BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Nov 2025

OXE-AugE: A Large-Scale Robot Augmentation of OXE for Scaling Cross-Embodiment Policy Learning

Guanhua Ji[*1,2], **Harsha Polavaram**[*1], **Lawrence Yunliang Chen**[*1], **Sandeep Bajamahal**[1], **Zehan Ma**[1], **Simeon Adebola**[1], **Chenfeng Xu**[1,3] **and Ken Goldberg**[1]

[1]Department of EECS, UC Berkeley, [2]Department of CIS, University of Pennsylvania, [3]Department of CS, UT Austin

Figure 1: We present OXE-AugE, a large-scale open-source dataset that augments the Open-X Embodiment (OXE) dataset [18] with 9 different robot embodiments across 16 datasets, covering 60% of the widely-used Octo pretraining mixture [69]. In total, OXE-AugE provides over 4 million trajectories, more than triple those in the original OXE. Robots in OXE-AugE include Panda, UR5e, Xarm7, Google robot, widowX, Sawyer, Kinova3, IIWA, and Jaco. We find that training on OXE-AugE improves OpenVLA [46] and π_0 [5] policy performance by up to 24-45% on previously unseen robot-gripper combinations across four real-world manipulation tasks.

Large and diverse datasets are the fuel for training generalist robot policies that have potential to control a variety of robot embodiments—robot arm and gripper combinations—across diverse tasks and environments. Because re-collecting demonstrations and retraining for each new hardware platform are prohibitively costly, we conjecture that existing robot data can be augmented for transfer and generalization. The Open X-Embodiment (OXE) dataset, which aggregates demonstrations from over 60 datasets, has been widely used as the foundation for training generalist policies. However, it is highly imbalanced: the top four robot types account for over 85% of its real data, which risks overfitting to robot–scene combinations. We present OXE-AugE, a high-quality open-source dataset that augments OXE with 9 different robot embodiments. OXE-AugE provides over 4.4 million trajectories, more than triple the size of the original OXE. We conduct a systematic study of how scaling robot augmentation impacts cross-embodiment learning. Results suggest that augmenting datasets with diverse arms and grippers improves policy performance not only on the augmented robots, but also on unseen robots and even the original robots under distribution shifts. In physical experiments, we demonstrate that state-of-the-art generalist policies such as OpenVLA and π_0 benefit from fine-tuning on OXE-AugE, improving success rates by 24–45% on previously unseen robot–gripper combinations across four real-world manipulation tasks. Project website: https://OXE-AugE.github.io/.

Correspondence: Lawrence Yunliang Chen (yunliang.chen@berkeley.edu), Chenfeng Xu (xuchenfeng@utexas.edu).

1. Introduction

Large and diverse datasets have been key to recent progress in general-purpose robot learning, where policies trained on broad experience can generalize to new tasks, objects, and embodiments [39, 10, 9, 40, 87, 88, 59, 91, 93, 92, 79, 76, 4, 15, 25]. Although performance improves with scale, collecting real-world robot data remains costly and time-consuming [49, 34, 42, 39, 10, 44, 29, 86]. The total data volume that can be collected over time is constrained by both the number of teleoperated robots and the duration required to execute and record each trajectory. While simulation offers a promising path to scale [67, 60, 20], the sim-to-real gap in dynamics and perception of manipulation remains a major challenge [72, 77, 57, 65, 7, 35].

This challenge is amplified as robotic hardware becomes more diverse. New robots with different kinematics, sizes, and grippers are regularly introduced, and policies trained on one platform often fail to transfer to others. Given the high cost of recollecting demonstrations for every new platform, it is desirable to reuse existing data across embodiments. Cross-embodiment generalization—the ability to transfer policies across different robot embodiments—is thus an important goal for scalable and practical robot learning [106, 24, 5].

The Open X-Embodiment (OXE) dataset [18], released in 2023, is a major step in this direction. It aggregates demonstrations from over 60 real-world robot datasets collected across different labs, platforms, and tasks. However, most constituent datasets are tied to a single robot in a fixed environment, risking overfitting to robot–scene combinations. Moreover, OXE is highly imbalanced: over 85% of real trajectories come from just four robots (Franka, xArm, Kuka iiwa, and Google Robot), while many others appear in only 1–2 datasets. Consequently, training a generalist policy on OXE relies on the hope that the policy will implicitly learn embodiment-agnostic features, without explicit mechanisms to mitigate robot bias. In practice, many generalist policies such as Octo [69], OpenVLA [46], GR00T [68], and π_0 [5] still require finetuning on new robots, even when they are visually or kinematically similar to those in the training data.

Chen et al. [13] propose robot embodiment augmentation—transforming demonstrations collected on one robot into synthetic versions as if performed by another embodiment—using a process called *cross-painting* [12]. In this work, we improve cross-painting and present AugE-Toolkit, which reduces visual artifacts, improves speed and scalability, while ensuring kinematically valid trajectories using a combination of simulation and learned models.

Using AugE-Toolkit, we generalize robot augmentation beyond pairwise transfer and apply it as a scalable data pipeline. Specifically, we study the effect of scaling robot augmentation and ask: (1) Can robot augmentation improve robustness on the original robot under visual perturbations? (2) Does increasing the number of robot augmentations improve performance on augmented robots? (3) Do policies trained on diverse augmentations generalize to unseen robots? Results suggest that scaling robot augmentation leads to consistent gains, particularly for generalization to unseen embodiments and visual perturbations. We conjecture that robot augmentation helps policies focus on the spatial geometry between gripper and object, rather than incidental visual features like arm shape or color.

We present **OXE-AugE**, a high-quality open-source dataset that augments 16 popular OXE datasets with 9 different robot embodiments. The resulting dataset provides over 4 million trajectories—more than triple the size of the original OXE—and covers 60% of the widely used Octo pretraining mixture. By varying the robot embodiment while preserving task and scene, OXE-AugE provides a new resource for training robust and transferable visuomotor policies.

We evaluate whether state-of-the-art generalist models such as π_0 [5] and OpenVLA-OFT [47] can benefit from OXE-AugE. We fine-tune each model on the augmented dataset and evaluate them on a real Franka robot equipped with two distinct grippers. Across four manipulation tasks, we observe 24–45% improvements in success on previously unseen robot–gripper configurations, suggesting the practical utility of large-scale robot augmentation.

This paper makes 4 contributions:

1. AugE-Toolkit, an improved and easy-to-use robot augmentation pipeline that enables scalable, high-quality augmentation.
2. OXE-AugE, a large open-source dataset that augments OXE with 9× more robot embodiments across

16 datasets, totaling over 4 million trajectories and covering 60% of the Octo pretraining mixture.
3. A simulation study of how scaling robot augmentation affects generalization to both seen and unseen embodiments, and robustness to visual perturbations.
4. Physical experiments suggesting that fine-tuning foundation models on OXE-AugE can improve zero-shot success by 24–45% on novel robot embodiments.

2. Related Work

2.1. Cross-Embodiment Robot Learning

A core challenge in generalist robot learning is how to generalize across robot embodiments without collecting new data for each platform. One common approach is domain randomization, where physical parameters of the robot (e.g., joint and link properties) are randomized in simulation to learn robot-conditioned policies [112, 14, 90, 104, 101, 82, 71, 38, 48]. Hu et al. [37] learn a world model with the robots masked out, and use visual MPC during execution time when deployed on a new robot. Another line of work explores using human data. Recent efforts have leveraged human videos [103, 3, 26, 52, 43] for robot manipulation, and motion retargeting methods [33, 56, 2, 111, 108] have been applied in locomotion settings. Other work has also explored pooling large and diverse data, including from different robots [41, 53, 22, 28, 29, 99, 8] and found that the resulting policies are generalizable to new tasks and embodiments [1, 39, 93, 40, 79, 76, 87, 106, 4, 10, 9, 15, 25].

The Open X-Embodiment project [18], in particular, aggregated more than 60 datasets and demonstrated the benefit of training on various embodiments through experiments in multiple labs. Many have leveraged the OXE dataset and developed "generalist policies" that can perform multiple tasks on a wide range of robots [69, 46, 5, 107, 6, 68, 102]. However, the robot types in OXE are severely unbalanced, and trained policies typically perform much better on robots that are well-represented in the datasets and still require a fair amount of finetuning data to transfer to a new robot. Mirage [12] proposes a test-time image inpainting pipeline to replace the new target robot in the image with the familiar source robot seen during training to achieve zero-shot cross-embodiment transfer. In this work, we inpaint the opposite direction and improve the pipeline for scalable data augmentation.

2.2. Augmenting Real Robot Data

Given the high cost of collecting real robot data, many approaches have explored augmenting existing datasets. Real-to-sim-to-real pipelines [57, 98, 97, 54, 110, 74, 21, 31, 119] reconstruct 3D object meshes from real videos and tune simulation parameters to build digital twins or "digital cousins" [20] for policy learning. Some works [113, 109] avoid simulating the physics but still require an extra step of taking multi-view images of the scene or objects to build 3D Gaussian Splatting (3DGS). Many works that leverage 3DGS for rerendering and trajectory synthesis [120, 70, 118] are most applicable to eye-in-hand images.

Another category of work applies 2D image or video generation models to augment directly in image space. Examples include editing backgrounds or objects [114, 16, 61, 4], or synthesizing novel views [13, 95]. For robot transfer, Shadow [50] masks out robots in images; RoVi-Aug [13] applies diffusion models to transform one robot appearance into another; and Phantom [52] and Masquerade [51] extend this idea to enable robot learning from human videos. However, most of these works focus on one-to-one transfer between a known source and target embodiment. In contrast, we study *scaling* robot augmentation across many target embodiments and investigate how such augmentation impacts generalization and robustness in both simulation and real-world settings.

2.3. Study of Scaling in Robot Learning

Several recent studies have analyzed how performance in robot learning scales with data volume and model capacity [83]. Models like VIMA [40], RT-1 [10], Octo [69], and HPT [100] report consistent gains when increasing dataset size and model scale. For data scaling, Lin et al. [58] show approximate power-law improvements in generalization as the number of environments and objects increases, with diversity often more valuable than additional demonstrations per setting. Saxena et al. [85] further dissect scaling effects by quantifying contributions from camera viewpoint, spatial layout, and object variety. For robot embodiment

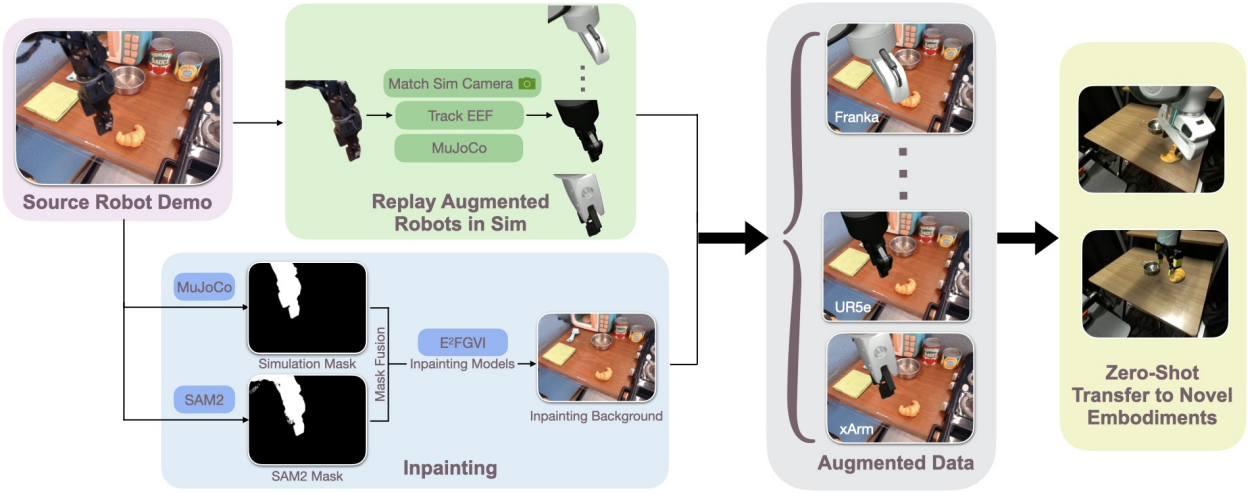

Figure 2: AugE-Toolkit pipeline. Given a source robot image and its corresponding robot poes, AugE-Toolkit fuses a learned SAM2 mask with a simulation-rendered mask to segment the robot, inpaints the background via E^{2}FGVI [55], and replays the same trajectory with another robot in simulation [116]. The augmented robot is composited into the reconstructed scene to form the augmented video. Policies are trained on both real and augmented data and evaluated on unseen embodiments.

scaling, OXE [18], RoboCat [8], and CrossFormer [24] find that training on multiple robot types leads to better transfer than training on a single target robot alone. Yang et al. [107] show benefits from jointly training across manipulation and navigation domains. In this work, rather than pooling additional real data from multiple robots, we study the effect of scaling *robot augmentation*—synthetically generating data from multiple robot embodiments—on performance across original, augmented, and entirely unseen robot configurations.

3. Problem Statement

We consider the standard imitation learning setting [75, 81], where we have a demonstration dataset $\mathcal{D}^{\mathcal{S}} = \{\tau_1^{\mathcal{S}}, \tau_2^{\mathcal{S}}, ..., \tau_n^{\mathcal{S}}\}$ consisting of n successful trajectories performed by a source robot $\mathcal{S}$. Each trajectory $\tau_i^{\mathcal{S}} = (o_{1:H_i}^{\mathcal{S}}, p_{1:H_i}^{\mathcal{S}}, a_{1:H_i}^{\mathcal{S}})$ contains RGB observations $o_t^{\mathcal{S}}$, corresponding gripper 6D poses $p_t^{\mathcal{S}}$, and actions $a_t^{\mathcal{S}}$ for timesteps $t = 1, \dots, H_i$.

We study *robot augmentation*—synthetically transforming each trajectory $\tau_i^{\mathcal{S}}$ into a corresponding trajectory $\tau_i^{\mathcal{R}}$ for a different robot embodiment $\mathcal{R}$ (arm and gripper) performing the same task. Given $\mathcal{D}^{\mathcal{S}}$ and the kinematic models of the robots (e.g., URDF), this yields a synthetic dataset $\mathcal{D}^{\mathcal{R}}$ with aligned images, poses, and actions $(o_{1:H_i}^{\mathcal{R}}, p_{1:H_i}^{\mathcal{R}}, a_{1:H_i}^{\mathcal{R}})$. Following prior work [13], we assume the grippers across robots are similar in shape and function (e.g., 2- or 3-jaw grippers), enabling all robots to perform the same task with a shared strategy. Similar to prior work [87, 12, 106, 107], we use Cartesian control and assume that the coordinate frames of all robots are known, allowing alignment through rigid transformations.

Let $\mathcal{R}_{\text{Aug}} = \{\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_N\}$ denote the set of robot embodiments used for augmentation. We train a policy on the resulting augmented dataset $\mathcal{D}^{\text{Aug}} == \bigcup_{i=1}^{N} \mathcal{D}^{\mathcal{R}_i}$ and evaluate it on a target robot $\mathcal{T}$ at test time. We examine how robot augmentation affects: (1) **Robustness:** $\mathcal{T} = \mathcal{S}$ (original source robot), (2) **Transfer:** $\mathcal{T} \in \mathcal{R}_{\text{Aug}}$ (performance on augmented robots), and (3) **Generalization:** $\mathcal{T} \notin \mathcal{R}_{\text{Aug}}$ (unseen robots). We further study how these outcomes vary as we scale the number of augmented embodiments, from a single target robot ($|\mathcal{R}_{\text{Aug}}| = 1$) to N distinct robot types.

4. Methods

4.1. Preliminaries: Cross-Painting Framework

Cross-painting [12, 13, 52] is a three-stage pipeline applied to each image in a robot trajectory: (i) *source-robot segmentation*, (ii) *background inpainting*, and (iii) *augmented-robot replay and compositing*. The goal is to replace the robot embodiment in each frame while preserving task and scene context.

Existing implementations differ mainly in how each stage is realized. Learning-based methods use diffusion models to modify the robot directly in pixel space [13], requiring no explicit calibration and producing visually realistic results. However, they lack kinematic guarantees and scale poorly to many target robots, since each new embodiment typically requires a separately trained model. Simulation-based methods [12] render the robot using known camera parameters, ensuring geometric fidelity. These methods were originally designed for test-time adaptation, where accurate calibration is feasible, but are difficult to apply to large-scale offline datasets that often lack such information.

We propose AugE-Toolkit, which builds on this framework to enable scalable robot augmentation. It retains the physical accuracy of simulation-based rendering but is also applicable to uncalibrated or coarsely aligned data. Through mask fusion and semi-automatic base tuning, AugE-Toolkit synthesizes physically consistent robot augmentations across many embodiments with minimal manual effort.

4.2. AugE-Toolkit: Scalable Robot Augmentation

AugE-Toolkit (Fig. 2) extends cross-painting in RoVi-Aug [13] with three key components.

(1) Fusion of Simulation and Learned Masks. To obtain accurate robot masks without camera calibration, we combine simulation-based and learned segmentation. We fine-tune SAM2 [78] on a small labeled subset (20 trajectories from each of 16 OXE datasets). While the learned masks align well with image appearance, they often over- or under-segment near the gripper. Simulation masks are geometrically accurate but may be globally misaligned due to unknown camera poses. We align and fuse them in three steps: (1) *Translation alignment*: shift the simulation mask within a small grid to maximize IoU with the learned mask; (2) *Distance pruning*: remove learned-mask pixels farther than a threshold τ from the aligned simulation boundary; (3) *Union and smoothing*: combine both masks and apply morphological closing. This fusion process corrects for calibration errors and enables accurate rendering even on uncalibrated datasets. Downstream training can also flag and filter data whose final masks' IoU are large.

(2) Automatic Base Position Tuning. To accommodate robots with different kinematic reach and scale, we automatically adjust the base position of each target robot in simulation to ensure all end-effector poses from the source trajectory are reachable. Starting from an initial base pose, we iteratively sample offsets $\pm\Delta$ along the (x, y, z) axes, compute tracking error, and halve the step size until the maximum error falls below 1 cm or a maximum iteration count is reached. Trajectories without feasible base positions are discarded.[1] This procedure enables consistent augmentation across compact arms (e.g., WidowX) and larger robots (e.g., Google Robot) while preserving motion fidelity.[2]

(3) Scalable Multi-Robot Deployment. We employ simulation rendering rather than generative synthesis to ensure temporal coherence and physical validity. Generative models often introduce flicker and geometric artifacts, whereas URDF-based rendering guarantees pose accuracy and realism. Our pipeline is implemented on MuJoCo Playground [116], which supports a large collection of robots [115]. Adding a new robot only requires registering its model; no retraining or calibration is needed.

Each stage of the pipeline is embarrassingly parallel. Since each target robot independently replays the same trajectory, augmentation across multiple robots can be performed concurrently. A 50-frame 640×480 clip completes in approximately 25 seconds per robot; with 32-way parallelism, throughput reaches up to 75 clips per minute.

5. Scaling Robot Augmentation: A Systematic Study in Simulation

We begin with a systematic simulation study to examine how robot augmentation scales in terms of transfer, generalization, and robustness. While prior work [13] has shown that augmenting demonstrations from a source robot to a known target enables zero-shot transfer, our goal is to investigate whether robot augmentation provides broader benefits when scaled across multiple target robots.

[1] In practice, two datasets (RT-1 Fractal and Language Table) have many trajectories that are unreachable by the WidowX robot; we exclude these from their augmentations. All other datasets are augmented into all 9 robot-gripper combinations, with >95% of trajectories achieving replay errors under 0.25 cm. See the appendix for details.

[2] For mobile robots, the optimized base translation can be interpreted as a movement action.

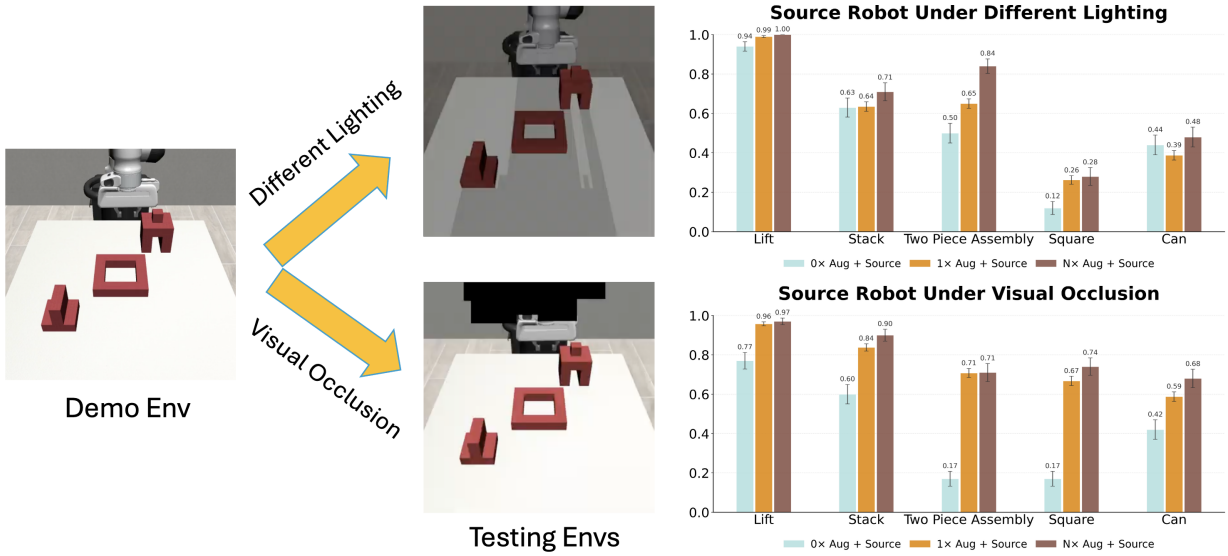

Figure 3: Robot augmentation improves robustness on the source robot under visual perturbations. Left: We consider two types of perturbations: different lighting conditions and visual occlusions. **Right:** Performance of policies trained on various augmented datasets on the Franka (source) robot. The performance of policies without augmentation severely degrades, while increasing the number of augmented robots makes policies more robust.

We follow the Mirage evaluation setup [12] and consider five Robosuite [124] tasks: LIFT, STACK, CAN, TWOPIECE ASSEMBLY, and SQUARE. For each task, we use 200 demonstrations from Robomimic [62] (LIFT, CAN, SQUARE) or MimicGen [63] (STACK, TWOPIECE ASSEMBLY), all collected on a Franka robot (source S = Franka). Using our pipeline, we augment the demonstrations into four additional robot embodiments: UR5e, Kinova Gen3, Sawyer, and Jaco (which has a 3-jaw gripper). We use Diffusion Policy [17] to train separate policies for each condition using RoboMimic [62]; see Appendix for training details.

Training configurations. We consider four data regimes:

- **No Augmentation ("0× Aug + Source"):** Train only on the original Franka demonstrations ($\mathcal{D}^{\text{Train}} = \mathcal{D}^{S}$).
- **1 Robot Augmentation ("1× Aug"):** Augment the source data into a single target robot $\mathcal{R}$ and train only on the augmented demonstrations ($\mathcal{D}^{\text{Train}} = \mathcal{D}^{\mathcal{R}}$). This is the standard setting studied in prior work.
- **1 Robot Augmentation Together with Source Data ("1× Aug + Source"):** Combine source and one target robot's data ($\mathcal{D}^{\text{Train}} = \mathcal{D}^{S} \cup \mathcal{D}^{\mathcal{R}}$).
- **Multi-Robot ("N× Aug + Source"):** Combine source data with all N augmented robots ($\mathcal{D}^{\text{Train}} = \mathcal{D}^{S} \cup \bigcup_{i=1}^{N} \mathcal{D}^{\mathcal{R}i}$, $\mathcal{R}_i \in \{\text{UR5e, Kinova Gen3, Sawyer, Jaco}\}$).

Evaluation protocols. Following Sec. 3, we evaluate each policy under three conditions:

- **Robustness (Source robot):** Evaluate on the original source robot ($\mathcal{T} = \mathcal{S}$) under visual perturbations (lighting shifts and occlusions). We compare policies trained on "0× Aug + Source," "1× Aug + Source," (averaged across 4 augmented robots) and "N× Aug + Source."
- **Transfer (Augmented robots):** Evaluate on robots used for augmentation ($\mathcal{T} \in \mathcal{R}_{\text{Aug}}$). We compare "1× Aug (Target)," "1× Aug (Target) + Source," and "N× Aug (incl. Target) + Source" settings.
- **Generalization (Unseen robots):** Evaluate on robots not used for augmentation ($\mathcal{T} \notin \mathcal{R}_{\text{Aug}}$). We compare "0× Aug + Source," "1× Aug (not Target) + Source," (averaged across 3 augmented robots) and "(N-1)× Aug (not incl. Target) + Source," where "N–1" excludes one hold-out robot from training and evaluates on it.

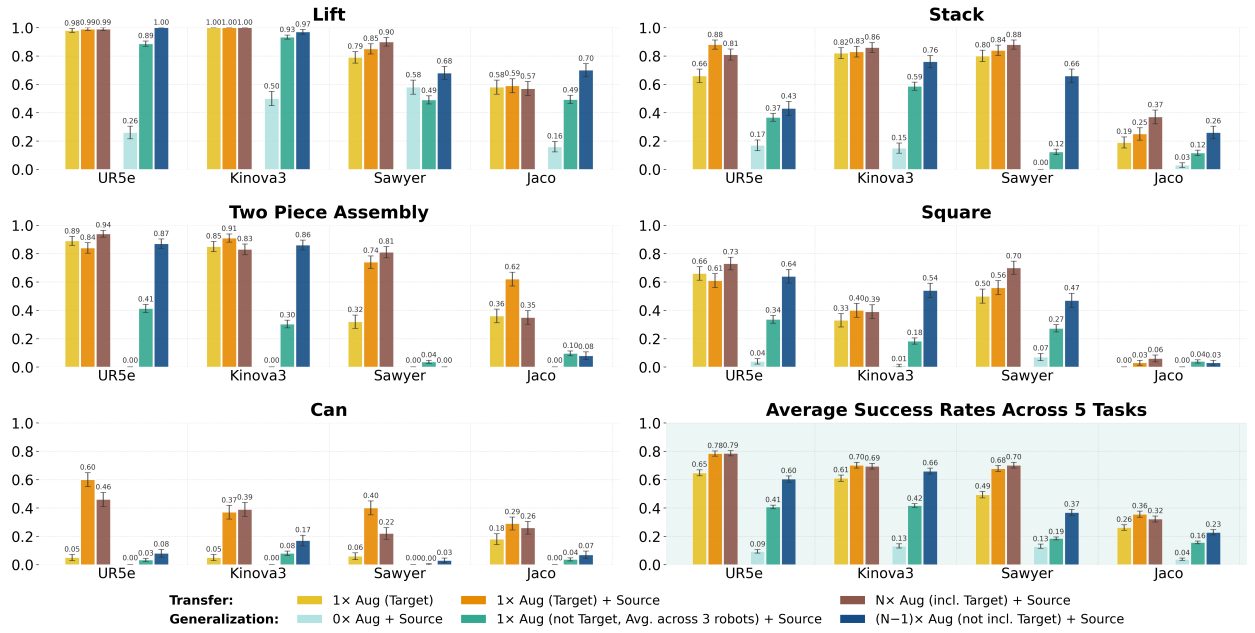

Figure 4: Simulation experiments on scaling robot augmentation. We evaluate how scaling the number of augmented robots impacts (1) **Transfer**: performance on augmented robots (orange), and (2) **Generalization**: performance on unseen robots not included in the augmented set (blue). For transfer, we compare policies trained only on the augmented target robot ("1× Aug (Target)"), on both the source data and that target robot ("1× Aug (Target) + Source"), and on the source data plus all N augmented robots ("N× Aug (incl. Target) + Source"). For generalization, we compare policies trained only on the source data ("0× Aug + Source"), on the source data and one non-target augmented robot ("1× Aug (not Target) + Source"), and on the source data plus $N-1$ augmented robots, leaving the target out ("(N-1)× Aug (not incl. Target) + Source"). In the transfer setting, "1× Aug + Source" substantially outperforms "1× Aug," and "N× Aug + Source" achieves comparable or slightly better performance. In the generalization setting, performance improves consistently with augmentation diversity, with "(N-1)× Aug (not incl. Target) + Source" often rivaling "1× Aug (Target)."

5.1. Study Findings

Fig. 4 reports the average success rates and standard errors for each of the policies across the five tasks. Each policy is evaluated with 100 trials per task.

Robustness to visual perturbations. Fig. 3 shows performance under test-time perturbations to the source robot's environment: (1) lighting shifts that introduce shadows, and (2) occlusions from randomly placed black rectangles. While policies trained on only the source robot degrade significantly, training with 1 and N augmented robots are significantly more robust, with N robot augmentation consistently achieving higher success than 1 robot augmentation. This suggests that robot augmentation enhances robustness on the original embodiment by encouraging the policy to focus on task-relevant structure (e.g., the spatial relationship between gripper and object) rather than incidental features such as arm texture or lighting cues.

Transfer to augmented robots. Among the transfer settings, training on the augmented target robot ("1× Aug (Target)") achieves 26–65% success across the four robots. Combining real and augmented data ("1× Aug (Target) + Source") improves performance by 9–19%. Training on all N augmented robots yields the best overall performance 60% of the time, though the gains over 1 augmented robot + source are small.

Generalization to unseen robots. In the generalization settings, policies without robot augmentation perform poorly, as expected. While 1× robot augmentation generalizes moderately well to some robots, $(N-1)$× augmentation achieves substantially higher success across all four target robots. In many cases, "$(N-1)$× Aug (not incl. Target) + Source" even rivals the performance of "1× Aug (Target)" policies trained directly on the augmented robot. This suggests that robot augmentation promotes embodiment-agnostic visual representations and spatial reasoning, and that generalization to unseen embodiments improves as we increase the diversity of robot augmentations.

Overall, these results suggest that robot augmentation improves not only pairwise transfer but also

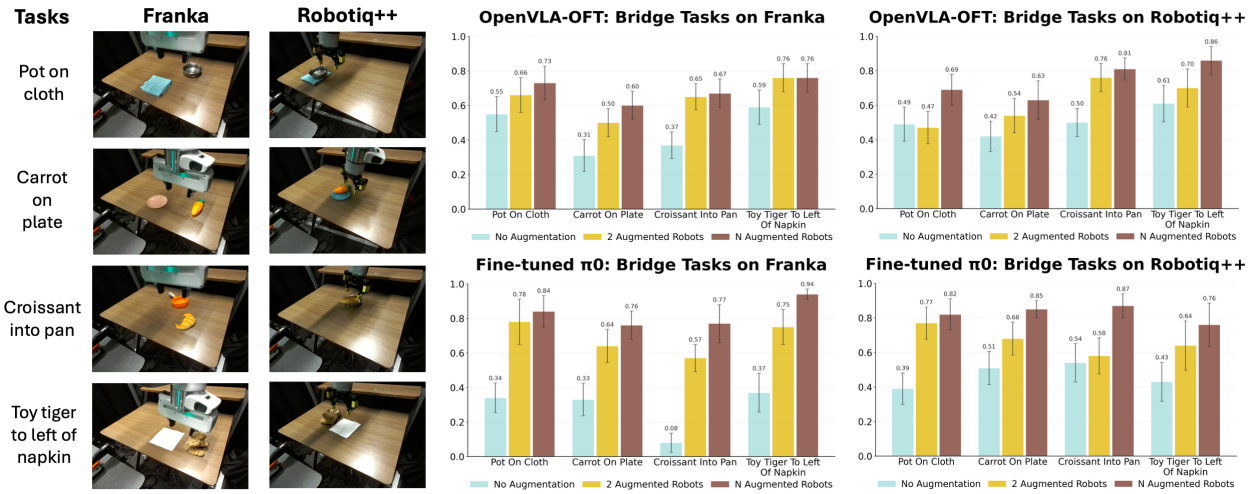

Figure 5: Physical Experiments. Left: Illustration of the 4 tasks and the 2 testing embodiments. "Franka" is a Franka robot equipped with the default Franka gripper, and "Robotiq++" is a Franka robot equipped with a custom modified Robotiq gripper with colorful padding. **Right:** Performance of fine-tuned OpenVLA and π_0 policies trained on the original Bridge and augmented Bridge data. Each policy is evaluated with 10 trials per task for each embodiment.

generalization and robustness. Scaling the number of augmented embodiments helps the policy learn invariances that transfer to unseen embodiments and visual conditions. These results support the view that robot augmentation is a broadly useful training strategy for generalist policies and not just a workaround for target-specific transfer.

6. OXE-AugE: A Large Open-Source Robot Augmentation Dataset

Motivated by the simulation results in Section 5, we present **OXE-AugE**, a large-scale robot-augmented dataset derived from the Open X-Embodiment (OXE) collection [18]. OXE-AugE is designed to scale the benefits of robot augmentation by applying cross-painting to a broad range of tasks, scenes, and robot embodiments.

We select 16 datasets from OXE that are commonly used in training robot foundation models [30, 18, 69, 46, 5, 68, 6]. The original demonstrations in these datasets were each collected using a single robot—one among Franka, UR5, xArm, WidowX, Google Robot, and Jaco. We augment each dataset with up to 9 different robots: the 6 aforementioned robots, as well as Sawyer, Kinova Gen3, and KUKA. Fig.1 shows example visualizations of cross-painted augmentations, and a detailed list of the source and available augmented robots and grippers for each dataset is in the appendix.

Overall, OXE-AugE contains over 4.4M trajectories—3× larger than the original OXE dataset. It spans diverse manipulation scenes and robot-task combinations, substantially increasing embodiment diversity. The AugE-Toolkit is also open-sourced to enable the community to extend augmentation to new datasets or robots.

6.1. Physical Experiments: Fine-tuning Generalist Policies on OXE-AugE

While Section 5 demonstrated that robot augmentation improves transfer and robustness in the single-task setting, in this section, we evaluate whether large-scale augmentation can also benefit pretrained foundation models.

We consider two state-of-the-art generalist policies—OpenVLA [46] and π_0 [5]—and fine-tune them using OXE-AugE. For evaluation, we use tasks from the Bridge dataset [27, 99], originally collected on a WidowX robot, and test on two embodiments: (1) a Franka arm with its default parallel-jaw gripper ("Franka"), which corresponds to one of the augmented robots in OXE-AugE, and (2) a Franka arm with a custom modified Robotiq gripper ("Robotiq++"), which features colored pads to simulate an unseen embodiment (see Fig. 5). This setup evaluates both transfer to augmented robots and generalization to unseen robot-gripper

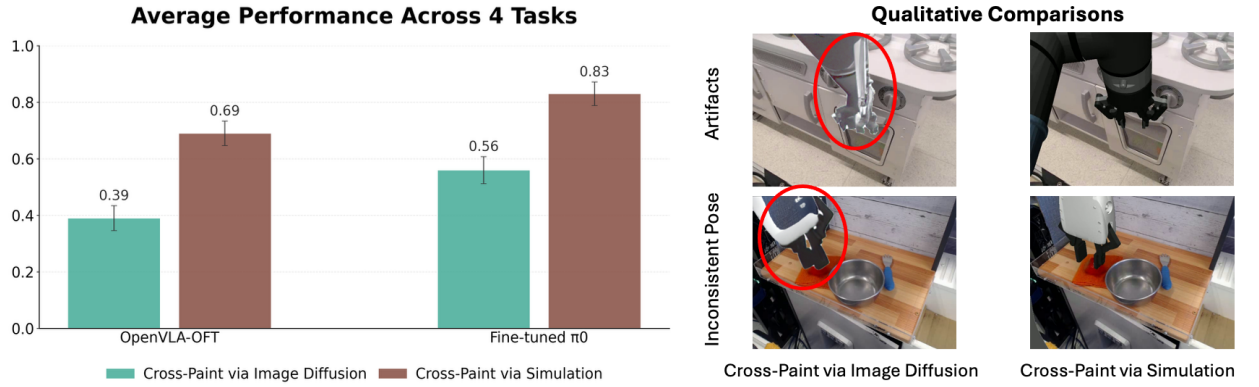

Figure 6: Comparison with RoVi-Aug [13]. We compare our simulator-based augmentation with the RoVi-Aug diffusion pipeline. **Left:** Diffusion-based augmentation leads to a 27–30% drop in policy success. **Right:** Qualitative examples show diffusion artifacts such as misaligned grippers and inconsistent geometry, which reduce policy performance.

configurations.

We evaluate 4 tasks: "Put the pot on the cloth," "Put the carrot on the plate," "Put the croissant into the pan," and "Put the toy tiger to the left of the napkin." The first three appear in Bridge, while the last is novel and absent from both Bridge and other OXE datasets, testing generalization.

For each base model, we compare "no augmentation," "2 augmented robots," and "N augmented robots." Specifically, "2 augmented robots" is the base model finetuned on the Franka and UR5e augmentation of OXE-AugE, and "N augmented robots" is the base model finetuned on all augmentations. All augmented robots use their default grippers. For fair comparison, "no augmentation" models are lightly fine-tuned on Bridge without augmentation to match the total number of fine-tuning steps. For OpenVLA, we follow OpenVLA-OFT [47] and perform LoRA fine-tuning. For π_0, we use full fine-tuning works as we find it works best. Both models use 256×256 third-person observations conditioned on language instructions. More details are in the Appendix.

6.2. Results

Each policy is evaluated over 10 trials per task, resulting in 40 trials per embodiment. Results are summarized in Fig. 5. Both OpenVLA-OFT and π_0's performances are relatively low when fine-tuned only on the original Bridge data, especially on Franka, due to the visual domain shift from the black WidowX gripper to the white Franka one. Fine-tuning on OXE-AugE significantly improves cross-embodiment performance: "2 augmented robots" improves success across all tasks, and $N\times$ augmentation yields the highest success overall. On average, $N\times$ augmentation improves performance by 24% for OpenVLA-OFT and 45% for π_0. On the novel Robotiq++ embodiment, fine-tuned policies reach 75% (OpenVLA-OFT) and 82% (π_0) average success, demonstrating strong generalization.

We further compare our simulator-based pipeline with the diffusion-based RoVi-Aug [13]. Following their setup, we generate 700K paired images between WidowX and target robots using MuJoCo [96] and train diffusion models (based on Stable Diffusion [80] and ControlNet [117]) to translate robot appearances. Fine-tuning with diffusion-generated data leads to a 27–30% drop in final policy performance. As shown in Fig. 6, diffusion-based augmentations sometimes produce misaligned gripper poses or geometric inconsistencies, highlighting the importance of physically accurate simulation in robot augmentation.

7. Conclusion

In this work, we generalize robot augmentation beyond pairwise transfer and develop it into a scalable data pipeline for large-scale robot learning. By improving the cross-painting process, we make augmentation both high-quality and applicable to many existing datasets at scale. Through systematic experiments in both simulation and the real world, we demonstrate that policy performance improves with the number and diversity of augmented embodiments, yielding stronger generalization to unseen robots and greater robustness to visual

perturbations.

We introduce **OXE-AugE**, a large-scale open-source extension of the Open X-Embodiment dataset that applies a scalable and high-quality augmentation pipeline to 16 widely used datasets, expanding them to over 4.4M trajectories and 9 robot embodiments. Fine-tuning foundation models such as OpenVLA and π_0 on OXE-AugE improves success rates by up to 45% on previously unseen robot–gripper combinations, demonstrating that explicit embodiment augmentation complements large pooled data training and promotes more robust, embodiment-agnostic reasoning.

Limitations and promising directions for future research: AugE-Toolkit performs augmentation in 2D image space using simulation replays, which is simpler and more scalable than full 3D scene reconstruction but does not model accurate object–robot occlusions. It also assumes that all augmented robots can perform the task with similar control strategies, neglecting interaction and dynamic differences across embodiments. Variations in robot size, shape, and material can lead to different feasible strategies, and our evaluation focuses mainly on pick-and-place manipulation tasks. Future work could incorporate 3D geometry and physics-aware augmentation, extend to dynamic or bimanual settings, and integrate complementary augmentations such as viewpoint, background, or object variations to enhance generalization across robots, scenes, and tasks.

Acknowledgments

This research was performed at the AUTOLab at UC Berkeley in affiliation with the Berkeley AI Research (BAIR) Lab, and the CITRIS "People and Robots" (CPAR) Initiative, and in collaboration with Google DeepMind. The authors are supported in part by donations from Google, Toyota Research Institute, and equipment grants from NVIDIA. L.Y. Chen was supported by the National Science Foundation (NSF) Graduate Research Fellowship Program under Grant No. 2146752. The authors thank Mehdi Khfifi for some early development of our simulator, and Pannag Sanketi, Ted Xiao, Ashwin Balakrishna, and Quan Vuong for helpful discussions on the research.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

[2] Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. *Conference on Robot Learning*, 2025.

[3] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *Robotics: Science and Systems (RSS)*, 2022.

[4] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4788–4795. IEEE, 2024.

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

[6] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y Galliker, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *9th Annual Conference on Robot Learning*, 2025.

[7] Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor, Kurt Konolige, et al. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4243–4250. IEEE, 2018.

[8] Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Manon Devin, Alex X Lee, Maria Bauza Villalonga, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, et al. Robocat: A self-improving generalist agent for robotic manipulation. *Transactions on Machine Learning Research*, 2023.

[9] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[10] 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. *Robotics: Science and Systems (RSS)*, 2023.

[11] Lawrence Yunliang Chen, Simeon Adebola, and Ken Goldberg. Berkeley UR5 demonstration dataset. https://sites.google.com/view/berkeley-ur5/home, 2023.

[12] Lawrence Yunliang Chen, Kush Hari, Karthik Dharmarajan, Chenfeng Xu, Quan Vuong, and Ken Goldberg. Mirage: Cross-embodiment zero-shot policy transfer with cross-painting. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[13] Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Muhammad Zubair Irshad, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. In *Conference on Robot Learning (CoRL)*, Munich, Germany, 2024.

[14] Tao Chen, Adithyavairavan Murali, and Abhinav Gupta. Hardware conditioned policies for multi-robot transfer learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[15] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali-x: On scaling up a multilingual vision and language model, 2023.

[16] Zoey Chen, Sho Kiami, Abhishek Gupta, and Vikash Kumar. Genaug: Retargeting behaviors to unseen situations via generative augmentation. *arXiv preprint arXiv:2302.06671*, 2023.

[17] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[18] Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller,

Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation*, 2024.

[19] Zichen Jeff Cui, Yibin Wang, Nur Muhammad, Lerrel Pinto, et al. From play to policy: Conditional behavior generation from uncurated robot data. *arXiv:2210.10047*, 2022.

[20] Tianyuan Dai, Josiah Wong, Yunfan Jiang, Chen Wang, Cem Gokmen, Ruohan Zhang, Jiajun Wu, and Li Fei-Fei. Automated creation of digital cousins for robust policy learning. In *Conference on Robot Learning*, 2024.

[21] Prithwish Dan, Kushal Kedia, Angela Chao, Edward Weiyi Duan, Maximus Adrian Pace, Wei-Chiu Ma, and Sanjiban Choudhury. X-sim: Cross-embodiment learning via real-to-sim-to-real. 2025. URL https://arxiv.org/abs/2505.07096.

[22] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019.

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

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

[25] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model, 2023.

[26] Jiafei Duan, Yi Ru Wang, Mohit Shridhar, Dieter Fox, and Ranjay Krishna. Ar2-d2: Training a robot without a robot. In *Conference on Robot Learning*, 2023.

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

[28] Clemens Eppner, Arsalan Mousavian, and Dieter Fox. ACRONYM: A large-scale grasp dataset based on simulation. In *2021 IEEE Int. Conf. on Robotics and Automation, ICRA*, 2020.

[29] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Junbo Wang, Haoyi Zhu, and Cewu Lu. RH20T: A robotic dataset for learning diverse skills in one-shot. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[30] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, 44(5):701–739, 2025.

[31] Haoran Geng, Feishi Wang, Songlin Wei, Yuyang Li, Bangjun Wang, Boshi An, Charlie Tianyue Cheng, Haozhe Lou, Peihao Li, Yen-Jen Wang, et al. Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning. *arXiv preprint arXiv:2504.18904*, 2025.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbab, Chaoyi Pan, et al. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. *Robotics: Science and Systems*, 2025.

[34] Alexander Herzog*, Kanishka Rao*, Karol Hausman*, Yao Lu*, Paul Wohlhart*, Mengyuan Yan, Jessica Lin, Montserrat Gonzalez Arenas, Ted Xiao, Daniel Kappler, Daniel Ho, Jarek Rettinghouse, Yevgen Chebotar, Kuang-Huei Lee, Keerthana Gopalakrishnan, Ryan Julian, Adrian Li, Chuyuan Kelly Fu, Bob Wei, Sangeetha Ramesh, Khem Holden, Kim Kleiven, David Rendleman, Sean Kirmani, Jeff Bingham, Jon Weisz, Ying Xu, Wenlong Lu, Matthew Bennice, Cody Fong, David Do, Jessica Lam, Yunfei Bai, Benjie Holson, Michael Quinlan, Noah Brown, Mrinal Kalakrishnan, Julian Ibarz, Peter Pastor, and Sergey Levine. Deep rl at scale: Sorting waste in office buildings with a fleet of mobile manipulators. In *Robotics: Science and Systems (RSS)*, 2023.

[35] Daniel Ho, Kanishka Rao, Zhuo Xu, Eric Jang, Mohi Khansari, and Yunfei Bai. Retinagan: An object-aware approach to sim-to-real transfer. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10920–10926. IEEE, 2021.

[36] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[37] Edward S Hu, Kun Huang, Oleh Rybkin, and Dinesh Jayaraman. Know thyself: Transferable visual control policies through robot-awareness. In *International Conference on Learning Representations*, 2022.

[38] Wenlong Huang, Igor Mordatch, and Deepak Pathak. One policy to control them all: Shared modular policies for agent-agnostic control. In *International Conference on Machine Learning*, pages 4455–4464. PMLR, 2020.

[39] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-Z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning (CoRL)*, pages 991–1002, 2021.

[40] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: General robot manipulation with multimodal prompts. *International Conference on Machine Learning (ICML)*, 2023.

[41] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on robot learning*, pages 651–673. PMLR, 2018.

[42] Dmitry Kalashnikov, Jake Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Scaling up multi-task robotic reinforcement learning. In *Conference on Robot Learning*, pages 557–575. PMLR, 2022.

[43] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video, 2024. URL https://arxiv.org/abs/2410.24221.

[44] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset, 2024.

[45] Minchan Kim, Junhyek Han, Jaehyung Kim, and Beomjoon Kim. Pre-and post-contact policy decomposition for non-prehensile manipulation with zero-shot sim-to-real transfer. 2023.

[46] 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. In *Conference on Robot Learning*, 2024.

[47] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Proceedings of Robotics: Science and Systems (RSS)*, Los Angeles, CA, USA, 2025.

[48] V Kurin, M Igl, T Rocktaschel, W Boehmer, and S Whiteson. My body is a cage: the role of morphology in graph- based incompatible control. In *Proceedings of the International Conference on Learning Representations*, 2021.

[49] Alex X Lee, Coline Manon Devin, Yuxiang Zhou, Thomas Lampe, Konstantinos Bousmalis, Jost Tobias Springenberg, Arunkumar Byravan, Abbas Abdolmaleki, Nimrod Gileadi, David Khosid, et al. Beyond pick-and-place: Tackling robotic stacking of diverse shapes. In *5th Annual Conference on Robot Learning*, 2021.

[50] Marion Lepert, Ria Doshi, and Jeannette Bohg. Shadow: Leveraging segmentation masks for zero-shot cross-embodiment policy transfer. In *Conference on Robot Learning (CoRL)*, Munich, Germany, 2024.

[51] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Masquerade: Learning from in-the-wild human videos using data-editing. *arXiv preprint arXiv:2508.09976*, 2025.

[52] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos. In *Conference on Robot Learning (CoRL)*, 2025.

[53] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5):421–436, 2018.

[54] Xinhai Li, Jialin Li, Ziheng Zhang, Rui Zhang, Fan Jia, Tiancai Wang, Haoqiang Fan, Kuo-Kun Tseng, and Ruiping Wang. Robogsim: A real2sim2real robotic gaussian splatting simulator. *arXiv preprint arXiv:2411.11839*, 2024.

[55] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[56] Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025.

[57] Vincent Lim, Huang Huang, Lawrence Yunliang Chen, Jonathan Wang, Jeffrey Ichnowski, Daniel Seita, Michael Laskey, and Ken Goldberg. Real2sim2real: Self-supervised learning of physical single-step dynamic actions for planar robot casting. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8282–8289. IEEE, 2022.

[58] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. *arXiv preprint arXiv:2410.18647*, 2024.

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

[60] Abhiram Maddukuri, Zhenyu Jiang, Lawrence Yunliang Chen, Soroush Nasiriany, Yuqi Xie, Yu Fang, Wenqi Huang, Zu Wang, Zhenjia Xu, Nikita Chernyadev, Scott Reed, Ken Goldberg, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Sim-and-real co-training: A simple recipe for vision-based robotic manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, Los Angeles, CA, USA, 2025.

[61] Zhao Mandi, Homanga Bharadhwaj, Vincent Moens, Shuran Song, Aravind Rajeswaran, and Vikash Kumar. Cacti: A framework for scalable multi-task multi-scene visual imitation learning. *arXiv preprint arXiv:2212.05711*, 2022.

[62] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning*, 2021.

[63] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. In *7th Annual Conference on Robot Learning*, 2023.

[64] Tatsuya Matsushima, Hiroki Furuta, Yusuke Iwasawa, and Yutaka Matsuo. Weblab xarm dataset, 2023.

[65] Marius Memmel, Andrew Wagenmaker, Chuning Zhu, Patrick Yin, Dieter Fox, and Abhishek Gupta. Asid: Active exploration for system identification in robotic manipulation. *arXiv preprint arXiv:2404.12308*, 2024.

[66] Soroush Nasiriany, Tian Gao, Ajay Mandlekar, and Yuke Zhu. Learning and retrieval from prior data for skill-based imitation learning. In *Conference on Robot Learning*, pages 2181–2204. PMLR, 2022.

[67] 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. In *Robotics: Science and Systems (RSS)*, 2024.

[68] NVIDIA, 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.

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

[70] Chuer Pan, Litian Liang, Dominik Bauer, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, and Shuran Song. One demo is worth a thousand trajectories: Action-view augmentation for visuomotor policies. In *Conference on Robot Learning (CoRL)*, 2025.

[71] Deepak Pathak, Christopher Lu, Trevor Darrell, Phillip Isola, and Alexei A Efros. Learning to control self-assembling morphologies: a study of generalization via modularity. *Advances in Neural Information Processing Systems*, 32, 2019.

[72] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.

[73] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017. URL https://arxiv.org/abs/1709.07871.

[74] Nicholas Pfaff, Evelyn Fu, Jeremy Binagia, Phillip Isola, and Russ Tedrake. Scalable real2sim: Physics-aware asset generation via robotic pick-and-place setups. *arXiv preprint arXiv:2503.00370*, 2025.

[75] Dean A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In D. Touretzky, editor, *Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1988.

[76] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, 2022.

[77] Fabio Ramos, Rafael Possas, and Dieter Fox. Bayessim: Adaptive domain randomization via probabilistic inference for robotics simulators. In *Proceedings of Robotics: Science and Systems*, Freiburgim-Breisgau, Germany, June 2019. doi: 10.15607/RSS.2019.XV.029.

[78] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

[79] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar

Bordbar, and Nando de Freitas. A generalist agent. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

[80] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[81] Stephane Ross, Geoffrey J Gordon, and J Andrew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[82] Alvaro Sanchez-Gonzalez, Nicolas Heess, Jost Tobias Springenberg, Josh Merel, Martin Riedmiller, Raia Hadsell, and Peter Battaglia. Graph networks as learnable physics engines for inference and control. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4470–4479. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/sanchez-gonzalez18a.html.

[83] Sebastian Sartor and Neil Thompson. Neural scaling laws in robotics. *arXiv preprint arXiv:2405.14005*, 2024.

[84] Saumya Saxena, Mohit Sharma, and Oliver Kroemer. Multi-resolution sensing for real-time control with vision-language models. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=WuBv9-IGDUA.

[85] Vaibhav Saxena, Matthew Bronars, Nadun Ranawaka Arachchige, Kuancheng Wang, Woo Chul Shin, Soroush Nasiriany, Ajay Mandlekar, and Danfei Xu. What matters in learning from large-scale datasets for robot manipulation. In *CoRL 2024 Workshop on Mastering Robot Manipulation in a World of Abundant Data*, 2024.

[86] Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home, 2023.

[87] Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. GNM: A general navigation model to drive any robot. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7226–7233. IEEE, 2023.

[88] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. ViNT: A Foundation Model for Visual Navigation. In *7th Annual Conference on Robot Learning (CoRL)*, 2023.

[89] Rutav Shah, Roberto Martín-Martín, and Yuke Zhu. MUTEX: Learning unified policies from multimodal task specifications. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=PwqiqaaEzJ.

[90] Lin Shao, Fabio Ferreira, Mikael Jorda, Varun Nambiar, Jianlan Luo, Eugen Solowjow, Juan Aparicio Ojea, Oussama Khatib, and Jeannette Bohg. Unigrasp: Learning a unified model to grasp with multifingered robotic hands. *IEEE Robotics and Automation Letters*, 5(2):2286–2293, 2020.

[91] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2021.

[92] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

[93] Austin Stone, Ted Xiao, Yao Lu, Keerthana Gopalakrishnan, Kuang-Huei Lee, Quan Vuong, Paul Wohlhart, Brianna Zitkovich, Fei Xia, Chelsea Finn, et al. Open-world object manipulation using pre-trained vision-language models. In *Conference on Robot Learning*, 2023.

[94] Bingjie Tang, Michael A Lin, Iretiayo Akinola, Ankur Handa, Gaurav S Sukhatme, Fabio Ramos, Dieter Fox, and Yashraj Narang. Industreal: Transferring contact-rich assembly tasks from simulation to reality. *arXiv preprint arXiv:2305.17110*, 2023.

[95] Stephen Tian, Blake Wulfe, Kyle Sargent, Katherine Liu, Sergey Zakharov, Vitor Guizilini, and Jiajun Wu. View-invariant policy learning via zero-shot novel view synthesis. *arXiv preprint arXiv:2409.03685*, 2024.

[96] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012.

[97] Marcel Torne, Arhan Jain, Jiayi Yuan, Vidaaranya Macha, Lars Ankile, Anthony Simeonov, Pulkit Agrawal, and Abhishek Gupta. Robot learning with super-linear scaling. *arXiv preprint arXiv:2412.01770*, 2024.

[98] Marcel Torne, Anthony Simeonov, Zechu Li, April Chan, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. In *Robotics: Science and Systems*, 2024.

[99] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale, 2023.

[100] 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.

[101] Tingwu Wang, Renjie Liao, Jimmy Ba, and Sanja Fidler. Nervenet: Learning structured policy with graph neural networks. In *International conference on learning representations*, 2018.

[102] Junjie Wen, Yichen Zhu, Minjie Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Xiaoyu Liu, Chaomin Shen, Yaxin Peng, and Feifei Feng. Diffusionvla: Scaling robot foundation models via unified diffusion and autoregression. In *Forty-second International Conference on Machine Learning*, 2025.

[103] Haoyu Xiong, Quanzhou Li, Yun-Chun Chen, Homanga Bharadhwaj, Samarth Sinha, and Animesh Garg. Learning by watching: Physical imitation of manipulation skills from human videos. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7827–7834. IEEE, 2021.

[104] Zhenjia Xu, Beichun Qi, Shubham Agrawal, and Shuran Song. Adagrasp: Learning an adaptive gripper-aware grasping policy. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4620–4626. IEEE, 2021.

[105] Ge Yan, Kris Wu, and Xiaolong Wang. ucsd kitchens Dataset. August 2023.

[106] Jonathan Yang, Dorsa Sadigh, and Chelsea Finn. Polybot: Training one policy across robots while embracing variability. *arXiv preprint arXiv:2307.03719*, 2023.

[107] Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, Quan Vuong, Chelsea Finn, Dorsa Sadigh, and Sergey Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[108] Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025.

[109] Sizhe Yang, Wenye Yu, Jia Zeng, Jun Lv, Kerui Ren, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Novel demonstration generation with gaussian splatting enables robust one-shot manipulation. *arXiv preprint arXiv:2504.13175*, 2025.

[110] Weirui Ye, Fangchen Liu, Zheng Ding, Yang Gao, Oleh Rybkin, and Pieter Abbeel. Video2policy: Scaling up manipulation tasks in simulation through internet videos. *arXiv preprint arXiv:2502.09886*, 2025.

[111] Shaofeng Yin, Yanjie Ze, Hong-Xing Yu, C Karen Liu, and Jiajun Wu. Visualmimic: Visual humanoid loco-manipulation via motion tracking and generation. *arXiv preprint arXiv:2509.20322*, 2025.

[112] Chen Yu, Weinan Zhang, Hang Lai, Zheng Tian, Laurent Kneip, and Jun Wang. Multi-embodiment legged robot control as a sequence modeling problem. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7250–7257. IEEE, 2023.

[113] Justin Yu, Letian Fu, Huang Huang, Karim El-Refai, Rares Andrei Ambrus, Richard Cheng, Muhammad Zubair Irshad, and Ken Goldberg. Real2render2real: Scaling robot data without dynamics simulation or robot hardware. *arXiv preprint arXiv:2505.09601*, 2025.

[114] Tianhe Yu, Ted Xiao, Austin Stone, Jonathan Tompson, Anthony Brohan, Su Wang, Jaspiar Singh, Clayton Tan, Jodilyn Peralta, Brian Ichter, et al. Scaling robot learning with semantically imagined experience. In *Robotics: Science and Systems*, 2023.

[115] Kevin Zakka, Yuval Tassa, and MuJoCo Menagerie Contributors. MuJoCo Menagerie: A collection of high-quality simulation models for MuJoCo, 2022. URL http://github.com/google-deepmind/mujoco_menagerie.

[116] Kevin Zakka, Baruch Tabanpour, Qiayuan Liao, Mustafa Haiderbhai, Samuel Holt, Jing Yuan Luo, Arthur Allshire, Erik Frey, Koushil Sreenath, Lueder A. Kahrs, Carlo Sferrazza, Yuval Tassa, and Pieter Abbeel. Demonstrating mujoco playground. In *Robotics: Science and Systems*, 2025.

[117] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[118] Xiaoyu Zhang, Matthew Chang, Pranav Kumar, and Saurabh Gupta. Diffusion meets dagger: Supercharging eye-in-hand imitation learning. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[119] Siheng Zhao, Jiageng Mao, Wei Chow, Zeyu Shangguan, Tianheng Shi, Rong Xue, Yuxi Zheng, Yijia Weng, Yang You, Daniel Seita, et al. Robot learning from any images. *arXiv preprint arXiv:2509.22970*, 2025.

[120] Allan Zhou, Moo Jin Kim, Lirui Wang, Pete Florence, and Chelsea Finn. Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2023.

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

[122] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. *6th Annual Conference on Robot Learning (CoRL)*, 2022.

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

[124] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. In *arXiv preprint arXiv:2009.12293*, 2020.

A. Supplementary Material

A.1. Author contributions:

G.J. implemented the core idea, developed the simulator, labeled, generated, and curated data, trained π_0 policies, and led the physical experiments.
H.P. implemented the core idea, developed the simulator, conducted the simulation study, generated and curated data, trained OpenVLA policies, and ran some physical experiments.
L.Y.C. proposed the research, set the vision and direction, steered and aligned the team, provided guidance on all aspects of the project including the core approach, systems, and evaluations, ran some physical experiments, and wrote the paper.
S.B. developed the simulator, trained the segmentation model, generated part of the OXE-AugE data, and ran some physical experiments.
Z.M. developed the simulator, generated part of the OXE-AugE data, and contributed to paper writing.
S.A. developed the simulator used to generate part of the OXE-AugE data.
C.X. co-proposed the research with L.Y.C., advised on the project, aligned the team, provided insights and guidance on the algorithm and experiment design, generated the RoVi-Aug baseline data, and gave feedback on the paper.
K.G. supervised the project, aligned the team, and provided feedback on the approach, evaluations, and paper.

A.2. OXE-AugE Dataset Details

Table 1 presents a list of the datasets in OXE-AugE. We select 16 datasets from OXE that are commonly used in training robot foundation models [30, 18, 69, 46, 5, 68, 6]. The original demonstrations in those datasets were collected using Franka, UR5, xArm, WidowX, Google Robot, and Jaco platforms. A filled circle (●) indicates the source robot, and a check mark (✓) indicates robots for which augmented demonstrations are available. For 14 out of the 16 datasets, all 9 robots are available. For RT-1 (Fractal) and Language Table datasets, we find most trajectories' range of the motion exceeds the workspace of WidowX, so we exclude the WidowX augmentation. Overall, OXE-AugE contains over 550,000 demonstrations per robot, totaling 4.4M trajectories—3× larger than the original OXE dataset.

Dataset	Panda	UR5e	Xarm7	Google	WidowX	Sawyer	Kinova3	IIWA	Jaco
Berkeley AUTOLab UR5 [11]	✓	●	✓	✓	✓	✓	✓	✓	✓
TACO Play [121]	●	✓	✓	✓	✓	✓	✓	✓	✓
Austin BUDS [123]	●	✓	✓	✓	✓	✓	✓	✓	✓
Austin Mutex [89]	●	✓	✓	✓	✓	✓	✓	✓	✓
Austin Sailor [66]	●	✓	✓	✓	✓	✓	✓	✓	✓
CMU Franka Pick-Insert [84]	●	✓	✓	✓	✓	✓	✓	✓	✓
KAIST Nonprehensile [45]	●	✓	✓	✓	✓	✓	✓	✓	✓
NYU Franka Play [19]	●	✓	✓	✓	✓	✓	✓	✓	✓
TOTO [121]	●	✓	✓	✓	✓	✓	✓	✓	✓
UTokyo xArm PickPlace [64]	✓	✓	●	✓	✓	✓	✓	✓	✓
UCSD Kitchen [105]	✓	✓	●	✓	✓	✓	✓	✓	✓
Austin VIOLA [122]	●	✓	✓	✓	✓	✓	✓	✓	✓
Bridge [99]	✓	✓	✓	✓	●	✓	✓	✓	✓
RT-1 Robot Action [10]	✓	✓	✓	●		✓	✓	✓	✓
Jaco Play [23]	✓	✓	✓	✓	✓	✓	✓	✓	●
Language Table [59]	✓	✓	●	✓		✓	✓	✓	✓

Table 1: Source and augmented robots in the OXE-AugE dataset. A filled circle (●) indicates the **source robot**, and a check mark (✓) indicates robots for which augmented demonstrations are available.

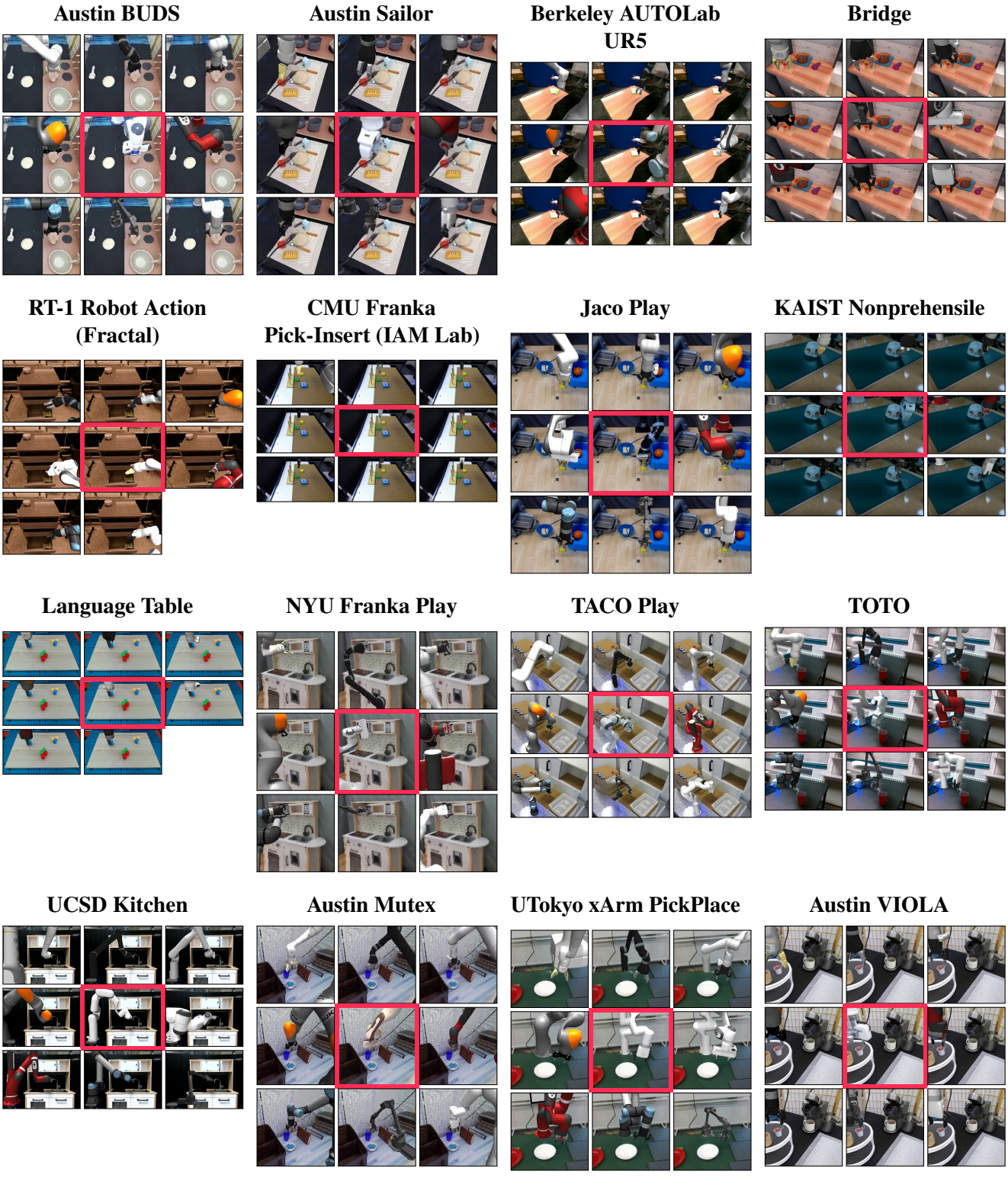

Figure 7: Example images in OXE-AugE. = Source robot (center cell, highlighted)

In both the original OXE dataset and OXE-AugE, the Franka robot uses the default Franka Hand, UR5e uses the Robotiq 2f-85 gripper, xArm7 uses UFACTORY xArm Gripper G2, Google Robot uses the custom 2-finger gripper, Jaco uses the Kinova KG-3 gripper, and WidowX 250 uses the default parallel-jaw gripper. Additionally, OXE-AugE uses the Robotiq 2f-85 gripper for KUKA iiwa and Kinova3, and the Rethink Gripper for the Sawyer robot. We use these grippers as they are the most common types used for each robot,

but switching grippers is also easy to do.

Figure 7 shows example images in OXE-AugE. Figure 8 illustrates the data sources of OXE-AugE and its relationship to OXE and the Octo training mixture. OXE v1.0 contains about 1.4M trajectories. V1.1 expands the total number of trajectories to 2.4M, however, only 1.4M of them are real robot data with manipulation or mobile manipulation (the remaining are sim, navigation, locomotion, human, or VQA data). Starting from there, we select 14 datasets out of the 25 datasets in the Octo Training Mix, collectively counting for 58% of the total weights. We also include 2 other high-quality datasets (UTokyo xArm PickPlace and KAIST Nonprehensile Objects) that are not used in Octo. Together, these 16 datasets include 0.55M trajectories and form the source of OXE-AugE. After augmenting each dataset with 8 or 9 robots, the total OXE-AugE consists of 4.44M trajectories.

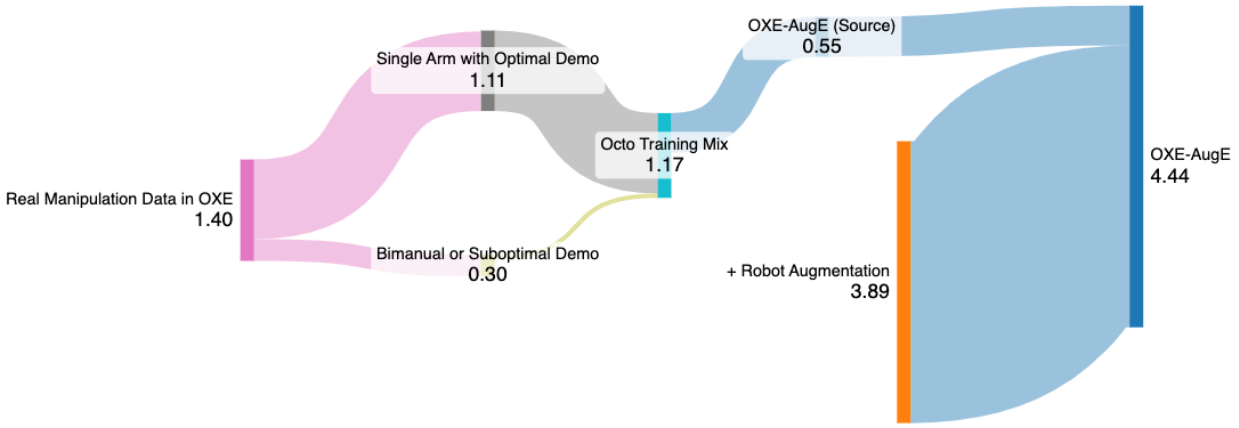

Figure 8: Sankey diagram illustrating the data sources of OXE-AugE (numbers in millions of trajectories). OXE v1.1 contains about 1.4M real robot manipulation trajectories. Starting from there, we select 14 datasets out of the 25 datasets in the Octo Training Mix, collectively counting for 58% of the total weights. We also include 2 other high-quality datasets (UTokyo xArm PickPlace and KAIST Nonprehensile Objects) that are not used in Octo. This set of 0.55M trajectories forms the source datasets for OXE-AugE. After robot augmentation, the total OXE-AugE consists of 4.44M trajectories.

A.3. Simulation Tasks and Robots Visualization

Fig. 9 illustrates the simulation tasks and robots. We use the Robosuite environment with 5 tasks: Lift, Stack, Two Piece Assembly, Square Peg Insertion tasks, and Can Pick-and-Place. The demonstration data is performed on a Franka robot, and we evaluate on 4 target robots: UR5e, Kinova Gen3, Sawyer, and Jaco. Jaco has a 3-jaw gripper while the other robots have 2-jaw grippers.

A.4. Policy Training Details

For simulation experiments, we train diffusion policies [17] using RoboMimic [62]. Each policy is trained on 200 demonstrations for a single task from scratch. The policy architecture consists of a non-pretrained ResNet18 [32] vision encoder and a 1D convolutional neural network (CNN) action denoiser, connected through FiLM [73]. All policies are trained with a learning rate of 1e-4, batch size of 16, and for 250k steps. The visual inputs are 84x84, with random crop data augmentation during training.

For physical experiments, we fine-tune OpenVLA and π_0. For OpenVLA, we follow OpenVLA-OFT [47] and perform LoRA fine-tuning [36] with a learning rate of 5e-4, batch size of 8, for 25k steps. For π_0, we perform full parameter fine-tuning with a learning rate of 5e-5, batch size of 32, for 20k steps. Both models take in a third-person observation of 256×256 resolution, and are conditioned on the language instructions.

A.5. Physical Experiment Details

In physical experiments, we use a Franka FR3 robot and a ZED 2 camera. We place the camera in roughly the same pose as that in the Bridge dataset, and crop the images to match the field of view of the Logitech camera used in Bridge. To address the challenges of different controller dynamics between the robots, we

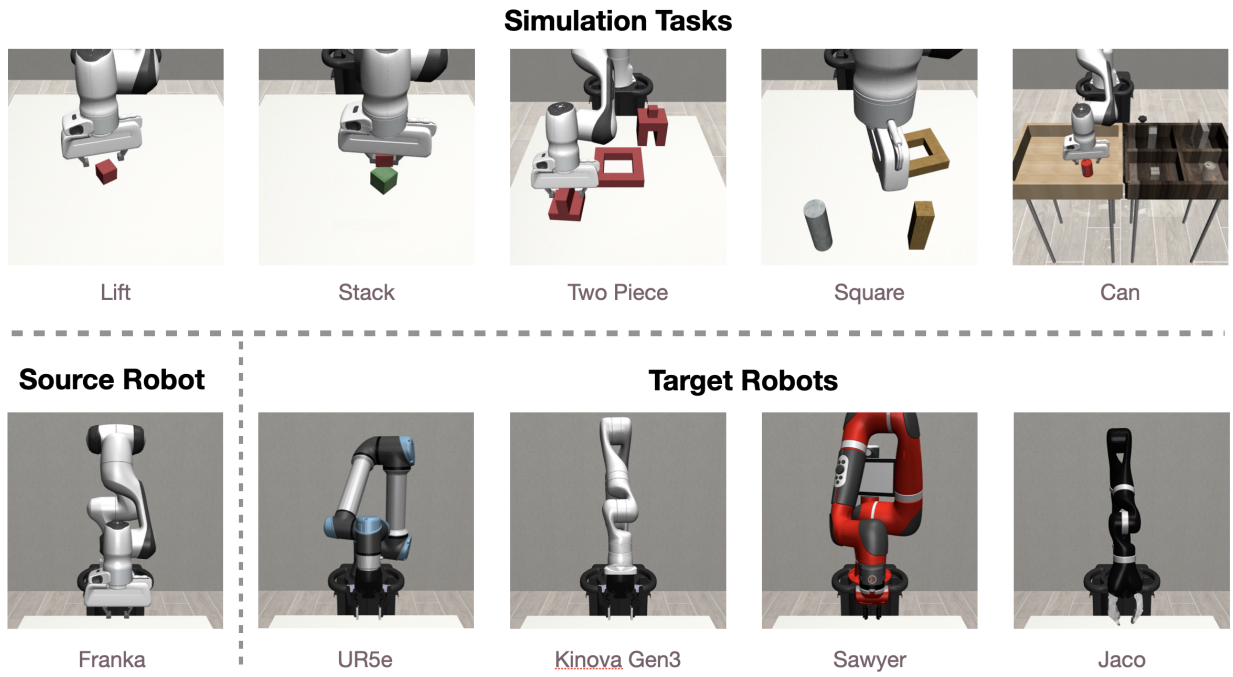

Figure 9: Simulation Tasks and Robots. We use the Robosuite environment with 5 tasks: Lift, Stack, Two Piece Assembly, Square Peg Insertion tasks, and Can Pick-and-Place. The demonstration data is performed on a Franka robot, and we evaluate on 4 target robots: UR5e, Kinova Gen3, Sawyer, and Jaco. Jaco has a 3-jaw gripper while the other robots have parallel-jaw grippers.

follow OpenVLA and train on the delta states of the WidowX robot instead of the action targets. During inference, we let the Franka controller reach the desired target state before executing the next action. To prevent compounding error, the policy takes in the predicted state instead of the actual state as inputs [94].

For the 4 tasks, we use the following metrics for measuring the policy performance: we give a score of 0.3 if the robot moves towards the right direction the object and attempts a grasp, a score of 0.5 if the robot successfully grasps the object, a score of 0.8 if the robot carries the object and moves towards the correct destination location, and a score of 1 if the robot successfully place the object in the target position. We perform 10 trials per task, resulting in 40 trials per policy for each robot embodiment. We compute the mean and standard error of the scores for each of the 6 policies on each target embodiment in Fig. 5.