TRANSLATION AS A BRIDGING ACTION: TRANSFERRING MANIPULATION SKILLS FROM HUMANS TO ROBOTS

Anonymous authors
Paper under double-blind review

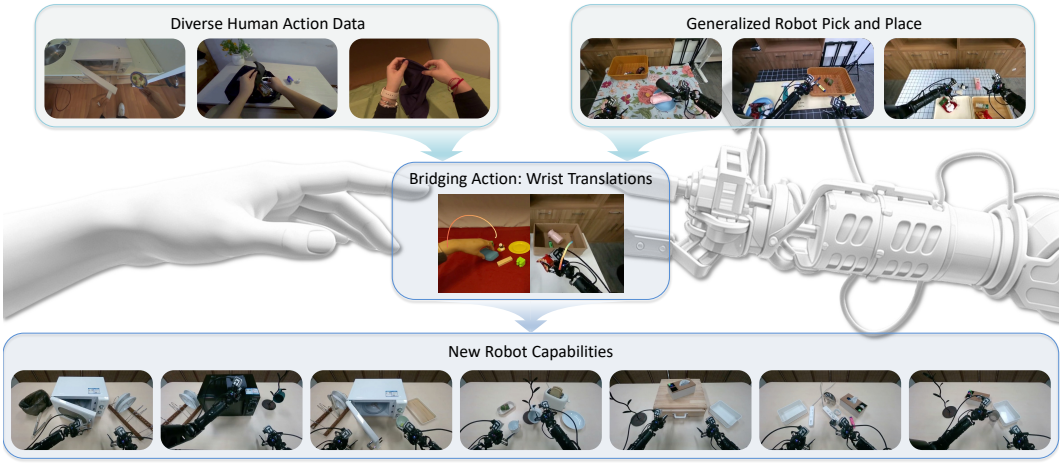


Figure 1: **Overview.** We study whether we can transfer human manipulation skills to bi-manual robots with grippers by learning relative wrist translation as the bridging action representations. We show that such bridging action not only enables robust and efficient manipulation knowledge transfer, but can benefit the model from large-scale human only pre-training.

ABSTRACT

We study whether we can learn novel manipulation skills from human actions to a bi-manual robot with parallel grippers. Benefiting from the robot-free setting, human action data becomes one of the most scalable sources for manipulation data. However, transferring manipulation skills from humans to robots remains hard: most prior work treats humans as just another bi-manual 6DoF embodiment, where hand-pose estimates are noisy and the contact patterns of human fingers differ fundamentally from those of a parallel gripper. We argue that learning rotation-inclusive action signals from human data is therefore sub-optimal and instead propose a bridging action representation, *i.e.*, the relative wrist translation within the head-camera frame, an action space shared by humans and robots. To handle the potential absence of the 6DoF end-effector action or gripper signal in human data, we build a π_0-like vision-language-action model with interleaved action tokens. On a suite of bi-manual manipulation tasks seen only in human action data, our bridging action enables a superior success rate of 22.50% over naively adopting noisy 6DoF actions (12.50%), and can be further improved to 38.33% with large-scale human-only pre-training.

1 INTRODUCTION

In this paper, we study whether we can learn novel robot manipulation skills from human action data on a bi-manual robotic platform. Benefiting from the robot-free setting, human manipulation data can be collected without dedicated robot hardware, calibrated tele-operation rigs, or well-trained

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.

operators; and can therefore be gathered in far more diverse environments and at a larger scale than tele-operated robot data (Team, 2025; Grauman et al., 2022). In terms of environmental and skill diversity, human data is considered one of the most scalable and promising ways to scale up robot learning (Hoque et al., 2025; Team, 2025; Zheng et al., 2026). Along with the growing popularity of ego-centric human data collection (Grauman et al., 2022; Hoque et al., 2025; Grauman et al., 2024), researchers have flocked to the exploration of how human data can improve embodied policies.

Prior research has proposed various types of human action representations for ego-centric human data. A large proportion of work adopts wrist poses (Kareer et al., 2025a;b) with optional fingertip positions (Cai et al., 2025), parameterized hands (Luo et al., 2026), or finger joints (Zheng et al., 2026). In addition to explicit action representation, researchers have also proposed latent action representations (Fan et al., 2025; Ye et al., 2024). Incorporating human actions when building robot policies has shown great potential for improving generalization (Cai et al., 2025) and steering pre-trained policies to manipulate novel objects and follow new instructions (Cheang et al., 2025; Kareer et al., 2025b; Yuan et al., 2025) via human-robot co-training.

Different from cross-embodiment robot data (O'Neill et al., 2024), learning from human data encounters unique challenges. Human action is typically derived from hand pose estimators (Kareer et al., 2025a), which is inevitably noisy even though it can be constrained by additional parameterization (Luo et al., 2025) or re-targeting (Zheng et al., 2026). Additionally, the contact patterns differ greatly between human hands and robot grippers (Wei et al., 2026): the additional degrees of freedom in fingers make wrist rotations less meaningful when reflecting manipulation behaviors.

To enable efficient manipulation knowledge transfer from human to robot, we propose to learn the wrist translations as a shared action space between humans and robots. We therefore end up with two different action spaces: 1) the rotation-inclusive end-effector action to control the robot (Zheng et al., 2026; Kareer et al., 2025b) and 2) the wrist-translation action that can be shared between humans and robots. We build our method upon a π_0-like (Black et al., 2024) vision-language-action model, which processes the observation and instruction inputs with pre-trained vision-language priors and generates actions via flow matching (Liu et al., 2022; Lipman et al., 2022). Meanwhile, to handle the missing action components for different embodiments in human-robot co-training, *i.e.*, the missing 6DoF end-effector actions and gripper signals for in-the-wild human data (Tab. 1), we adopt interleaved action tokens to model: wrist translation, 6DoF end-effector actions, and gripper signals. By masking tokens corresponding to missing action components, the Transformer model infers the targeted action components by reusing the inductive bias of positional encodings and its native capability to process variable-length sequences, removing the need for additional design.

We first pre-train the model on a larger-scale in-the-wild human action data with only wrist translation actions. Then, we co-train the model on skill-specific human action data with robot pick-and-place data (Cheang et al., 2025). We show the bridging action also enables more stable manipulation skill transfer than naively using 6DoF human actions (Fig. 5). Empirically, we find that the human to robot manipulation skill transfer only takes effect when the shared bridging action is explicitly supervised on both robot and human action data: removing the bridging action component while keeping the model exposed to the identical human and robot data drops the overall success rate from 38.33% to 12.50% (Sec. 5.3). Experiments also show that the bridging action is a scalable action representation that it can benefit the model from large-scale pre-training on in-the-wild human action data. With extensive experiments:

- We validate that we can transfer manipulation skills from human action data to a robot with a shared translation-only bridging action. Given the noisy human actions, our proposed translation-only bridge also transfers manipulation skills more robustly than directly using noisy 6DoF human wrist actions.
- We build an interleaved action sequence that reuses the inductive bias in Transformer's position encoding and variable length to handle the potential missing action components in different manipulation data. We find that manipulation skill transfer only takes effect when the shared bridging action is explicitly supervised on both robot and human action data.
- We also perform a diagnostic study to confirm that the bridging action itself is an effective medium for manipulation skill transfer by applying our bridging action directly to robot data, which removes the observation gap between robots and humans and reduces the action noise caused by hand pose estimators.

2 Related Works

Learning from human. Along with the growing number and size of ego-centric datasets (Zhou et al., 2018; Goyal et al., 2017a; Miech et al., 2019; Damen et al., 2018; Grauman et al., 2022; 2024; Hoque et al., 2025), researchers have explored diverse approaches to harvest physical knowledge from large-scale human data. Some works learn generative visual representations from human videos (Cheang et al., 2024; Wu et al., 2023; Hu et al., 2024). Another line of work learns intermediate representations, *e.g.*, affordance (Bahl et al., 2023; 2022), keypoints (Wang et al., 2023; Wen et al., 2023; Bharadhwaj et al., 2024), and latent actions (Bruce et al., 2024; Ye et al., 2024; Bjorck et al., 2025). There are also works that collect explicit human actions directly with gloves (Xu et al., 2025; Tao et al., 2025) or using hand pose estimators (Kareer et al., 2025b;a). Human data is not only regarded as the pretraining data source (Zheng et al., 2026; Cheang et al., 2024; Bjorck et al., 2025), but is also widely adopted during human-robot alignment for compositional generalization (Kareer et al., 2025a; Zhu et al., 2026; Cheang et al., 2025; Kareer et al., 2025b). In our project, we are mainly interested in the compositional generalization from human-robot co-training. Considering the noise in explicit human actions and the semantic misalignment of contact behaviors between robots and humans (Wei et al., 2026), we argue that human wrist rotation is not always a reliable supervision signal for embodied policies. Instead, we establish a translation-based bridging action that is shared between humans and robots to enable effective human to robot manipulation knowledge transfer.

Cross embodiment learning aims to learn the shared manipulation knowledge across different robot configurations within a unified model (O'Neill et al., 2024; Black et al., 2024; Intelligence et al., 2025). Meanwhile, learning from robot and human actions can also be considered cross embodiment learning (Kareer et al., 2025a; Zheng et al., 2026; Zhu et al., 2026). To handle the different degrees of freedoms in different embodiments, some works define a unified action space by concatenating (Liu et al., 2024; Cheang et al., 2025; Kareer et al., 2025a), padding missing dimensions (Black et al., 2024; Pertsch et al., 2025; Intelligence et al., 2025), or simply building different projectors (Wang et al., 2024; Bjorck et al., 2025) for different action representations. Our work explores the compositional generalization of human and robot actions by presenting a translation-based bridging action that can be shared between robots and humans. To handle the potential missing action components in different manipulation data sources, we break down the action components and adopt an interleaved action sequence. By masking tokens corresponding to missing action components, we reuse the Transformer's inductive bias of positional encoding and its native capability to process variable-length sequences to distinguish different data sources.

Data pyramid and co-training are widely discussed when developing robust and generalizable robot learning systems (Bjorck et al., 2025; Barreiros et al., 2025; Lin et al., 2026; Team, 2026; Cheang et al., 2025). Instead of resorting to the costly practice of collecting real robot data, cross-embodiment data (O'Neill et al., 2024; Kim et al., 2024; Liu et al., 2024; Black et al., 2024; Intelligence et al., 2025), human data (Kareer et al., 2025a; Zheng et al., 2026; Cheang et al., 2025), vision-language corpora (Goyal et al., 2017b; Chen et al., 2015; Li et al., 2024a; Zitkovich et al., 2023; Wu et al., 2026), and videos (Bu et al., 2025; Bjorck et al., 2025; Cheang et al., 2024; Wu et al., 2023; Hu et al., 2024; Ye et al., 2026; Li et al., 2026) serve as more cost-effective alternatives for scaling up robot learning. By exposing the model to a more diverse data sources and supervision, the robot is more robust to environmental distortions (Cheang et al., 2025; Intelligence et al., 2025; Lin et al., 2026) and can follow more generalized instructions (Zitkovich et al., 2023; Cheang et al., 2025). In this paper, our work focuses specifically on how to learn new manipulation skills from human manipulation data, and whether it can benefit from large-scale human action pre-training.

3 Hardware and Platform

Robot platform. We experiment with a bi-manual mobile manipulation platform with two 7-DoF arms, parallel grippers, and three RGB-D cameras mounted on the head and both wrists (Cheang et al., 2025) (Fig. 3). During data collection, we randomly adjust the robot's height, position, and rotation for diversity. Meanwhile, during autonomous rollouts, we keep the robot base static and control the arms with 6DoF end-effector poses and discrete gripper signals.

Human action platform. We collect our in-lab and in-the-wild human action data with PICO 4 Ultra Enterprise, where we can directly read the 6DoF wrist pose $\mathbf{W}_w^t$ and the head-mounted camera pose $\mathbf{T}_{w\leftarrow c}^t$ with respect to the world frame at each timestep t from PICO's built-in hand-tracking and headset SLAM modules. During our in-lab data collection, we ask the operators to imitate the

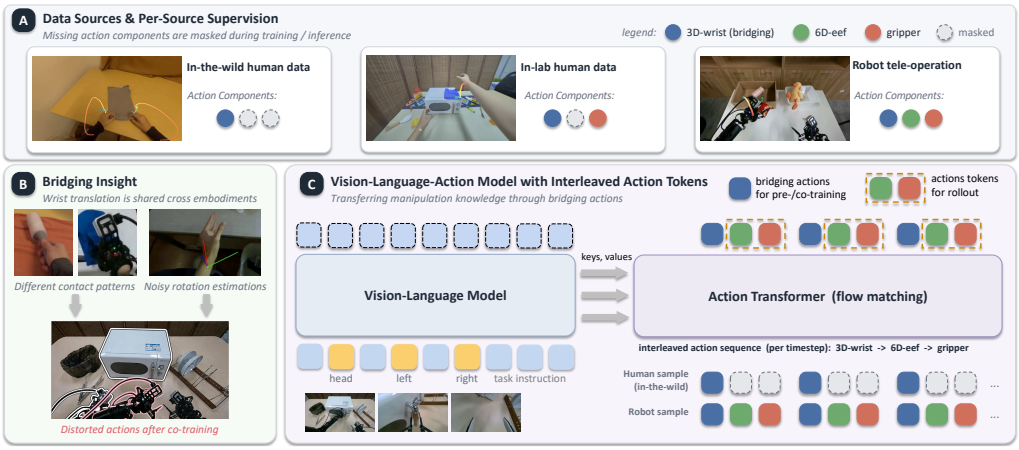


Figure 2: **Architecture Overview.** We train our model on (a) the mixture of human and robot action data. However, (b) learning from 6DoF human action is challenging and suboptimal because of different contact patterns and noisy human actions, thus, we propose the translation-based bridging action, and adopt (c) a π_0-like (Black et al., 2024) vision-language-action model with interleaved action sequences to handle potential missing action components for manipulation skill transfer.

robot grippers with their hand postures and to keep their hands within the head camera's field of view, which narrows the contact-behavior discrepancy between humans and the robot. However, it remains challenging for us to strictly control the contact behavior during the in-the-wild dataset collection.

4 Method

We treat wrist translation as the bridging action, which can be extracted from both human and robot actions (Sec. 4.1). Then, we introduce how we deal with the potential missing action components from different action data sources (Sec. 4.2) and how we train our model (Sec. 4.3).

4.1 Translation as a Bridging Action

Why not just 6DoF human wrist actions? Though the mainstream practice treats humans as another embodiment by extracting relative 6DoF wrist actions from hand-pose estimators (Kareer et al., 2025a; Yuan et al., 2025; Zheng et al., 2026), we argue this is sub-optimal: 1) the wrist rotations for human data are noisy due to the cumulative errors from hand pose estimators (Wei et al., 2026), and 2) the contact patterns of fingers and parallel grippers are different, decoupling the wrist rotation from the actual manipulation behavior. Empirically, we also find that directly co-training on the extracted 6DoF human wrist actions on robots often leads to distorted behaviors (Fig. 5).

From wrist poses to a shared bridging action $\mathbf{a}^{\text{3D-wrist}}$. We aim to find a shared action signal that is reasonably reliable and can be shared among humans and robots. Since humans and robots both act upon what they perceive, we treat the head camera frame as the reference frame and treat the relative wrist translations as our bridging actions. Specifically, we use $\mathbf{W}_w^t \in \mathbb{SE}(3)$ to represent the wrist pose in the world frame at time t, and $\mathbf{T}_{w\leftarrow c}^t \in \mathbb{SE}(3)$ for the head-camera pose, whose camera frame we abbreviate as c_t. During human action data collection, both $\mathbf{W}_w^t$ and $\mathbf{T}_{w\leftarrow c}^t$ are directly obtained from the headset's built-in hand-tracking and SLAM modules (Sec. 3). Applying $(\mathbf{T}_{w\leftarrow c}^t)^{-1}$ maps the wrist pose into c_t, yielding $\mathbf{W}_{\mathrm{c}_t}^{t+i} = (\mathbf{T}_{w\leftarrow c}^t)^{-1}\,\mathbf{W}_w^{t+i}$. Our bridging action $\mathbf{a}_t^{\text{3D-wrist}}$ describes the relative wrist translation over a k-sized window ($k = 30$) with respect to the camera frame at t:

$$\boldsymbol{a}_{t+i}^{\text{3D-wrist}} = \Delta\mathbf{W}^{\text{3D}} = \boldsymbol{t}\big(\mathbf{W}_{\mathrm{c}_t}^{t+i}\big) - \boldsymbol{t}\big(\mathbf{W}_{\mathrm{c}_t}^{t}\big)\,, \qquad i = 1, \ldots, k. \tag{1}$$

Here, $\boldsymbol{t}(\cdot)$ extracts the 3×1 translation components in $\mathbb{SE}(3)$. To model actions for both bi-manual robots and humans, we concatenate wrist translations for both arms, yielding $\mathbf{a}_t^{\text{3D-wrist}} \in \mathbb{R}^{k\times 6}$.

Robot end-effector action $\mathbf{a}^{\text{6D-eef}}$. We use the relative 6DoF end-effector action to control the robot, which is defined as the relative wrist motion with respect to the initial end-effector pose:

$$\boldsymbol{a}_{t+i}^{\text{6D-eef}} = \Delta\mathbf{W}^{\text{6D}} = \big(\mathbf{W}_w^t\big)^{-1}\,\mathbf{W}_w^{t+i}, \qquad i = 1, \ldots, k. \tag{2}$$

Being a relative pose between two $\mathbb{SE}(3)$ elements, this action representation is invariant to the absolute camera poses and is physically meaningful for describing arm motions (Zheng et al., 2026). We further transform $\mathbb{SE}(3)$ into Cartesian coordinates and Euler angles for robot actions, resulting in $\mathbf{a}_t^{\text{6D-eef}} \in \mathbb{R}^{k\times 12}$ by concatenating the end-effector actions for both arms.

Gripper action $\mathbf{a}^{\text{gripper}} \in \mathbb{R}^{k\times 2}$ is defined as a chunk of binary signal $a_i^{\text{gripper}} \in \{0, 1\}$ each gripper, where we use 1 for close and 0 for open. For our in-lab human action collection, we also annotate hand closures as gripper signals.

Unified action supervision. We define a unified action space via the combination of different action components $\mathbf{a}_t = (\mathbf{a}_t^{\text{3D-wrist}}, \mathbf{a}_t^{\text{6D-eef}}, \mathbf{a}_t^{\text{gripper}})$. Differences in data-collection hardware, together with variable difficulty in ensuring that data collectors closely follow the predefined manipulation standards, pose significant challenges for us in ensuring the quality of all action components across different data sources. Therefore, we seek comparably reliable action components for different data sources within the unified action representation. As shown in Table 1, we learn the shared $\mathbf{a}^{\text{3D-wrist}}$ for human actions while additionally grounding $\mathbf{a}^{\text{3D-wrist}}$ into the executable $\mathbf{a}^{\text{6D-eef}}$ for robot actions.

Table 1: **Per-embodiment action supervision.** Based on reliably extractable action components, we learn distinct action representations for various embodiments under a unified action space.

Data source	$\mathbf{a}^{\text{3D-wrist}}$	$\mathbf{a}^{\text{6D-eef}}$	$\mathbf{a}^{\text{gripper}}$
In-the-wild human action data	✓	–	–
In-lab human action data	✓	–	✓
Robot tele-operation	✓	✓	✓

4.2 VLA Model with Interleaved Action Sequence

Architecture. We build our model upon a π_0-like (Black et al., 2024; Cheang et al., 2025) end-to-end vision-language-action (VLA) model, $\pi_\theta(l, o_t)$, which generates an action chunk $\mathbf{a}_t = a_{t:t+k}$ conditioned on the language instruction l and the observations o_t from the head and two wrist cameras. We pad blank images for the missing wrist views in human action data.

The VLA model formulates vision, language, and action tokens within a sequence, sharing the self-attention operation, but adopting two separate sets of parameters to balance different training objectives. The vision and language tokens (o_t, l) are processed via weights initialized from a pre-trained VLM (Bai et al., 2025), and the resulting vision-language KV-cache serves as the context condition for the action chunk generation via flow matching (Liu et al., 2022; Lipman et al., 2022).

Interleaved action tokens. Given previous introduction on our action definitions (Sec. 4.1), the action chunk can be represented as a subset of $\{\mathbf{a}^{\text{3D-wrist}}, \mathbf{a}^{\text{6D-eef}}, \mathbf{a}^{\text{gripper}}\}$ depending on the data source. Since the Transformer itself is able to handle variable-length inputs via attention masks and position ids (Vaswani et al., 2017), we organize the action tokens in the $\boldsymbol{a}_{t+i}^{\text{3D-wrist}} \rightarrow \boldsymbol{a}_{t+i}^{\text{6D-eef}} \rightarrow \boldsymbol{a}_{t+i}^{\text{gripper}}$ order, and mask the corresponding action tokens to handle possible absence of certain action types, minimizing the introduction of additional inductive bias to the model architecture. The ordering of action components is established based on two priors: 1) the shared bridging signal should be attended to by the 6DoF action tokens, enabling explicit manipulation knowledge transfer from human to robot within the attention layer; and 2) the gripper signal is typically triggered after the end-effector reaches its target.

Action flow-matching. Our model generates an action chunk via flow matching (Lipman et al., 2022; Liu et al., 2022). The model learns to estimate the velocity field $\hat{v}$ from $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to $\mathbf{a}_t$ based on the observation o_t, text instruction l, and noisy action chunk $\mathbf{a}_t^\tau = \tau\,\epsilon + (1-\tau)\,\mathbf{a}_t$, where $\tau \in (0, 1]$. We write the velocity estimation as $\hat{v}(\mathbf{a}_t^\tau, o_t, l, \tau)$ and the ground-truth velocity as $v^* = \epsilon - \mathbf{a}_t$. Therefore, the velocity prediction is supervised by minimizing the below MSE loss:

$$\mathcal{L}_{\text{FM}} = \|\hat{v}(\mathbf{a}_t^\tau, o_t, l, \tau) - v^*\|_2^2 . \tag{3}$$

For each training sample, $\mathcal{L}_{\text{FM}}$ is computed on the provided action components. During inference, we generate only $\mathbf{a}^{\text{6D-eef}}$ and $\mathbf{a}^{\text{gripper}}$ for robot control by integrating the velocity from $\tau = 1$ to 0 with $\Delta\tau = 0.2$ using the Euler method, *i.e.*, $\mathbf{a}_t^{\tau+\Delta\tau} = \mathbf{a}_t^\tau + \Delta\tau \cdot \hat{v}(\mathbf{a}_t^\tau, o_t, l, \tau)$, across all the experiments.

Vision-language co-training. To prevent feature collapse on action data, we co-train the VLA model on the action data with a collection of vision-language data (Cheang et al., 2025; Intelligence et al., 2025; Zitkovich et al., 2023) using the standard next-token prediction (NTP) objective, *i.e.*, $\mathcal{L}_{\text{NTP}} = -\frac{1}{|s|}\sum_{i=1}^{|s|} \log P\big(s^i \mid s^{[1,\dots,i-1]};\, v, l\big)$, where s is the text response to the vision (v) language (l) query, and $|s|$ is the length of s. Each training sample adopts either $\mathcal{L}_{\text{FM}}$ or $\mathcal{L}_{\text{NTP}}$, depending on the types of input query. Therefore, the total loss is a weighted sum of $\mathcal{L}_{\text{FM}}$ and $\mathcal{L}_{\text{NTP}}$ over the batch.

4.3 Data and Training Strategies

Stage I: Pre-training on human actions. Being one of the most scalable manipulation data sources (Hoque et al., 2025; Grauman et al., 2022), human data offers unmatched skill and scene diversity. Accordingly, we pre-train our VLA model on approximately 600 hours of human actions: covering ∼70 hours from selected tasks in EgoDex (Hoque et al., 2025), ∼500 hours of out-sourced free-form household manipulation, and ∼45 hours of in-lab human actions. As we cannot standardize the contact behaviors at this scale, we supervise only the bridging signal $\mathcal{L}_{\text{FM}}^{\text{3D-wrist}}$ for pre-training.

Stage II: Human-robot co-training. We expect the model to transfer the knowledge learned from the bridging action toward executable 6-DoF robotic actions. Through mixed training on human and robot action, we aim to achieve certain levels of "*compositional generalization*": enabling the robot to execute skills that have appeared only within human demonstration data. To isolate the skills seen in robot data from those present in human action data, we adopt a relatively aggressive data strategy by using only ∼72 hours of **robot pick-and-place data** over 100 objects, annotated with the fixed template "put {object} into {container}" (Cheang et al., 2025). Paired with robot data, we collect ∼3 hours per task of task-specific **in-lab human actions** across 15 tasks, e.g., "open the microwave door", *etc.* (Fig. 3), which is identical to the 45 hours of in-lab human actions in stage I. To minimize contact-pattern mismatch, we ask in-lab operators to mimic the gripper with their hand posture, and annotate hand closure as the gripper control signal. The model is now supervised with $\mathcal{L}_{\text{FM}}^{\text{3D-wrist}}$, $\mathcal{L}_{\text{FM}}^{\text{6D-eef}}$, and $\mathcal{L}_{\text{FM}}^{\text{gripper}}$. In practice, we randomly drop either $\mathbf{a}_t^{\text{3D-wrist}}$ or $\mathbf{a}_t^{\text{6D-eef}}$ during training. Dropping $\mathbf{a}_t^{\text{3D-wrist}}$ enables us to generate solely $\mathbf{a}_t^{\text{6D-eef}}$ for inference, whereas supervision over $\mathbf{a}_t^{\text{3D-wrist}}$ remains most important for robot data to inherit manipulation knowledge from human data through bridging actions (Sec. 5.3).

Stage III: Robot post-training. We collect an additional set of robot data, containing 100 tele-operated robot trajectories for each of the 15 tasks listed in Fig. 3. We will further detail how we use this data for performance analysis in Sec. 5 to study the effectiveness of our bridging action.

Implementation. Our VLA model is basically a Mixture-of-Transformer (Liang et al., 2024; Black et al., 2024) with approximately 4B parameters, and is initialized from a pre-trained VLM (Bai et al., 2025). Throughout the training process, we co-train the model with VLM data, and unfreeze all parameters. We adopt a global batch size of 1024 for 400k iterations in **Stage I** (human-only pre-training), and a global batch size of 256 for another 120k iterations in **Stage II** (human-robot co-training), and further 25k iterations for **Stage III** (robot post-training). Additionally, we follow prior practice (Li et al., 2024b; Cheang et al., 2025) to increase the effective batch size for action generation by repeating the VLM's KV-cache by $4\times$ to accelerate convergence.

5 Experiments

We conduct real-world evaluations on 15 manipulation tasks (Fig. 3), for which we collect corresponding human actions to study whether and how the bridging action helps transfer manipulation skills from human actions to robots. Concretely, we aim to answer the following questions:

- **Q1.** Is the bridging action able to transfer manipulation skills beyond generic pick-and-place, and does it scale with large-scale human-only pre-training? (Sec. 5.1)
- **Q2.** How does the bridging action compare to 6DoF human actions? (Sec. 5.2)
- **Q3.** What training objective is essential during human-robot co-training? (Sec. 5.3)
- **Q4.** After removing observation gaps and action noise, how good is the bridging action? (Sec. 5.4)

Evaluation protocol. We evaluate model variants across 15 manipulation tasks (Fig. 8), where we perform four rollouts for two different layouts, resulting in eight trials per task. For a fair comparison, we reset the robot's initial pose and object layouts according to a pre-recorded mask, and report both the success rate and the average progress (Fig. 8) as metrics.

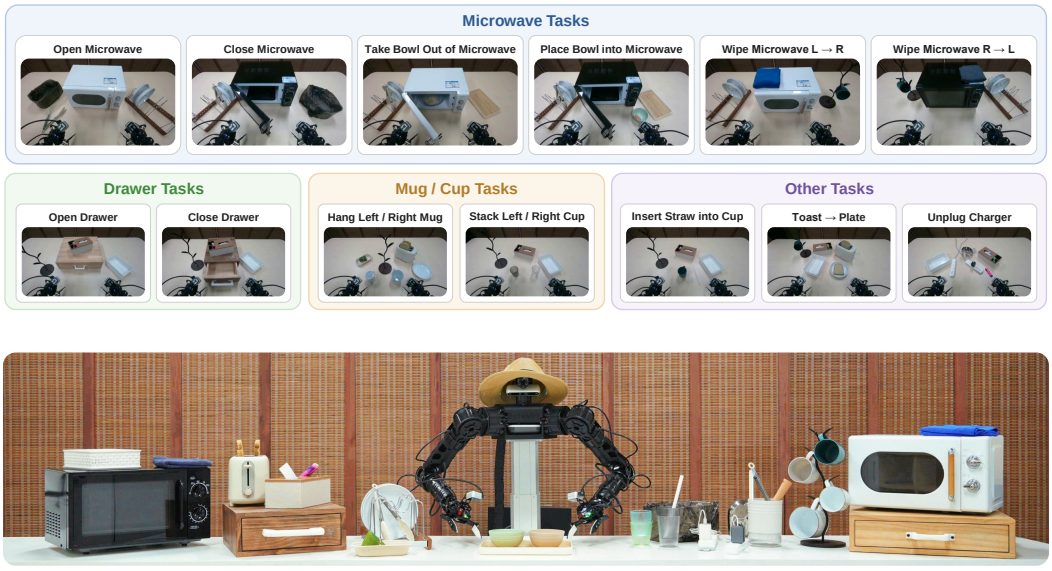


Figure 3: **Evaluation setups.** We show one of the initialized setups for each task (top), and the objects used throughout policy evaluations (bottom).

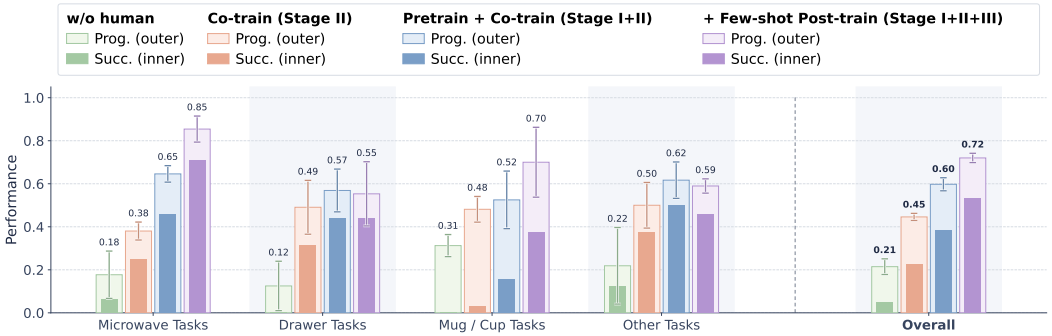


Figure 4: **Main results.** Co-training on human and robot actions with bridging action (orange) enables manipulation knowledge transfer from humans to robots. The bridging action is a scalable action representation that can benefit from large-scale human-only pre-training (blue). We can further improve the performance with a relative small amount of robot demonstrations (10 traj./task, purple).

5.1 The bridging action is a scalable objective for manipulation skill transfer

Finding 1: The bridging representation transfers skills beyond pick-and-place. An alternative hypothesis is that, the robot itself is already able to solve the evaluation tasks simply by exposing to robot pick-and-place data and vision language corpus. To rule this out, we first compare two checkpoints trained on generalized robot pick-and-place data with and without human data in Fig. 4. Both checkpoints are exposed to the same vision-language data. We find that training the model without any human actions (green) achieves very low performance on our evaluation tasks, while there are substantial improvements (0.21 to 0.45 average progress) after co-training the model with human actions (orange) in Fig. 4. Such improvement validates that our proposed bridging action helps to learn the human-to-robot motion transfer on manipulation tasks exceeding simple pick-and-place.

Finding 2: The bridging representation scales with large-scale human-only pre-training. After our preliminary verification that human data enables skill transfer via the bridging action, we aim to leverage large-scale human-only pre-training to further validate the scalability of this representation. We compare human-robot co-training with (blue) and without (orange) human-only pre-training in Fig. 4. Though the policy is pre-trained with only the bridging action $\mathbf{a}_t^{\text{3D-wrist}}$ instead of the actual robot actions $\mathbf{a}_t^{\text{6D-eef}}$, this large-scale human-only pre-training yields substantial performance gains from 0.45 to 0.60, which confirms both the scalability and effectiveness of the bridging action.



Figure 5: **Qualitative comparison on co-training with different human actions.** We show that when co-training on 6DoF human actions (left), the robot yields a distorted, off-target wrist pose; while our bridging actions (right) yield a natural pose that aligns with the microwave door handle.

5.2 Why translation-only bridging actions instead of 6DoF human actions?

In our early attempts, we adopt relative 6DoF human actions as in Zheng et al. (2026). However, we observe that co-training with 6DoF human actions will lead to noisy and twisted behaviors (Fig. 5, left). We also quantitatively compare this baseline against our bridging action $\mathbf{a}_t^{\text{3D-wrist}}$ in Tab. 2, both co-trained from scratch. We observe that our bridging action is more robust to action noises, and leads to better performance than the 6DoF human action baseline (22.50 vs. 12.50% success).

Table 2: **Comparison on human action representation.** When co-trained from scratch, learning from bridging actions is more efficient than naive 6DoF human actions.

Human Actions	Microwave Tasks		Drawer Tasks		Mug / Cup Tasks		Other Tasks		Overall	
	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)
$\mathbf{a}^{\text{6D-eef}}$	25.00	4.17	**55.00**	31.25	28.13	0.00	49.17	33.33	34.67	12.50
$\mathbf{a}^{\text{3D-wrist}}$ (Ours)	**38.02**	**25.00**	49.06	31.25	**48.13**	**3.13**	**50.00**	**37.50**	**44.58**	**22.50**

5.3 The shared bridging signal, not just co-training, drives the transfer

We here study how the bridging action transfers manipulation knowledge from human actions to robots. Remind that during human-robot co-training, we supervise the model to also generate bridging action for robot data. Here, we test an isolation alternative by never exposing the model to $\mathbf{a}^{\text{3D-wrist}}$ for robot actions after human-only pre-training. In this case, the baseline can be treated as cross-embodiment learning with different action input/output projectors while sharing the Transformer backbone. We show that removing the bridging action for robot data greatly degrades the performance across all task groups, dropping the overall success rate from 38.33% to 12.50% (Tab. 3). This confirms that the supervision signal on the shared bridging action carries the transferable manipulation knowledge from humans to robots. Simply sharing the Transformer backbone with separate embodiment projectors, without explicitly binding the bridging actions to executable robot actions, is insufficient.

Table 3: **Ablation of the bridging objective on robot data.** During co-training, it is essential to also supervise the bridging action on robot data for manipulation skill transfer.

Robot Actions	Microwave Tasks		Drawer Tasks		Mug / Cup Tasks		Other Tasks		Overall	
	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)
Ours w/o $\mathbf{a}^{\text{3D-wrist}}$	35.73	10.42	39.38	12.50	39.38	0.00	48.13	33.33	39.67	12.50
Ours	**64.58**	**45.83**	**56.88**	**43.75**	**52.50**	**15.63**	**61.67**	**50.00**	**59.75**	**38.33**

5.4 Disentangling the observation gap and action noise

Though we can narrow the manipulation gap between human and robot actions by asking in-lab human data operators to imitate grippers with their hand postures, there are still gaps including the action noise caused by hand pose estimators, and observation gaps like missing wrist cameras. Here, we construct a diagnostic variant by co-training the model on all action components for robot pick-and-place data while generating only bridging actions ($\mathbf{a}^{\text{3D-wrist}}$) for **task-specific robot demonstrations** (100 traj./task). Such variant isolates the two confounds present in the human-to-robot setting, *i.e.*,

the observation gap (no wrist camera, differing viewpoint) and human action noise, while holding the bridging action representation and training recipe fixed. As reported in Tab. 4, this gap-free variant outperforms the model co-trained with human data, raising the success rate from 38.33% to 55.83%. This confirms that 1) the bridging representation itself is an effective medium for manipulation skill transfer, and 2) when narrowing the embodiment gap and reducing action noises, our bridging action enables better manipulation skill transfer.

Table 4: **Disentangling the observation gap and action noise with task-specific robot bridging actions** further shows the bridging action 1) is an effective medium for manipulation skill transfer, and 2) takes better effect when narrowing the embodiment gap and reducing action noises.

Model	Microwave Tasks		Drawer Tasks		Mug / Cup Tasks		Other Tasks		Overall	
	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)
Ours (Human→Robot)	64.58	45.83	56.88	43.75	52.50	15.63	61.67	50.00	59.75	38.33
Gap-Free Ref. (Robot→Robot)	**68.75**	**54.17**	**75.94**	**62.50**	**81.25**	**53.13**	**71.25**	**58.33**	**73.54**	**55.83**

5.5 Failure Case Analysis

Failure cases mainly arise in tasks requiring a precise end-effector contact. The most representative failures arise in tasks like "*inserting the straw into the cup*" and "*opening the drawer*", where the robot often exhibits clear task intent yet fails at critical steps, such as a near-miss grasp or a failure to rotate the wrist for proper grasping, as shown in Fig. 6. These failures are consistent with our design choice: instead of naively learning noisy 6-DoF human actions, we sought to learn strong task intent and a robust action prior to enable more resilient transfer of manipulation knowledge.

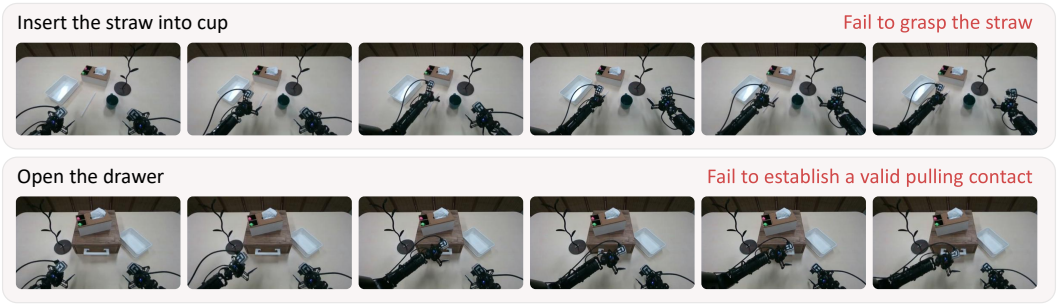


Figure 6: **Failure case analysis.** Representative failures exhibits clear task intent yet fails at critical steps, such as a near-miss grasp or a failure to rotate the wrist for proper grasping.

6 Conclusions

Limitations and Future Work. Though our translation-based bridging action enables a more robust manipulation skill learning from noisy human actions, it is somewhat limited to tasks that do not require fine rotational adjustments (Sec. 5.5). We attribute much of the effectiveness of our translation-only bridging action to the limitations of current capture devices, which cannot provide wrist-pose estimates reliable enough at scale. While high-precision in-lab human-action capture devices can indeed yield more reliable 6-DoF human actions, our flexible architecture and unified and interleaved action sequence enable swift switching among different combinations of action spaces for different data sources. We attribute this line of work to possible future directions.

Conclusions. In this work, we investigate whether we can learn manipulation skills from human actions. We propose a translation-based bridging action representation that is shared between human and robot action data, and further introduce an interleaved action-sequence representation to handle the action components that may be missing from either data source. Across 15 real-world bi-manual manipulation tasks, our translation-only bridging action transfers manipulation knowledge from humans to robots more robustly than naively supervising noisy 6DoF human actions (22.50% vs. 12.50% success), and it scales effectively with large-scale human-only pre-training, further raising the overall success rate to 38.33%. These results suggest that a simple, carefully chosen shared action space, rather than a faithful reconstruction of full human hand motion, is key to unlocking scalable manipulation-skill transfer from human data.

AI USE STATEMENT

In this work, we used generative AI for basic translation and sentence polishing for readability, including reorganizing sentence order and logical flow. We also used image segmentation models to assist in producing teaser figures (Fig. 1) and highlighting specific elements in qualitative comparisons (Fig. 5). We did not use generative AI tools to generate synthetic datasets; to develop theoretical models or conceptual frameworks; to formulate mathematical claims or assist in writing proofs; to propose or refine hypotheses; to design or provide feedback on our research methodology or experiments; to implement methods; to clean or reformat datasets. All AI-assisted text and figures were reviewed and verified by the authors, who take full responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

This work studies manipulation skill transfer from human action data to a bi-manual robot. The intended use is robot-learning research on tabletop tasks. It is not a surveillance system, a person-identification system, or a method aimed at harmful physical applications.

Human action data is part of the training dataset. Adult operators performed manipulation while wearing a PICO 4 Ultra Enterprise headset. Both the in-lab and the in-the-wild data were captured in laboratory rooms furnished as household environments rather than private residences. The in-lab data come from one laboratory room with fixed lighting and a fixed layout. The in-the-wild data were collected across multiple laboratory rooms, with more varied furniture layouts and lighting. Before collection, every operator was told what the headset would record and that the recordings would be used only for this research. We use these recordings only for the work reported in this paper, and we will not release raw videos, pose trajectories, or other records that could identify an operator or a recording site. We also use about 70 hours of selected tasks from the public EgoDex dataset, under its license.

REPRODUCIBILITY STATEMENT

In the main paper, we fully specify how raw human motion is captured and converted into our action space, including the acquisition setup and collection standard instructions on operators (Sec. 3), and the exact transformations from raw wrist/camera poses to the bridging action and the 6DoF robot action (Sec. 4.1, Eq. 1–2, Table 1). We also detail our model design and training recipe, including the Mixture-of-Transformer architecture, the interleaved action-token design, the flow-matching objective and its inference-time step size $\Delta\tau$, and the VLM-action co-training objective (Sec. 4.2), as well as batch sizes, running iterations, and the KV-cache-repeat training trick for each training stage (Sec. 4.3). Finally, Appendix. A gives the definitions, scoring rubrics, and scene layouts for all 15 evaluation tasks, enabling our evaluation protocol to be replicated.

REFERENCES

Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13778–13790, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, et al. A careful examination of large behavior models for multitask dexterous manipulation. *arXiv preprint arXiv:2507.05331*, 2025.

Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision*, pp. 306–324. Springer, 2024.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

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.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

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.

Xiongyi Cai, Ri-Zhao Qiu, Geng Chen, Lai Wei, Isabella Liu, Tianshu Huang, Xuxin Cheng, and Xiaolong Wang. In-n-on: Scaling egocentric manipulation with in-the-wild and on-task data. *arXiv preprint arXiv:2511.15704*, 2025.

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.

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.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 720–736, 2018.

Shichao Fan, Kun Wu, Zhengping Che, Xinhua Wang, Di Wu, Fei Liao, Ning Liu, Yixue Zhang, Zhen Zhao, Zhiyuan Xu, et al. Xr-1: Towards versatile vision-language-action models via learning unified vision-motion representations. *arXiv preprint arXiv:2511.02776*, 2025.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017a.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017b.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

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

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.

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.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233. IEEE, 2025a.

Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. Emergence of human to robot transfer in vision-language-action models. *arXiv preprint arXiv:2512.22414*, 2025b.

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.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

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.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024b.

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.

Fanqi Lin, Kushal Arora, Jean Mercat, Haruki Nishimura, Paarth Shah, Chen Xu, Mengchao Zhang, Mark Zolotas, Maya Angeles, Owen Pfannenstiehl, et al. A systematic study of data modalities and strategies for co-training large behavior models for robot manipulation. *arXiv preprint arXiv:2602.01067*, 2026.

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

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, 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.

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.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640, 2019.

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.

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.

Tony Tao, Mohan Kumar Srirama, Jason Jingzhou Liu, Kenneth Shaw, and Deepak Pathak. Dexwild: Dexterous human interactions for in-the-wild robot policies. *Robotics: Science and Systems (RSS)*, 2025.

Generalist AI Team. Gen-0: Embodied foundation models that scale with physical interaction. *Generalist AI Blog*, 2025. https://generalistai.com/blog/nov-04-2025-GEN-0.

Generalist AI Team. Gen-1: Scaling embodied foundation models to mastery. *Generalist AI Blog*, 2026. https://generalistai.com/blog/apr-02-2026-GEN-1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023.

Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *Advances in neural information processing systems*, 37:124420–124450, 2024.

Zhenyu Wei, Yunchao Yao, and Mingyu Ding. One hand to rule them all: Canonical representations for unified dexterous manipulation. *arXiv preprint arXiv:2602.16712*, 2026.

Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023.

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.

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.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.

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.

Chengbo Yuan, Rui Zhou, Mengzhen Liu, Yingdong Hu, Shengjie Wang, Li Yi, Chuan Wen, Shanghang Zhang, and Yang Gao. Motiontrans: Human vr data enable motion-level learning for robotic manipulation policies. *arXiv preprint arXiv:2509.17759*, 2025.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, et al. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026.

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Lawrence Y Zhu, Pranav Kuppili, Ryan Punamiya, Patcharapong Aphiwetsa, Dhruv Patel, Simar Kareer, Sehoon Ha, and Danfei Xu. Emma: Scaling mobile manipulation via egocentric human data. *IEEE Robotics and Automation Letters*, 2026.

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

We here provide additional details for the main paper, including more details for real-world evaluation (Sec. A), and additional results and visualizations (Sec. B).

A EVALUATION DETAILS

A.1 EVALUATION TASKS

We evaluate our policy on 15 manipulation tasks, which are grouped according to their manipulated objects (microwave, drawer, mug/cup, and others), as shown in Fig. 3. For each task, we further visualize the desired manipulation procedure using keyframes in Fig. 7, where we primarily show the *left→right* variants and omit the *right→left* tasks, as they are symmetrical.

1. **Open the microwave door.** The robot reaches for the microwave door handle, grasps it, and pulls it outward till the door is opened.
2. **Close the microwave door.** The robot pushes the microwave door until it is fully closed.
3. **Take the bowl out of the microwave.** The robot reaches inside the open microwave, picks the bowl up, and places it on the table while avoiding collisions.
4. **Place the bowl into the microwave.** The robot picks up the bowl from the table and places it inside the open microwave while avoiding collisions.
5. **Wipe the microwave top from left to right.** The robot grasps the towel on the left top of the microwave and pushes it towards the right top.
6. **Wipe the microwave top from right to left.** The robot grasps the towel on the right top of the microwave and pushes it towards the left top.
7. **Open the drawer.** The robot reaches for the drawer handle, grasps it, and pulls it outward.
8. **Close the drawer.** The robot pushes the drawer until it is fully closed.
9. **Hang the left mug on the mug holder.** The robot picks up the left mug on the table and hangs it onto the corresponding side of the mug holder.
10. **Hang the right mug on the mug holder.** The robot picks up the right mug on the table and hangs it onto the corresponding side of the mug holder.
11. **Stack the left cup onto the other cup.** The robot picks up the left cup from the table and stacks it onto the right cup.
12. **Stack the right cup onto the other cup.** The robot picks up the right cup from the table and stacks it onto the left cup.
13. **Insert the straw into the cup.** The robot picks up a straw from the table and inserts it vertically into a cup.
14. **Take the toast from the toaster and set it on a plate.** The robot picks the toast from the toaster and places it on a plate.
15. **Unplug the charger.** The robot grasps the charger from the power strip, pulls it out, and places it on the table.

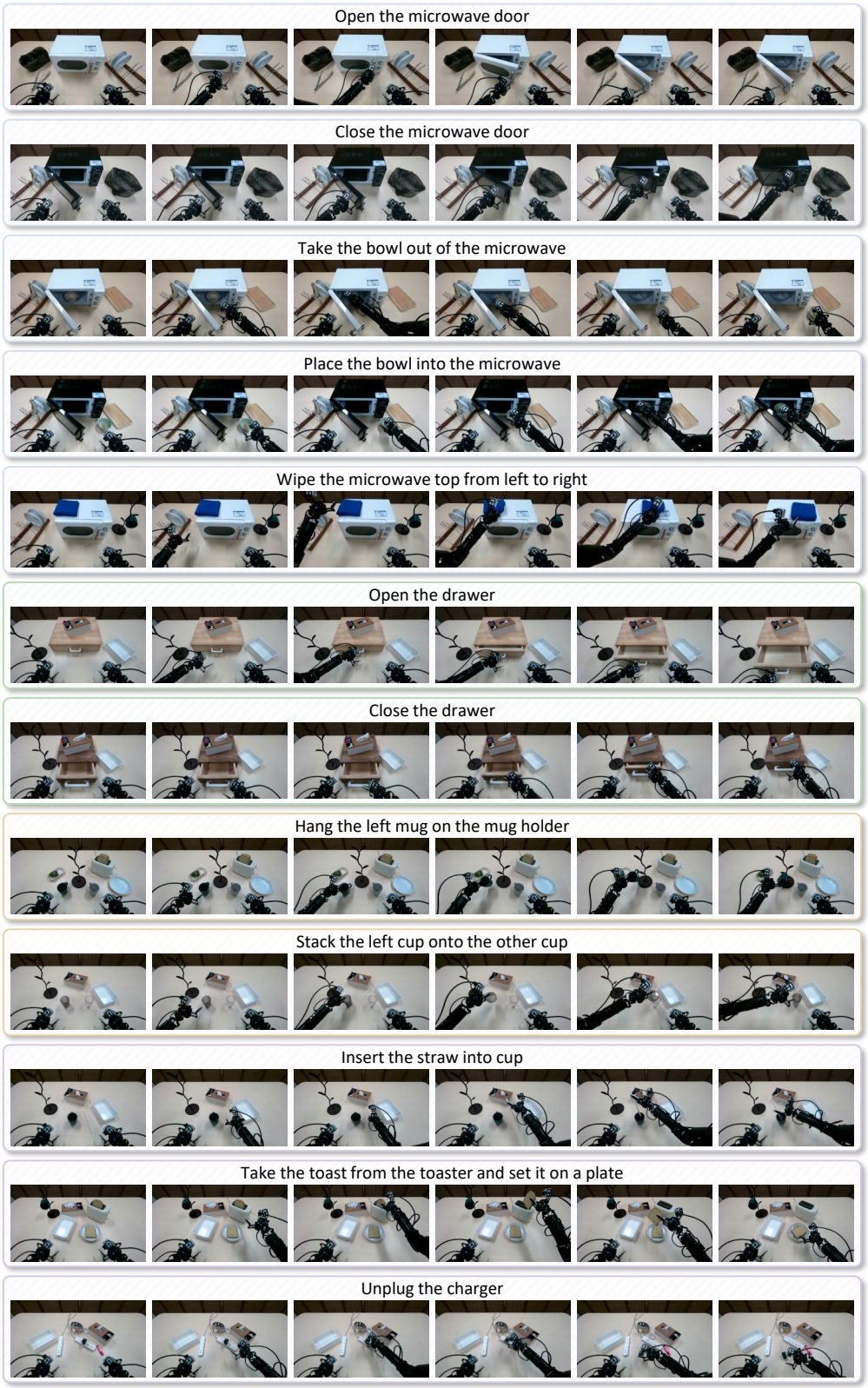


Figure 7: **Keyframe visualizations of desired manipulation procedure.** We primarily show the *left→right* variants and omit the *right→left* tasks, as they are symmetrical.

A.2 PROGRESS RUBRICS

We also provide per-task progress rubrics in Fig. 8 for fine-grained evaluation, and additionally illustrate the two distinct test layouts for each task.

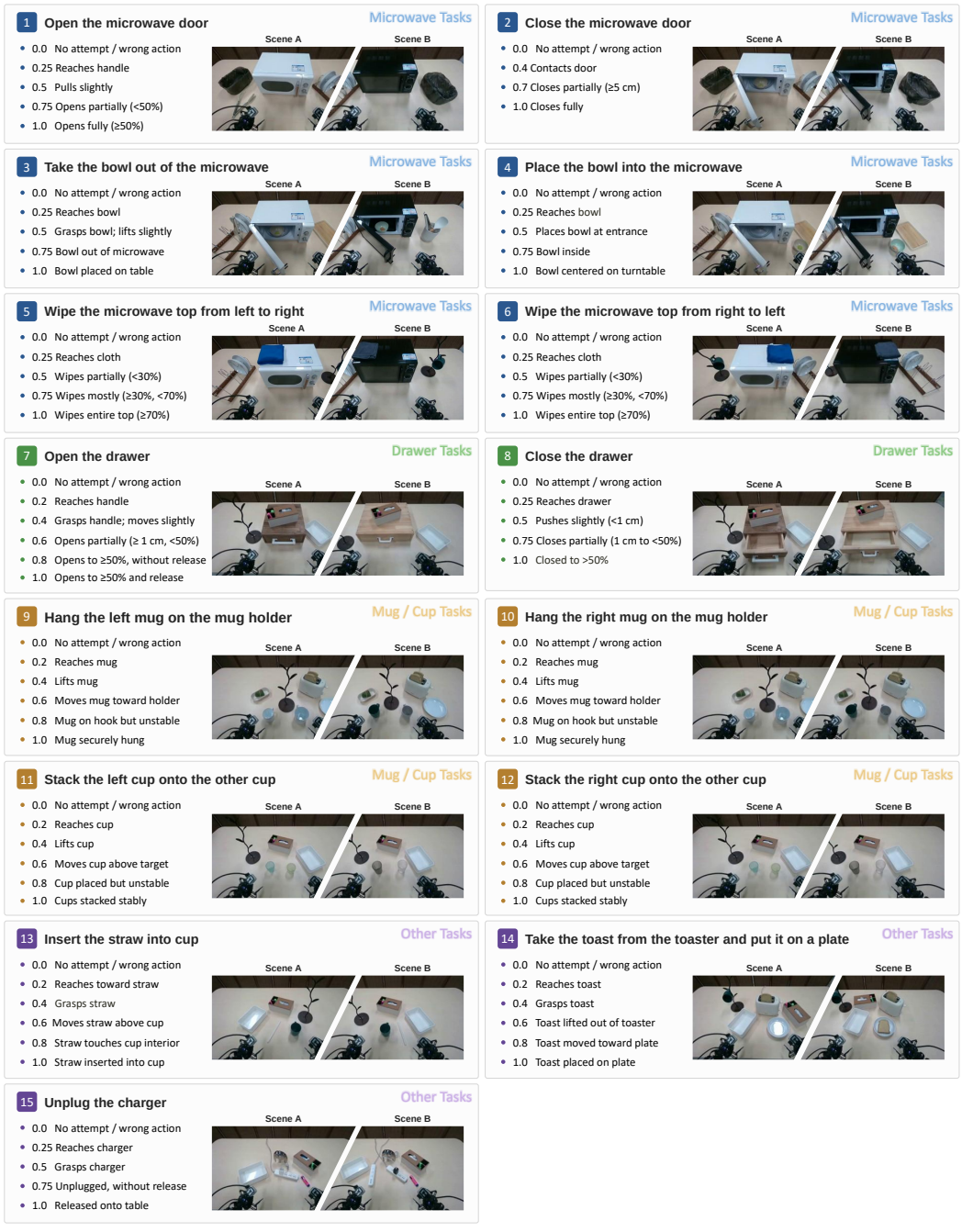


Figure 8: **Per-task progress rubrics.** We experiment with 15 manipulation tasks, categorized with respect to their manipulation objects. For each task, we adopt two distinct testing layouts and a detailed progress scoring criterion.

B Additional Evaluation Results

In addition to the experiment section in the main paper, we provide two additional studies:

- **Q5.** Does the bridging action pre-training align with 6DoF robot action generation? (Sec. B.1)
- **Q6.** Does human-only pre-training on bridging actions improve robot post-training? (Sec. B.2)

We also show per-task performance for Sec. 5.1 and 5.4, and further qualitative results for Sec. 5.2.

B.1 The bridging action aligns with 6DoF robot actions

We analyze why the manipulation knowledge can transfer from translation-only bridging actions to the full 6DoF robot action from both learning and representation perspective.

Pre-training on bridging action accelerates 6DoF action learning. We first present the training loss of each action component during human-robot co-training in Fig. 9, comparing a model initialized from scratch (red) against one initialized from human-only pre-training (blue). Although human-only pre-training only exposes the model to the non-executable bridging action, it yields a lower training loss for both $\mathbf{a}_t^{\text{6D-eef}}$ and the $\mathbf{a}_t^{\text{gripper}}$ during co-training. This implies a similar objective landscape between the bridging action $\mathbf{a}^{\text{3D-wrist}}$ and the robot actions, which is also consistent with the findings in Sec. 5.1 that human-only pre-training improves downstream task success and progress.

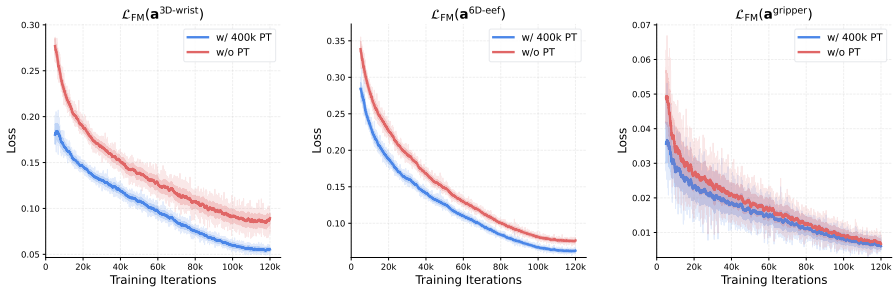


Figure 9: **Training loss comparison during human-robot co-training.** Although human-only pre-training supervises only wrist translations, it accelerates the convergence of the other action components, *i.e.*, the 6DoF end-effector action and the gripper action.

The predicted bridging action aligns with the predicted 6DoF action. We further visualize both the predicted bridging action $\mathbf{a}^{\text{3D-wrist}}$ (yellow) and the 6DoF end-effector action $\mathbf{a}^{\text{6D-eef}}$ (orange) by prompting the model with the same language and observation inputs and then projecting both onto the head camera in Fig. 10. We observe that the two types of actions align closely across diverse tasks for both arms, indicating that $\mathbf{a}^{\text{3D-wrist}}$ can serve as a reliable proxy for the robot action when learning transferable manipulation knowledge.

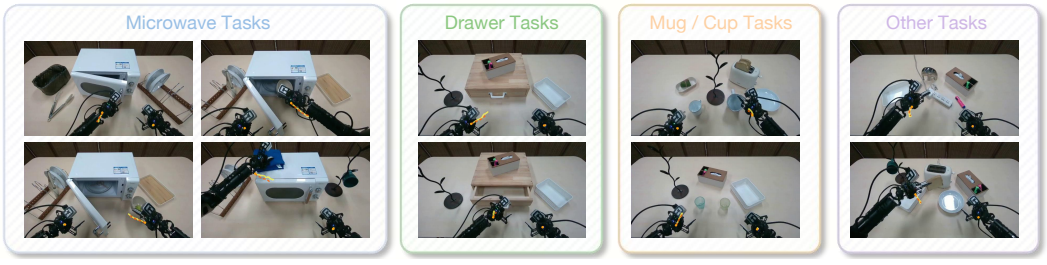


Figure 10: **Visualization of the predicted bridging action (yellow) and 6DoF action (orange).** Based on the same vision and language input, we prompt the model to produce bridging actions and end-effector actions, and then project both actions onto the head camera. The two types of actions align closely, confirming the bridging action to be a reliable proxy for real robot action learning.

B.2 HUMAN-ONLY PRE-TRAINING IMPROVES ROBOT POST-TRAINING

On a subset of 10 robot trajectories (same post-training subset in Fig. 4, purple) for each of the 15 evaluation tasks collected in Sec. 4.3, we compare few-shot post-training with and without human-only pre-training in Tab. 5. Same as before, we only supervise the model with $\mathcal{L}_{\text{FM}}^{\text{3D-wrist}}$ during human-only pre-training (w/ PT.). Results show that human-only pre-training with bridging actions substantially improves the few-shot robot learning efficiency, increasing success rates from an average of 35.83% to 55.00%.

Table 5: **Few-shot robot post-training.** Compared to training from scratch (w/o PT.), human-only pre-training (w/ PT.) on bridging actions substantially improves the few-shot robot learning efficiency.

Method	Microwave Tasks		Drawer Tasks		Mug / Cup Tasks		Other Tasks		Overall	
	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)	Prog.(%)	Succ.(%)
w/o PT.	71.77	58.33	**56.88**	**43.75**	37.50	6.25	37.50	25.00	53.79	35.83
w/ PT.	**80.73**	**68.75**	44.38	25.00	**71.25**	**46.88**	**70.00**	**58.33**	**71.21**	**55.00**

B.3 PER-TASK PERFORMANCE BREAKDOWN FOR MAIN RESULTS

We provide further per-task breakdown results corresponding to Fig. 4 in Fig. 11. We show that training only on robot pick-and-place data (green) is not enough for our downstream evaluation tasks, achieving near zero success. However, our human-robot co-training with bridging actions (orange) shows an effective manipulation knowledge transfer from human data to robot actions. The bridging action further benefits from large-scale human-only pre-training (blue) and from additional 10-shot robot trajectories (purple).

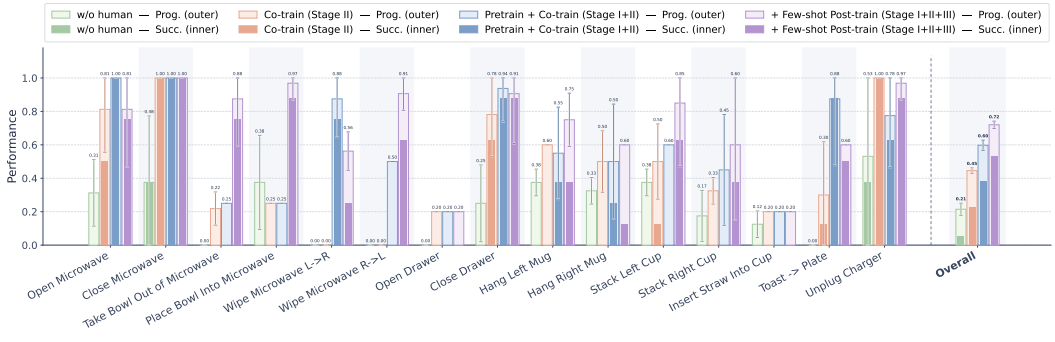


Figure 11: **Per-task performance breakdown across training settings.** Training only on robot pick-and-place data (green) achieves near zero success, while co-training on human and robot actions with bridging action (orange) enables manipulation knowledge transfer from humans to robots. The bridging action is a scalable action representation that can benefit from large-scale human-only pre-training (blue) and from additional few-shot robot demonstrations (purple).

B.4 PER-TASK PERFORMANCE BREAKDOWN FOR THE GAP-FREE DIAGNOSTIC

In the main paper, we construct a diagnostic variant by co-training the model on all action components for robot pick-and-place data while generating only bridging actions ($\mathbf{a}^{\text{3D-wrist}}$) for **task-specific robot demonstrations** (100 traj./task), isolating the observation gap (no wrist camera, differing viewpoint) and human action noise to analyze the effectiveness of bridging action.

Here, we provide per-task performance breakdown in Fig. 12, corresponding to the grouped results in Tab. 4. Applying the same training objective employed for human actions to task-specific in-lab robot actions eliminates observation gaps and action noise, yielding performance gains on tasks that demand precise manipulation and where the workspace can be occluded by end-effectors, *e.g.* "*stack cups*", "*hang mug*", and *etc.*

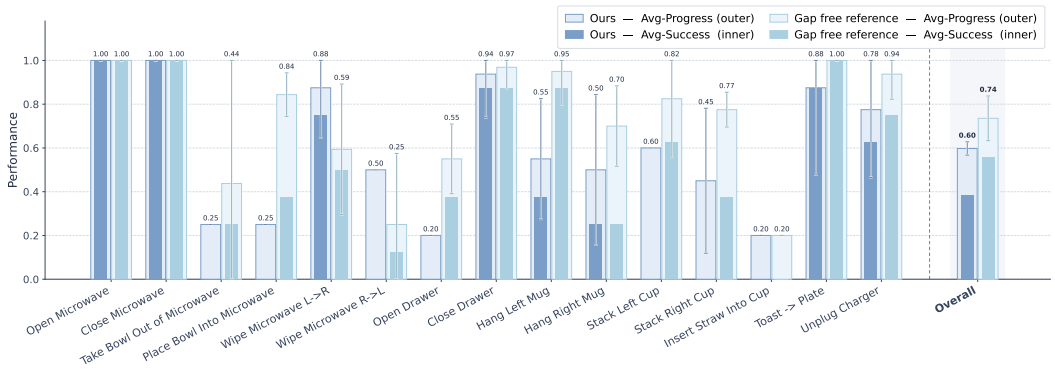


Figure 12: **Per-task comparison for the gap-free reference diagnostic.** After isolating the observation gap and human action noise by training bridging actions ($\mathbf{a}^{\text{3D-wrist}}$) for task-specific robot demonstrations (100 traj./task), we see performance gains on tasks that demand precise manipulation and where the workspace can be occluded by end-effectors.

B.5 Additional qualitative comparisons with 6DoF human actions

We here provide additional qualitative comparisons on robot rollouts for 1) human-robot co-trained on bridging data, and 2) naive 6DoF human actions in Fig. 13. In addition to the quantitative results in Sec. 5.2, given the noisy human actions, learning from bridging action leads to more robust manipulation skill transfer.



Figure 13: **Qualitative comparisons on co-training with different human actions.** Co-training with bridging actions (green) enables more stable manipulation behaviors than the 6DoF human action baseline (orange).