Xiaomi-Robotics-1: Scaling Vision-Language-Action Models with over 100K Hours of Real-World Trajectories

Xiaomi Robotics[1]

Abstract

We present `Xiaomi-Robotics-1`, a foundational vision-language-action (VLA) model capable of (1) following diverse language instructions to perform a wide range of mobile manipulation tasks in unseen environments out-of-the-box, and (2) efficiently adapting to novel downstream tasks with minimal fine-tuning data. We propose a two-stage training recipe consisting of pre-training and post-training. During pre-training, we imbue the model with broad and generalizable action-generation capabilities by training on over 100k hours of real-world manipulation trajectories, collected via UMI devices across a massive scale of environments and tasks. Crucially, we develop a scalable auto-labeling pipeline that annotates trajectory clips with natural languages describing scene state transitions, providing rich and precise conditioning for action learning. During post-training, we aim to align these capabilities with robot embodiments and imperative task instructions that humans naturally use to prompt robots, effectively mapping descriptive state transition understanding into actionable task prompts. Extensive experiments demonstrate strong scaling behavior. `Xiaomi-Robotics-1` consistently improves with increased data scales and model sizes during pre-training. This scaling behavior directly transfers to post-training, where a stronger pre-training model yields better out-of-the-box performance in real-robot evaluations within unseen environments. Furthermore, `Xiaomi-Robotics-1` serves as a strong robot foundation policy that can be efficiently fine-tuned on complex, dexterous tasks with high data efficiency. Across multiple simulation benchmarks, `Xiaomi-Robotics-1` outperforms state-of-the-art methods. Notably, it establishes a new state-of-the-art with a 57.6% success rate on RoboCasa365, surpassing the previous best of 46.6%. Furthermore, it achieves an average score of 20.07 on RoboDojo, significantly outperforming the prior state-of-the-art (13.07). Code and model checkpoints will be released. Project page: https://robotics.xiaomi.com/xiaomi-robotics-1.html

1 Introduction

The remarkable capabilities of modern large models are fundamentally driven by scale, where massive and diverse training corpora have underpinned unprecedented leaps in performance for both large language models [7, 22, 27, 45] and vision-language models [1, 14, 62, 63]. Recent work on vision-language-action (VLA) models [4, 5, 24, 25, 54, 71, 76] and world-action models (WAM) [36, 81, 83] has produced increasingly promising results in robot manipulation, with early evidence that policies become more capable and generalizable as the training data grows in scale and diversity. Following the same scaling trajectory of large models is therefore a natural and appealing direction for robotics. However, robotics is hindered by a unique bottleneck of data. The dominant data collection paradigm, real-robot teleoperation, is slow, costly, and hardware-bound, making

[1]See Contributions section for full author list. Please send correspondence to mi-robotics@xiaomi.com.

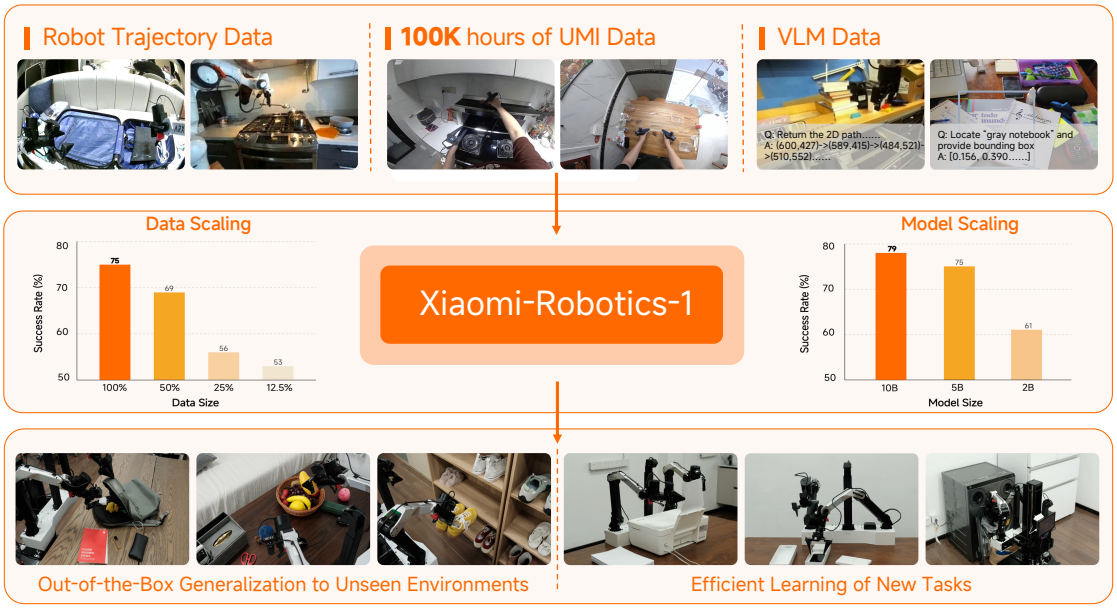


Figure 1 Overview. `Xiaomi-Robotics-1` is pre-trained on over 100k hours of real-world UMI trajectories with auto-labeled state-transition language prompts. It is then aligned to robot embodiments and imperative instruction prompts via cross-embodiment post-training. `Xiaomi-Robotics-1` scales effectively with data and model size. It is able to perform multiple tasks in unseen environment out-of-the-box and learn new tasks efficiently.

it difficult to scale. Furthermore, teleoperated data tends to be highly redundant, concentrated on a narrow slice of tasks and environments, limiting the diversity of the data.

We present `Xiaomi-Robotics-1` (Fig. 1), a foundational vision-language-action (VLA) model trained on a massive scale of real-world manipulation trajectories. Drawing inspiration from the training paradigms of large language models, we propose a two-stage training recipe comprising pre-training and post-training. During pre-training, we endow the model with robust and generalizable action-generation capabilities by leveraging data sources that scale readily in both volume and diversity. Specifically, we curate a dataset of over 100k hours of real-world manipulation trajectories with UMI devices [17], spanning a wide range of environments and tasks. Traditional trajectory labeling typically requires manual segmentation by task semantics and language annotations—a labor-intensive process that becomes prohibitive at this scale. To address this challenge, we develop a scalable auto-labeling pipeline that leverages a pre-trained vision-language model (VLM) [70] to annotate fixed-length trajectory segments with language descriptions detailing scene state transitions. These annotations provide precise and sufficient semantic supervision. Trained on these data, the model learns to generate actions that transform the scene from its current state to the language-specified target state (Fig. 6). In the post-training phase, we utilize over 10k hours of cross-embodiment data to align the strong action-generation capabilities acquired during pre-training. This stage bridges two gaps: adapting the model from generating actions for UMI grippers to actions for robot embodiments, and transitioning from state-transition prompts to imperative instructions typically used by humans to prompt robots. After post-training, `Xiaomi-Robotics-1` is able to follow instructions and perform a wide range of tasks in unseen environments. Furthermore, it serves as a strong robot foundation policy that can be efficiently fine-tuned to learn new tasks.

We perform extensive experiments to study the scaling properties of `Xiaomi-Robotics-1`. Results show that `Xiaomi-Robotics-1` scales effectively during the pre-training phase, achieving lower validation action errors as data and model scale up. Moreover, the scaling behavior observed in pre-training directly transfers to post-training, where stronger pre-training models yield better post-training success rates in out-of-the-box real-robot evaluation in unseen environments. These results are encouraging, as they indicate that we are able to continue improving performance as we further scale data and model size. In addition, we fine-tune the model

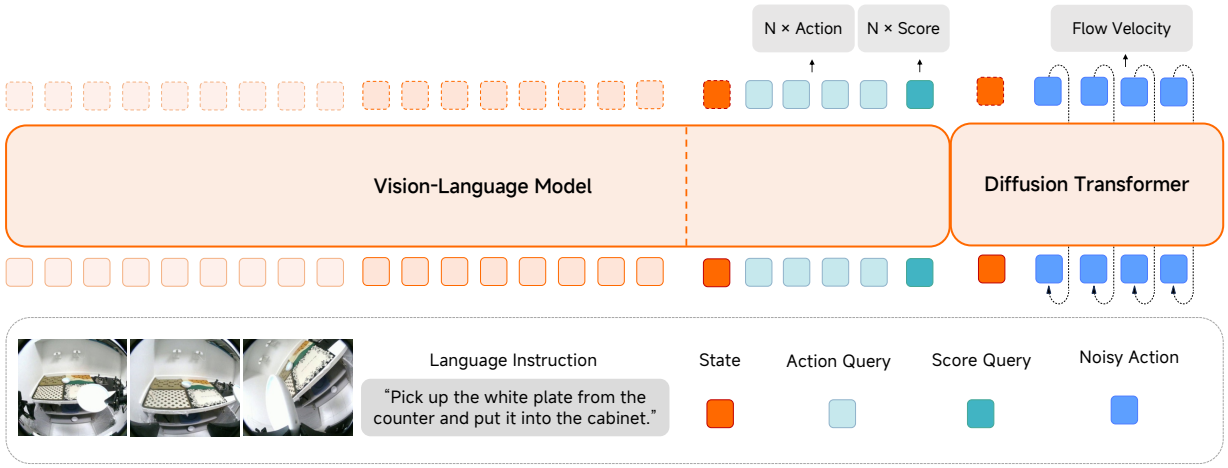


Figure 2 Model Architecture. `Xiaomi-Robotics-1` adopts a Mixture-of-Transformers [44] architecture that couples a pre-trained VLM with a DiT. The VLM encodes the observation and language instruction, and additionally predicts action chunks via Choice Policies [59] to accelerate training convergence. Conditioned on the robot state and the VLM's KV cache of the observation and language tokens, the DiT generates the action chunk via flow matching. Note that the action-related tokens from the VLM are excluded from the DiT's attention computation.

on multiple complex dexterous tasks with minimal data. `Xiaomi-Robotics-1` achieves an average success rate of 75% across four challenging tasks given less than 10 hours of data per task on average, outperforming $\pi_{0.5}$ [5] which obtains 40%. In addition, we evaluate `Xiaomi-Robotics-1` on four challenging simulation benchmarks, *i.e.*, RoboCasa [52], RoboCasa365 [53], VLABench [87], and RoboDojo [12]. `Xiaomi-Robotics-1` achieves state-of-the-art results across all four benchmarks. Notably, it sets a new state-of-the-art with a 57.6% success rate on RoboCasa365, a substantial leap from the previous best of 46.6%. On RoboDojo, it delivers an average score of 20.07, significantly outperforming the prior state-of-the-art of 13.07. Finally, `Xiaomi-Robotics-1` enables the robot to autonomously accomplish a long-horizon, room-level mobile manipulation task of suitcase packing that spans over 10 minutes (see the project page for the video). Code and model checkpoints will be released. Project page: https://robotics.xiaomi.com/xiaomi-robotics-1.html

2 Xiaomi-Robotics-1

`Xiaomi-Robotics-1` is an end-to-end vision-language-action (VLA) model trained at scale on heterogeneous data sources, including UMI trajectories, cross-embodiment robot trajectories, and vision-language data. Given an observation $\mathbf{o}_t$ and a language instruction l, the model π_θ is trained to predict an action chunk $\mathbf{a}_{t:t+H}$ by maximizing the log-likelihood over the training dataset $\mathcal{D}$:

$$\max_\theta \mathbb{E}_{(\mathbf{o}_t, l, \mathbf{a}_{t:t+H}) \sim \mathcal{D}} \log \pi_\theta(\mathbf{a}_{t:t+H} \mid \mathbf{o}_t, l)$$

We adopt a two-stage training recipe consisting of pre-training and post-training. Pre-training leverages a scalable non-robot dataset with rich open-world diversity to endow the model with broad and generalizable representations for action generation. Post-training then aligns these representations to robot embodiments and instruction-conditioned action generation, using a high-quality set of cross-embodiment data. In the following sections, we describe the details of model architecture, data curation, and training recipe.

2.1 Model

As illustrated in Fig. 2, `Xiaomi-Robotics-1` adopts a Mixture-of-Transformers (MoT) [44] architecture consisting of a pre-trained vision-language model (VLM) (*i.e.*, Qwen3-VL [3]) and a diffusion transformer (DiT) [57]. The DiT matches the VLM in the number of layers but employs a smaller hidden size for faster inference speed. The model parameters for different scaling variants of `Xiaomi-Robotics-1` are detailed in

Table 1 Model configurations for different scaling variants of `Xiaomi-Robotics-1`.

Model	# Layers	VLM		DiT		Total Params
		Hidden Size	Params	Hidden Size	Params	
`Xiaomi-Robotics-1-2B`	28	2048	2.1B	1024	470M	2.6B
`Xiaomi-Robotics-1-5B`	36	2560	4.4B	1024	604M	5.1B
`Xiaomi-Robotics-1-10B`	36	4096	8.8B	2048	1.5B	10.5B

Tab. 1. The VLM takes the current observation $\mathbf{o}_t$ and language instruction l as inputs. Conditioned on the robot proprioceptive state $\mathbf{s}_t$ and the KV cache produced by the VLM, the DiT generates the action chunk via flow-matching [49]:

$$L_{\text{Flow}}(\theta) = ||\mathbf{v}_\theta(\mathbf{o}_t, l, \mathbf{s}_t, \tilde{\mathbf{a}}^\tau_{t:t+H}, \tau) - \mathbf{u}(\tilde{\mathbf{a}}^\tau_{t:t+H}, \mathbf{a}_{t:t+H}, \tau)||_2^2$$

τ is the flow-matching timestep. $\tilde{\mathbf{a}}^\tau_{t:t+H} = \tau \mathbf{a}_{t:t+H} + (1-\tau)\boldsymbol{\epsilon}$ is the noisy action where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Following [4], we sample timestep τ from a Beta distribution, placing more weight on noisier timesteps during training:

$$u \sim \text{Beta}(1.5, 1), \quad \tau = (1-u) * 0.999 \in [0, 0.999]$$

Similar to [8], we leverage adaptive normalization layers (adaLN) [57] to inject the flow-matching timestep condition into the DiT for action generation. During inference, we initialize the predicted action chunk from a random noise $\mathbf{a}^{\tau=0}_{t:t+H} \sim \mathcal{N}(\mathbf{0}, I)$. The clean action chunk is recovered via a 5-step Euler integration, $\mathbf{a}^{\tau+\Delta\tau}_{t:t+H} = \mathbf{a}^{\tau}_{t:t+H} + \Delta\tau \cdot \mathbf{v}_\theta(\mathbf{o}_t, l, \mathbf{s}_t, \mathbf{a}^\tau_{t:t+H}, \tau)$, where the step size is set to $\Delta\tau = 0.2$.

To accelerate convergence [58], we introduce an auxiliary action-generation supervision on the VLM. Specifically, we leverage Choice Policies [59] to enable action generation directly within the VLM framework [8]. We encode the robot state into a token using a multi-layer perceptron (MLP) and append it, along with the action and score query tokens, to the end of the vision-language token sequence. The outputs corresponding to the action and score query tokens predict K candidate action chunks and their associated K scores, respectively. We adopt a winner-takes-all paradigm as in [59], where only the candidate with the smallest L_1 loss is included in action loss computation:

$$L_{\text{Regression}}(\theta) = ||\hat{\mathbf{a}}^*_{t:t+H} - \mathbf{a}_{t:t+H}||_1 + \sum_k^K ||\hat{s}_k - s_k||_2^2$$

Let $\hat{\mathbf{a}}^k_{t:t+H}$ denotes the k-th predicted candidate action chunk, then $\hat{\mathbf{a}}^*_{t:t+H}$ is the one with the smallest L_1 distance to the ground truth $\mathbf{a}_{t:t+H}$. $\hat{s}_k$ is the predicted score for the k-th candidate, and its regression target s_k is defined as $s_k = ||\hat{\mathbf{a}}^k_{t:t+H} - \mathbf{a}_{t:t+H}||_1$. That is, the L_1 distances between the K predicted action chunks and the ground-truth action chunk serve as the target labels for score prediction.

Applying action-generation supervision directly on the VLM steers its representations toward features that better support action generation, thereby making the DiT learning more effective. However, we empirically observe that letting the DiT tokens attend to the KV cache of these action-related tokens degrades performance. We hypothesize that this arises from a shortcut in which the DiT simply copies the actions generated by the VLM rather than effectively grounding its own generation in the visual and textual context. To mitigate this issue, we exclude these action-related tokens from the DiT's attention computation, constraining the DiT tokens to attend solely to the representations of the language instruction and visual observations.

2.2 Training & Data

2.2.1 Pre-training

During pre-training, our primary objective is to endow the model with broad and generalizable representations that transfer across diverse manipulation scenarios. To this end, we curate a dataset of over 100,000 hours of real-world manipulation trajectories, captured with Universal Manipulation Interface (UMI) handheld

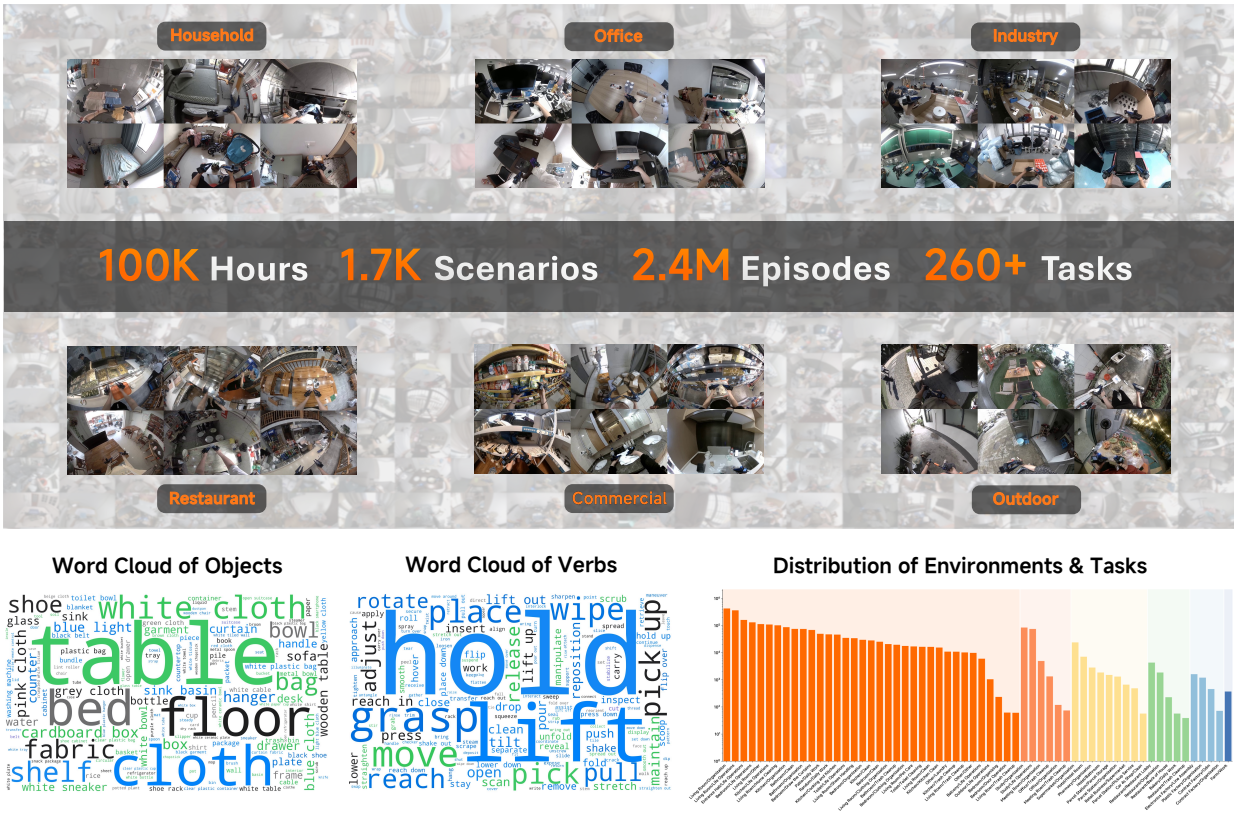


Figure 3 Pre-training Dataset. The pre-training dataset of `Xiaomi-Robotics-1` contains over 100k hours of real-world manipulation trajectories collected with UMI devices.

grippers [17] and egocentric cameras (Fig. 3). The dataset spans a diverse array of tasks collected across a massive scale of environments, including households, commercial premises, industrial sites, offices, and outdoor spaces. Traditional robot trajectory annotation requires manually segmenting trajectories according to task semantics and labeling each segment with a language instruction—a labor-intensive process that becomes prohibitive at this scale.

To scale language annotation, we develop an auto-labeling pipeline that first divides each trajectory into equal-length segments and leverage Qwen3.5-27B [70] to caption the state transitions of both the grippers and the interacting objects in the scene within each segment (see Fig. 11 for examples). To accelerate the annotation process, we develop a producer–consumer pipeline that decouples clip segmentations from caption labeling: while CPU worker threads cut per-segment clips into an in-memory filesystem, client threads keep hundreds of captioning requests in flight. This highly effcient pipeline allows us to label the entire corpus of over 100k hours in roughly two weeks. Trained on this dataset, the model learns to generate actions that drive the scene from the state in the current observation to the target state described by the language annotation.

The model is optimized to predict actions by jointly minimizing the flow-matching loss L_{Flow} of the DiT and the regression loss $L_{\text{Regression}}$ of the VLM choice policy. To preserve the vision-language capabilities of the pre-trained VLM, we further co-train the model on a high-quality vision-language dataset curated in our previous work [8] under the next-token prediction objective L_{NTP}. The overall training objective is formulated as:

$$L = L_{\text{Flow}} + L_{\text{Regression}} + \lambda L_{\text{NTP}} \tag{1}$$

where λ is set to 0.1 in our experiments. Vision-language data and UMI trajectories are sampled at a ratio of 1:9. To maximize training throughput, we pack all vision-language tokens within a batch into a single sequence for a VLM forward pass. Since the VLM is computationally more expensive than the DiT, we

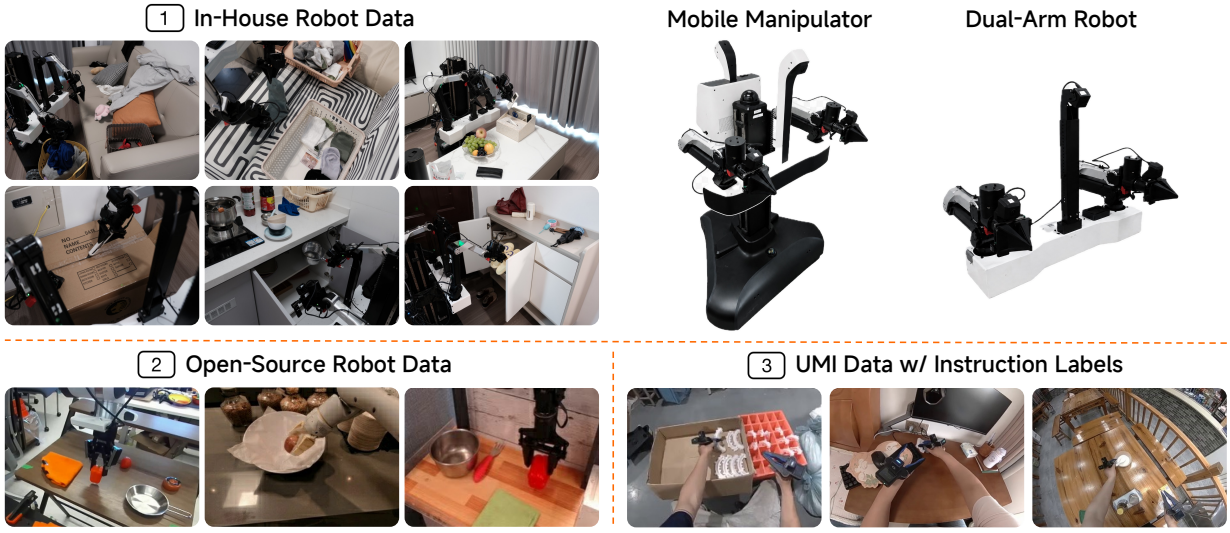


Figure 4 Post-training Dataset. The post-training dataset of `Xiaomi-Robotics-1` comprises about 10k hours of cross-embodiment trajectories, including over 7.2k hours of in-house robot data collected with mobile manipulators and dual-arm robots, over 1k hours of instruction-labeled UMI data, and open-source robot datasets.

amortize its cost by sampling four flow-matching timesteps per sample. The resulting four DiT inputs are similarly packed and processed in one DiT pass, conditioned on the corresponding unpacked VLM KV cache.

2.2.2 Post-training

The goal of post-training is twofold. First, we transfer the action-generation capabilities of UMI grippers acquired during pre-training to robot embodiments. Second, we shift the language conditioning from the state-transition descriptions used in pre-training to the imperative instructions humans typically issue when prompting robots to perform tasks.

We curate the post-training dataset with cross-embodiment manipulation trajectories collected using UMI devices, static robot arms, and mobile manipulators. Specifically, we collect over 7,200 hours of robot data using mobile manipulators and dual-arm robots across a diverse range of household environments and tasks (Fig. 4). We leverage Qwen3.5 [70] to annotate human-segmented video clips with language instructions. In addition, we incorporate over 1,000 hours of human-annotated UMI data labeled with both temporal segments and language instructions. Unlike the state-transition descriptions used in pre-training, these language instructions closely mirror how humans prompt robots to perform tasks, directly matching our alignment objective in the post-training phase (see Fig. 11 and 12 for comparison). Finally, we include open-source robot datasets, including Bridge V2 [74], RT-1 [6], and DROID [28]. We filter out idle segments within trajectories to prevent the model from learning uninformative or noisy signals. In total, our post-training dataset comprises about 10,000 hours of trajectory data.

For arm actions, we adopt relative delta end-effector (EE) poses with respect to the current state:

$$\mathbf{a}_{t+i} = ({}^{\text{EE}}_{\text{Base}}T_t)^{-1} \, {}^{\text{EE}}_{\text{Base}}\hat{T}_{t+i}$$

where ${}^{\text{EE}}_{\text{Base}}T_t$ denotes the pose of the end-effector with respect to the base at the current timestep t, and ${}^{\text{EE}}_{\text{Base}}\hat{T}_{t+i}$ represents the target end-effector pose at timestep $t+i$. To align the arm action spaces across different embodiments, we unify the orientation of the end-effector frames across all robot data and UMI data in both the pre-training and post-training datasets. Consequently, similar arm motions (*e.g.*, moving forward or backward with respect to the end-effector frame) yield consistent action values regardless of the underlying hardware platform. For mobile robot data, we represent the base and waist actions using the base velocity and the relative delta of the waist position, respectively. To accommodate heterogeneous embodiments, we adopt

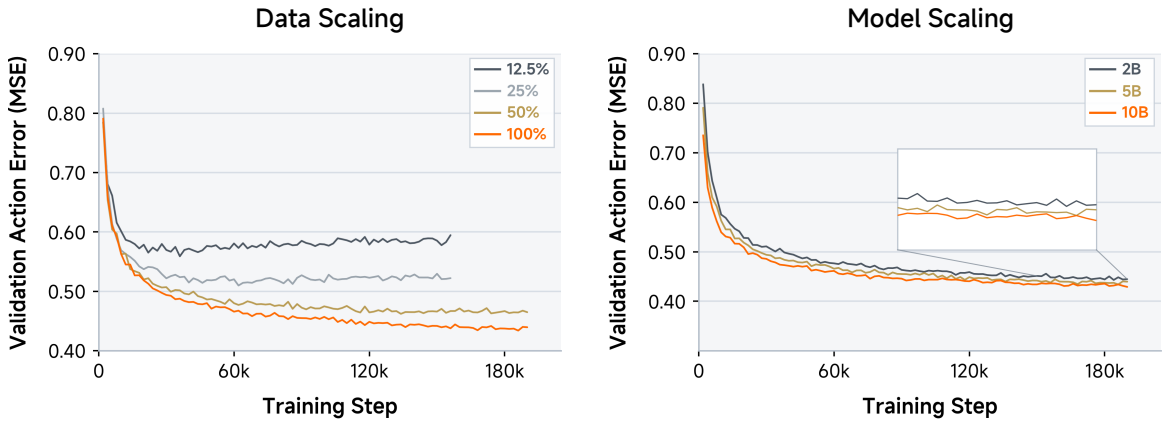


Figure 5 Scaling of Pre-training. We show the validation action errors (MSE) from the data-scaling and model-scaling pre-training experiments. We terminate the training for 12.5% and 25% data in the data-scaling experiment early as the validation loss indicates overfitting.

a unified action vector for all trajectory data. Although the arm actions are aligned across embodiments, the action spaces of different robots still differ in dimensionality. We mask out the dimensions corresponding to missing action components during loss computation.

We train the model with the same objective as in pre-training (Eq. 1). Vision-language data, open-source robot data, instruction-labeled UMI data, and our in-house robot data are sampled at a ratio of 0.5:0.5:0.5:8.5. After post-training, the model can be prompted with language instructions to perform a wide range of tasks in unseen environments out-of-the-box. In addition, it can efficiently adapt to novel downstream tasks with minimal amount of data.

3 Experiments

We design `Xiaomi-Robotics-1` with *scaling* in mind. In this section, we investigate its scaling properties through extensive experiments. Specifically, we design experiments to answer the following questions:

- Does `Xiaomi-Robotics-1` scale effectively with increasing data scale and model size during pre-training?
- Does a stronger pre-trained model translate to better post-training performance when evaluated out-of-the-box in novel environments?
- Can `Xiaomi-Robotics-1` adapt to challenging new tasks with a minimal amount of data?
- How does `Xiaomi-Robotics-1` compare to other robot foundation models in real-robot experiments and simulation benchmarks?

3.1 Pre-training: Data and Model Scaling

Data Scaling. We perform data-scaling experiments with `Xiaomi-Robotics-1`-5B. Due to compute budget limit, we pre-train the model on 12.5%, 25%, 50%, and 100% of about 20k hours of UMI data, respectively. Each model is evaluated on a held-out validation set. We use the mean-squared error (MSE) between the action predicted by flow-matching and the ground truth as the evaluation metric. As shown in Fig. 5, `Xiaomi-Robotics-1` attains lower validation action errors with the increase of data scale. With 12.5% and 25% of data, the validation action errors first decrease and then increase during training, indicating overfitting. In contrast, the 50% and 100% data settings yield a monotonic decrease in loss, with the 20k setting exhibiting a steeper descent. We show qualitative results of action prediction on validation data in Fig. 6.

Model Scaling. We perform model-scaling experiments on three size variants of `Xiaomi-Robotics-1` (2B, 5B, and 10B) as specified in Tab. 1. All three models are trained on the same 20k hours of data as in the

Figure 6 Qualitative Results of Pre-training. After pre-training, `Xiaomi-Robotics-1` is able to predict action trajectories for UMI grippers on a held-out validation set according to the language description of state transitions.

data-scaling experiments and then evaluated on the same held-out validation set. As illustrated in Fig. 5, `Xiaomi-Robotics-1` exhibits consistent improvements in action prediction precision as the model size scales up. However, the performance gap among different model sizes are less pronounced than those observed across different data scales. This result suggests that model capacity at the billions-parameter scale may already be sufficient to capture the current dataset's distribution, thereby making data volume the primary bottleneck for further generalization. These findings do not diminish the value of model scaling, but rather highlight the critical importance of prioritizing the collection of large-scale, diverse datasets to unlock further performance gains.

3.2 Post-training: Out-of-the-Box Evaluation in Novel Environments

In this section, we perform post-training experiments on the cross-embodiment post-training dataset and study its out-of-the-box performance in novel environments that are unseen during training. In particular, we are interested in understanding whether the data scaling and model scaling properties from pre-training can transfer to post-training. To mitigate overfitting, for the in-house robot data, we sample a diverse subset from the whole dataset for post-training. Models are evaluated out-of-the-box in unseen environments after post-training without any per-task or per-environment fine-tuning. Specifically, we evaluate on 4 tasks (Fig. 7), *i.e.*, shoe storage, bag packing, table organization, sofa tidying. These tasks are seen in the post-training dataset but the environments and object instances during evaluation are unseen.

3.2.1 Effectiveness of Scaling Pre-training Data

We first examine whether the benefits of scaling pre-training data transfer to post-training with the 5B variant of `Xiaomi-Robotics-1`. Using an identical training recipe, we post-train models initialized from checkpoints pre-trained on 12.5%, 25%, 50%, and 100% of 20k pre-training data (Sec. 3.1), alongside a baseline initialized from the Qwen3-VL [3] pre-trained weight without any action pre-training. Out-of-the-box evaluation results are shown in Fig. 8. The overall success rate increases monotonically with the scale of pre-training data, rising from 26% without action pre-training to 75% with 100% of pre-training data. The gains from scaling pre-training data are particular pronounced on tasks that demand contact-rich manipulation. For instance, the baseline without pre-training fails completely on shoe tidying, whereas the model pre-trained on 100% of the data reaches a 75% success rate. Notably, utilizing only 12.5% of the pre-training data more than doubles the baseline's overall success rate (26% vs. 53%). While the marginal gains gradually moderate as the pre-training corpus grows, the performance shows no sign of saturation: doubling the data from 50% to

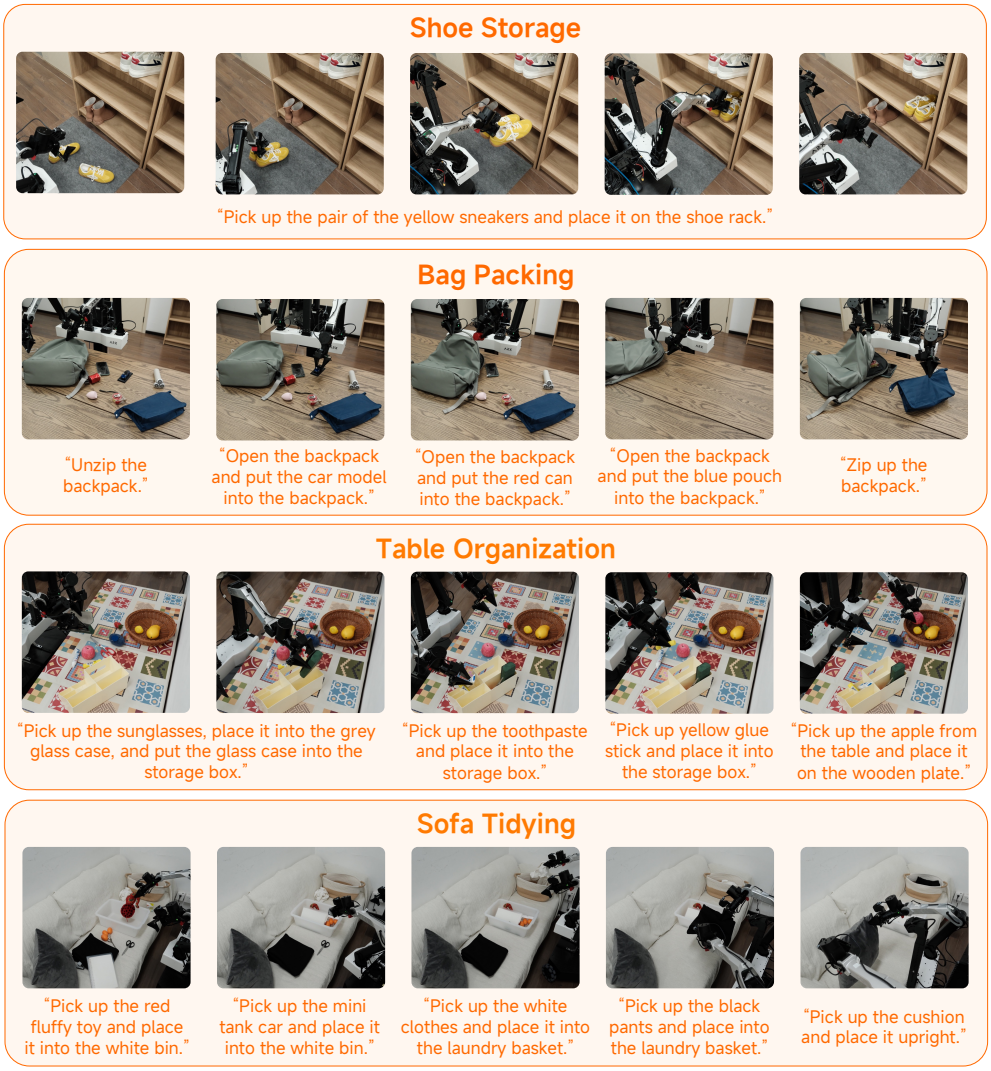


Figure 7 Post-training Evaluation. We evaluate the post-trained model out-of-the-box across four tasks in novel environments. Crucially, both the environments and object instances are unseen during training.

100% yields an additional 6 percentage point improvement. These findings suggest that further scaling robot pre-training data remains a highly promising avenue for achieving stronger out-of-the-box performance in unseen environments.

3.2.2 Effectiveness of Scaling Model Size

We further investigate the impact of model scale during post-training. Specifically, We post-train the three size variants of `Xiaomi-Robotics-1` (2B, 5B, and 10B) specified in Tab. 1. These variants are initialized from checkpoints pre-trained on 20k hours of UMI pre-training data. As shown in Fig. 8, the overall success rate increases monotonically with model size, rising from 61% for the 2B variant to 75% and 79% for the 5B and 10B variants, respectively. Similar to data scaling, the gains from model scaling are most pronounced on shoe tidying, where the success rate climbs from 58% (2B) to 75% (5B) and further to 92% (10B). Performance improves consistently with model size across three out of four tasks, with the 5B and 10B variants performing comparably (80% and 77%) on sofa tidying. Combined with the results in Sec. 3.2.1, these findings suggest that pre-training data scale and model size constitute two complementary axes for improving out-of-the-box performance in out-of-distribution settings. And a stronger pre-trained model is able to translate to better

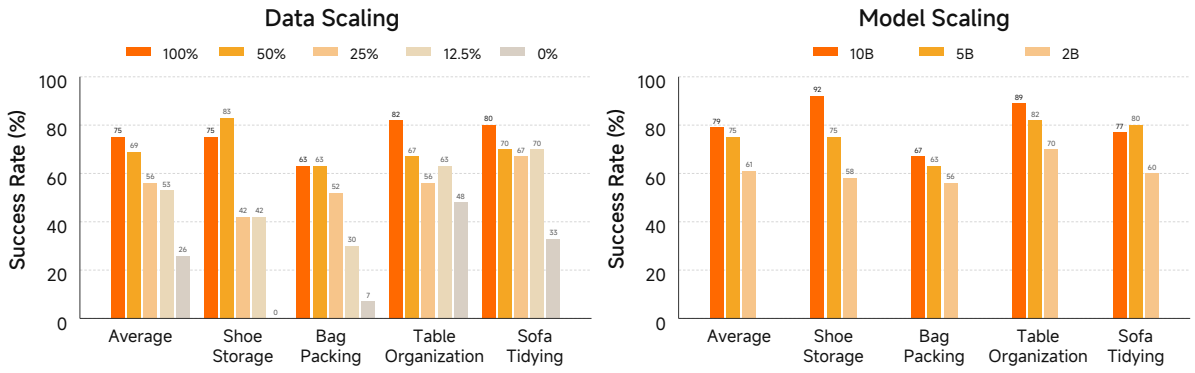


Figure 8 Quantitative Results of Post-training. We showcase the success rates of post-trained models across different pre-training data scales and model sizes.

out-of-the-box real-robot performance after post-training.

3.3 Downstream Fine-tuning: Efficient Adaptation to New Tasks

A key desideratum of robot foundation models is their ability to efficiently adapt to novel tasks with minimal data. To investigate how `Xiaomi-Robotics-1` performs in this setting, we fine-tune our post-trained model on a suite of four novel challenging tasks: phone packing, laundry loading, printer refilling, and box packing (Fig. 9). Crucially, these tasks are entirely held out from the in-house robot dataset. Each task introduces different complexities: phone packing requires bimanual coordination; laundry loading is a long-horizon mobile manipulation task involving multi-step instruction following; printer refilling demands handling of highly deformable sheets of paper; and box packing evaluates language grounding across multiple objects.

To evaluate data efficiency, we fine-tune the model under two settings. For the high-data setting, we leverage a total of 144 hours of data across all tasks. For the low-data setting, we sample 25% of the data for each task from the high-data setting, resulting in a subset of 36 hours in total. The average data per task for the low-data setting is less than 10 hours, with the maximum-data task, printer refilling, containing only 10.3 hours. This poses a significant challenge for policy learning. We fine-tune `Xiaomi-Robotics-1` on the data of all four tasks using the asynchronous training recipe proposed in [8]. We compare our method against two baselines: $\pi_{0.5}$ [5] and Xiaomi-Robotics-0 [8]. For $\pi_{0.5}$, we follow the official OpenPi[1] fine-tuning protocol and fine-tune the base model on these tasks. For Xiaomi-Robotics-0 [8], we fine-tune its pre-trained model using the asynchronous setting. Each model is evaluated for 10 trials per task. We report both the average success rate and progress, where the progress measures partial task completion based on the task-specific milestones (Tab. 6). Results are shown in Fig. 10.

`Xiaomi-Robotics-1` outperforms the two baseline methods in terms of average success rates and progress in both low-data and high-data settings. In the low-data setting with less than 10 hours per task on average, it achieves an average success rate of 75% and an average progress of 90% across all four tasks, significantly outperforming $\pi_{0.5}$ with a 40% success rate and a progress of 66%. The advantage of our method is most substantial on tasks requiring dexterous manipulation and mobile manipulation. In phone packing, `Xiaomi-Robotics-1` outperforms both baseline methods substantially in the two data settings. In printer refilling, `Xiaomi-Robotics-1` improves the success rate of the best baseline from 20% to 70% in the low-data setting, showcasing powerful capabilities in manipulating deformable objects. In laundry loading, our method shows strong robustness over the full task horizon where failure may occur at any stage, achieving an 80% success rate and a progress of 96% in the low-data setting. In this task, $\pi_{0.5}$ struggles with opening the washing machine while Xiaomi-Robotics-0 fails to complete the task in the low-data setting. In box packing, all methods perform relatively well compared to other tasks, reaching 100% success rates when more task-specific data are available. Overall, all methods benefit from increasing the amount of fine-tuning

[1] https://github.com/Physical-Intelligence/openpi

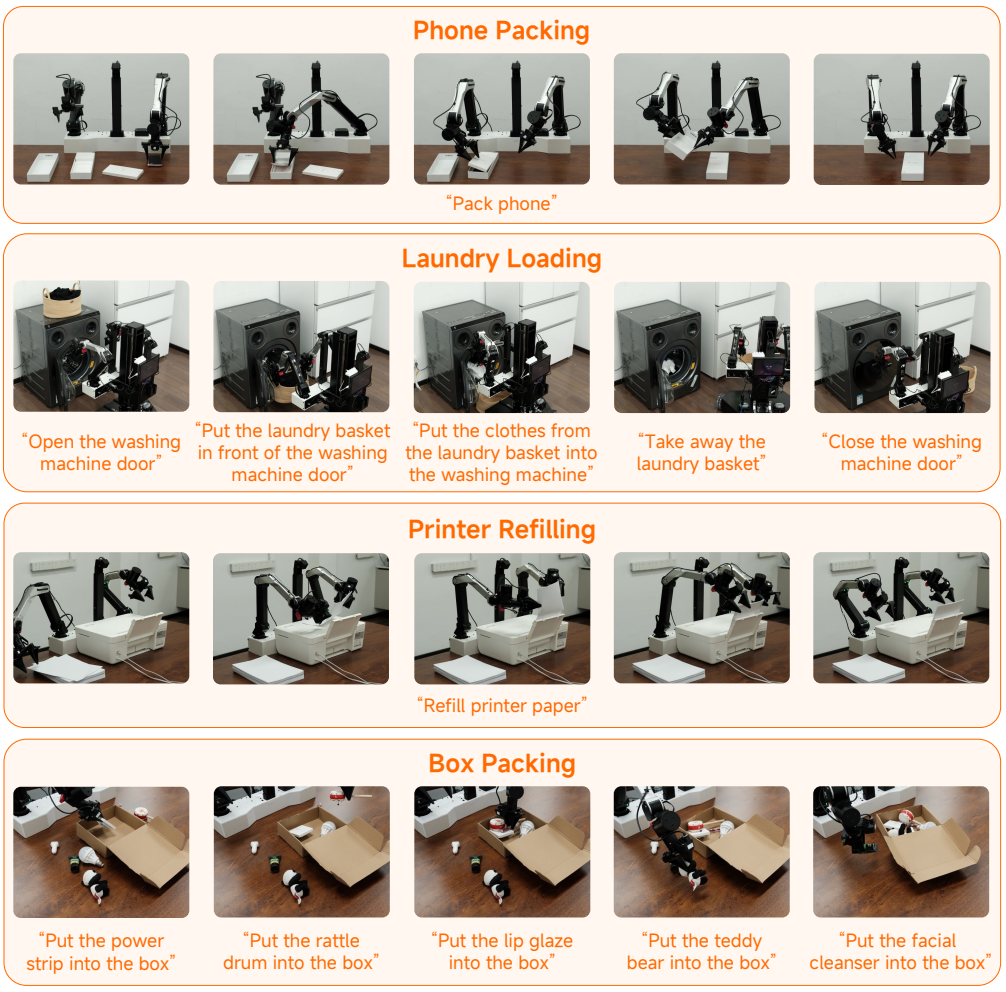


Figure 9 Downstream Fine-tuning Evaluation. We fine-tune the post-trained model on four new challenging tasks with a minimal amount of data.

data, but `Xiaomi-Robotics-1` is substantially more data-efficient. We attribute this advantage to large-scale pre-training which exposes the model to diverse environments and tasks, and careful post-training alignment that aligns the strong manipulation capabilities acquired during pre-training to robot embodiments and instruction prompts. These results demonstrate that `Xiaomi-Robotics-1` can serve as a strong foundation robot policy for efficient adaptation to novel tasks.

3.4 Simulation Benchmarks

In this section, we evaluate `Xiaomi-Robotics-1` on four simulation benchmarks.

- **RoboCasa** [52]: RoboCasa features single-arm manipulation in realistic kitchen environments. The benchmark contains 24 everyday kitchen manipulation tasks spanning pick-and-place, articulated-object interaction, appliance operation, and coffee-making. In order to test generalization capabilities, it evaluates policies on unseen object instances and includes two scenes of which the styles were unseen in the training data among the five evaluation scenes. For training, we use the official set of 300 synthetic demonstrations provided by the benchmark. Following the standard evaluation protocol, we report the average success rate over 100 evaluation episodes per task across the five evaluation scenes. Results are shown in Tab. 2.
- **RoboCasa365** [53]: RoboCasa365 extends RoboCasa into a large-scale simulation benchmark for evaluating general-purpose robot manipulation, with a particular emphasis on generalization beyond the training

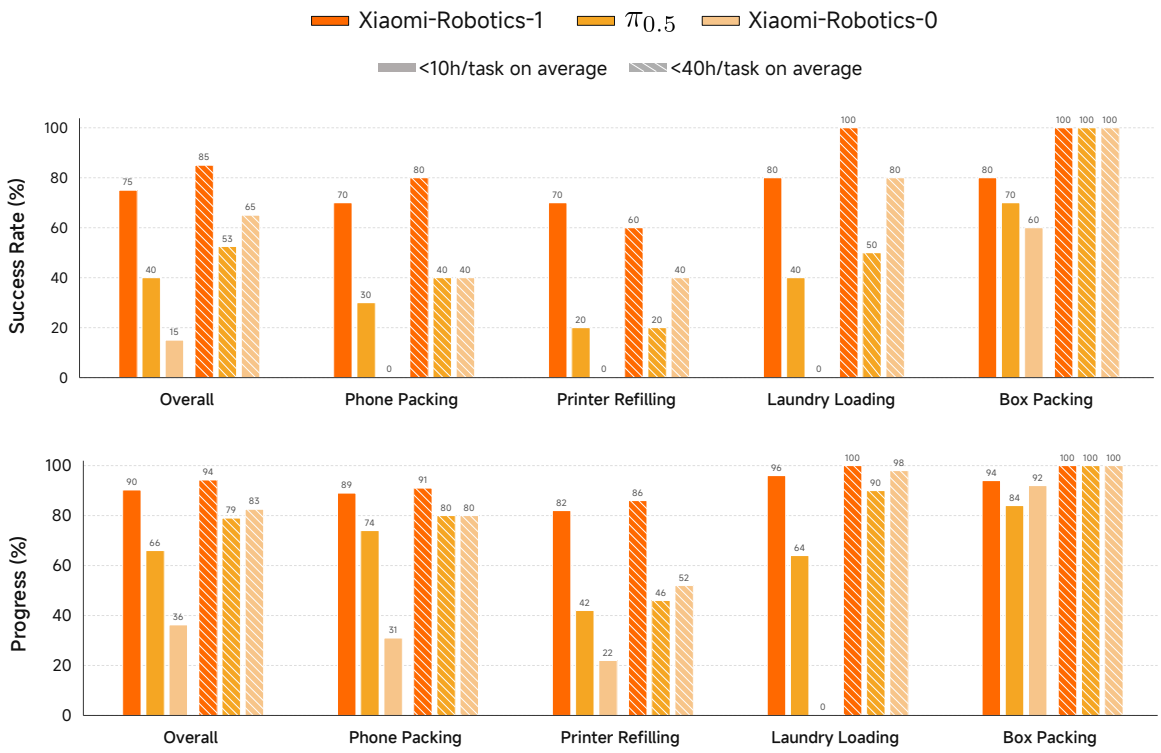


Figure 10 Quantitative Results of Downstream Fine-tuning. We report the success rates and progresses of different models across the four different tasks.

task distribution. It expands the original 24-task benchmark to 365 tasks across over 2,500 procedurally generated kitchen scenes and 3,200 object instances, covering both short-horizon manipulation and long-horizon mobile manipulation. This diversity introduces substantial variations in kitchen layouts, object appearances, and spatial relationships, testing whether policies remain robust across unseen scene and object configurations rather than overfitting. More importantly, RoboCasa365 explicitly evaluates generalization on task composition via a split featuring task templates that are unseen during training. We train our policy using the officially released dataset, which provides 100 demonstrations for each task. We follow the standard evaluation protocol and evaluate across 50 benchmark tasks, comprising 18 seen atomic tasks, 16 seen composite tasks, and 16 unseen composite tasks. The unseen composite tasks provide a zero-shot evaluation of whether the policy can recombine previously learned atomic skills and semantic knowledge to solve novel long-horizon tasks. Results are shown in Tab. 3.

- **VLABench** [87]: VLABench is a large-scale benchmark designed for comprehensive evaluation of language-conditioned manipulation. It comprises 100 task categories with over 2,000 object instances, emphasizing challenging scenarios involving semantic and distribution shifts. Specifically, VLABench introduces instructions with implicit human intentions, long-horizon tasks requiring multi-step reasoning, and evaluation settings that demand commonsense reasoning, category-level generalization, and robustness to unseen object appearances. These challenges are distributed across five evaluation tracks—In-distribution, Cross-Category, Commonsense, Instruction, and Texture—to thoroughly assess policy generalization. Beyond success rate (SR), VLABench introduces progress score (PS) and intention score (IS) to measure task completion quality and instruction understanding. Following the standard evaluation protocol, we train our policy using only the official training set from the In-distribution track, which contains 10 tasks with 500 demonstrations per task. We leverage chain-of-thought (CoT) labeling as in ERVLA [61] and train our model with a 50% probability on the next-token-prediction loss of CoT alongside the action loss. During evaluation, each task is tested across all five tracks with 50 evaluation episodes per track, resulting

Table 2 Results on the RoboCasa Benchmark. We report the average success rate (%). The best and second-best results are highlighted in **bold** and underline, respectively.

Method	Avg. Success (%)
UVA [41]	50.0
UWM [96]	60.8
$\pi_{0.5}$ [5]	62.1
π_0-FAST [58]	63.6
GR00T N1.6 [54]	66.2
Cosmos Policy [32]	67.1
RLDX-1 [29]	70.6
World2Act [73]	72.6
Xiaomi-Robotics-1 (Ours)	**74.5**

Table 3 Results on the RoboCasa365 benchmark. We report task success rates (%). The best and second-best results are highlighted in **bold** and underline, respectively.

Method	Average	Atomic	Comp.-Seen	Comp.-Unseen
Diffusion Policy [16]	6.1	15.7	0.2	1.3
$\pi_{0.5}$ [5]	16.9	39.6	7.1	1.2
GigaWorld-Policy 0.1 [79]	20.7	44.4	11.8	2.9
GR00T-N1.6 [54]	21.9	51.1	9.4	1.7
WorldDreamer [75]	35.3	66.3	26.7	9.0
Qwen-RobotManip [71]	35.9	68.6	20.1	14.9
RLDX-1 [29]	36.0	67.6	27.9	8.5
ABot-M0.5 [11]	40.4	75.9	38.3	2.7
ABot-M0.6 [11]	46.6	79.4	48.3	7.9
Xiaomi-Robotics-1 (Ours)	**57.4**	**80.2**	**57.1**	**32.1**

in 250 rollouts for each task and 2,500 rollouts in total. Results are shown in Tab. 4.

- **RoboDojo** [12]: RoboDojo is a unified simulation and real-world benchmark for comprehensively evaluating generalist robot manipulation policies. Unlike existing benchmarks that focus primarily on individual skills, RoboDojo provides a diverse and challenging suite comprising over 42 simulation tasks with varied object configurations, scene layouts, and language objectives. These tasks are organized into five core capability dimensions: Generalization (robustness to object/layout changes), Precision (fine-grained control), Long-Horizon (multi-step execution), Memory (state-dependent manipulation), and Open (open-ended instruction following). Following the official protocol, we evaluate `Xiaomi-Robotics-1` on the RoboDojo simulation benchmark and report the average score and success rate of each capability dimensions. Results are shown in Tab. 5.

`Xiaomi-Robotics-1` achieves state-of-the-art results across all the four challenging benchmarks. It achieves an average success rate of 74.5% in RoboCasa. In RoboCasa365, it significantly outperforms the previous best methods by 10.8 percentage points, obtaining an average success rate of 57.4%. Specifically, it delivers the largest improvement on the most challenging split of *Composite-Unseen*, showcasing powerful generalization capabilities to task compositions that were unseen during training. In VLABench, `Xiaomi-Robotics-1` achieves the highest average success rate and progress score, while maintaining a competitive intention score. This highlights its strong robustness to semantic and distribution shifts as well as superior long-horizon reasoning and language-conditioned manipulation capabilities. Notably, under cross-category and texture shifts, `Xiaomi-Robotics-1` surpasses the strongest baseline by 6.0 and 15.2 percentage points in terms of success rate, respectively, demonstrating effective generalization to unseen objects and high resilience to visual appearance and background perturbations. In RoboDojo, `Xiaomi-Robotics-1` significantly outperforms existing baseline methods, achieving absolute improvements of 7% and 5.13% in terms of average score and average success rate, respectively, over the second-best method. It ranks first across four out of five

Table 4 Results on the VLABench Benchmark. SR, PS, and IS denote success rate, progress score, and intention score, respectively. The best and second-best results are highlighted in **bold** and underline, respectively.

Method	In-dist.			Cross Category			Commonsense		
	SR ↑	PS ↑	IS ↑	SR ↑	PS ↑	IS ↑	SR ↑	PS ↑	IS ↑
π_0-FAST [58]	56.2	66.8	72.4	31.0	38.2	47.8	38.0	48.6	56.8
X-VLA [92]	–	67.8	–	–	25.1	–	–	48.2	–
ACoT-VLA [94]	–	66.1	79.8	–	38.9	54.1	–	37.8	52.3
$\pi_{0.5}$ [5]	65.4	77.8	80.4	38.2	49.7	52.0	43.9	57.3	**60.0**
ERVLA [61]	69.7	81.1	**84.2**	47.0	61.0	66.4	44.0	55.0	57.2
Xiaomi-Robotics-1 (Ours)	**75.6**	**85.0**	79.8	**53.0**	**66.6**	**66.4**	**48.4**	**58.3**	58.2

Method	Instruction			Texture			Avg.		
	SR ↑	PS ↑	IS ↑	SR ↑	PS ↑	IS ↑	SR ↑	PS ↑	IS ↑
π_0-FAST [58]	35.0	45.0	59.4	39.0	49.0	56.8	39.8	49.5	58.6
X-VLA [92]	–	63.1	–	–	–	–	–	51.1	–
ACoT-VLA [94]	–	39.6	56.8	–	54.6	74.6	–	47.4	63.5
$\pi_{0.5}$ [5]	48.2	64.2	67.0	44.9	62.3	65.0	48.1	62.3	64.9
ERVLA [61]	**58.0**	**70.2**	**73.8**	47.4	62.3	70.6	53.2	65.9	**70.4**
Xiaomi-Robotics-1 (Ours)	55.8	66.8	70.2	**62.6**	**74.9**	**74.8**	**59.1**	**70.3**	69.9

dimensions, indicating the model's outstanding overall performance across diverse task types. Notably, we do not incorporate history observation in the evaluation of the RoboDojo benchmark. Consequently, it scores lower on the memory dimension than Hy-Embodied-0.5-VLA [85] which explicitly models memory, yet it still significantly outperforms all other models.

4 Related Work

Scaling for Robot Learning. Research on the scaling laws of large language models (LLMs) demonstrates that performance improves predictably when data, compute, and model capacity are scaled in tandem [22, 27]. Recent advances in large language models [7, 72] and multi-modal foundation models [1–3, 63] further demonstrate substantial capability gains driven by scaling of data and model size. Motivated by these advancements, robot learning has increasingly embraced this scaling paradigm, enlarging both data and model sizes in pursuit of foundational generalist policies [4–6, 8, 31, 42, 54, 58, 65–67, 97]. Scaling robot learning, however, differs fundamentally from scaling models trained on web-scale data. Collecting real-robot trajectories requires costly and labor-intensive teleoperation, and the resulting data are often confined to a narrow slice of environments and tasks, severely constraining the data diversity. To alleviate this data bottleneck, recent work leverages portable UMI devices [17], enabling the collection of real-world manipulation trajectories without requiring physical robot embodiments [17, 46, 65, 66, 77, 90]. This lowers the barrier to data collection by facilitating in-the-wild collection across a highly diverse spectrum of unconstrained, real-world environments. Complementarily, another line of research leverages egocentric human manipulation trajectories to capture diverse behaviors across various tasks, objects, and environments, and subsequently converts them to robot data via representation alignment or motion retargeting [15, 40, 50]. In this work, we leverage over 100k hours of real-world manipulation trajectories collected from UMI devices. By varying data scale and model size, we investigate the scaling behavior of our model during pre-training and whether scaling translates to real-robot performance after post-training.

Robot Foundation Models. Recently, robot foundation models have emerged as a new paradigm for generalizable robot learning. By leveraging large-scale datasets, these models enable robust generalization across diverse environments and facilitate efficient adaptation to novel downstream tasks. World-action models (WAMs) [21,

Table 5 Results on the RoboDojo Simulation Benchmark. Each entry reports score / success rate (%). The best and second-best results are highlighted in **bold** and underline, respectively.

Method	Generalization	Precision	Long-Horizon
GalaxeaVLA (G0) [26]	4.53 / 2.83%	8.10 / 3.83%	12.60 / 5.58%
GigaWorld-Policy [79]	5.34 / 2.89%	6.15 / 1.83%	15.51 / 8.92%
StarVLA-α [80]	3.93 / 2.33%	9.90 / 4.33%	14.15 / 6.50%
Xiaomi-Robotics-0 [8]	7.43 / 5.56%	8.42 / 4.58%	13.51 / 6.92%
X-WAM [21]	7.39 / 3.33%	6.72 / 1.83%	17.47 / 9.08%
X-VLA [92]	10.48 / 6.78%	18.32 / 12.00%	16.53 / 9.75%
$\pi_{0.5}$ [5]	13.37 / 8.17%	12.40 / 5.50%	23.54 / 14.67%
Spatial Forcing [35]	14.12 / 9.33%	17.33 / 10.58%	23.26 / 14.58%
Hy-Embodied-0.5-VLA [85]	11.77 / 8.39%	13.81 / 8.00%	25.74 / 14.92%
Xiaomi-Robotics-1 (Ours)	**23.55** / **17.00**%	**26.69** / **18.83**%	**38.39** / **23.67**%

Method	Memory	Open	Average
GalaxeaVLA (G0) [26]	3.17 / 1.89%	0.70 / 0.67%	5.82 / 2.96%
GigaWorld-Policy [79]	3.46 / 2.22%	0.54 / 0.50%	6.20 / 3.27%
StarVLA-α [80]	3.34 / 2.44%	0.68 / 0.58%	6.40 / 3.24%
Xiaomi-Robotics-0 [8]	5.07 / 3.67%	0.22 / 0.17%	6.93 / 4.18%
X-WAM [21]	6.32 / 4.67%	0.57 / 0.25%	7.69 / 3.83%
X-VLA [92]	4.76 / 3.56%	0.55 / 0.50%	10.13 / 6.52%
$\pi_{0.5}$ [5]	5.78 / 4.56%	1.98 / 1.67%	11.41 / 6.91%
Spatial Forcing [35]	5.43 / 4.11%	1.78 / 1.58%	12.38 / 8.04%
Hy-Embodied-0.5-VLA [85]	**13.37** / **12.11**%	0.65 / 0.58%	13.07 / 8.80%
Xiaomi-Robotics-1 (Ours)	7.81 / 6.56%	**3.94** / **3.58**%	**20.07** / **13.93**%

32, 36, 39, 51, 56, 68, 78, 81, 83, 86] and vision-language-action (VLA) models [4, 5, 8–10, 30, 38, 42, 60, 82, 92] represent two popular paradigms of robot foundation models. WAMs are generally built upon pre-trained video models. They explicitly model future observations or environment dynamics and use these predictions to facilitate action generation. Early approaches typically generate visual plans from language-specified goals and subsequently translate them into executable robot actions [18, 33, 37, 43, 95]. More recent work extends this paradigm to several directions, including jointly modeling future prediction and action generation [23, 32, 36, 41, 96], incorporating explicit 3D geometric structure [21, 39, 88, 91], and training on large-scale heterogeneous video-action data to improve zero-shot generalization and cross-embodiment transfer [81, 86]. In contrast, VLA models leverage pre-trained vision-language models as backbone, aiming to harness their rich, general semantic knowledge to facilitate robust action prediction [4, 5, 8–10, 24, 25, 30, 47, 54, 97]. Recent advances improve VLA models along three main axes. First, reasoning-oriented methods introduce intermediate representations (*e.g.*, embodied reasoning tokens [13, 19, 34, 64, 84] and visual chains of thought [89, 93]) to support semantic, spatial, and temporal reasoning. Second, advances in action representation address the limitations of naive action discretization through learned tokenizers and flow matching [4, 20, 48, 58, 82]. Third, large-scale pre-training across heterogeneous datasets with different robot embodiments enables strong cross-embodiment transfer and robust downstream performance [8, 30, 54, 55, 69, 82]. Our work follows the VLA paradigm but focuses on a complementary question: the scaling behavior of robot foundation models. To support this investigation, we develop a scalable infrastructure to curate massive trajectory datasets with precise language annotations. Leveraging this data for large-scale training, we conduct extensive experiments to systematically explore the scaling properties of foundational VLA models.

5 Conclusions

In this work, we present `Xiaomi-Robotics-1`, a foundational vision-language-action (VLA) model that is able to follow instructions to perform a wide range of mobile manipulation tasks out-of-the-box in unseen environments, and efficiently adapt to novel challenging tasks with a minimal amount of data. During pre-training, we leverage over 100,000 hours of real-world manipulation trajectories, endowing the model with broad and generalizable manipulation capabilities. To scale training effectively, we propose an auto-labeling pipeline that annotates the large-scale dataset with detailed descriptions of scene state transitions as language prompts. In the post-training phase, we align these strong capabilities acquired during pre-training with robot embodiments and imperative instruction prompts using a cross-embodiment dataset. Extensive experiments demonstrate that the performance of `Xiaomi-Robotics-1` consistently improves with increasing data scale and model size during pre-training. More importantly, the scaling property directly translates to out-of-the-box performance in unseen environments after post-training. `Xiaomi-Robotics-1` can also serve as a powerful robot foundation model that is able to adapt to novel challenging real-robot tasks with minimal data. In addition, it achieves strong state-of-the-art results on four challenging simulation benchmarks. We hope this work can serve as a foundation for future exploration of scalable robot policies that can be deployed out-of-the-box in the real world.

Contributions

Authors are listed in alphabetical order.

Core Contributors

- Jun Guo
- Piaopiao Jin
- Jason Li
- Peiyan Li
- Yingyan Li
- Futeng Liu
- Wanli Peng
- Optimus Qin
- Yifei Su
- Nan Sun
- Qiao Sun
- Runze Suo
- Heyun Wang
- Yunhong Wang
- Rujie Wu
- Caoyu Xia
- Lina Zhang
- Jack Zhao

Contributors

- Guoliang Chen
- Wenlong Chen
- Xinze He
- Bin Li
- Qing Li
- Zhuorong Li
- Heng Qu
- Wenxuan Song
- Diyun Xiang
- Yifan Xie
- Peiran Xu
- Hangjun Ye
- Wen Ye
- Han Zhao
- Quanyun Zhou

Acknowledgment

We express our sincere appreciation to the broader team for their tremendous support, including those not listed above: Li Jiang, Xiaohan Yu, Meichen Mu, Xiaoke Xilinjueluo, Qingyi Li, Qi Liu, Yayun Liu, Jun Xia, Feng Qiu, Donghao Wang, Yan Hou, Dong Wang, Liangliang He, Jiaxin Liu, Kang Zhou, Rui Cai, Shuoxue Bi, Yingchao Zhou, Kun Ma, Yiwei Zhou and Dongsheng Li.

References

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

[2] Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] 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.

[5] 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.

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

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Rui Cai, Jun Guo, Xinze He, Piaopiao Jin, Jie Li, Bingxuan Lin, Futeng Liu, Wei Liu, Fei Ma, Kun Ma, et al. Xiaomi-robotics-0: An open-sourced vision-language-action model with real-time execution. *arXiv preprint arXiv:2602.12684*, 2026.

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

[10] Chilam Cheang, Sijin Chen, Zhongren Cui, Yingdong Hu, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, et al. Gr-3 technical report. *arXiv preprint arXiv:2507.15493*, 2025.

[11] Ronghan Chen, Yandan Yang, Zuojin Tang, Dongjie Huo, Tong Lin, Haoning Wu, Haoyun Liu, Yuzhi Chen, Lulu Zheng, Botai Yuan, Tianlun Li, Mingxin Wang, Dekang Qi, Bin Hu, Wei Mei, Yuze Xuan, Haolong Yang, Yanqing Zhu, Mu Xu, Zhiheng Ma, and Xinyuan Chang. Abot-m0.5: Unified mobility-and-manipulation world action model. *arXiv preprint arXiv:2607.00678*, 2026.

[12] Tianxing Chen, Yue Chen, Zixuan Li, Junyuan Tang, Kailun Su, Haoran Lu, Weijie Wan, Baijun Chen, Songling Liu, Haowen Yan, Honghao Su, Zhiyang Dou, Kaixuan Wang, Dandan Zhang, Yunze Liu, Yan Qin, Qiwei Liang, Qiwei Wu, Zijian Lin, Wenwei Lin, Yuran Wang, Minghua He, Tianshu Wu, Ruihai Wu, Jingquan Zhou, Kai-Chong Lei, Haibao Yu, Yuanfeng Ji, Weiyang Jin, Guanyu Lin, Xiaofan Li, Qi Xiong, Renjing Xu, Zhongyu Li, Wenhao Chai, Enze Xie, Ziwei Wang, Yao Mu, Hao Dong, Wojciech Matusik, Mingyu Ding, Wenbo Ding, Ping Luo, and Masayoshi Tomizuka. Robodojo: A unified sim-and-real benchmark for comprehensive evaluation of generalist robot manipulation policies, 2026. URL https://arxiv.org/abs/2607.04434.

[13] William Chen, Suneel Belkhale, Suvir Mirchandani, Oier Mees, Danny Driess, Karl Pertsch, and Sergey Levine. Training strategies for efficient embodied reasoning. *arXiv preprint arXiv:2505.08243*, 2025.

[14] Xi Chen, Xiao Wang, Soravit Changpinyo, Anthony J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[15] Yangtao Chen, Zixuan Chen, Peiyang Wang, Yong-Lu Li, Jing Huo, Jieqi Shi, and Yang Gao. Wh0: Generative world models as scalable sources of egocentric human hand manipulation data. *arXiv preprint arXiv:2606.22136*, 2026.

[16] 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*, 2024.

[17] 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 preprint arXiv:2402.10329*, 2024.

[18] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

[19] Haoquan Fang, Jiafei Duan, Donovan Clay, Sam Wang, Shuo Liu, Weikai Huang, Xiang Fan, Wei-Chuan Tsai, Shirui Chen, Yi Ru Wang, et al. Molmoact2: Action reasoning models for real-world deployment. *arXiv preprint arXiv:2605.02881*, 2026.

[20] Galaxea Team. Galaxea g0.5 technical report. 2026. URL https://opengalaxea.github.io/G05/.

[21] Jun Guo, Qiwei Li, Peiyan Li, Zilong Chen, Nan Sun, Yifei Su, Heyun Wang, Yuan Zhang, Xinghang Li, and Huaping Liu. Unified 4d world action modeling from video priors with asynchronous denoising. *arXiv preprint arXiv:2604.26694*, 2026.

[22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10, 2022.

[23] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

[24] Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, et al. $\pi^*_{0.6}$: a vla that learns from experience. *arXiv preprint arXiv:2511.14759*, 2025.

[25] Physical Intelligence, Bo Ai, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Greg Balke, Kevin Black, George Bokinsky, Shihao Cao, Thomas Charbonnier, et al. $\pi_{0.7}$: a steerable generalist robotic foundation model with emergent capabilities. *arXiv preprint arXiv:2604.15483*, 2026.

[26] 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.

[27] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[28] 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.

[29] Dongyoung Kim, Huiwon Jang, Myungkyu Koo, Suhyeok Jang, Taeyoung Kim, Beomjun Kim, Byungjun Yoon, Changsung Jang, Daewon Choi, Dongsu Han, et al. Rldx-1 technical report. *arXiv preprint arXiv:2605.03269*, 2026.

[30] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

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

[32] Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

[33] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to act from actionless videos through dense correspondences. In *International Conference on Learning Representations*, volume 2024, pages 40938–40958, 2024.

[34] Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.

[35] Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. *arXiv preprint arXiv:2510.12276*, 2025.

[36] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

[37] Peiyan Li, Hongtao Wu, Yan Huang, Chilam Cheang, Liang Wang, and Tao Kong. Gr-mg: Leveraging partially-annotated data via multi-modal goal-conditioned policy. *IEEE Robotics and Automation Letters*, 10(2):1912–1919, 2025.

[38] Peiyan Li, Yixiang Chen, Hongtao Wu, Xiao Ma, Xiangnan Wu, Yan Huang, Liang Wang, Tao Kong, and Tieniu Tan. Bridgevla: Input-output alignment for efficient 3d manipulation learning with vision-language models. *Advances in Neural Information Processing Systems*, 38:63635–63673, 2026.

[39] Peiyan Li, Yixiang Chen, Yuan Xu, Jiabing Yang, Xiangnan Wu, Jun Guo, Nan Sun, Long Qian, Xinghang Li, Xin Xiao, et al. Multi-view video diffusion policy: A 3d spatio-temporal-aware video action model. *arXiv preprint arXiv:2604.03181*, 2026.

[40] Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025.

[41] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

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

[43] Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. *arXiv preprint arXiv:2406.16862*, 2024.

[44] Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

[45] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[46] Fangchen Liu, Chuanyu Li, Yihua Qin, Jing Xu, Pieter Abbeel, and Rui Chen. Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface. *arXiv preprint arXiv:2504.06156*, 2025.

[47] 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. In *International Conference on Learning Representations*, volume 2025, pages 29982–30009, 2025.

[48] 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. *arXiv preprint arXiv:2602.03310*, 2026.

[49] 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.

[50] Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025.

[51] Teli Ma, Jia Zheng, Zifan Wang, Chunli Jiang, Andy Cui, Junwei Liang, and Shuo Yang. Dit4dit: Jointly modeling video dynamics and actions for generalizable robot control. *arXiv preprint arXiv:2603.10448*, 2026.

[52] 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.

[53] Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. Robocasa365: A large-scale simulation framework for training and benchmarking generalist robots. *arXiv preprint arXiv:2603.04356*, 2026.

[54] NVIDIA, Johan Bjorck, Nikita Cherniadev Fernando Castañeda, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. In *ArXiv Preprint*, March 2025.

[55] 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)*, pages 6892–6903. IEEE, 2024.

[56] Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025.

[57] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

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

[59] Haozhi Qi, Yen-Jen Wang, Toru Lin, Brent Yi, Yi Ma, Koushil Sreenath, and Jitendra Malik. Coordinated humanoid manipulation with choice policies. *arXiv preprint arXiv:2512.25072*, 2025.

[60] 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.

[61] Nan Sun, Yuan Zhang, Yongkun Yang, Wentao Zhao, Peiyan Li, Jun Guo, Wenxuan Song, Pengxiang Ding, Runze Suo, Yifei Su, et al. Revisiting embodied chain-of-thought for generalizable robot manipulation. *arXiv preprint arXiv:2606.03784*, 2026.

[62] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[63] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[64] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[65] Generalist Team. Gen-0: Embodied foundation models that scale with physical interaction. *Generalist AI Blog*, 2025. https://generalistai.com/blog/gen-0.

[66] Generalist Team. Gen-1: Scaling embodied foundation models to mastery. *Generalist AI Blog*, 2026. https://generalistai.com/blog/gen-1.

[67] Genesis AI Team. Gene-26.5: Advancing robotic manipulation to human level. *Genesis AI Blog*, May 2026. URL https://genesis.ai/blog/gene-26-5-advancing-robotic-manipulation-to-human-level.

[68] MotuBrain Team, Chendong Xiang, Fan Bao, Haitian Liu, Hengkai Tan, Hongzhe Bi, James Li, Jiabao Liu, Jingrui Pang, Kiro Jing, et al. Motubrain: An advanced world action model for robot control. *arXiv preprint arXiv:2604.27792*, 2026.

[69] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[70] Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

[71] Qwen Team. Qwen-robotmanip technical report: Alignment unlocks scale for robotic manipulation foundation models. 2026.

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

[73] An Dinh Vuong, Tuan Van Vo, Abdullah Sohail, Haoran Ding, Liang Ma, Xiaodan Liang, Anqing Duan, Ivan Laptev, and Ian Reid. World2act: Latent action post-training from world model dynamics. *arXiv preprint arXiv:2603.10422*, 2026.

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

[75] Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985*, 2024.

[76] Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, et al. A pragmatic vla foundation model. *arXiv preprint arXiv:2601.18692*, 2026.

[77] Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation. *arXiv preprint arXiv:2505.21864*, 2025.

[78] Sizhe Yang, Juncheng Mu, Tianming Wei, Chenhao Lu, Xiaofan Li, Linning Xu, Zhengrong Xue, Zhecheng Yuan, Dahua Lin, Jiangmiao Pang, et al. Memorywam: Efficient world action modeling with persistent memory. *arXiv preprint arXiv:2606.20562*, 2026.

[79] Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, Min Cao, Peng Li, Qiuping Deng, Wenjun Mei, Xiaofeng Wang, Xinze Chen, Xinyu Zhou, Yang Wang, Yifan Chang, Yifan Li, Yukun Zhou, Yun Ye, Zhichao Liu, and Zheng Zhu. Gigaworld-policy: An efficient action-centered world-action model. *arXiv preprint arXiv:2603.17240*, 2026.

[80] Jinhui Ye, Ning Gao, Senqiao Yang, Jinliang Zheng, Zixuan Wang, Yuxin Chen, Pengguang Chen, Yilun Chen, Shu Liu, and Jiaya Jia. Starvla-α: Reducing complexity in vision-language-action systems. In *European Conference on Computer Vision (ECCV)*, 2026.

[81] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

[82] Ryan Yu, Pushi Zhang, Starrick Liu, Brae Liu, Miracle Kang, Shalfun Li, Lights Shi, Ellie Ma, Ping Yang, Chris Pan, et al. Wall-oss-0.5 technical report. *arXiv preprint arXiv:2605.30877*, 2026.

[83] Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

[84] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

[85] He Zhang, Lingzhu Xiang, Haitao Lin, Zeyu Huang, Minghui Wang, Dingyan Zhong, Yubo Dong, Yihao Wu, Yongming Rao, Dongsheng Zhang, et al. Hy-embodied-0.5-vla: From vision-language-action models to a real-world robot learning stack. *arXiv preprint arXiv:2606.14409*, 2026.

[86] Qihang Zhang, Lin Li, Luyao Zhang, Shuai Yang, Yiming Luo, Shuaiting Li, Ruilin Wang, Junke Wang, Jiahao Shao, Gangwei Xu, et al. Native video-action pretraining for generalizable robot control. *arXiv preprint arXiv:2607.08639*, 2026.

[87] Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11142–11152, 2025.

[88] Haoyu Zhao, Xingyue Zhao, Siteng Huang, Xin Li, Deli Zhao, and Zhongyu Li. Rynnworld-4d: 4d embodied world models for robotic manipulation. *arXiv preprint arXiv:2607.06559*, 2026.

[89] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713, 2025.

[90] Zhaxizhuom Zhaxizhuoma, Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Pengan Chen, Pingrui Zhang, et al. Fastumi: A scalable and hardware-independent universal manipulation interface with dataset. In *Conference on Robot Learning*, pages 3069–3093. PMLR, 2025.

[91] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

[92] Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. *arXiv preprint arXiv:2510.10274*, 2025.

[93] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. In *International Conference on Learning Representations*, volume 2025, pages 54277–54296, 2025.

[94] Linqing Zhong, Yi Liu, Yifei Wei, Ziyu Xiong, Maoqing Yao, Si Liu, and Guanghui Ren. Acot-vla: Action chain-of-thought for vision-language-action models. *arXiv preprint arXiv:2601.11404*, 2026.

[95] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.

[96] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025.

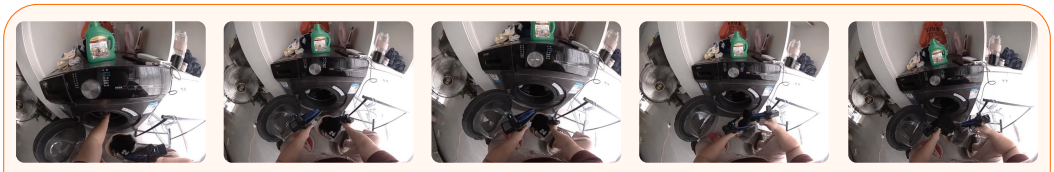
[97] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

Table 6 Progress Definition for Evaluation on Efficient Adaptation to New Tasks. Each rollout is assigned a progress score from 0 to 100% according to completed task milestones.

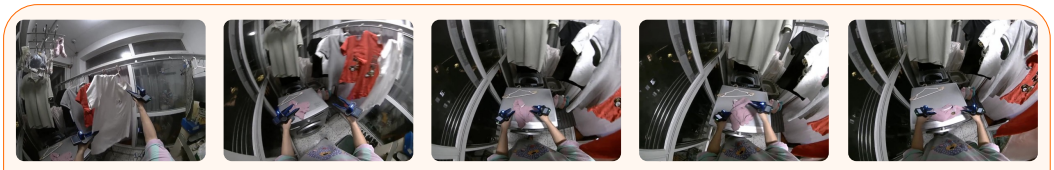
Task	Progress milestones	Progress (%)
Phone Packing	Grasp the phone; place the phone into the box; grasp the instruction manual; place the manual into the box; grasp the lid; successfully close the lid.	10, 10, 30, 10, 10, 30
Printer Refilling	Grasp the paper; complete the handover; successfully insert one end of the paper into the printer tray; fully insert the paper stack into the printer tray; return both robot arms to the resting pose.	20, 20, 20, 30, 10
Laundry Loading	Open the washing machine door; move the laundry basket to the door; transfer the clothes into the washing machine; remove the laundry basket; close the washing machine door.	20 each
Box Packing	Grasp and place each specified target object into the box according to the language instruction. Five target objects are evaluated in each rollout.	20 each

Gripper: Open the refrigerator door, move the food bag, then close the door.
Object: {
refrigerator door: Opened, then closed.
plastic bag: Moved inside the refrigerator.
}

Gripper: Flip the white patterned quilt and spread it flat on the bed.
Object: {
white patterned quilt: Flipped over and spread flat.
bed: Covered by the quilt.
}



Gripper: Move the black garment from the washing machine to the beige basin.
Object: {
black garment: Moved from the washing machine to the beige basin.
washing machine drum: No longer contains the garment.
beige basin: Contains the garment.
}



Gripper: Hang up the white shirt, then place the pink shirt on the washing machine.
Object: {
white shirt: Hung up.
pink shirt: Placed on the washing machine.
}

Figure 11 Examples of UMI data in the Pre-training Dataset.

Pick up the brown bottle on the table and put it on the shelf in the refrigerator with the left hand; Pick up the plastic bag containing food on the table and put it on the refrigerator shelf with the right hand.

Pick up the pink pillow from the left side of the sofa and place it on the left side of the first pillow.

Take the silver-packaged snacks from the white basket in front and place them in the nearest transparent box on the fourth layer of the front shelf.

Take the blue and white box at the lower right of the sink and place it to the right of the green box on the right side of the faucet with the right hand.

Pick up a white sports shoe on the left foot from the ground with the left hand; Wipe the white sports shoes on the left foot clean with the right hand.

Figure 12 Examples of UMI data in the Post-training Dataset.