Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

LAP: Language-Action Pre-training Enables Zero-Shot Cross-Embodiment Transfer

Lihan Zha[1], Asher J. Hancock[1*], Mingtong Zhang[1*], Tenny Yin[1], Yixuan Huang[1], Dhruv Shah[1], Allen Z. Ren[2†], Anirudha Majumdar[1†]
[1]Princeton University [2]Physical Intelligence
[*]Equal contribution. [†]Equal advising.
https://lap-vla.github.io

Fig. 1: We introduce Language-Action Pre-training (LAP), a general VLA pre-training recipe that represents low-level actions directly in natural language to supervise a vision–language backbone, and instantiate it as LAP-3B, the first VLA to demonstrate strong *zero-shot transfer* to novel embodiments. Compared to state-of-the-art VLAs, LAP-3B learns more generalizable embodiment representations and exhibits favorable scaling behavior.

***Abstract*—A long-standing goal in robotics is a generalist policy that can be deployed zero-shot on new robot embodiments without per-embodiment adaptation. Despite large-scale multi-embodiment pre-training, existing Vision–Language–Action models (VLAs) remain tightly coupled to their training embodiments and typically require costly fine-tuning. We introduce *Language-Action Pre-training* (LAP), a simple recipe that represents low-level robot actions directly in natural language, aligning action supervision with the pre-trained vision–language model's input–output distribution. LAP requires no learned tokenizer, no costly annotation, and no embodiment-specific architectural design. Based on LAP, we present LAP-3B, which to the best of our knowledge is the first VLA to achieve substantial zero-shot transfer to previously unseen single-arm manipulators without any embodiment-specific fine-tuning. Across multiple novel robots and manipulation tasks, LAP-3B attains over 50% average zero-shot success, delivering roughly a $2\times$ improvement over the strongest prior VLAs. We further show that LAP enables efficient adaptation and favorable scaling, while unifying action prediction and VQA in a shared language-action format that yields additional gains through co-training.**

I. Introduction

The ultimate promise of robot foundation policies is the ability to control *any* robot embodiment to perform human-level intelligent tasks. Vision–Language–Action models (VLAs) have made remarkable progress toward this goal, leveraging pre-trained Vision–Language models (VLMs) to obtain strong visual–semantic priors for broad scene and even task generalization [1–9]. However, despite large-scale multi-embodiment training [10–14], state-of-the-art VLAs still rarely function *zero-shot on new robots* with even minor differences to training embodiments (e.g., an altered gripper or wrist camera location), as shown in Fig. 1. In practice, deploying a VLA on a new robot often requires extensive per-embodiment adaptation, which is costly and undermines the goal of a single broadly deployable foundation model.

At its core, zero-shot cross-embodiment transfer is difficult because the embodiment space is vast and fundamentally expensive to cover with data. While task and environment diversity can scale via in-the-wild robot data [11, 15, 16] and internet-scale egocentric video [17–22], embodiment diversity remains constrained by hardware availability and data-collection logistics; for example, the Open X-Embodiment dataset aggregates data from only a few dozen robot types [10]. Achieving zero-shot transfer therefore requires extracting maximal generalization from limited embodiment diversity, rather than relying on exhaustive coverage.

Our key observation is that zero-shot cross-embodiment transfer depends critically on how we adapt a pre-trained VLM for motor control. Concretely, the model must acquire precise control capabilities while retaining the semantic and visual representations that enable transfer across embodiments. When

this balance is mismanaged, embodiment-specific action learning can override the very features needed for generalization. However, because VLMs are not pre-trained to interpret or generate motor-level, high-frequency control signals, standard VLA training can induce distributional mismatch and degrade pre-trained knowledge [23–25].

Motivated by this observation, we propose *Language-Action Pre-training (LAP)*, a simple pre-training recipe for VLAs that adapts a pre-trained VLM backbone for motor control by training it to predict *actions described in natural language*, herein referred to as *language-actions*. Language-actions express the net effect of a raw end-effector action (e.g., "move left 5 cm", "tilt forward 45 degrees") in the same modality the VLM is trained to model, as shown in Fig. 1. They can be parsed from raw actions under a *fixed* coordinate convention and template without additional data annotation.

This representation choice directly targets the distributional mismatch that arises in standard VLA training recipes, which fine-tune VLMs to predict *raw actions*—continuous joint or end-effector commands, or discretizations thereof—represented by arbitrary or abstract tokens [2, 7, 26]. Such representations carry little semantic structure that could unify behaviors across embodiments [27–30]. Notably, this limitation persists even when a unified end-effector action space is adopted: existing VLAs still exhibit little to no zero-shot generalization to novel embodiments [5, 7] (see Section IV-A). Together, these results suggest that cross-embodiment transfer cannot be achieved by action space unification alone, but instead requires a fundamentally different action representation. Language-actions provide such a representation by keeping the supervision signal close to the VLM's pre-training distribution [23], promoting semantically grounded features that transfer across embodiments.

With this representation in place, we next consider how to instantiate LAP in a practical VLA architecture. In principle, LAP can be applied to any VLM-based VLA by treating action generation as a language modeling problem. However, such formulations typically rely on auto-regressive language-action token sampling at inference time, which is inefficient and ill-suited for robot control [7, 23, 25, 26]. We therefore adopt a hybrid design that combines a LAP-trained VLM backbone with a lightweight diffusion-based action expert for continuous control, following $\pi_{0.5}$ [1]. This Mixture-of-Transformers architecture [31] preserves the representation benefits of language-action supervision while enabling efficient execution.

Statement of contributions. We introduce Language-Action Pre-training (LAP), a general pre-training recipe for Vision–Language–Action models that represents actions as natural language to enable cross-embodiment generalization. We instantiate LAP in LAP-3B, which pairs a VLM backbone with a lightweight action expert for efficient real-time control. Trained on existing open-sourced robot datasets [2, 10, 11], LAP-3B achieves substantial zero-shot generalization to previously single-arm manipulators, delivering roughly a 2× relative improvement (approximately 30% absolute) over prior pre-training recipes based on alternative action representations. *To the best of our knowledge, this is the first VLA system to operate zero-shot on novel real robot embodiments.*[1] In addition, LAP-3B can be fine-tuned to new embodiments and more complex dexterous tasks using significantly fewer demonstrations and gradient steps, matching prior performance with up to 2.5× less data. Finally, we analyze why LAP learns more transferable embodiment representations and exhibits more stable training dynamics, demonstrate its favorable scaling behavior with increasing model size, and show that the language-action interface naturally supports co-training with auxiliary VQA objectives (e.g., motion prediction), yielding further performance gains.

II. Related Work

Vision-Language-Action Models. Recent work has developed generalist robot policies trained on large and diverse robotic datasets [10–14, 32–36]. Among these, Vision–Language–Action models (VLAs) have emerged as a particularly promising paradigm [3, 4, 7, 9, 10, 23, 26, 37–44]. These models are typically fine-tuned from Vision–Language models (VLMs) pre-trained on internet-scale data. The pre-trained representations equip VLAs with the ability to follow natural language instructions and understand diverse visual cues. As a result, VLAs often not only perform well within their training distribution, but also show promise for generalization to out-of-distribution environments [1, 7, 39, 45–47]. Nevertheless, despite being trained on data spanning multiple robot embodiments, existing VLAs do not generalize zero-shot to novel embodiments without extensive fine-tuning (Section IV-A). In this work, we introduce a new action representation that significantly improves VLA's embodiment generalization capability.

Cross-Embodiment Policy Learning. Large robot policies are typically trained on heterogeneous, cross-embodiment datasets, and a common strategy for handling varying state and action spaces is to manually define a unified representation. The most straightforward approach pads states and actions to the maximum dimensionality or uses shared action spaces like end-effector pose [5, 7, 39]. Other works learn embodiment-specific state or action projectors, or employ separate action heads for different robot types [4, 38, 42, 48, 49]. A parallel line of work learns shared latent actions or skills across embodiments [28, 29, 50–53]. More recently, human hand motion has been used as a unified action space bridging human and robot data [54, 55], enabling task generalization to seen embodiments. While effective, these methods typically require embodiment-specific fine-tuning to adapt to new embodiments, whereas our method achieves *zero-shot embodiment transfer* without any additional training.

Another line of work facilitates cross-embodiment learning by conditioning robot policies on embodiment features. Early approaches infer such features from kinematic structure [56–60], and several works focus on human-to-robot transfer

[1] All code, checkpoints, and data configurations are open-sourced at https://lap-vla.github.io

by exploiting structural similarity [61–64]. For example, X-VLA [5] conditions policies on per-embodiment soft prompts, and DexVLA [65] introduces an explicit embodiment-specific training stage. RDT2 [66] similarly reports zero-shot embodiment transfer, but does so under the assumption that new robots share the same gripper design and wrist-mounted camera configuration as those used during data collection, effectively constraining embodiment variation to matched hardware interfaces rather than arbitrary morphologies. In parallel, a broader class of approaches infers embodiment implicitly from long interaction histories, as commonly explored in navigation and embodied control [49, 67–69]. Taken together, these methods primarily model embodiment variation within the training distribution—through prompts, explicit embodiment conditioning, hardware alignment, or historical context—and typically require additional assumptions or adaptation when deployed on previously unseen embodiments.

By contrast, LAP is motivated by the complementary goal of true zero-shot embodiment transfer. Rather than introducing explicit embodiment-conditioning mechanisms or relying on matched hardware interfaces across robots (e.g., identical grippers and wrist-mounted cameras), LAP-3B demonstrates that cross-embodiment generalization can emerge directly from large-scale VLA pre-training with appropriately structured action representations. Without any embodiment-specific design or embodiment-aligned data collection assumptions, LAP-3B, to the best of our knowledge, is the first VLA to achieve substantial *zero-shot generalization to previously unseen embodiments*, while also enabling efficient adaptation to more challenging tasks when additional data are available.

	Grover et al. [70]	VLA-0 [6]	VLM2VLA [23]	**LAP**
Large-Scale Training	✓	✗	✗	✓
Inference Frequency	∼0.5Hz	4Hz	0.2Hz	**25Hz**
Cross-Embodiment	✗	✗	✗	✓

TABLE I: Comparison of VLAs that use (variants of) language-actions along key design and performance axes. LAP uniquely enables cross-embodiment generalization through large-scale pre-training and language-action representation.

Action Representations for VLAs. How actions are represented determines how effectively a pre-trained VLM can be adapted for motor control [2, 6, 23, 26, 28, 70–72]. Early approaches represent actions as code or language-level subtasks [73–77]. These formulations are closely related in spirit to our use of linguistic supervision, but they typically operate at coarse temporal abstraction and rely on separate low-level controllers to execute each sub-task. In contrast, LAP uses *fine-grained* language-actions parsed directly from continuous end-effector motion and trains the backbone to model these motor-level effects, yielding representations tailored for low-level control. More recent works instead encode actions as 2D keypoints or visual traces [2, 43, 72, 78–80], which enhance grounding but are typically still used as intermediate representations rather than direct control.

Another line of work maps actions to unused or non-linguistic tokens in VLM vocabularies [7, 26, 41, 71], which can induce distributional misalignment and degrade pre-trained VLM representations [23]. More recently, approaches have begun representing actions directly in natural language [23] or as raw string tokens [6, 70], yielding improved semantic grounding, language following, and task-level generalization. As summarized in Table I, LAP builds on this emerging paradigm with two key distinctions: it introduces LAP-3B as a *practical* language-action VLA for real-time control, and demonstrates that language-action representations can scale to large robot pre-training corpora—thereby enabling zero-shot cross-embodiment transfer.

III. LAP: Language-Action Pre-Training for VLAs

We now introduce *Language-Action Pre-training* (LAP), a general pre-training recipe for VLAs. We first formalize LAP's language-action learning objective (Section III-A), then describe how to obtain language-actions at scale without manual annotation (Section III-B). We then present LAP-3B, a vision-language-action model trained with LAP that demonstrates substantial cross-embodiment transfer (Section III-C), followed by specific implementation details (Section III-D).

A. Formulations

LAP reframes action learning in VLAs as a standard conditional language modeling problem. Given an observation $o_t = \{I_t^1, \dots, I_t^n, s_t\}$ consisting of multi-view RGB images and proprioceptive state, and a task instruction l, we deterministically derive a structured language-action sequence $\hat{a}^{\text{lang.}}_{t:t+H}$, as shown in Fig 2$(b)$, summarizing the corresponding continuous action chunk $a_{t:t+H}$.

The VLM backbone is trained to auto-regressively predict the language-action tokens conditioned on (o_t, l) using the standard cross-entropy objective

$$\mathcal{L}_{\text{CE.}} = -\mathbb{E}_{(\hat{a}^{\text{lang.}}_{t:t+H}, o_t, l)\sim\mathcal{D}} \sum_i \log p_\theta\big(\hat{a}^{\text{lang.}}_{t,i} \mid o_t, l, \hat{a}^{\text{lang.}}_{t,<i}\big), \quad (1)$$

which is identical to standard sequence-to-sequence language modeling and corresponds to maximum likelihood estimation of language-actions, analogous to behavioral cloning for continuous actions.

B. Language-Action Dataset Curation

A key requirement for LAP is the access to language-action supervision at scale. Rather than relying on manual annotation, we leverage the structured nature of robotic datasets to derive language-actions directly from end-effector trajectories.

Specifically, we represent language-actions as natural language descriptions of end-effector delta motions. Under a fixed coordinate convention and a small set of structured templates, continuous actions can be deterministically converted into language descriptions with arbitrary numerical resolution.

Coordinate System. We adopt a convention commonly used in robotics datasets, where $+x$ is forward, $+y$ is left, and $+z$ is up; rotational components are expressed as Euler angles following the right-hand rule. Since most manipulation datasets already follow this convention, language-actions can be parsed directly without additional frame transformations.

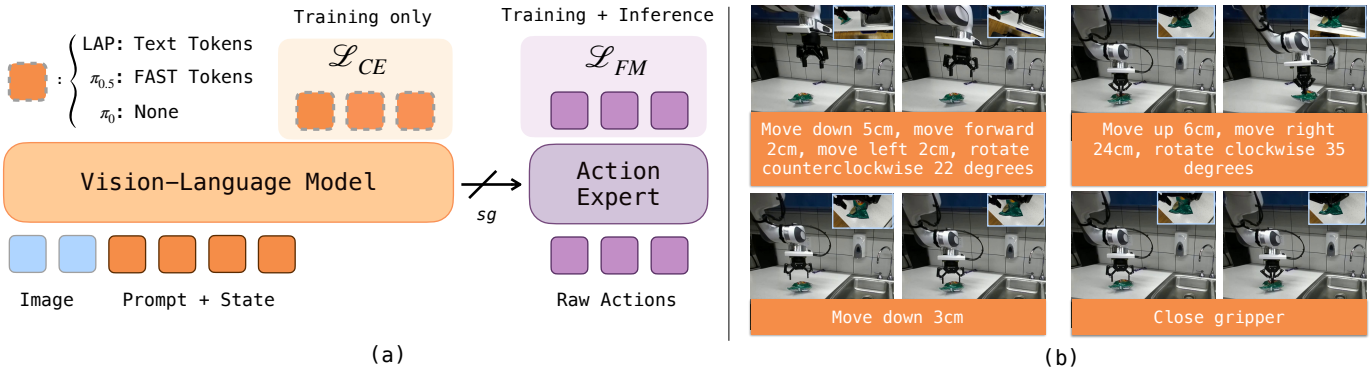

Fig. 2: (a) A unified view comparing LAP-3B with prior VLAs in terms of action representation. A VLM backbone predicts discrete language-action tokens using a cross-entropy objective, while a lightweight action expert predicts continuous actions via flow matching. Gradients from the action expert are blocked from the VLM through knowledge insulation [25], ensuring that the VLM is trained purely via language supervision. At test time, the action expert is rolled out for fast inference. (b) Visualizations of language-actions from the DROID dataset [11].

To encourage embodiment-agnostic reasoning and consistent use of multi-view visual observations, we additionally express language-actions in multiple reference frames. Specifically, during training we randomize the reference frame used to generate each language-action, expressing 50% in the robot base frame and 50% in the end-effector frame, and include the name of the reference frame in the model prompt.

Templates. Language-actions are generated using a small set of structured natural language templates that describe translational and rotational end-effector motions:

$$\hat{a}_t^{\text{lang.}} = \text{“}\langle\text{verb}\rangle\ \langle\text{direction}\rangle\ \langle\text{magnitude}\rangle\ \langle\text{unit}\rangle\text{”}, \qquad (2)$$

where the verb determines the action category: *move* denotes translational end-effector displacement, *tilt* denotes roll and pitch rotations, and *rotate* denotes yaw rotation. For translation, directions correspond to $\{\pm x, \pm y, \pm z\}$ (e.g., forward/backward, left/right, up/down). For roll and pitch, we use “left/right” and “back/forward” following the right-hand rule convention. For yaw, “clockwise” and “counterclockwise” represent negative and positive rotations, respectively.

Rather than generating language-actions at every timestep, we represent $\hat{a}^{\text{lang.}}_{t:t+H}$ as a single description summarizing the net displacement of the action chunk $a_{t:t+H}$. Since actions are expressed as delta end-effector poses, this corresponds to the cumulative translation and rotation between timesteps t and $t+H$. This temporal abstraction produces a low-frequency supervision signal that is easier for VLMs to model while abstracting away high-frequency control. We also discretize magnitudes to integer values and express translations in centimeters (e.g., *“move forward 5 cm”*), yielding language-actions that are both structured and precise. Visual examples are shown in Fig. 2(b).

Compared to prior action tokenization approaches that discretize continuous actions into arbitrary or out-of-vocabulary symbols, language-actions remain well supported by the VLM’s natural language pre-training distribution, mitigating the distributional mismatch known to degrade representation quality [23, 25]. In contrast to learned tokenizers such as FAST [26], which introduce a separate representation learning stage to compress actions into discrete non-linguistic tokens, language-actions are produced purely through deterministic parsing, eliminating the need for any learned action tokenizer. Moreover, unlike conventional action representations, language-actions naturally align with the input–output format of Visual–Question–Answering (VQA) tasks, enabling synergistic co-training. For instance, in a “motion prediction” VQA task, the model is prompted with image pairs (I^i_t, I^i_{t+H}) and trained to predict the corresponding language-action $\hat{a}^{\text{lang.}}_{t:t+H}$ describing the observed displacement.

C. Model Architecture and Training

LAP is agnostic to the underlying VLA architecture and can, in principle, be integrated into any VLM-based policy by treating action prediction as a language modeling problem. However, auto-regressive generation of language-actions at inference time is slow and impractical for real-time robot control. To obtain an efficient system, we instantiate LAP as LAP-3B, which adopts a Mixture-of-Transformers architecture [31] combining a LAP-trained VLM backbone with a lightweight flow-matching action expert for high-frequency motor execution, following $\pi_{0.5}$ [1]. This design enables a controlled comparison: LAP-3B differs from $\pi_{0.5}$ only in the action representation used to supervise the VLM (language-actions versus FAST tokens), isolating the downstream effect of the representation itself, which we evaluate in Sec. IV.

During training, the VLM backbone is optimized to predict structured language-actions using the cross-entropy loss in Eq. (1), while the action expert predicts continuous action chunks $a_{t:t+H}$ via a flow-matching objective

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{(a_{t:t+H}, o_t, l)\sim\mathcal{D}}\left[\left\|v_\phi(o_t, l, a_{t:t+H}, \tau) - \dot{x}_\tau\right\|^2\right], \qquad (3)$$

where v_ϕ denotes the velocity field parameterized by the action expert and $\dot{x}_\tau$ is the target flow corresponding to the ground-truth action chunk (see Appendix B for the full formulation).

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda \mathcal{L}_{\text{CE}}. \tag{4}$$

Under this architecture, the VLM and action expert communicate solely through cross-attention. We apply causal masking to ensure that the raw action and the language-action do not attend to each other. Following [25], we further block gradients from the action expert propagating back into the VLM backbone. While not required by LAP, this knowledge insulation helps preserve pre-trained VLM representations and stabilizes joint training. At inference time, only the action expert is rolled out for control, enabling real-time execution at 25 Hz on an NVIDIA RTX 4090 GPU.

D. Implementation Details

LAP-3B's VLM backbone is initialized as the PaliGemma-3B [81] model. We train on the Open X-Embodiment (OXE) dataset [10] and MolmoAct dataset [2] (see full data mixture details in Appendix B). We use a shuffle buffer size of 16 million samples to ensure near uniform mixing, which improves training stability. For proprioception, we represent state s_t using Cartesian end-effector pose: position, a continuous 6D rotation representation [82], and a binary gripper state. Since open-source datasets use inconsistent rotation conventions, we canonicalize rotations to extrinsic XYZ before converting to the 6D representation. The proprioceptive state is discretized into 255 bins and appended to the prompt as text tokens, following [1]. Actions are represented as delta end-effector poses. We normalize states and actions using the 1st and 99th global quantiles computed over the full training dataset, and reuse the same statistics at inference on unseen embodiments.

Training uses a global batch size of 2048 across 64 TPU v6e chips for 15k gradient steps (*approximately only 10 wall-clock hours*), corresponding to roughly 0.65 epoch over the full dataset, at which point the checkpoint already performs well in real-world experiments. Exponential moving average (EMA) updates begin after 5k steps. Image inputs are resized to 224×224, with at most two images per sample. We use a fixed learning rate of 1×10^{-4} with linear warmup over the first 5,000 steps.

IV. Experiments

Our experiments aim to evaluate whether language-actions provide a superior action representation for VLA pre-training. In particular, we focus on **cross-embodiment generalization**, which remains a significant limitation of state-of-the-art VLAs. Concretely, we seek to answer:

1) Does LAP enable zero-shot cross-embodiment transfer?
2) Does LAP support efficient fine-tuning to new embodiments on challenging tasks?
3) How does LAP's language-action representation shape the learned representations and enable strong cross-embodiment generalization?
4) What is the effect of co-training LAP-3B with VQA datasets?
5) Does LAP scale favorably with model size?

To address these questions, we evaluate LAP across **four robot embodiments**, **ten real-world manipulation tasks**, and the LIBERO simulation benchmark [84]. We first test out-of-the-box performance on one seen and three unseen embodiments, then study fine-tuning efficiency on new robots and tasks. We investigate the impact of LAP's language-action representation on training dynamics and conclude with an analysis of co-training with Visual–Question–Answering (VQA) datasets.

Baselines. We evaluate two categories of baselines below. All replicated baselines are carefully tuned (learning rate, batch size, loss weights, etc.) for fair comparison; full details are in Appendix B.

a) Open-Sourced VLAs.: We evaluate several publicly released VLA checkpoints to assess whether existing methods already exhibit zero-shot cross-embodiment transfer: (1) **$\pi_{0.5}$-DROID** [1], currently the strongest VLA on the DROID benchmark; (2) **$\pi_{0.5}$-Base** [1], the strongest publicly available VLA base model to our knowledge; (3) **X-VLA** [5], which explicitly targets embodiment generalization; (4) **MolmoAct** [2]; and (5) **OpenVLA** [7].

b) Action-Representation Comparison.: To isolate the impact of language-action supervision, we train controlled baselines with identical architectures and data mixtures, differing only in action representation:

1) **$\pi_{0.5}$-replicated**: supervises the VLM using FAST tokens [26]. Since the FAST tokenizer is pre-trained on substantially more robot data than our model, this forms a particularly strong baseline.
2) **π_0-replicated**: removes action supervision from the VLM entirely and only supervises the action expert, measuring the benefit of language-action pre-training.
3) **VLA-0-replicated**: supervises the VLM following Goyal et al. [6], where continuous actions are represented as digit tokens.

Across all experiments, we conduct **over 1300 real-robot evaluation trials**.

A. Zero-Shot Cross-Embodiment Generalization

Embodiments and Tasks. We evaluate on four robot embodiments (Fig. 3): the DROID setup [11] which is included in the pre-training mixture, three unseen embodiments—a custom 7-DoF Franka Panda robot with a novel gripper and wrist-mounted camera configuration, the 6-DoF YAM robot, and the 7-DoF Kinova robot.

We evaluate five manipulation task types (Fig. 3, top row):

1) **Pick and Place**: pick up an object and place it into a container, testing generalization across gripper geometry, camera placement, and robot appearance.
2) **Sort**: sort all objects on the table into a basket, evaluating long-horizon manipulation.
3) **Tissue**: pull tissue from a box and place it on the table, requiring adaptive grasping of deformable objects and height reasoning.
4) **Put Towel**: pick up a flat towel and place it into a basket, requiring precise grasp localization and large vertical motion.

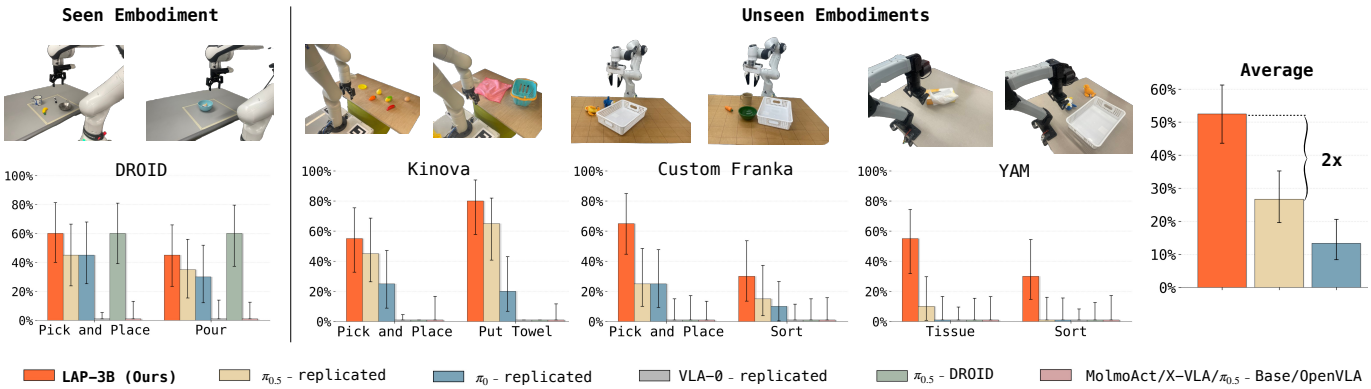

Fig. 3: **Zero-shot cross-embodiment generalization performance.** LAP-3B achieves performance comparable to the $\pi_{0.5}$-DROID on the seen embodiment. Across three previously unseen embodiments and six real-world manipulation tasks, LAP-3B attains over **50% average zero-shot success**, delivering approximately a $2\times$ improvement over the strongest baselines, while all open-sourced VLAs collapse to zero success rate. Error bars denote 95% finite-sample–valid confidence intervals that control the Type-I error (miscoverage) probability, following [83].

5) **Pour**: pour contents from a bowl, testing rotational control and primitive sequencing.

These tasks require full 6-DoF manipulation, e.g., rotating the gripper to avoid collisions during sorting, and probe complementary primitive skills.

For each embodiment, we evaluate on two of these tasks with 20 trials per task. We report mean success rates with 95% confidence intervals following the method of [83], which provides statistically valid finite-sample coverage. Unlike heuristic standard deviations, these intervals explicitly control the Type-I error (miscoverage) probability. Success metric is binary: a trial is marked successful only if the full task is completed. Results are summarized in Fig. 3.

Seen Embodiment Evaluation. On the in-distribution DROID setup, we observe that nearly all open-sourced VLAs fail to achieve non-trivial task success, with the sole exception of $\pi_{0.5}$ – DROID [1], which has been explicitly fine-tuned on the DROID dataset [11]. This highlights a critical limitation of current VLAs: despite large-scale multi-embodiment pre-training, effective deployment even on seen embodiments typically requires embodiment-specific adaptation.

For our replicated comparison baselines, $\pi_{0.5}$-replicated and π_0-replicated achieve moderate and comparable performance. However, VLA-0-replicated consistently collapses to near-zero actions. We find this failure mode stems from overfitting to trivial predictions, which manifests as predicting only small-magnitude action tokens. This issue may be less of a concern on smaller, cleaner benchmarks such as LIBERO [84] but becomes detrimental at large scale, where noise and behavioral diversity are unavoidable.

Notably, LAP-3B consistently outperforms $\pi_{0.5}$-replicated and π_0-replicated by approximately 15 percentage points across tasks, despite sharing the same backbone architecture and training data; this highlights the superiority of the language-action representation. Moreover, LAP-3B achieves performance on par with $\pi_{0.5}$-DROID—without any embodiment-specific fine-tuning on the DROID dataset itself. This indicates that language-action supervision enables effective learning from large-scale, heterogeneous robot data, leading to strong in-distribution performance.

Unseen Embodiment Evaluation. We next evaluate zero-shot transfer to three previously unseen robot embodiments. Across all settings, LAP-3B demonstrates a substantial and consistent advantage, with an *average success rate exceeding 50%*, representing an approximate **2× improvement over the strongest baseline**. Importantly, LAP-3B is the only model to achieve consistently high performance across all novel embodiments.

In contrast, all open-sourced VLAs, including $\pi_{0.5}$-DROID, which performed strongly in-distribution, fail to generate meaningful behavior on unseen robots. This sharp collapse suggests that existing VLA training recipes remain tightly coupled to embodiment-specific control conventions, preventing transfer even when visual and task semantics remain similar.

By contrast, LAP-3B exhibits both *semantically grounded movements* and *fine-grained spatial control* across embodiments. We hypothesize that the language-action interface provides a shared, embodiment-agnostic abstraction that better preserves the VLM's generalizable visual-semantic representations while enabling precise execution under novel robot kinematics. This decoupling allows LAP-3B to retain coherent task intent while avoiding the embodiment-specific overfitting that hampers existing VLAs.

B. Fine-Tuning Efficiency on New Embodiments

While LAP-3B exhibits non-trivial zero-shot performance on novel embodiments for simple tasks, more challenging manipulation scenarios still require embodiment-specific adaptation at the current training scale. We therefore investigate whether LAP improves *fine-tuning efficiency* along two axes: *data efficiency*—requiring fewer demonstrations to reach a given performance level, and *compute efficiency*—converging with fewer gradient steps.

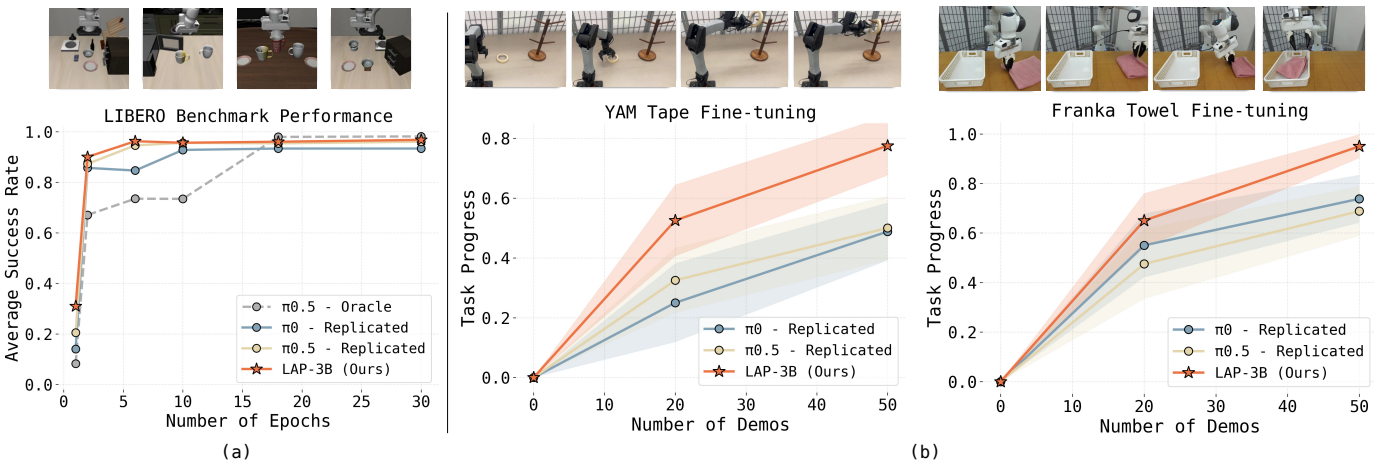

Fig. 4: **Fine-tuning efficiency in simulation (LIBERO) and real-world manipulation tasks**. Across both domains, LAP-3B adapts substantially faster than baseline policies, reaching high performance with significantly fewer epochs and demonstrations. In simulation, LAP-3B converges to near-optimal success within a fraction of the training steps required by baselines. On real robots, LAP-3B achieves comparable task performance using approximately **2.5× fewer demonstrations**, demonstrating substantially improved data and compute efficiency when transferring to new embodiments.

Simulation Fine-Tuning Experiments. We first fine-tune LAP-3B on the LIBERO simulation benchmark [84] and compare against $\pi_{0.5}$-replicated and π_0-replicated baselines trained on the same pre-training mixture. We additionally report an oracle baseline, $\pi_{0.5}$-Oracle, initialized from the original $\pi_{0.5}$ weights pre-trained on substantially larger datasets.

As shown in Fig. 4(a), LAP-3B converges substantially faster than all baselines. With only a single epoch of fine-tuning, LAP-3B already reaches 78% success, and achieves near-maximum performance (96.8%) within six epochs—significantly fewer updates than both replicated baselines and the oracle model [1]. While $\pi_{0.5}$-Oracle eventually attains slightly higher asymptotic performance after extended training, LAP-3B demonstrates markedly superior compute efficiency. These results indicate that language-action pre-training yields a more transferable initialization that adapts rapidly to new embodiments. We summarize LIBERO benchmark results for LAP-3B and baseline methods in Appendix A.

Real-World Fine-Tuning Experiments. We next fine-tune on two challenging real-robot tasks (Fig. 4 top row, right):

- **Hang Tape on Rack**: pick up the tape by its edge and hang it on the top rack, requiring dexterity and precision. The task is evaluated on the YAM robot.
- **Fold Towel and Place in Basket**: fold the towel in half from left to right, then place the folded towel into a basket, requiring both fine-grained dexterity and long-horizon manipulation. The task is evaluated on the Custom Franka robot.

Results are shown in Fig. 4. We evaluate task progress metrics over 20 trials per data point (see Appendix B-K for details).

Across both tasks, LAP-3B achieves strong performance with substantially fewer demonstrations. On YAM, LAP-3B reaches approximately 50% task progress using only 20 demonstrations—roughly **2.5× demos fewer than baselines**. On Franka, LAP-3B consistently outperforms $\pi_{0.5}$-replicated and π_0-replicated across all data regimes. Together, these results demonstrate that language-action pre-training produces a more adaptable VLA, improving both sample efficiency and fine-tuning speed when transferring to new embodiments and tasks.

C. Analyzing LAP's Cross-Embodiment Generalization

T-SNE visualization of embodiment representations. We visualize learned embodiment features using t-SNE in Fig. 5(a) by embedding the average prelogits of action tokens, which capture internal control representations across robot embodiments. LAP-3B exhibits substantially improved alignment between unseen embodiments and training data, with features from unseen embodiments overlapping closely with those from training embodiments rather than forming separate clusters. In this representation space, stronger overlap between training and unseen embodiments indicates that the model has learned more transferable, embodiment-agnostic representations, helping explain why language-action pre-training enables strong zero-shot embodiment transfer.

Low action prediction error on unseen embodiments. We next evaluate the quality of representations learned during pre-training by measuring action prediction error on held-out datasets from *unseen* robot embodiments described in Sec. IV-B. Prediction error is measured as the ℓ_2 distance between predicted and ground-truth actions.

As shown in Fig. 5(b), LAP-3B consistently achieves lower action prediction error on unseen embodiments than baselines trained with alternative action representations, indicating that language-action supervision leads to more generalizable control representations across embodiment shifts. In addition, the prediction error decreases smoothly and monotonically over training, further suggesting that LAP facilitates generalizable

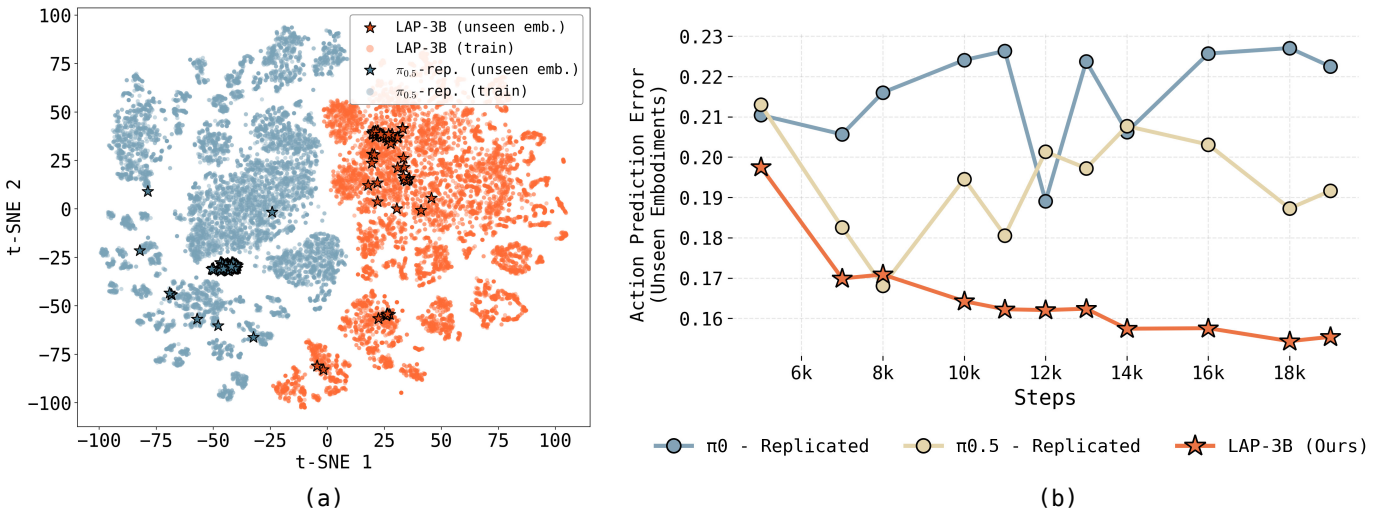

Fig. 5: **(a) T-SNE visualizations of learned embodiment representations** for LAP-3B and $\pi_{0.5}$-replicated. LAP-3B exhibits substantial overlap between training and unseen embodiments, whereas $\pi_{0.5}$-replicated shows limited alignment, indicating that LAP-3B learns more transferable, embodiment-agnostic control representations. **(b) Action prediction error on unseen embodiments during pre-training.** LAP-3B achieves consistently lower action prediction error on held-out unseen embodiments throughout training, compared to $\pi_{0.5}$-replicated and π_0-replicated baselines. This indicates that language-action supervision enables the model to learn control representations that generalize across embodiments, allowing more accurate action prediction on novel robots as well as smoother training dynamics.

representation learning and promotes more stable training dynamics.

TABLE II: Best prediction errors on held-out datasets from *unseen robot embodiments* for LAP-3B and baseline methods. Best per row is bolded.

	LAP-3B	$\pi_{0.5}$-replicated	π_0-replicated
Prediction Error (Unseen)	**0.151**	0.168	0.189
Validation Loss (Unseen)	**0.049**	0.051	0.052

We additionally report the best validation loss of the (flow-matching) action expert in Table II. Taken together, these results suggest that language-action pre-training does not merely stabilize optimization, but fundamentally improves how control representations transfer across embodiments—allowing LAP-3B to predict more accurate actions on unseen robots and explaining both its strong zero-shot performance and superior fine-tuning efficiency.

D. Co-Training with Vision-Language Datasets

We further investigate whether co-training with additional Visual-Question-Answering (VQA) tasks benefits LAP-3B. We consider a motion-prediction task in which the model is prompted with two consecutive images and asked to predict the corresponding language-action describing the change between frames, analogous to an "inverse dynamics" prediction (Fig. 1, bottom left). This co-training objective naturally bridges standard VQA-style supervision and our language-action prediction by sharing a unified, language-based output format (Fig. 6). We evaluate the resulting models on two embodiments, Custom Franka and YAM, with results shown in Incorporating this motion-prediction VQA task leads to more precise action generation and improved spatial generalization, yielding higher success rates across embodiments. In addition, co-training produces checkpoints that are easier to adapt to new tasks, enabling faster fine-tuning with fewer gradient steps and achieving higher final convergence performance (see Appendix A).

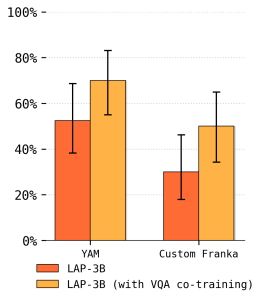

Fig. 6: **VQA co-training** with a motion-prediction task improves cross-embodiment performance.

E. Scaling with Model Size

We next investigate whether LAP benefits from increased model capacity, and whether its advantages persist in the large-model regime.

Using Gemma3 [85] as the VLM backbone, we train LAP models at 4B, 12B, and 27B parameter scales, and compare them against corresponding $\pi_{0.5}$-replicated variants, which represent the current state of the art in VLAs.

We evaluate two metrics on a held-out split comprising 2.5% of the full data mixture: (i) token validation loss of the VLM backbone, and (ii) validation loss of the flow-matching-based action expert. For token validation loss, we report the percentage reduction relative to the 4B model, as absolute loss values are not directly comparable across different action representations.

As shown in Fig. 7, LAP consistently attains lower validation loss across all model sizes and exhibits favorable scaling behavior, with performance improving monotonically as model capacity increases. In contrast, the $\pi_{0.5}$-replicated baseline shows early saturation and even degradation at larger scales. These results indicate that language-action supervision remains effective in the large-model regime and increasingly benefits from additional representational capacity.

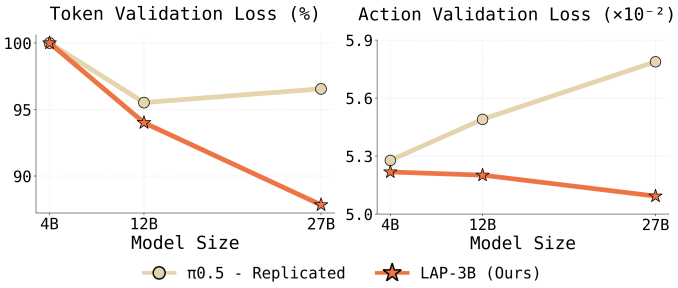

Fig. 7: **Model scaling behavior of LAP compared to $\pi_{0.5}$-replicated.** Left: Token validation loss, reported as percentage drop relative to the 4B model. Right: Continuous action validation loss of the diffusion-based action expert. LAP-3B improves consistently with scale, whereas the baseline saturates and degrades.

V. Conclusions and Discussions

We introduce Language-Action Pre-training (LAP), a scalable VLA pre-training recipe that predicts actions in natural language, yielding representations that transfer effectively across robot embodiments. We further present LAP-3B, which achieves substantial zero-shot generalization on real robots—delivering roughly a 30% absolute improvement over prior action representations and, to our knowledge, the first substantial zero-shot success rate on novel embodiments without fine-tuning. LAP-3B also exhibits stable training, favorable scaling, and strong fine-tuning efficiency.

Broader implications. LAP opens the possibility of training VLMs *explicitly for robotics*, enabling robot (and potentially non-robot) interaction data to be seamlessly mixed with standard VLM corpora to learn models that reason, communicate, and act within a unified representation space.

Limitations and future work. Despite these promising results, several limitations remain. While LAP can be applicable to substantially more heterogeneous robot systems and data sources, this paper focuses only on zero-shot transfer across single-arm manipulators. For example, LAP can be easily adapted to bimanual robots, which actions can be represented through coordinated dual-arm descriptions (e.g., Left Arm: move forward 5 cm; Right Arm: rotate clockwise 20 degrees). Another key advantage of language-actions is that they naturally operate at varying levels of precision and are more tolerant to less accurate action labels, enabling the use of data such as human pose tracking, UMI, or internet video where precise low-level control signals are difficult to obtain. Systematically investigating these directions at scale remains an important avenue for future work.

Second, although LAP-3B adapts efficiently to moderately dexterous tasks such as *Fold Towel and Put in Basket* and *Hang Tape on Rack*, we have not yet evaluated regimes requiring substantially higher control frequency or extreme precision (e.g., fast reactive control, or fine-grained deformable object manipulation). Extending language-action supervision to these settings—potentially through hierarchical or multi-scale action representations—remains an open challenge.

Acknowledgments

We thank Google's TPU Research Cloud (TRC) for providing access to Cloud TPUs. We also thank Maria Attarian, Samuel Bateman and Julian Mendes for their valuable feedback on this project. The authors were partially supported by the NSF CAREER Awards #2044149, the Office of Naval Research under Grant N00014-23-1-2148, the Army Research Laboratory under DCIST CRA W911NF-17-2-0181, a Sloan Fellowship, and Gemini Academic Program. Asher J. Hancock was supported by the NSF Graduate Research Fellowship (DGE-2146755).

References

[1] Physical Intelligence et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv:2504.16054*, 2025.

[2] Jason Lee et al. Molmoact: Action reasoning models that can reason in space. *arXiv:2508.07917*, 2025.

[3] Gemini Robotics Team et al. Gemini robotics: Bringing ai into the physical world. *arXiv:2503.20020*, 2025.

[4] NVIDIA et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv:2503.14734*, 2025.

[5] Jinliang Zheng et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. *arXiv:2510.10274*, 2025.

[6] Ankit Goyal, Hugo Hadfield, Xuning Yang, Valts Blukis, and Fabio Ramos. Vla-0: Building state-of-the-art vlas with zero modification. *arXiv:2510.13054*, 2025.

[7] Moo Jin Kim et al. Openvla: An open-source vision-language-action model. *arXiv:2406.09246*, 2024.

[8] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv:2502.19645*, 2025.

[9] 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:2410.07864*, 2025.

[10] Embodiment Collaboration et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv:2310.08864*, 2025.

[11] Alexander Khazatsky et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv:2403.12945*, 2025.

[12] AgiBot-World-Contributors et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv:2503.06669*, 2025.

[13] Galaxea Team. Galaxea g0: Open-world dataset and dual-system vla model. *arXiv preprint arXiv:2509.00576*, 2025.

[14] Kun Wu et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv:2412.13877*, 2025.

[15] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv:2402.10329*, 2024.

[16] Tony Tao, Mohan Kumar Srirama, Jason Jingzhou Liu, Kenneth Shaw, and Deepak Pathak. Dexwild: Dexterous human interactions for in-the-wild robot policies. *arXiv:2505.07813*, 2025.

[17] Kristen Grauman et al. Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv:2110.07058*, 2022.

[18] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. *arXiv:2401.08399*, 2024.

[19] Prithviraj Banerjee et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. *arXiv:2411.19167*, 2025.

[20] Xin Wang et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. *arXiv:2309.17024*, 2023.

[21] Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv:2505.11709*, 2025.

[22] Dima Damen et al. The epic-kitchens dataset: Collection, challenges and baselines. *arXiv:2005.00343*, 2020.

[23] Asher J. Hancock, Xindi Wu, Lihan Zha, Olga Russakovsky, and Anirudha Majumdar. Actions as language: Fine-tuning vlms into vlas without catastrophic forgetting. *arXiv:2509.22195*, 2025.

[24] Zhongyi Zhou et al. ChatVLA: Unified multimodal understanding and robot control with vision-language-action model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, November 2025.

[25] Danny Driess et al. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *arXiv:2505.23705*, 2025.

[26] Karl Pertsch et al. Fast: Efficient action tokenization for vision-language-action models. *arXiv:2501.09747*, 2025.

[27] Jinliang Zheng et al. Universal actions for enhanced embodied foundation models. *arXiv:2501.10105*, 2025.

[28] Seonghyeon Ye et al. Latent action pretraining from videos. *arXiv:2410.11758*, 2025.

[29] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv:2505.06111*, 2025.

[30] Anzhe Chen, Yifei Yang, Zhenjie Zhu, Kechun Xu, Zhongxiang Zhou, Rong Xiong, and Yue Wang. Toward embodiment equivariant vision-language-action policy. *arXiv:2509.14630*, 2025.

[31] Weixin Liang et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

[32] Frederik Ebert et al. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv:2109.13396*, 2021.

[33] Homer Rich Walke et al. Bridgedata v2: A dataset for robot learning at scale. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*. PMLR, 06–09 Nov 2023.

[34] 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. *arXiv:2307.00595*, 2023.

[35] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *arXiv:1910.11215*, 2020.

[36] 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 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 991–1002. PMLR, 08–11 Nov 2022.

[37] Anthony Brohan et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv:2212.06817*, 2023.

[38] Octo Model Team et al. Octo: An open-source generalist robot policy. *arXiv:2405.12213*, 2024.

[39] Kevin Black et al. π_0: A vision-language-action flow model for general robot control. *arXiv:2410.24164*, 2024.

[40] Qixiu Li et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv:2411.19650*, 2024.

[41] Anthony Brohan et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv:2307.15818*, 2023.

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

[43] Ruijie Zheng et al. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv:2412.10345*, 2025.

[44] Qingqing Zhao et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1702–1713, June 2025.

[45] Jensen Gao, Annie Xie, Ted Xiao, Chelsea Finn, and Dorsa Sadigh. Efficient data collection for robotic manipulation via compositional generalization. *arXiv:2403.05110*, 2024.

[46] Lihan Zha, Apurva Badithela, Michael Zhang, Justin

Lidard, Jeremy Bao, Emily Zhou, David Snyder, Allen Z. Ren, Dhruv Shah, and Anirudha Majumdar. Guiding data collection via factored scaling curves. *arXiv:2505.07728*, 2025.

[47] Yingdong Hu, Fanqi Lin, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. *arXiv:2410.18647*, 2025.

[48] Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. *arXiv:2408.11812*, 2024.

[49] 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:2402.19432*, 2024.

[50] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22508–22519, June 2025.

[51] Yi Chen et al. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19752–19763, October 2025.

[52] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery. *arXiv:2307.09955*, 2023.

[53] Tianyu Wang, Dwait Bhatt, Xiaolong Wang, and Nikolay Atanasov. Cross-embodiment robot manipulation skill transfer using latent space alignment. *arXiv:2406.01968*, 2024.

[54] Hao Luo et al. Being-h0.5: Scaling human-centric robot learning for cross-embodiment generalization. *arXiv:2601.12993*, 2026.

[55] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, Hongxu Yin, Sifei Liu, Song Han, Yao Lu, and Xiaolong Wang. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv:2507.12440*, 2025.

[56] Wenlong Huang, Igor Mordatch, and Deepak Pathak. One policy to control them all: Shared modular policies for agent-agnostic control. *arXiv:2007.04976*, 2020.

[57] Vitaly Kurin, Maximilian Igl, Tim Rocktäschel, Wendelin Boehmer, and Shimon Whiteson. My body is a cage: the role of morphology in graph-based incompatible control. *arXiv:2010.01856*, 2021.

[58] Agrim Gupta, Linxi Fan, Surya Ganguli, and Li Fei-Fei. Metamorph: Learning universal controllers with transformers. *arXiv:2203.11931*, 2022.

[59] Austin Patel and Shuran Song. Get-zero: Graph embodiment transformer for zero-shot embodiment generalization. *arXiv:2407.15002*, 2024.

[60] Nico Bohlinger, Grzegorz Czechmanowski, Maciej Krupka, Piotr Kicki, Krzysztof Walas, Jan Peters, and Davide Tateo. One policy to run them all: an end-to-end learning approach to multi-embodiment locomotion. *arXiv:2409.06366*, 2025.

[61] Lawrence Yunliang Chen et al. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. *arXiv:2409.03403*, 2024.

[62] Tyler Ga Wei Lum, Olivia Y. Lee, C. Karen Liu, and Jeannette Bohg. Crossing the human-robot embodiment gap with sim-to-real rl using one human demonstration. *arXiv:2504.12609*, 2025.

[63] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos. *arXiv:2503.00779*, 2025.

[64] Juntao Ren, Priya Sundaresan, Dorsa Sadigh, Sanjiban Choudhury, and Jeannette Bohg. Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning. *arXiv:2501.06994*, 2025.

[65] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv:2502.05855*, 2025.

[66] Songming Liu, Bangguo Li, Kai Ma, Lingxuan Wu, Hengkai Tan, Xiao Ouyang, Hang Su, and Jun Zhu. Rdt2: Exploring the scaling limit of umi data towards zero-shot cross-embodiment generalization, 2026. URL https://arxiv.org/abs/2602.03310.

[67] Noriaki Hirose, Catherine Glossop, Dhruv Shah, and Sergey Levine. Omnivla: An omni-modal vision-language-action model for robot navigation. *arXiv:2509.19480*, 2025.

[68] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. Vint: A foundation model for visual navigation. *arXiv:2306.14846*, 2023.

[69] Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. Gnm: A general navigation model to drive any robot. *arXiv:2210.03370*, 2023.

[70] Shresth Grover, Akshay Gopalkrishnan, Bo Ai, Henrik I. Christensen, Hao Su, and Xuanlin Li. Enhancing generalization in vision-language-action models by preserving pretrained representations. *arXiv:2509.11417*, 2025.

[71] Yicheng Liu et al. Faster: Toward efficient autoregressive vision language action modeling via neural action tokenization. *arXiv:2512.04952*, 2025.

[72] Dantong Niu, Yuvan Sharma, Giscard Biamby, Jerome Quenum, Yutong Bai, Baifeng Shi, Trevor Darrell, and Roei Herzig. Llarva: Vision-action instruction tuning enhances robot learning. *arXiv:2406.11815*, 2024.

[73] Michael Ahn et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv:2204.01691*, 2022.

[74] Danny Driess et al. Palm-e: an embodied multimodal

language model. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[75] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. *arXiv:2209.07753*, 2023.

[76] Lihan Zha, Yuchen Cui, Li-Heng Lin, Minae Kwon, Montserrat Gonzalez Arenas, Andy Zeng, Fei Xia, and Dorsa Sadigh. Distilling and retrieving generalizable knowledge for robot manipulation via language corrections. *arXiv:2311.10678*, 2024.

[77] Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language. *arXiv:2403.01823*, 2024.

[78] Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-world robotic manipulation through mark-based visual prompting. *arXiv:2403.03174*, 2024.

[79] Norman Di Palo and Edward Johns. Keypoint action tokens enable in-context imitation learning in robotics. *arXiv:2403.19578*, 2024.

[80] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv:2409.01652*, 2024.

[81] Lucas Beyer et al. Paligemma: A versatile 3b vlm for transfer. *arXiv:2407.07726*, 2024.

[82] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019.

[83] Joseph A. Vincent, Haruki Nishimura, Masha Itkina, Paarth Shah, Mac Schwager, and Thomas Kollar. How generalizable is my behavior cloning policy? a statistical approach to trustworthy performance evaluation. *arXiv:2405.05439*, 2024.

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

[85] Gemma Team et al. Gemma 3 technical report. *arXiv:2503.19786*, 2025.

[86] Homer Walke et al. Bridgedata v2: A dataset for robot learning at scale. *arXiv:2308.12952*, 2024.

[87] Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. Grounding language with visual affordances over unstructured data. *arXiv:2210.01911*, 2023.

[88] Shivin Dass, Jullian Yapeter, Jesse Zhang, Jiahui Zhang, Karl Pertsch, Stefanos Nikolaidis, and Joseph J. Lim. Clvr jaco play dataset, 2023.

[89] Minho Heo, Youngwoon Lee, Doohyun Lee, and Joseph J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. *arXiv:2305.12821*, 2023.

[90] Rutav Shah, Roberto Martín-Martín, and Yuke Zhu. Mutex: Learning unified policies from multimodal task specifications. In *7th Annual Conference on Robot Learning*, 2023.

[91] 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. 2023.

[92] Jianlan Luo et al. Fmb: a functional manipulation benchmark for generalizable robotic learning. *arXiv:2401.08553*, 2024.

[93] Lawrence Yunliang Chen, Simeon Adebola, and Ken Goldberg. Berkeley UR5 demonstration dataset.

[94] Yifeng Zhu, Peter Stone, and Yuke Zhu. Bottom-up skill discovery from unsegmented demonstrations for long-horizon robot manipulation. *arXiv:2109.13841*, 2022.

[95] Soroush Nasiriany, Tian Gao, Ajay Mandlekar, and Yuke Zhu. Learning and retrieval from prior data for skill-based imitation learning. In *Conference on Robot Learning (CoRL)*, 2022.

[96] 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. *arXiv:2211.08416*, 2023.

[97] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. *arXiv:2210.11339*, 2023.