RETHINKING VISION-LANGUAGE-ACTION MODEL SCALING: ALIGNMENT, MIXTURE, AND REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–Language–Action (VLA) models show strong promise for generalist robot control. However, it remains unclear whether—and under what conditions—the standard "scale data" recipe translates to robotics, given that robotic training data is inherently heterogeneous across embodiments, sensors, and action spaces. We present a systematic, controlled study of VLA scaling that revisits core training choices for pretraining across diverse robots. Using a representative VLA framework combining a vision–language backbone with flow-matching, we ablate key design decisions under matched conditions in extensive simulation and real-robot experiments. To improve the reliability of real-world evaluation, we introduce a Grouped Blind Ensemble protocol that blinds operators to model identity and randomizes evaluation order within groups. Our analysis targets three dimensions of VLA scaling: (1) Physical alignment: Using SE(3) as a shared action space, we find that end-effector (EEF)-relative actions support positive cross-embodiment transfer. (2) Embodiment mixture: Adding heterogeneous robot datasets can reduce downstream transfer performance, even with the same training budget. (3) Training regularization: We observe that common strategies, such as sensory dropout and multi-stage training, do not consistently improve performance at scale. Together, this study challenges common assumptions about embodied scaling and provides practical, evidence-based guidance for training large-scale VLA policies.

1 INTRODUCTION

Vision–Language–Action (VLA) models (Brohan et al., 2022; Intelligence et al., 2025) have become a promising direction for general-purpose embodied AI. Following the scaling trends of vision–language models (VLMs) (Comanici et al., 2025; Bai et al., 2025), robotics is moving from single-task policies to generalist policies that aim to solve many tasks across different environments. This shift is supported by recent large robotic datasets (O'Neill et al., 2024; Bu et al., 2025) with thousands of hours from diverse robot platforms, collected in both simulation and the real world.

A common belief is that, as in language modeling, improved generalization in robotics will primarily emerge from scaling data and model size. However, scaling in robotics introduces physical and system-level challenges that do not arise in text or image generation. Robotic data is inherently heterogeneous: robots differ in kinematics, joint limits, control frequencies, sensing modalities, and action spaces, which are often incompatible by design. While joint-space control yields strong performance in homogeneous setups, its tight coupling to specific degrees of freedom and joint limits makes it difficult to define a shared action space across embodiments. Therefore, this study focuses on the SE(3) task space—which inherently shares uniform physical semantics—exploring its potential and bottlenecks as a universal protocol for scaling. As a result, it remains unclear whether scaling heterogeneous robotic data within this space reliably leads to positive transfer, or whether representation misalignments introduce interference that limits performance. These challenges lead to several open questions: *How do action representations affect the alignment of heterogeneous data? What are the effects of mixing data from different robot datasets?* And *do common training strategies remain effective when applied at scale in these settings?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, we address these questions through a controlled empirical study of VLA scaling. Instead of proposing a new architecture, we build a controlled testbed based on a representative VLA framework that combines a VLM backbone with flow matching (Liu et al., 2022; Black et al., 2024; Luo et al., 2026). We treat the training pipeline itself as the primary object of study and systematically ablate key design choices under matched conditions. Our study is organized around three pillars: **(1) Physical Alignment.** Focusing on the SE(3) task space, we examine how different spatial reference frames (e.g., EEF-relative vs. world-relative actions) affect general control performance and cross-embodiment transfer. **(2) Embodiment Mixture.** We study how performance changes with different mixtures of heterogeneous data, specifically investigating whether cross-embodiment variation promotes positive transfer or introduces interference. **(3) Training Regularization.** We evaluate the scalability of training modifications, such as sensory dropout and multi-stage curricula, to determine if these approaches provide tangible benefits in a large-scale pretraining regime.

Reliable evaluation is critical for drawing robust conclusions about scaling. Real-world testing is highly susceptible to experimenter bias, including unconscious adjustments to execution speed or initial object placement. To systematically mitigate this, we introduce a Grouped Blind Ensemble protocol: operators execute randomized, anonymized policies and record subtask outcomes without access to model identities.

Our main findings provide practical guidance for large-scale VLA pretraining: 1) EEF-relative actions support positive cross-embodiment transfer; 2) Simply mixing heterogeneous data frequently degrades performance; 3) Standard regularization heuristics (sensory dropouts, multi-stage curricula) do not consistently improve performance at scale; 4) Our grouped blind ensemble protocol combines operator blinding and randomized evaluation to reduce experimenter bias in real-world robotic testing.

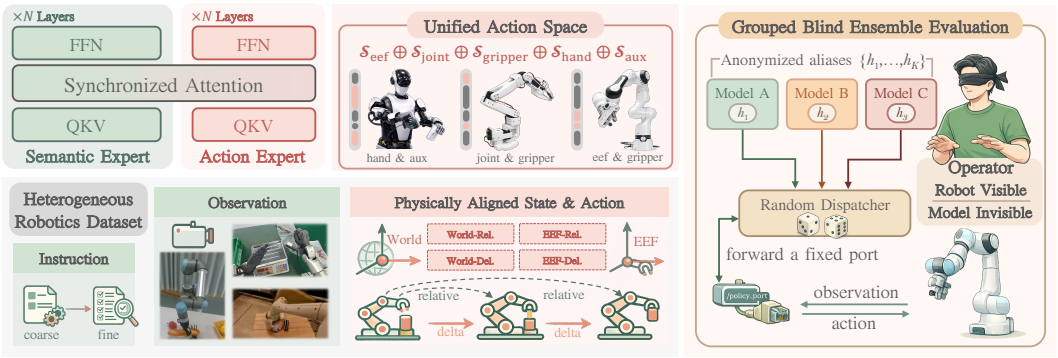

Figure 1: Overview of systematic VLA analysis framework, comprising the Mixture-of-Transformers architecture, physically aligned action spaces, and the Grouped Blind Ensemble protocol.

2 Related Work

Vision-Language-Action Models. Research in robotic manipulation has evolved significantly, transitioning from narrowly defined, single-task policies to more versatile systems that are trained on extensive and varied datasets (Berscheid et al., 2019; Dasari et al., 2019; Fang et al., 2023; Shafiullah et al., 2023). Vision–Language–Action (VLA) models (Brohan et al., 2022; Zitkovich et al., 2023; Kim et al., 2024; Janner et al., 2022; Chi et al., 2025) exemplify this trend by transferring the perceptual and reasoning priors of pre-trained vision-language models (VLMs) (Steiner et al., 2024; Li et al., 2025; Wang et al., 2024; Zhang et al., 2025b;a; Hao et al., 2025; Feng et al., 2025) into embodied control through fine-tuning on robot demonstrations. Although many VLAs share similar backbone architectures, their action generation choices vary widely. Early works often cast continuous control as discrete tokens and decode actions autoregressively to better match VLM-style objectives (Kim et al., 2024; Qu et al., 2025); this design eases transfer but can increase inference cost and struggle with high-precision, high-DoF execution. In contrast, more recent approaches adopt diffusion-style objectives (Ho et al., 2020; Li et al., 2024; Janner et al., 2022; Liu et al., 2024; Liang et al., 2025) or flow-matching formulations (Lipman et al., 2022; Nvidia et al., 2025; Black

et al., 2024; Intelligence et al., 2025; Luo et al., 2026; Cai et al., 2026) to generate robot action chunks. Despite these advances, it remains challenging to consistently translate high-level VLM priors into fine-grained control across diverse embodiments. In contrast to proposing a new architecture, our work focuses on how training choices such as action parameterization, cross-embodiment data mixture, and regularization affect VLA scaling under controlled conditions.

Robotic Data for Pre-Training. The efficacy of Vision-Language-Action (VLA) training is closely linked to the scale and diversity of the robotic data utilized. Numerous initiatives, such as Open X-Embodiment (O'Neill et al., 2024), Droid (Khazatsky et al., 2024) and BridgeData (Walke et al., 2023), compile significant demonstration hours across various tasks, environments, and platforms. Community-driven datasets and benchmarking efforts further enhance accessibility and coverage in this domain (Shukor et al., 2025). Recent datasets such as AgiBot World (Bu et al., 2025), RoboCOIN (Wu et al., 2025), and RoboMIND (Wu et al., 2024; Hou et al., 2025) broaden coverage of bimanual coordination and dexterous manipulation. Open Galaxea (Jiang et al., 2025) focuses on mobile manipulation. Nevertheless, much teleoperated data is still collected in constrained settings. Synthetic data (Ye et al., 2025; Chen et al., 2025b) can alleviate data scarcity, but the sim-to-real gap remains a persistent obstacle. More fundamentally, these datasets are heterogeneous by construction, spanning different kinematic structures, sensing/control conventions, and action spaces, which complicates naive pooling and transfer. Our study directly examines this heterogeneity and provides controlled evidence on when mixing data across embodiments helps or hinders, together with a bias-reduced real-robot evaluation protocol.

3 Methodology and Evaluation Protocol

To study VLA scaling with heterogeneous robot data, we build a controlled testbed that combines a generalist architecture, a unified action space, and a blinded real-world evaluation protocol (Fig. 1).

3.1 Mixture-of-Transformers with Flow-Matching Control

Our goal is to combine high-frequency, precise control with strong semantic reasoning. Following prior work, we adopt a Mixture-of-Transformers (MoT) design (Deng et al., 2025; Black et al., 2024).

Dual Experts with Layerwise Shared Attention. The model uses two transformer backbones in parallel. The **Semantic Expert** ($\mathcal{E}_{\text{sem}}$) is initialized from a pretrained VLM to preserve visual and language priors. The **Action Expert** ($\mathcal{E}_{\text{act}}$) is randomly initialized. For efficiency, it uses a smaller hidden dimension, but matches the Semantic Expert in depth. We split the input sequence into four token groups: $\mathbf{X} = \{\mathbf{X}_V, \mathbf{X}_L, \mathbf{X}_S, \mathbf{X}_A\}$, corresponding to visual tokens, instruction tokens, proprioceptive state tokens, and action-latent tokens. The Semantic Expert processes the $\{\mathbf{X}_V, \mathbf{X}_L\}$, while the Action Expert is dedicated to the kinematic subset $\{\mathbf{X}_S, \mathbf{X}_A\}$. To support strong cross-modal fusion, the two experts interact at every layer through a shared causal self-attention mechanism. Although the hidden sizes of the two streams differ, we use the same number of attention heads and the same per-head dimension in both experts. At each layer, we compute attention by concatenating the projected Queries, Keys, and Values from the two streams as follows:

$$\mathbf{T}^{(l)} = \left[\, \mathbf{T}^{(l)}_{\text{sem}} \;\; \mathbf{T}^{(l)}_{\text{act}} \,\right], \quad \mathbf{T} \in \{\mathbf{Q}, \mathbf{K}, \mathbf{V}\}, \tag{1}$$

where square brackets $[\,\cdot\,]$ denote concatenation along the sequence dimension. This lets the Action Expert attend to the full visual–semantic context directly, injecting high-frequency control information without first compressing it through a discrete text-embedding space.

Continuous Action Chunking via Flow Matching. We use flow matching (Lipman et al., 2022) to model action generation as a conditional distribution over an action chunk $p(\mathbf{a}_{t:t+H}|\mathbf{o}_t)$ where H is a short prediction horizon. Concretely, the conditioning context $\mathbf{c}$ combines features from the Semantic Expert with proprioceptive inputs. The Action Expert learns a time-dependent vector field $v_\theta(\mathbf{x}, \tau, \mathbf{c})$ that transports samples from the Gaussian distribution $\pi_0(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ to the action-chunk data distribution $\pi_1(\mathbf{x} \mid \mathbf{c})$. During training, given a data sample $\mathbf{x}_1$, a noise sample $\mathbf{x}_0 \sim \pi_0$, and a timestep $\tau \sim \mathcal{U}[0, 1]$, we form the interpolated point $\mathbf{x}_\tau = (1-\tau)\mathbf{x}_0 + \tau\mathbf{x}_1$ and minimize the flow-matching objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\tau, \mathbf{x}_0, \mathbf{x}_1}\Big[\big\| v_\theta(\mathbf{x}_\tau, \tau, \mathbf{c}) - (\mathbf{x}_1 - \mathbf{x}_0)\big\|^2\Big]. \tag{2}$$

At inference time, we integrate the resulting Ordinary Differential Equation (ODE) with an explicit Euler solver to generate the action sequence.

Physically Grounded Unified Action Space. To support scalable pre-training across heterogeneous robots—from single-arm grippers to bimanual dexterous hands—we define a physically grounded unified action space $\mathcal{A}_{\text{uni}} \subseteq \mathbb{R}^D$. This space acts as a superset of all supported physical degrees of freedom (DoF), and is partitioned into semantically aligned subspaces:

$$\mathcal{A}_{\text{uni}} = \mathcal{S}_{\text{eef}} \oplus \mathcal{S}_{\text{joint}} \oplus \mathcal{S}_{\text{gripper}} \oplus \mathcal{S}_{\text{hand}} \oplus \mathcal{S}_{\text{aux}} \tag{3}$$

Here, $\mathcal{S}_{\text{eef}}$ represents end-effector poses (translation and axis–angle rotation) for Cartesian control; $\mathcal{S}_{\text{joint}}$ supports direct joint-space commands (e.g., 7-DoF arms); $\mathcal{S}_{\text{gripper}}$ encodes parallel-jaw gripper states; $\mathcal{S}_{\text{hand}}$ provides higher-dimensional slots for dexterous hands; and $\mathcal{S}_{\text{aux}}$ covers auxiliary mechanisms. To study which action parameterization scales best, we allow flexible action parameterization within $\mathcal{S}_{\text{eef}}$. Let $\mathbf{T}_\tau \in SE(3)$ denote the target end-effector pose at horizon step τ, and $\mathbf{T}_0$ be the pose at the start of the action chunk. We define a coordinate mapping Ψ with four modes:

$$\Psi(\mathbf{T}_\tau) = \begin{cases} \mathbf{T}_\tau \ominus \mathbf{T}_0 & \text{(World-Rel: start-relative)} \\ \mathbf{T}_\tau \ominus \mathbf{T}_{\tau-1} & \text{(World-Delta: step-relative)} \\ \mathbf{T}_0^{-1} \circ \mathbf{T}_\tau & \text{(EEF-Rel: start-relative)} \\ \mathbf{T}_{\tau-1}^{-1} \circ \mathbf{T}_\tau & \text{(EEF-Delta: step-relative)} \end{cases} \tag{4}$$

where $\ominus$ computes the pose difference in the world frame (translation difference and rotational displacement), and $\circ$ denotes $SE(3)$ composition in the local (end-effector) frame. *World/EEF-Rel* measure displacements from the chunk start, while *World/EEF-Delta* capture step-to-step changes. For both Relative and Delta actions, translation and rotation magnitudes are similar between the World and EEF frames. Their means and standard deviations differ by less than 2% across our training data. For each robot embodiment r with native action space $\mathcal{A}_r$, we define an embedding mapping $\phi_r : \mathcal{A}_r \to \mathcal{A}_{\text{uni}}$ that places robot-specific actions into the appropriate semantic slots of the unified space. Dimensions not used by embodiment r are disabled with a binary mask. This construction encourages the model to learn shared physical priors in subspaces that overlap across robots, while still allowing embodiment-specific control in non-overlapping subspaces.

Algorithm 1 Grouped Blind Ensemble Protocol

Require: Model Pool $\mathcal{M}$, Task Set $\mathcal{T}$, Group Size K, Trials per model N

Divide $\mathcal{M}$ into random non-overlapping groups $\{G_1, \ldots, G_M\}$
for each task $\tau \in \mathcal{T}$ **do**
  for each model group G_j **do**
    $\mathcal{H} \leftarrow \text{Anonymize}(G_j)$ ▷ Map models to aliases
    $\mathcal{Q} \leftarrow \text{Shuffle}(\{\underbrace{h_1, \ldots, h_1}_{N}, \ldots, \underbrace{h_K, \ldots, h_K}_{N}\})$ ▷ Create randomized trial queue
    Initialize results $\mathcal{R}_j \leftarrow \emptyset$
    while $\mathcal{Q}$ is not empty **do**
      $h \leftarrow \mathcal{Q}.\text{pop}()$ ▷ Get next anonymous model
      Operator executes task τ with policy π_h
      Record each subtask's binary outcome in $\mathcal{R}_j$
    end while
    Allow the operator to rest between groups
  end for
end for
return Deanonymized aggregate statistics

3.2 Grouped Blind Ensemble Evaluation

Real-world robotic evaluations are prone to experimenter bias, where prior knowledge of model identity can unconsciously influence physical resets or success judgments (Rosenthal, 1976). To mitigate these expectancy effects, we introduce the Grouped Blind Ensemble Protocol (Algorithm 1). The model pool is partitioned into anonymized groups of size K (typically 4–8), within which evaluation order is randomized for each task. The operator acts as a blinded executor, performing

rollouts and recording the success or failure of each subtask without access to model identities. Furthermore, its structured group-wise design enforces rest intervals that mitigate fatigue and ensure evaluation consistency across large-scale experiments.

4 Pre-training Data and Implementation

4.1 Large-scale Heterogeneous Robot Data

Training a generalist VLA requires data that spans diverse embodiments, environments, and control modalities. We aggregate a large-scale collection of robot manipulation trajectories and organize them along two axes: domain (real vs. simulation) and control space (end-effector vs. joint space).

Dataset Composition. Serving as the universal bridge for cross-embodiment transfer, high-quality real-world end-effector (SE(3)) trajectories form the core of our pre-training mixture, sourced from Open X-Embodiment (OXE) (O'Neill et al., 2024), Agibot (Bu et al., 2025), and RoboMind (Wu et al., 2024; Hou et al., 2025). To broaden kinematic coverage and supplement embodiment-specific fine-grained control, we additionally incorporate simulation and joint-space data from InternData (Chen et al., 2025b), SO-100 teleoperation logs (Shukor et al., 2025), and humanoid datasets (Tian et al., 2025).

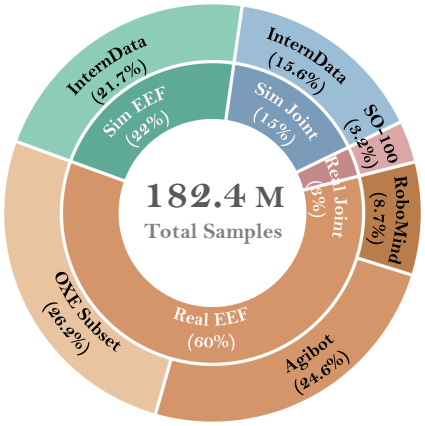

Figure 2: Composition of the balanced pre-training data.

Data Balancing Strategy. A key challenge in co-training is the severe imbalance in dataset scales, with raw frame counts ranging from hundreds of thousands to hundreds of millions. Naively mixing them would cause the largest sources to dominate the gradient. We mitigate this by applying a dynamic, dataset-specific downsampling strategy, aggressively compressing dense simulation streams and high-frequency logs while preserving the density of smaller datasets. As illustrated in Figure 2, this yields a balanced effective corpus of approximately 182.5 million transitions. Comprehensive dataset statistics are provided in Appendix A.

4.2 Pretraining Implementation Details

Architecture and Action Space. The semantic expert $\mathcal{E}_{\text{sem}}$ is initialized from InternVL-3.5-2B (Wang et al., 2025) (hidden size 2048), and the action expert $\mathcal{E}_{\text{act}}$ is a 0.7B parameter Transformer decoder trained from scratch (hidden size 1024). To enable the synchronized attention mechanism across mismatched hidden sizes, we use a shared per-head dimension in both experts. The model operates in the unified action space. We rely on relative end-effector (EEF) commands as the shared representation for cross-embodiment scaling, reserving absolute joint control strictly for embodiment-specific datasets.

Regularization and Optimization. To assess the effect of regularization in large-scale VLA pre-training, we integrate two stochastic techniques: masking proprioception by zeroing the input vector with probability $p = 0.2$, and independently dropping each camera view with $p = 0.2$ (retaining at least one view to ensure valid visual inputs). We train the model using a two-stage curriculum to investigate whether gradual adaptation outperforms direct joint optimization: we first freeze the VLM backbone to train only the action expert (Stage 1), and subsequently unfreeze all parameters for joint training (Stage 2). Detailed hyperparameters, including optimizers, learning rate schedules, and downstream fine-tuning settings, are deferred to Appendix C.

5 Experiments

We study three design dimensions of VLA scaling with heterogeneous robot data: **(1) Physical Alignment**: investigating how coordinate frames and action representations affect control precision; **(2) Embodiment Mixture**: assessing whether aggregating heterogeneous data sources yields

Table 1: Pre-training transfer ablation on Libero 5-shot benchmark under different action space.

Action Space	Init.	Unfrozen VLM					Frozen VLM				
		Spat	Obj	Goal	Long	**Avg**	Spat	Obj	Goal	Long	**Avg**
World-Relative	Scratch	90.8	93.6	78.2	74.8	84.4	72.2	85.0	75.4	43.6	69.1
	Pretrain	89.8 (-1.0)	86.8 (-6.8)	88.4 **(+10.2)**	68.8 (-6.0)	83.5 (-0.9)	81.2 **(+9.0)**	89.8 **(+4.8)**	75.0 (-0.4)	52.6 **(+9.0)**	74.7 **(+5.6)**
World-Delta	Scratch	90.4	93.0	82.8	73.6	85.0	74.2	75.0	75.0	42.6	66.7
	Pretrain	86.2 (-4.2)	92.0 (-1.0)	88.8 **(+6.0)**	71.0 (-2.6)	84.5 (-0.5)	82.6 **(+8.4)**	89.6 **(+14.6)**	73.8 (-1.2)	53.2 **(+10.6)**	74.8 **(+8.1)**
EEF-Relative	Scratch	87.6	93.0	76.4	70.6	81.9	73.0	84.2	70.6	39.6	66.9
	Pretrain	86.0 (-1.6)	90.0 (-3.0)	88.6 **(+12.2)**	73.4 **(+2.8)**	84.5 **(+2.6)**	79.8 **(+6.8)**	90.6 **(+6.4)**	76.6 **(+6.0)**	53.2 **(+13.6)**	75.1 **(+8.2)**
EEF-Delta	Scratch	87.6	93.0	72.0	67.0	79.9	73.2	79.8	70.4	41.0	66.1
	Pretrain	85.6 (-2.0)	89.8 (-3.2)	87.4 **(+15.4)**	66.2 (-0.8)	82.3 **(+2.4)**	77.8 **(+4.6)**	88.2 **(+8.4)**	73.6 **(+3.2)**	48.4 **(+7.4)**	72.0 **(+5.9)**

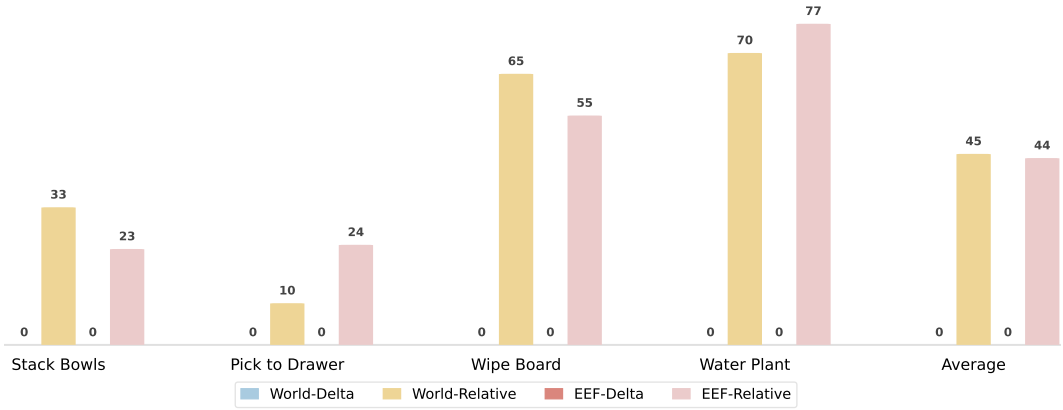

Figure 4: Real-world blinded evaluation of action space parameterizations.

positive transfer or introduces interference; and **(3) Training Regularization**: evaluating the effectiveness of sensory dropout and multi-stage training strategies at scale.

Simulation Benchmarks. We evaluate downstream transfer using multi-task fine-tuning on all four LIBERO (Liu et al., 2023) suites with 5 demonstrations per task, and all 24 RoboCasa (Nasiriany et al., 2024) kitchen tasks with 50 demonstrations per task.

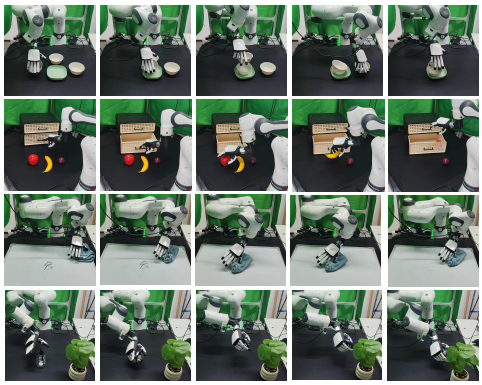

Figure 3: Real-world task sequences.

Real-Robot Protocol. Our setup features a Franka Panda arm with dual RGB cameras. As illustrated in Figure 3 (tasks ordered top-to-bottom), we evaluate four tasks probing complementary physical capabilities. *Stack Bowls* (spatial precision) requires placing three bowls into a plate. *Pick-to-Drawer* (multi-step sequential manipulation) involves opening a drawer, placing three objects inside, and closing it. *Wipe Board* (continuous dynamic motion) includes picking up a cloth, wiping the surface, and returning the cloth. *Water Plant* (precision functional tool use and trigger actuation) requires picking up a spray bottle, aiming, and pressing the trigger. For each task, we conduct 10 trials using the Grouped Blind

Table 3: LIBERO 5-shot transfer performance across incremental pre-training mixtures.

Pre-train Mixture	Frozen VLM					Unfrozen VLM				
	Spat	Obj	Goal	Long	**Avg**	Spat	Obj	Goal	Long	**Avg**
Scratch	73.0	84.2	70.6	39.6	66.9	87.6	93.0	76.4	70.6	81.9
$\mathcal{D}_1$: OXE Only	**87.6** (+14.6)	88.2 (+4.0)	78.8 (+8.2)	**54.6** (+15.0)	**77.3** (+10.4)	**89.2** (+1.6)	89.2 (-3.8)	**92.8** (+16.4)	**74.4** (+3.8)	**86.4** (+4.5)
⌞ $\mathcal{D}_2$: $\mathcal{D}_1$ + Real EEF	84.4 (-3.2)	86.0 (-2.2)	77.4 (-1.4)	47.4 (-7.2)	73.8 (-3.5)	86.2 (-3.0)	**94.6** (+5.4)	88.8 (-4.0)	66.4 (-8.0)	84.0 (-2.4)
⌞ $\mathcal{D}_3$: $\mathcal{D}_2$ + Sim EEF	76.0 (-8.4)	82.2 (-3.8)	**79.0** (+1.6)	51.2 (+3.8)	72.1 (-1.7)	87.4 (+1.2)	91.4 (-3.2)	86.2 (-2.6)	69.6 (+3.2)	83.7 (-0.3)
⌞ $\mathcal{D}_4$: $\mathcal{D}_3$ + Joint	79.8 (+3.8)	**90.6** (+8.4)	76.6 (-2.4)	53.2 (+2.0)	75.1 (+3.0)	86.0 (-1.4)	90.0 (-1.4)	88.6 (+2.4)	73.4 (+3.8)	84.5 (+0.8)

Ensemble protocol to reduce operator bias. Success is reported as the percentage of total points achieved; scoring rubrics are detailed in Appendix D.

5.1 Analysis of Physical Alignment

Standardizing action representations is critical for unifying heterogeneous kinematics and base placements. We evaluate four parameterizations: **World-Relative**, **World-Delta**, **EEF-Relative**, and **EEF-Delta**, comparing scratch-trained policies against those initialized from our heterogeneous pre-training corpus.

Table 2: Pre-training transfer ablation on RoboCasa benchmark under different action space.

Action Space	Init.	Pick/ Place	Doors/ Drawers	Other	**Avg**
World-Relative	Scratch	20.0	60.0	40.0	38.3
	Pretrain	22.5 (+2.5)	58.3 (-1.7)	45.0 (+5.0)	40.8 (+2.5)
World-Delta	Scratch	21.0	62.3	46.0	41.8
	Pretrain	23.8 (+2.8)	60.0 (-2.3)	45.0 (-1.0)	41.7 (-0.1)
EEF-Relative	Scratch	33.0	57.3	47.4	45.1
	Pretrain	35.3 (+2.3)	62.3 (+5.0)	54.4 (+7.0)	50.0 (+4.9)
EEF-Delta	Scratch	26.0	63.7	44.8	43.3
	Pretrain	36.3 (+10.3)	56.7 (-7.0)	49.0 (+4.2)	46.7 (+3.4)

Simulation Results. As shown in Table 1, the benefits of pre-training depend on the action representation. On LIBERO, World-relative actions allow scratch-trained policies to perform well within a fixed environment. However, with the VLM backbone unfrozen, pre-training lowers average success by 0.9 and 0.5 percentage points for World-Relative and World-Delta, respectively. In contrast, EEF-Relative and EEF-Delta provide consistent physical semantics across datasets, yielding gains of 2.6 and 2.4 percentage points, respectively. When the VLM backbone is frozen (Table 1, Frozen), performance relies entirely on the pre-trained Action Expert and vision-language alignment. In this regime, **EEF-Relative** achieves an average success of 75.1% (8.2 percentage points above scratch), confirming its effectiveness in preserving physical priors from heterogeneous data. On RoboCasa (Table 2), pre-training increases EEF-Relative's average success from 45.1% to 50.0%, the highest among the four pre-trained variants. The gain on door and drawer tasks is 5.0 percentage points.

Real-World Results. Both Delta variants score 0% in blinded real-robot evaluations (Figure 4) and exhibit high-frequency jitter during execution, while both Relative variants achieve nonzero scores with stable control.

5.2 Analysis of Embodiment Mixture

With the action space fixed to EEF-Relative, we investigate scaling behavior using four cumulative mixtures: $\mathcal{D}_1$ **(OXE Only)**; $\mathcal{D}_2$ **(+Real EEF)**: adding heterogeneous real-world EEF data; $\mathcal{D}_3$ **(+Sim**

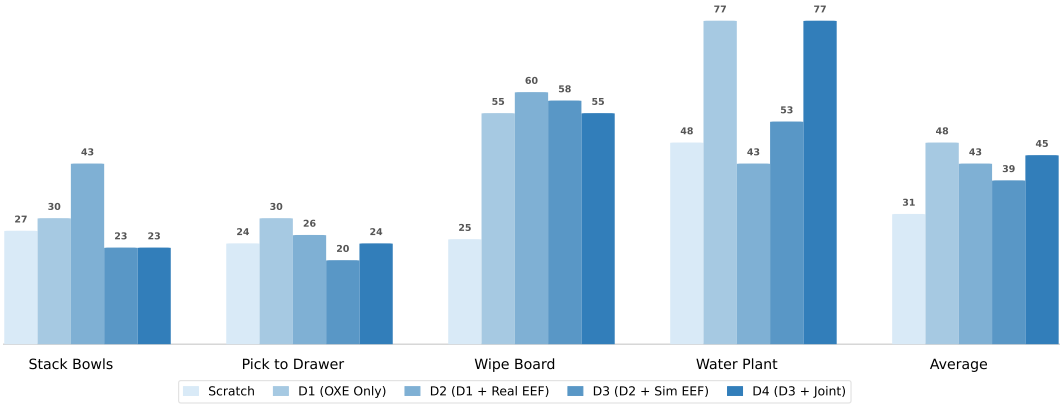

Figure 5: Real-world blind evaluation of different pre-training data mixtures.

EEF): incorporating simulation EEF data; and $\mathcal{D}_4$ **(+Joint)**: including joint-space data projected into the unified space. All four mixtures use the same number of pre-training steps and the same batch size, keeping the total number of sampled training examples fixed.

Simulation Results. Expanding the robotic pre-training corpus does not yield monotonic gains (Table 3 and 4). On LIBERO, $\mathcal{D}_1$ provides the strongest transfer, reaching 77.3% average success (frozen VLM), 10.4 percentage points above Scratch. However, average success drops by 3.5 percentage points in $\mathcal{D}_2$ and a further 1.7 points in $\mathcal{D}_3$. In $\mathcal{D}_4$, we observe a slight recovery (+3.0 percentage points over $\mathcal{D}_3$). Since joint-space data is mapped to separate semantic slots, it may reduce direct action-level interference in the SE(3) subspace and act as auxiliary supervision for the shared vision-language backbone.

Table 4: RoboCasa 50-shot transfer performance across incremental pre-training mixtures.

Pre-train Mixture	Pick/ Place	Doors/ Drawers	**Other**	**Avg**
Scratch	33.0	57.3	47.4	45.1
$\mathcal{D}_1$: OXE Only	**42.0** (+9.0)	**72.3** (+15.0)	54.2 (+6.8)	**54.7** (+9.6)
⌞ $\mathcal{D}_2$: +Real	38.5 (-3.5)	60.5 (-11.8)	47.4 (-6.8)	47.7 (-7.0)
⌞ $\mathcal{D}_3$: +Sim	36.3 (-2.2)	56.7 (-3.8)	**56.0** (+8.6)	49.6 (+1.9)
⌞ $\mathcal{D}_4$: +Joint	35.3 (-1.0)	62.3 (+5.6)	54.4 (-1.6)	50.0 (+0.4)

On RoboCasa, $\mathcal{D}_1$ also achieves the highest average success, at 54.7% (9.6 percentage points above Scratch). In $\mathcal{D}_2$, average success falls to 47.7%, with the Doors/Drawers category 11.8 percentage points below $\mathcal{D}_1$. These results show that adding heterogeneous data can reduce average success relative to OXE-only pre-training.

Real-World Results. Blinded evaluations (Figure 5) show that all pre-trained variants achieve higher average scores than training from scratch, with $\mathcal{D}_1$ scoring highest. Possible explanations for the performance gap relative to $\mathcal{D}_1$ include: (1) Kinematic Inconsistency: divergent TCP frame definitions across robots; (2) Temporal Mismatch: unaligned control frequencies (3–30Hz) creating conflicting per-step magnitudes; and (3) Semantic Drift: inconsistent instruction granularities.

Table 5: OOD and RoboTwin results.

Method	**OOD 5-shot**	**OOD 50-shot**	**Robo Twin**
Scratch	16.0	22.1	4.1
D_4 (Full Mix)	13.7	29.1	14.1
D_1 (OXE Only)	**22.8**	**46.4**	**15.2**

Generalization and Joint-Space Transfer. We evaluate generalization to unseen environments and joint-space control (Table 5). $\mathcal{D}_1$ improves success on LIBERO-Plus (Fei et al., 2025) (16.0% $\rightarrow$ 22.8%) and RoboTwin (Chen et al., 2025a) (4.1% $\rightarrow$ 15.2%). In contrast, $\mathcal{D}_4$ achieves success rates of 13.7% on LIBERO-Plus (5-shot) and 14.1% on RoboTwin, both lower than $\mathcal{D}_1$. These results show that EEF pre-training on $\mathcal{D}_1$ improves generalization to new scenarios and transfers to joint-space control tasks.

5.3 Analysis of Training Regularization

We evaluate the impact of sensory dropout ($p_{\text{state}}, p_{\text{view}}$) and multi-stage curricula on model performance. Results in Table 6 suggest that these common practices provide limited benefits at scale.

Table 6: Impact of training regularization.

Mask Ratio		Schedule	Success Rate (%)				
State	View		Spat	Obj	Goal	Long	**Avg**
0.2	0.2	2-Stage	86.0	90.0	88.6	73.4	84.5
0	0.2	2-Stage	88.2	**96.0**	87.6	69.0	85.2
0.5	0.2	2-Stage	**88.8**	90.4	86.8	70.0	84.0
0.2	0	2-Stage	87.0	92.2	**89.2**	**74.0**	85.6
0.2	0.5	2-Stage	85.8	93.2	86.6	72.2	84.5
0.2	0.2	Stage 2 Only	88.2	93.8	87.0	**74.0**	**85.8**

Explicit Regularization. Stochastic modality dropout does not consistently improve performance. Disabling visual dropout ($p_{\text{view}} = 0$) increases success to 85.6% compared to the balanced baseline (84.5%), while proprioceptive masking leads to performance drops. The inherent diversity of the large-scale pre-training corpus may already function as an effective implicit regularizer, making additional noise injection redundant or counterproductive.

Optimization Strategy. Direct end-to-end training ("Stage 2 Only") achieves 85.8% average success, compared with 84.5% for the two-stage curriculum. These results support a simpler single-stage training schedule.

5.4 Comparison with Generalist Policies

We benchmark our base model against representative generalist VLAs, including π_0 (Black et al., 2024), $\pi_{0.5}$ (Intelligence et al., 2025), and GR00T-N variants (Nvidia et al., 2025), using a 50-shot fine-tuning protocol (Table 7). On LIBERO, our model achieves 97.8% average success, exceeding both $\pi_{0.5}$ (96.9%) and GR00T-N1.6 (97.0%). On RoboCasa, our model achieves 50.0% success, exceeding $\pi_{0.5}$ (41.4%). GR00T-N1.6 reaches 66.6%, but includes RoboCasa data in pre-training, whereas our model does not. These results indicate that our testbed provides a strong and representative foundation for studying VLA scaling behaviors.

Table 7: Benchmarking results on LIBERO and RoboCasa.

Model	LIBERO					RoboCasa			
	Spat	Obj	Goal	Long	**Avg**	Pick/ Place	Doors/ Draw	Other	**Avg**
GR00T-N1	94.4	97.6	93.0	90.6	93.9	18.6	50.2	39.1	36.0
GR00T-N1.6	97.7	**98.5**	97.5	<u>94.4</u>	<u>97.0</u>	**43.2**	**84.2**	**81.3**	**66.6**
π_0	98.0	96.8	94.4	88.4	94.4	14.0	53.1	<u>58.5</u>	42.4
$\pi_{0.5}$	**98.8**	<u>98.2</u>	<u>98.0</u>	92.4	96.9	21.5	57.8	44.9	41.4
Ours (Base)	<u>98.4</u>	96.2	**98.4**	**98.2**	**97.8**	<u>35.3</u>	<u>62.3</u>	54.4	<u>50.0</u>

6 Conclusion

In this work, we present a systematic study of VLA scaling using heterogeneous robotic data. To ensure reliable evaluation, we introduce a Grouped Blind Ensemble protocol to minimize human bias in real-world experiments, alongside a few-shot fine-tuning protocol for simulation benchmarks. Our results show that broadening the pre-training corpus does not automatically improve performance; effective transfer depends on how heterogeneous robot data is represented and combined. While EEF-Relative actions facilitate cross-embodiment transfer, naively pooling heterogeneous datasets can still degrade performance. Furthermore, sensory dropout and multi-stage training do not consistently improve performance at scale. Overall, this study suggests that data alignment, rather than raw scale alone, is the primary driver for high-performance general-purpose VLA models.

AI use statement

We use generative AI tools to assist with language editing and LaTeX formatting.

Ethics statement

This work studies VLA scaling for robotic manipulation. Advances in general-purpose robot learning may support industrial automation, assistive robotics, and household tasks, but deploying such systems may also raise concerns about labor displacement and safety. We encourage responsible deployment and further research on safety constraints for robotic systems.

Reproducibility statement

We will release the training and evaluation code to support reproducibility. We provide the data composition, training hyperparameters, evaluation protocols, and scoring criteria in the paper and appendices.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Lars Berscheid, Pascal Meißner, and Torsten Kröger. Robot learning of shifting objects for grasping in cluttered environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 612–618. IEEE, 2019.

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.

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.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Junhao Cai, Zetao Cai, Jiafei Cao, Yilun Chen, Zeyu He, Lei Jiang, Hang Li, Hengjie Li, Yang Li, Yufei Liu, et al. Internvla-a1: Unifying understanding, generation and action for robotic manipulation. *arXiv preprint arXiv:2601.02456*, 2026.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025a.

Xinyi Chen, Yilun Chen, Yanwei Fu, Ning Gao, Jiaya Jia, Weiyang Jin, Hao Li, Yao Mu, Jiangmiao Pang, Yu Qiao, et al. Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy. *arXiv preprint arXiv:2510.13778*, 2025b.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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 preprint arXiv:1910.11215*, 2019.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

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 preprint arXiv:2307.00595*, 2023.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Yicheng Feng, Yijiang Li, Wanpeng Zhang, Sipeng Zheng, Hao Luo, Zihao Yue, and Zongqing Lu. Videoorion: Tokenizing object dynamics in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20401–20412, 2025.

Luo Hao, Yue Zihao, Zhang Wanpeng, Feng Yicheng, Zheng Sipeng, Ye Deheng, and Lu Zongqing. OpenMMEgo: Enhancing egocentric understanding for LMMs with open weights and data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

Chengkai Hou, Kun Wu, Jiaming Liu, Zhengping Che, Di Wu, Fei Liao, Guangrun Li, Jingyang He, Qiuxuan Feng, Zhao Jin, et al. Robomind 2.0: A multimodal, bimanual mobile manipulation dataset for generalizable embodied intelligence. *arXiv preprint arXiv:2512.24653*, 2025.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea open-world dataset and g0 dual-system vla model. *arXiv preprint arXiv:2509.00576*, 2025.

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.

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.

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.

Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025.

Zhixuan Liang, Yizhuo Li, Tianshuo Yang, Chengyue Wu, Sitong Mao, Tian Nian, Liuao Pei, Shunbo Zhou, Xiaokang Yang, Jiangmiao Pang, et al. Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies. *arXiv preprint arXiv:2508.20072*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36: 44776–44791, 2023.

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.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, et al. Being-h0. 5: Scaling human-centric robot learning for cross-embodiment generalization. *arXiv preprint arXiv:2601.12993*, 2026.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

J Bjorck Nvidia, Fernando Castaneda, N Cherniadev, X Da, R Ding, L Fan, Y Fang, D Fox, F Hu, S Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

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: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

Robert Rosenthal. *Experimenter effects in behavioral research*. Irvington, 1976.

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.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

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.

Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, et al. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. *arXiv preprint arXiv:2511.16651*, 2025.

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024.

Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, Bowen Yang, Zhe Li, Kai Zhu, Hongyu Wu, Yiheng Liu, et al. Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025.

Jianglong Ye, Keyi Wang, Chengjing Yuan, Ruihan Yang, Yiquan Li, Jiyue Zhu, Yuzhe Qin, Xueyan Zou, and Xiaolong Wang. Dex1b: Learning with 1b demonstrations for dexterous manipulation. *arXiv preprint arXiv:2506.17198*, 2025.

Wanpeng Zhang, Yicheng Feng, Hao Luo, Yijiang Li, Zihao Yue, Sipeng Zheng, and Zongqing Lu. Unified multimodal understanding via byte-pair visual encoding. *arXiv preprint arXiv:2506.23639*, 2025a.

Wanpeng Zhang, Zilong Xie, Yicheng Feng, Yijiang Li, Xingrun Xing, Sipeng Zheng, and Zongqing Lu. From pixels to tokens: Byte-pair encoding on quantized visual modalities. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=3TnLGGHhNx.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

APPENDIX

This appendix provides extended analyses, implementation details, and supplementary experiments that support the main paper. We first present detailed statistics of the heterogeneous pre-training corpus in §A, followed by an analysis of the numerical scales of different SE(3) action representations in §B. Training and fine-tuning hyperparameters are described in §C, while real-world robotic evaluation tasks and scoring protocols are detailed in §D. Additional results on few-shot generalization and adaptation settings are provided in §E. Finally, we discuss the limitations of the current study in §F.

A DATA MIXTURE STATISTICS

Table 8 details the composition of our balanced pre-training corpus. We calculate the effective frame count N_{eff} using the raw frame count N_{raw} and the dataset-specific sampling step size S:

$$N_{\text{eff}} = \lfloor N_{\text{raw}}/S \rfloor \tag{5}$$

The final balanced dataset contains approximately 182.5 million frames. We assign larger step sizes ($S \in \{3, 4, 7, 8\}$) to the massive Agibot and InternData datasets. This effectively prevents them from dominating the gradients and keeps them balanced with the diverse but smaller Open X-Embodiment (OXE) subsets.

Table 8: Detailed statistics of the heterogeneous pre-training data mixture.

Category	**Dataset Source**	S	N_{raw}	N_{eff}
Real EEF	OXE - DROID	1	27.04M	27.04M
	OXE - Bridge	1	1.89M	1.89M
	OXE - BC-Z	1	5.47M	5.47M
	OXE - Language Table	1	7.05M	7.05M
	OXE - Fractal	1	3.79M	3.79M
	OXE - Kuka	1	2.46M	2.46M
	Subtotal (OXE)		*47.70M*	**47.70M**
	Agibot Gripper	7	249.04M	35.58M
	Agibot Dexterous	1	9.29M	9.29M
	RoboMind AgileX	1	5.10M	5.10M
	RoboMind UR	1	1.34M	1.34M
	RoboMind Franka	1	3.34M	3.34M
	RoboMind Tienkung (Gello)	1	2.99M	2.99M
	RoboMind Tienkung (Xsens)	1	3.18M	3.18M
	Subtotal		*321.98M*	**108.52M**
Sim EEF	InternData-M1 (Franka)	4	93.39M	23.35M
	InternData-A1 (Franka)	3	48.82M	16.27M
	Subtotal		*142.21M*	**39.62M**
Sim Joint	InternData-A1 (Lift2)	8	92.29M	11.54M
	InternData-A1 (Split-Aloha)	8	90.54M	11.31M
	InternData-A1 (Genie)	1	5.56M	5.56M
	Subtotal		*188.39M*	**28.41M**
Real Joint	SO-100	1	5.02M	5.02M
	InternData-A1 (Real Genie)	1	0.92M	0.92M
	Subtotal		*5.94M*	**5.94M**
Total			*658.52M*	**182.49M**

B Action Representation Scale Analysis

Table 9 presents the statistical scales of the four SE(3) representations across the pre-training corpus. The magnitude differences between paired representations (World-Rel vs. EEF-Rel, and World-Delta vs. EEF-Delta) are strictly less than 2%, indicating that their numerical scales are essentially consistent.

Table 9: Statistical scale of the action representations computed across the pre-training corpus.

	World-Rel	**EEF-Rel**	**Diff.**	**World-Delta**	**EEF-Delta**	**Diff.**
Mean $\|\|$Pos$\|\|$	0.1043	0.1032	$< 1.1\%$	0.0014	0.0014	$< 0.8\%$
Std $\|\|$Pos$\|\|$	0.2060	0.2043	$< 0.8\%$	0.0069	0.0069	$< 0.1\%$
Mean $\|\|$Rot$\|\|$	0.2479	0.2469	$< 0.4\%$	0.0030	0.0030	$< 0.1\%$
Std $\|\|$Rot$\|\|$	0.4350	0.4426	$< 1.8\%$	0.0109	0.0109	$< 0.1\%$

C Training and Evaluation Hyperparameters

Pre-training. We conduct pre-training in two stages. In the first stage, we freeze the VLM backbone and train the action expert for 40,000 steps with a learning rate of 1×10^{-4}. In the second stage, we unfreeze the full model and train for 200,000 steps with a learning rate of 2×10^{-5}. Throughout pre-training, we use a global batch size of 256 spread across 8 NVIDIA A800 GPUs. We use the AdamW optimizer with a cosine learning rate scheduler, a warmup ratio of 0.05, and a weight decay of 1×10^{-5}.

Fine-tuning. For downstream fine-tuning, we update the full model with a global batch size of 128 and a learning rate of 1×10^{-4}. We train for 30,000 steps on the LIBERO 5-shot benchmark, and for 60,000 steps on the full LIBERO benchmark and RoboCasa.

D Real-World Task Details

We evaluate policies on a Franka Panda robot with dual RGB cameras. Each task consists of 10 evaluation trials with randomized object initialization. We record performance using the following multi-stage criteria:

- **Stack Bowls (Max Score: 3)**
  Three bowls are initialized at random positions. The robot must identify, grasp, and stack them into a target plate sequentially.
  - **Point 1:** Successfully grasp the first bowl and place it into the plate.
  - **Point 2:** Grasp the second bowl and stack it inside the first.
  - **Point 3:** Grasp the third bowl and stack it inside the second.
- **Pick-to-Drawer (Max Score: 5)**
  A multi-step sequential manipulation task involving an articulated drawer and three distinct objects.
  - **Point 1:** Grasp the handle and fully open the drawer.
  - **Point 2–4:** Pick up Objects A, B, and C and place them inside the drawer (1 point each).
  - **Point 5:** Push the drawer back to a fully closed state.
- **Wipe Board (Max Score: 4)**
  This task evaluates continuous dynamic motion and contact-rich manipulation.
  - **Point 1:** Successfully grasp the cloth from the table.
  - **Point 2:** Establish contact between the cloth and the board.
  - **Point 3:** Execute a continuous wiping motion across the target area.
  - **Point 4:** Return the cloth to the designated area.

- **Water Plant (Max Score: 3)**
  Evaluates precision functional tool use and trigger actuation.
  - **Point 1:** Grasp the spray bottle handle with correct functional orientation.
  - **Point 2:** Reorient the nozzle to aim at the target plant.
  - **Point 3:** Actuate the trigger mechanism.

Table 10: LIBERO multi-shot and adaptation protocol results.

Method	1-shot	5-shot	10-shot	25-shot	50-shot	Co-training	LoRA
Scratch	27.9	55.4	58.6	78.5	81.8	59.0	5.5
D_4 (Full Mixture)	31.9	64.5	71.3	85.9	90.5	61.2	16.9
D_1 (OXE Only)	**37.8**	**72.8**	**76.6**	**87.6**	**92.2**	**72.2**	**33.7**

E Few-shot Generalization and Protocols

Table 10 reports results from separate fine-tuning and evaluation runs. In the 5-shot setting, we use a newly sampled set of five training demonstrations per task. For D_1 and D_4, we reuse the pre-trained checkpoints from the main experiments. D_1 consistently maintains a 9.1–18.0 percentage point performance lead over the *Scratch* baseline across all few-shot counts (1 to 50 shots). Notably, the full mixture (D_4) fails to surpass D_1 in any regime despite having a $4\times$ larger pre-training corpus. These trends remain invariant across different adaptation protocols, including **Co-training** and **LoRA** ($r = 16$). For Co-training, we jointly fine-tune on target tasks while replaying a 10% fraction of the OXE data per batch to mitigate catastrophic forgetting and preserve pre-trained priors.

F Limitations

Despite the systematic analysis of VLA scaling under controlled conditions, several limitations remain. First, our experiments primarily focus on single-arm end-effector control scenarios, offering insufficient coverage of more complex robotic embodiments such as bimanual coordination or mobile manipulation. Second, although we introduced the Grouped Blind Ensemble protocol to mitigate experimenter bias, the scale of our evaluation remains relatively constrained due to the significant labor and time costs inherent in real-robot testing. Due to the computational cost of large-scale pre-training, we also do not report variability across training seeds.

Furthermore, while this work identifies the phenomenon where mixing heterogeneous data leads to negative transfer, the analysis of its root causes—including kinematic inconsistency, temporal mismatches in control frequencies, and semantic drift—remains largely qualitative and lacks fine-grained quantitative attribution. Future research could further explore targeted data alignment methodologies, such as more granular alignment within the perception-action space, to effectively translate increased data scale into reliable performance gains.