denotes partitioning spatial grids from the new dataset Gaussian distribution and updating the spatial feature embedding with the grids.

On large-scale datasets (#1*v.s.*#2), models fine-tuned with spatial embedding adaptation yield marginal gains of +2.9% in visual matching on Move Near), as the large-scale dataset dis-

tribution closely matches the pre-training distribution, allowing the model to learn fine-grained features thereby limiting the benefits of the adaption. While, on the LIBERO small dataset tasks (#4*v.s.*#5), initializing the feature grid with the new distribution boosts model performance by +4.6%, +5.1%, +2.2%, and +5.4% on LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-Long, respectively. As shown in Fig. 8, *feature adaptation from the new distribution aligns pre-trained spatial features with the target fine-tuned model*, improving initialization and accelerating convergence. Moreover, LoRA fine-tuning outperforms full-parameter fine-tuning on small dataset tasks (#3*v.s.*#4), making LoRA the preferred method for small datasets.

V. Discussion, Limitations, and Future Work

In this paper, we present SpatialVLA, an innovative vision-language-action model to explore efficient spatial representations for generalist robot policy. SpatialVLA introduces Ego3D position encoding and adaptive action grids to inject 3D awareness into robot observation representation and spatial action tokenization through robot-agnostic designs, equipping the VLA models with the spatial understanding ability of the 3D physical world. After pre-training on large-scale heterogeneous robot datasets, we find that SpatialVLA is a more generalizable and transferrable generalist policy for zero-shot robot control. Our extensive real-world and simulated robot experiments show that SpatialVLA leads to dramatically improved performance over the previous VLA models, especially on tasks that require precise spatial understanding. We also show that the pre-trained SpatialVLA model can effectively adapt to new robot setups and tasks via action grids re-discretization, which offers a new way for robot-specific post-training. In the following, we discuss our limitations of SpatialVLA and potential solutions, hoping to inspire further innovative works.

More Generalizable Distribution Fitting. In this paper, SpatialVLA fits action signals with Gaussian distributions to encode actions as spatial grids, demonstrating remarkable generalizability and flexible adaptation to new robot setups through re-initialized grids and token embeddings. However, this raises a crucial question: Is modeling data distributions as Gaussian optimal? We argue that Gaussian modeling is sub-optimal, as it can lead to grid clustering on specific coordinate axes in extreme robot operation scenarios, such as single-axis motion, resulting in lost motion capabilities on other axes. Moreover, dataset noises can further distort the spatial grid distribution. One future solution is to combine implicit data distribution modeling techniques, such as Variational Auto-Encoder-based high-dimensional feature space mapping, with explicit grid partitioning, enhancing action presentation efficiency and noise robustness.

More Flexible VLA architectures. In our implementation, we predict spatial action tokens through the autoregressive paradigm and further decode them into actions, resulting in each action being represented by 3 tokens. Although SpatialVLA achieves 21Hz inference speed, it is slower than diffusion decoding [5, 12, 32], which decodes tokens into multiple consecutive actions. In the future, integrating diffusion decoding with spatial grid action presentation and exploring dynamic token numbers for action mapping will be valuable. Furthermore, as the model relies solely on current frame observations and history tokens for action prediction, it faces challenges in long-horizon tasks, similar to other generalizable policies [12, 48]. Future work should focus on designing efficient historical information perception mechanisms to enhance the model's long-sequence modeling capabilities, enabling seamless task switching in real-time manipulation scenarios.

Higher-Quality Diverse Data. SpatialVLA is pre-trained on OXE and RH20T, but the variable quality of OXE data can hinder training. Therefore, future work exploring optimal data composition and distilling high-quality subsets from the heterogeneous robot data collections is vital for boosting model efficiency and generalizability.

Acknowledgements. This work is supported by the Shanghai AI Laboratory and the National Natural Science Foundation of China (624B2044).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Proceedings of the Conference on Neural Information Processing System (NeurIPS)*, 2022.

[2] Suneel Belkhale, Yuchen Cui, and Dorsa Sadigh. Hydra: Hybrid robot actions for imitation learning. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

[3] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[5] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[7] 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.

[8] 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.

[9] Lawrence Yunliang Chen, Simeon Adebola, and Ken Goldberg. Berkeley UR5 demonstration dataset. https://sites.google.com/view/berkeley-ur5/home.

[10] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[11] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[12] 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. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[13] Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[14] Zichen Jeff Cui, Yibin Wang, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. From play to policy: Conditional behavior generation from uncurated robot data. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2023.

[15] Shivin Dass, Jullian Yapeter, Jesse Zhang, Jiahui Zhang, Karl Pertsch, Stefanos Nikolaidis, and Joseph J. Lim. Clvr jaco play dataset, 2023. URL https://github.com/clvrai/clvr_jaco_play_dataset.

[16] Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2024.

[17] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. In *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[18] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[19] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024.

[20] Charles R Gallistel. *The organization of learning*. The MIT Press, 1990.

[21] Siddhant Haldar and Lerrel Pinto. Polytask: Learning unified policies through behavior distillation. *arXiv preprint arXiv:2310.08573*, 2023.

[22] Siddhant Haldar, Zhuoran Peng, and Lerrel Pinto. Baku: An efficient transformer for multi-task policy learning. In *Proceedings of the Conference on Neural Information Processing System (NeurIPS)*, 2024.

[23] Minho Heo, Youngwoon Lee, Doohyun Lee, and Joseph J Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[24] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *Proceedings of the Conference on Neural Information Processing System (NeurIPS)*, 2023.

[25] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[26] 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 *Proceedings of the Conference on Robot Learning (CoRL)*, 2022.

[27] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2018.

[28] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

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

[30] 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.

[31] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024.

[32] Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. Towards generalist robot policies: What matters in building vision-language-action models. *arXiv preprint arXiv:2412.14058*, 2024.

[33] Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. Towards generalist robot policies: What matters in building vision-language-action models. *arXiv preprint arXiv:2412.14058*, 2024.

[34] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024.

[35] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2024.

[36] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023.

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the Conference on Neural Information Processing System (NeurIPS)*, 2024.

[38] Huihan Liu, Soroush Nasiriany, Lance Zhang, Zhiyao Bao, and Yuke Zhu. Robot learning on the job: Human-in-the-loop autonomy and learning during deployment. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[39] 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.

[40] Robert H Logie. *Visuo-spatial working memory*. Psychology Press, 2014.

[41] Jianlan Luo, Charles Xu, Xinyang Geng, Gilbert Feng, Kuan Fang, Liam Tan, Stefan Schaal, and Sergey Levine. Multi-stage cable routing through hierarchical imitation learning. *IEEE Transactions on Robotics*, 40:1476–1491, 2024.

[42] Jianlan Luo, Charles Xu, Fangchen Liu, Liam Tan, Zipeng Lin, Jeffrey Wu, Pieter Abbeel, and Sergey Levine. Fmb: a functional manipulation benchmark for generalizable robotic learning. *The International Journal of Robotics Research*, 2024.

[43] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.

[44] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2018.

[45] Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. Grounding language with visual affordances over unstructured data. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[46] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

[47] Soroush Nasiriany, Tian Gao, Ajay Mandlekar, and Yuke Zhu. Learning and retrieval from prior data for skill-based imitation learning. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

[48] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[49] Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Actor-mimic: Deep multitask and transfer reinforcement learning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2016.

[50] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[51] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

[52] Jean Piaget. *Child's Conception of Space: Selected Works vol 4*. Routledge, 2013.

[53] Delin Qu, Qizhi Chen, Pingrui Zhang, Xianqiang Gao, Junzhe Li, Bin Zhao, Dong Wang, and Xuelong Li. Livescene: Language embedding interactive radiance fields for physical scene rendering and control. *arXiv*

preprint arXiv:2406.16038, 2024.

[54] Gabriel Quere, Annette Hagengruber, Maged Iskandar, Samuel Bustamante, Daniel Leidner, Freek Stulp, and Jörn Vogel. Shared control templates for assistive robotics. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[56] Erick Rosete-Beas, Oier Mees, Gabriel Kalweit, Joschka Boedecker, and Wolfram Burgard. Latent plans for task-agnostic offline reinforcement learning. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2022.

[57] Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2016.

[58] Saumya Saxena, Mohit Sharma, and Oliver Kroemer. Multi-resolution sensing for real-time control with vision-language models. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

[59] Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.

[60] Rutav Shah, Roberto Martín-Martín, and Yuke Zhu. Mutex: Learning unified policies from multimodal task specifications. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

[61] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2022.

[62] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024.

[63] Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.

[64] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

[65] Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. In *Proceedings of the Conference on Neural Information Processing System (NeurIPS)*, 2024.

[66] Ge Yan, Kris Wu, and Xiaolong Wang. ucsd kitchens dataset. https://github.com/geyan21/rlds_dataset_builder/tree/main/ucsd_kitchens, 2023.

[67] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, 2024.

[68] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[69] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[70] Jinliang Zheng, Jianxiong Li, Dongxiu Liu, Yinan Zheng, Zhihao Wang, Zhonghong Ou, Yu Liu, Jingjing Liu, Ya-Qin Zhang, and Xianyuan Zhan. Universal actions for enhanced embodied foundation models. *arXiv preprint arXiv:2501.10105*, 2025.

[71] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024.

[72] Gaoyue Zhou, Victoria Dean, Mohan Kumar Srirama, Aravind Rajeswaran, Jyothish Pari, Kyle Hatch, Aryan Jain, Tianhe Yu, Pieter Abbeel, Lerrel Pinto, et al. Train offline, test online: A real robot learning benchmark. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[73] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024.

[74] Xinghao Zhu, Ran Tian, Chenfeng Xu, Mingxiao Huo, Wei Zhan, Masayoshi Tomizuka, and Mingyu Ding. Fanuc manipulation: A dataset for learning-based manipulation with fanuc mate 200id robot. https://sites.google.com/berkeley.edu/fanuc-manipulation, 2023.

[75] Yifeng Zhu, Peter Stone, and Yuke Zhu. Bottom-up skill discovery from unsegmented demonstrations for long-horizon robot manipulation. *IEEE Robotics and Automation Letters*, 7(2):4126–4133, 2022.

[76] Yifeng Zhu, Zhenyu Jiang, Peter Stone, and Yuke Zhu. Learning generalizable manipulation policies with object-centric 3d representations. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

[77] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.