MonoDuo: Using One Robot Arm to Learn Bimanual Policies

†Sandeep Bajamahal[1*], Lawrence Yunliang Chen[1*], Toru Lin[1], Zehan Ma[1], Jitendra Malik[1], Ken Goldberg[1]

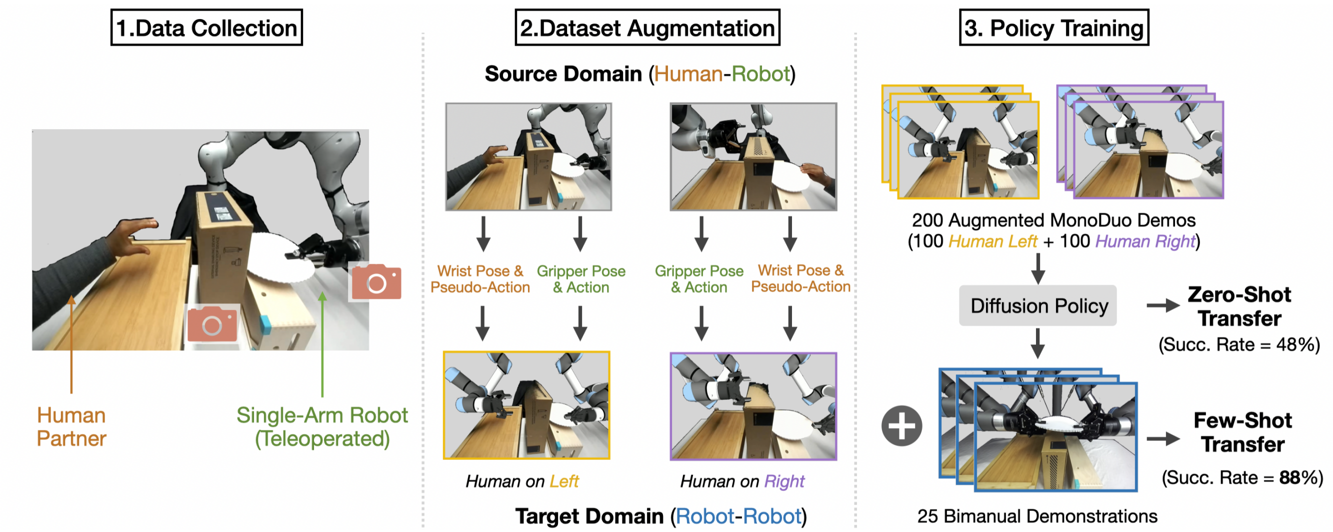

Fig. 1: **Overview of MonoDuo.** The teleoperation system uses a fixed RGB-D camera and a wrist-mounted camera. We begin by teleoperating a single-arm robot to collaborate with a human arm on a bimanual task, alternating left-right arm roles across episodes. This results in complementary interaction data covering both sides of the task. These human-robot bimanual demonstrations are then augmented into synthetic robot-robot bimanual demonstrations using segmentation and inpainting techniques, creating a visually and physically grounded dataset for training bimanual robots.

***Abstract*— Bimanual coordination is essential for many real-world manipulation tasks, yet learning bimanual robot policies is limited by the scarcity of bimanual robots and datasets. Single-arm robots, however, are widely available in research labs. Can we leverage them to train bimanual robot policies? We present *MonoDuo*, a framework for learning bimanual manipulation policies using single-arm robot demonstrations paired with human collaboration. MonoDuo collects data by teleoperating a single-arm robot to perform one side of a bimanual task while a human performs the other, then swapping roles to cover both sides. RGB-D observations from a wrist-mounted and fixed camera are augmented into synthetic demonstrations for target bimanual robots using state-of-the-art hand pose estimation, image and point cloud segmentation, and inpainting. These synthetic demonstrations, grounded in real robot kinematics, are used to train bimanual policies. We evaluate MonoDuo on five tasks—box lifting, backpack packing, cloth folding, jacket zipping, and plate handover. Compared to approaches relying solely on human bimanual videos, MonoDuo enables zero-shot deployment on unseen bimanual robot configurations, achieving success rates up to 70%. With only 25 target robot demonstrations, few-shot finetuning further boosts success rates by 65–70% over training from scratch, demonstrating MonoDuo's effectiveness in efficiently transferring knowledge from single-arm robot data to bimanual robot policies.**

I. Introduction

Bimanual robotic systems offer the potential to perform complex, coordinated manipulation tasks that are difficult or impossible for single-arm robots to execute. Many industrial and home tasks require two arms working in concert, with precise timing, spatial awareness, and physical coordination. However, a majority of available datasets and research infrastructure uses single-arm robots. This creates a bottleneck for learning bimanual policies, where the scarcity of bimanual robots significantly limits scalability.

We address this gap by proposing *MonoDuo*, a framework that democratizes bimanual robot learning by enabling training from single-arm demonstrations augmented with human collaboration. MonoDuo builds on recent advances in cross-embodiment learning—techniques for transferring behaviors across different robot morphologies—and extends them to the challenging setting of single-arm to bimanual transfer. In our setup, a human teleoperates a single-arm robot to perform one side of a bimanual task while coordinating with a second human arm, alternating left-right roles across episodes to ensure both human and robot contribute equally to the dataset. This data is then augmented into synthetic demonstrations for specified bimanual robot hardware using state-of-the-art hand pose estimation, image and point cloud segmentation, and inpainting. To ensure both fidelity and coverage, MonoDuo introduces a novel, structured augmentation strategy that preserves real robot actions for accurate supervision while filling in the missing arm with retargeted end-effector actions derived from the human collaborator. This structured mixing produces balanced datasets that maintain the action distribution when only a single-arm robot is available,

[1]University of California, Berkeley
†Corresponding author: sandeep24@berkeley.edu

while explicitly enabling cross-embodiment transfer for both human-to-robot and robot-to-robot generalization.

We evaluate MonoDuo on five challenging bimanual tasks—box lifting, backpack packing, cloth folding, jacket zipping, and plate handover. It achieves 35%–70% zero-shot success on unseen robot configurations. We additionally study a practical few-shot learning scenario with limited target demonstrations. In this setting, we show that MonoDuo improves sample efficiency significantly, increasing success rates by 65∼70% compared to policies without MonoDuo. This highlights our approach as both a transferable knowledge source and a complement to limited real-world bimanual data. This paper makes four contributions:

1) MonoDuo, a novel framework for collecting demonstration data using one robot arm in collaboration with a human arm and synthesizing bimanual robot demonstrations.
2) A data transformation pipeline that converts human-robot arm demonstrations into bimanual robot demonstrations using pose estimation, segmentation, and inpainting, and introduces novel mixing of human retargeted end-effector actions with real robot actions to enhance policy learning in comparison to robot-only policies.
3) Experiments suggesting that policies trained with MonoDuo can generalize zero-shot to previously unseen bimanual robot configurations, evaluated on a set of 5 bimanual tasks.
4) Experiments suggesting that MonoDuo significantly improves sample efficiency when finetuned with 25 bimanual robot demonstrations in comparison to policies trained from scratch.

II. RELATED WORKS

A. Learning-Based Approaches to Bimanual Manipulation

Existing learning-based approaches to equip robots with bimanual manipulation skills can be broadly classified into three categories: learning from demonstrations [1]–[8], sim-to-real reinforcement learning [9]–[11], and learning from human videos [12]–[18] or human motion data [19], [20]. In learning from demonstrations, a human teleoperates the robot arms or end-effectors to directly collect sensorimotor data from the bimanual robot. The collected data can then be used to train bimanual manipulation policies in a straightforward way, using state-of-the-art imitation learning policies [21]. Advances in bimanual teleoperation systems [3], [5], [6], [8] and imitation learning algorithms [22], [23] in recent years have lowered the barrier for adopting this approach, making it a popular choice among both industry and academic labs. On the other hand, sim-to-real RL avoids real robot data by training policies on a digital twin of the target robot [9]–[11], then transferring to hardware, but faces challenges in reward design and the sim-to-real gap. Learning from human video or motion data leverages human action priors to ease bimanual policy learning but lacks directly usable robot actions. Our work is most related to this line, especially human video.

B. Cross-Embodiment Robot Learning

Cross-embodiment robot learning [24] aims to learn or transfer policies across robots with different physical embodiments. This is crucial for generalizing learning across hardware platforms, reducing the need to retrain for every new robot configuration. A growing line of research tackles this problem using domain randomization to learn robot-conditioned policies [25]–[33] or training on large real robot data [34]–[41] to make policies more robust and generalizable [42]–[56]. RoVi-Aug [57] introduces a diffusion-based augmentation pipeline that replaces the robot in demonstration images and generates new camera views, producing synthetic demonstrations with varied robots and viewpoint. In contrast to offline augmentation, Mirage [58] performs test-time image editing to create an illusion that the original robot is performing the task. This "cross-painting" technique decouples visual differences from control and achieved successful zero-shot transfer of manipulation policies between different single-arm embodiments. We build on these cross-painting techniques in this work. A related approach, Shadow [59], simplifies cross-embodiment image editing by overlaying segmentation masks of the source and target robots on training and test images. Learning from human video is a form of cross-embodiment robot learning. Recent works leverage human demonstrations as robot training data: for example, Phantom [60] trains robots without any robot demos by converting human videos into robot-like observations, while EgoMimic [61] co-trains on egocentric human videos and teleoperated robot demonstrations using cross-domain alignment. Masquerade [62] edits in-the-wild human videos to create synthetic bimanual robot demonstrations and co-trains policies for robust zero-shot transfer to unseen scenes. These methods show that visual editing and alignment can make human video demonstrations viable for robot policy learning. Other existing frameworks [14], [42], [63], [64] also explore translating human videos into robot actions via learned correspondences.

C. Learning Bimanual Manipulation with A Single-Arm Robot

Bimanual manipulation poses challenges for cross-embodiment learning due to coordinated, high-dimensional actions.

Prior multi-embodiment studies avoid direct visual retargeting, instead using specialized methods: CrossFormer [65] adds a bimanual action head, DexMimicGen [66] expands limited demos via simulation synthesis, and AnyBimanual [67] composes single-arm skills with task-level reasoning.

In contrast, our approach learns end-to-end coordinated bimanual policies directly from synthetic demonstrations generated using only single-arm robot data paired with human interaction. The only closely related work that we have found is LfDT [68], which uses human-robot interaction videos to learn dual-arm policies by learning a CycleGAN to transform human-robot images into robot-robot images. However, LfDT requires robot-robot target domain videos for training the CycleGAN [69] and is only validated on relatively simple

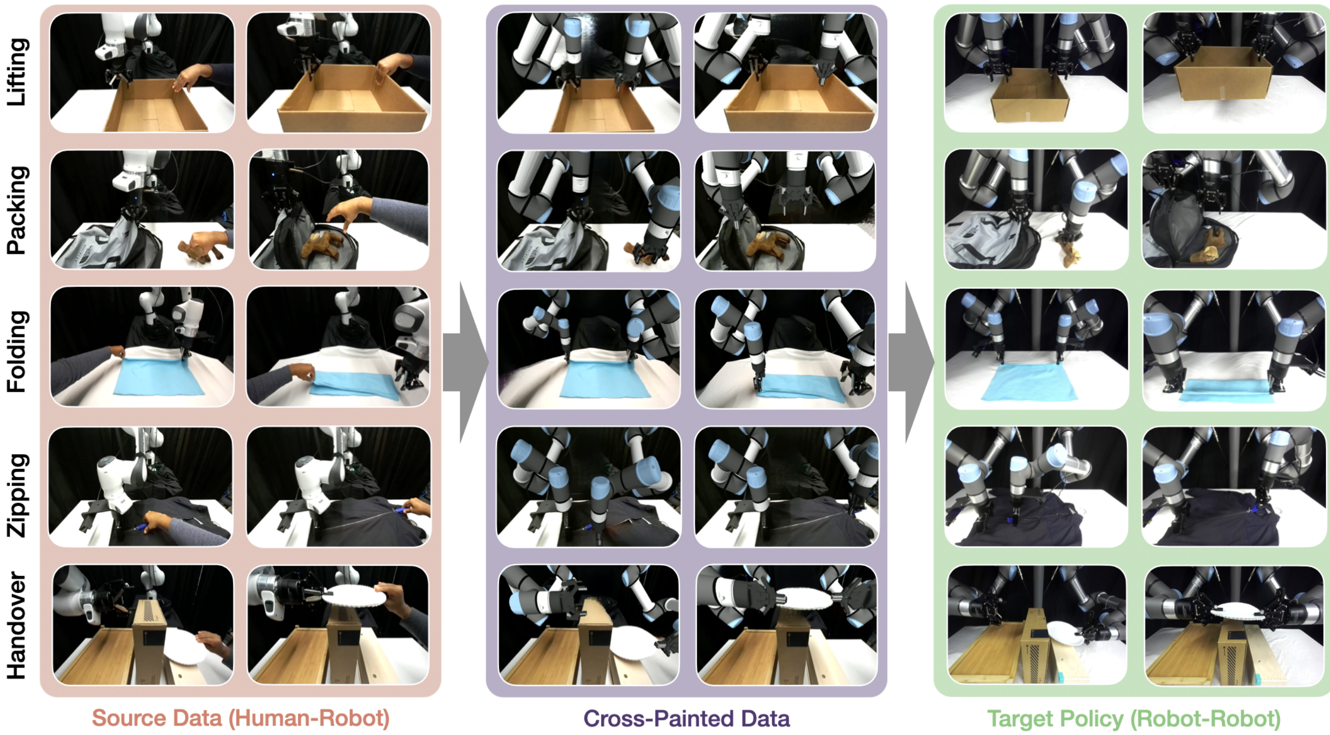

Fig. 2: **From Human-Robot Demonstrations to Robot-Robot Policies.** Given collaborative demonstration trajectories between a single-arm robot and a human, MonoDuo uses state-of-the-art diffusion models to augment the image data and generate synthetic dataset tailored to a specified bimanual robot. Policies trained with the augmented dataset can be deployed on this target bimanual robot zero-shot. The same dataset can also be used to improve sample efficiency for few-shot learning.

tasks such as pushing. To the best of our knowledge, this work is the first to learn bimanual manipulation policies using only a single-arm robot, and to demonstrate success on complex, contact-rich tasks with zero-shot success, as shown in Section V.

III. Problem Statement

As described in Figure 1, MonoDuo collects a demonstration dataset $\mathcal{D}^{\mathcal{S}} = \{\tau_1^{\mathcal{S}}, \tau_2^{\mathcal{S}}, ..., \tau_n^{\mathcal{S}}\}$ consisting of $2N$ successful trajectories of a source robot-human pair $\mathcal{S} = (\mathcal{S}_r, \mathcal{S}_h)$ performing some task. Each trajectory $\tau_i^{\mathcal{S}} = (\{o^{\mathcal{S}}_{1..H_i}\}, \{p^{\mathcal{S}_r}_{1..H_i}\}, \{p^{\mathcal{S}_h}_{1..H_i}\}, \{a^{\mathcal{S}_r}_{1..H_i}\}, \{a^{\mathcal{S}_h}_{1..H_i}\})$, where $\{o_1^{\mathcal{S}}, ..., o_{H_i}^{\mathcal{S}}\}$ is a sequence of RGB-D camera observations, $\{p_1^{\mathcal{S}_r}, ..., p_{H_i}^{\mathcal{S}_r}\}$ is the sequence of corresponding robot state observations, $\{p_1^{\mathcal{S}_h}, ..., p_{H_i}^{\mathcal{S}_h}\}$ is the sequence of corresponding human hand state observations, $\{a_1^{\mathcal{S}_r}, ..., a_{H_i}^{\mathcal{S}_r}\}$ is the sequence of corresponding robot actions, and $\{a_1^{\mathcal{S}_h}, ..., a_{H_i}^{\mathcal{S}_h}\}$ is the sequence of corresponding retargeted end-effector actions derived from the human. Robot state observations include gripper pose and opening width, while human hand states come from a hand pose estimation algorithm. Each robot or human-retargeted action consists of gripper pose and opening width. To account for morphological differences between the human hand and a parallel-jaw gripper, MonoDuo includes a module that translates estimated human hand poses into gripper actions. Details on obtaining human-retargeted actions are provided in Section IV.

MonoDuo then augments $\mathcal{D}^{\mathcal{S}}$ into $\mathcal{D}^{\text{Aug}}$ to train a bimanual robot policy that can be deployed on a specified target bimanual robot $\mathcal{T}$ without test-time modification. This is illustrated in Figure 2. We assume the grippers of robot $\mathcal{S}$ and robot $\mathcal{T}$ are both parallel-jaw grippers, and that each single-arm robot with gripper has kinematics that can be approximated with a human arm and hand. We also assume fixed and known camera poses for both the source and target domains. This allows us to render robots with known URDFs in ways that are within the training image distribution. Similar to prior work [46], [58], [70], [71], we use Cartesian control and assume known inverse kinematics of the end-effector coordinate frames with respect to robot bases, such that we can use a rigid transformation $T_{\mathcal{T}}^{\mathcal{S}}$ to preprocess the data and align all end-effector poses $p^{\mathcal{S}} = T_{\mathcal{T}}^{\mathcal{S}} p^{\mathcal{T}}$ and actions $a^{\mathcal{S}} = T_{\mathcal{T}}^{\mathcal{S}} a^{\mathcal{T}}$ into the same vector space. Thus, for notational convenience, we omit the superscript differentiating end-effector poses and actions between $\mathcal{S}$ and $\mathcal{T}$.

After data augmentation, we learn a policy $\pi(a_t | o_t^{\mathcal{T}}, p_t)$ on $\mathcal{D}^{\text{Aug}}$ using a behavior cloning algorithm of choice. At test time, this policy takes as inputs the observations from the target robot and outputs actions that can be deployed on the target robot. In a second set of experiments, we co-train $\mathcal{D}^{\text{Aug}}$ with a small number of demonstration data $\mathcal{D}^{\mathcal{T}}$ directly obtained from the target bimanul robot, and study how this leads to improvement on few-shot generalization.

IV. MonoDuo

In this section, we describe more details of how MonoDuo enables the learning of bimanual robot policies when only a single-arm robot is available. An overview is shown in

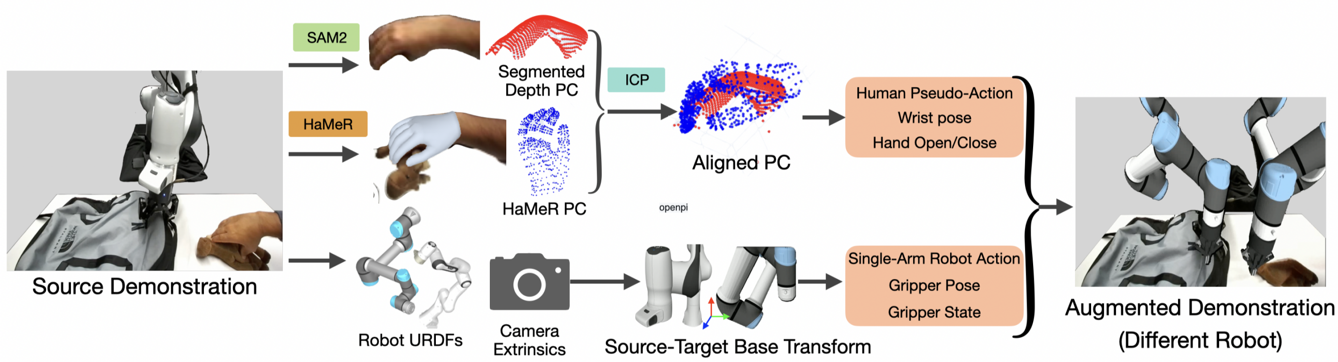

Fig. 3: **Data Collection and Dataset Augmentation.** *Left*: We apply HaMeR [72] to estimate the hand pose at each frame and refine with ICP [73], [74]. The refined hand pose is then retargeted into robot end-effector actions in the source dataset. *Right*: We perform cross-painting from both the source robot and the human arm to the target robot.

Figure 1.

A. Data Collection

For each bimanual task, we collect a source dataset $\mathcal{D}^{\mathcal{S}}$ using a human to teleoperate a single-arm robot to collaborate with a human partner on the task. To ensure a balanced data distribution, the roles of left-arm and right-arm are alternated across episodes. Specifically, human arm and robot collect N trajectories on each side, where the total number of trajectories is $2N$. Data is collected in the format outlined in Section III.

We resolve the morphology gap between human and robot by retargeting the human arm-hand motions into end-effector actions. This is feasible based on two observations: (1) human wrist pose can be approximated as robot end-effector pose; (2) human hand pose can be approximated as gripper state. We begin by estimating the 3D human hand pose at each timestep—specifically, by applying HaMeR [72] to each RGB image from camera observation $o_t^{\mathcal{S}}$. HaMeR predicts 21 keypoints, $\hat{\mathbf{X}}_t \in \mathbb{R}^{21\times3}$, corresponding to anatomical landmarks following the MANO [75] model. Since HaMeR struggles to estimate the absolute 3D pose due to its reliance on a monocular image, we incorporate depth to refine this estimate. In the RGB image observation, we use SAM2 [76] to obtain a segmentation mask of the hand; then, we extract a partial point cloud of the hand by applying the segmentation mask on the aligned depth image. Next, we align the HaMeR-predicted mesh $\hat{\mathbf{V}}_t$ with the segmented hand point cloud $\mathbf{P}_t$ via Iterative Closest Point (ICP) registration [73], obtaining the optimal rigid transformation $\mathbf{T}_t \in SE(3)$ such that $\mathbf{P}_t \approx \mathbf{V}_t = \mathbf{T}_t\hat{\mathbf{V}}_t$. Since $\hat{\mathbf{V}}_t$ and $\hat{\mathbf{X}}_t$ are internally consistent, we can apply $\mathbf{T}_t$ to the predicted keypoints to refine their positions: $\mathbf{X}_t = \mathbf{T}_t\hat{\mathbf{X}}_t$. Once the keypoints $\mathbf{X}_t$ are refined, we define the human retargeted actions $a_t^{\mathcal{S}_h}$ in $\mathcal{D}^{\mathcal{S}}$ as follows: the end-effector pose is set as the estimated wrist pose, and the gripper opening is computed as a binary variable based on the scalar angle defined by three MANO [75] landmarks: thumb fingertip, index finger fingertip, and index proximal frame. We set a threshold value for the scalar angle value, such that angle below which is translated to a closed gripper.

B. Dataset Augmentation

Given the source dataset $\mathcal{D}^{\mathcal{S}}$, we aim to augment it into $\mathcal{D}^{\text{Aug}}$ to learn a bimanual policy that can be deployed on the target bimanual robot $\mathcal{T}$. To this end, we apply "cross-painting"—which in prior works [57], [58] means replacing the source robot with the target robot in the camera observations at test time so that it appears to the policy as if the source robot were performing the task. In this work, we extend cross-painting to also include human as a data source. We describe the details below, and illustrate the cross-painting procedure in Figure 2.

Source Robot to Target Robot Cross-Painting. We leverage knowledge of the source and target robot URDFs and camera poses to perform robot-robot cross-painting at training time, as illustrated in Figure 3. First, given known camera extrinsics, we re-project the images from the source domain to the target domain given that depth sensing is available. Next, given the RGB image observation and joint angles of the source robot, we use a renderer to determine which image pixels correspond to the source robot and mask out these pixels. Then, we inpaint the missing pixels using a video inpainting model E^{2}FGVI [77]. Finally, we use the URDF of the target robot to solve for the joint angles that would put its end effector at the same pose as that of the source robot, render it using a simulator, and overlay it onto the source image. For the gripper, we similarly compute and set the joints of the target robot gripper in the renderer so that its width would roughly match that of the source robot's gripper. To prevent the trained policy on the augmented data from overfitting to the synthetic robot visuals, we perform random brightness augmentation to the generated robot before pasting it. Previous work [57] has demonstrated this random brightness augmentation to significantly help improve the performance of trained policies.

Human to Target Robot Cross-Painting. Cross-painting from human to target robot largely follows the same process as robot-robot cross-painting, except that we segment out the pixels corresponding to the human arm using SAM2 [76] before replacing the human embodiment with a robot. The target model is similarly rendered, with its end effector pose and gripper opening width corresponding to the human

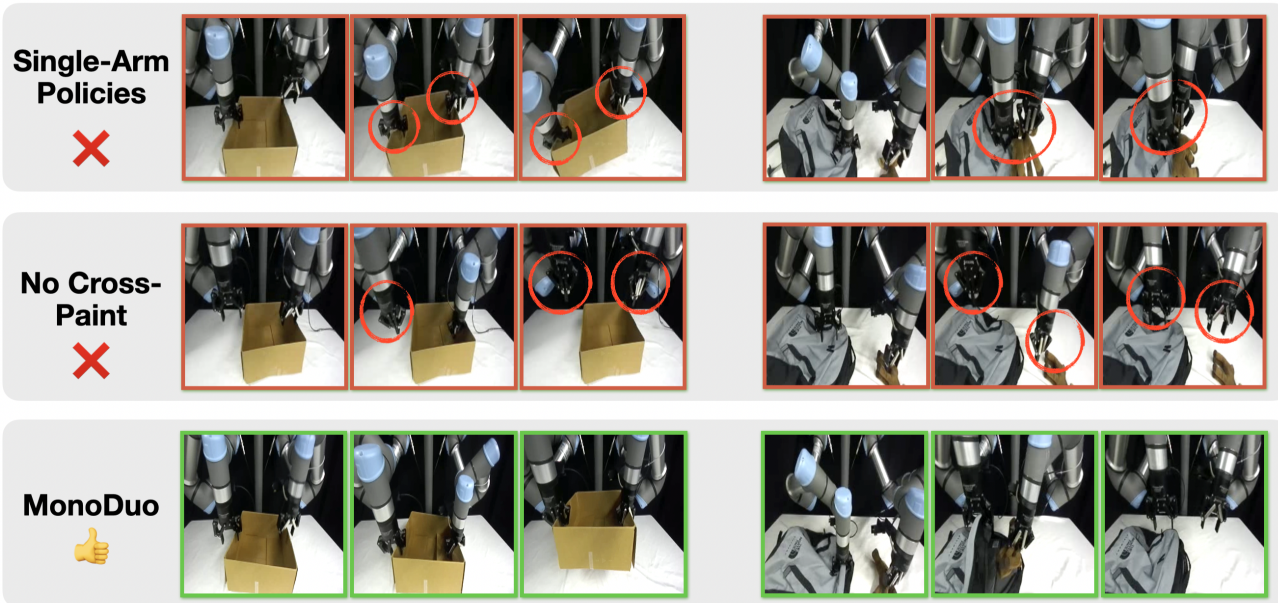

Fig. 4: **Examples of zero-shot rollout on the target bimanual UR5e.** *Left:* Lift Box; *Right:* Pack Bag. Single-Arm policies do not coordinate the actions well, leading to asynchronous movements as shown in the Lift Box task and collision in the Pack Bag task. Policies trained without cross-painting is less robust and misgrasps often. MonoDuo exhibits coordinated behaviors while being precise.

retargeted end-effector action extracted.

C. Policy Training

After applying dataset augmentation, we can train a policy π based on the Diffusion Policy architecture [23] on the augmented dataset $\mathcal{D}^{Aug}$ and zero-shot deploy the policy on the target robot $\mathcal{T}$. The policy input is RGB image observations and bimanual robot state observations; policy output is bimanual robot actions. For challenging tasks or when there is a large difference in the dynamics between the robots, we can also collect a small demonstration dataset $\mathcal{D}^{\mathcal{T}}$ on the target robot directly and few-shot finetune π on $\mathcal{D}^{\mathcal{T}}$ to further improve policy performance.

V. Experiments

A. Hardware Setup and Task Definition

We use a Franka arm as the single-arm source robot, and a pair of UR5e arms setup as the bimanual target robot. For RGB-D data collection, we use a ZED2 stationary fixed camera and a ZED-mini wrist-mounted camera. We design five bimanual tasks for policy evaluation: (1) **Box Lifting**: the robot needs to coordinate the two grippers to lift up a box; (2) **Backpack Packing**: the robot needs to use one gripper to open a backpack, pick up a toy using the other gripper, put the toy into the backpack, and finally close the backpack with the first arm; (3) **Jacket Zipping**: the robot uses one arm to pin the jacket and the other arm to grasp and zip up the zipper of a jacket; (4) **Plate Handover**: the robot uses one gripper to pick up a plate and hands it over to the other gripper, while the other gripper needs to come to the waiting pose, stably grasp the plate, and put it down; (5) **Cloth Folding**: the robot needs to coordinate the two grippers to fold a piece of cloth by half. All tasks require highly coordinated behaviors of two arms and cannot be accomplished with a single-arm robot.

B. Implementation Details

We collect 200 demonstration trajectories on the source robot for each task, half of which has human on the left side and the other half the right side. Starting object positions and starting robot positions are randomized for each demonstration. For the generalization experiment, we additionally collect 25 trajectories on the target robot. On average, collecting a collaborative demonstration took 37 seconds, whereas a full teleoperated bimanual demonstration required $\sim$3 minutes, thus our collaborative setup reduced collection time by $\sim$79%. This efficiency enables us to scale dataset collection more quickly compared to prior bimanual teleoperation setups. On both setups, we use Meta Quest as the teleoperation device. We use the UNet-based Diffusion Policy as outlined in Chi et al. [23], with a ResNet encoder for visual observations. Policies take Cartesian proprioception, 2 image observations, and predict Cartesian end-effector actions.

C. Results

Zero-Shot Bimanual Policies. We report success rates of zero-shot deployed MonoDuo policies on each evaluation task in Table I and visualize their qualitative behaviors in Figure 4. These results show that MonoDuo is able to effectively bridge both the visual and physical gaps among different robots and human, allowing one to learn bimanual policies when only a single-arm robot is available.

Few-Shot MonoDuo. We study the finetuned performance of MonoDuo in comparison with bimanual policy training from scratch. We collect an additional 25 trajectories per task from direct teleoperation of the target bimanual UR5e robot. Using

Policies	Policy Attributes				Task Success Rates				
	Use Robot	Cross Paint	Retargeted Action	Weight Sharing	Lift Box	Pack Bag	Zip Jacket	Handover Plate	Fold Cloth
Robot-Only Policies (Naive)	✓	✓			15%	10%	15%	0%	5%
Pure Human Videos		✓	✓	✓	10%	0%	0%	0%	0%
Ablation: No Cross-Paint	✓		✓	✓	40%	30%	15%	10%	15%
Ablation: No Retargeted Action	✓	✓		✓	50%	20%	20%	30%	25%
MonoDuo	✓	✓	✓	✓	**70%**	**55%**	**45%**	**35%**	**35%**

TABLE I: **Zero-shot experiments comparing MonoDuo with baselines.** Each policy is evaluated on five manipulation tasks in a zero-shot transfer setting from Franka-human demos to a bimanual UR5e. We collect 200 Franka-human demos per task. Most of MonoDuo's failures are due to missed grasps, with some two-arm coordination failures for the *Plate Handover* and *Cloth Folding* tasks. In contrast, the baseline policies have a large number of two-arm coordination failures on all tasks.

	No WristCam	*With WristCam*
Lift Box	60%	**70%**
Pack Bag	40%	**55%**
Zip Jacket	25%	**45%**

TABLE II: **Impact of Wrist Camera on Zero-Shot Performance.** Using only a third-person camera yields strong results, but wrist-mounted cameras improve precision in tasks requiring fine manipulation, such as zipper grasping.

	Scratch	*Few-Shot MonoDuo*
Lift Box	30%	**100%**
Pack Bag	25%	**90%**
Zip Jacket	5%	**75%**

TABLE III: **Few-Shot Learning with MonoDuo.** Incorporating 25 bimanual robot demonstrations enables MonoDuo to significantly outperform policies trained from scratch, demonstrating improved sample efficiency.

these 25 trajectories, we finetune MonoDuo and also train new policies from scratch on the same data. This situation corresponds to a common realistic scenario, where only a small number of demonstrations on the target bimanual robot is available. Results in Table III demonstrate that MonoDuo significantly improves learning efficiency in comparison to policy training from scratch, achieving higher performance with the same number of real demonstrations. Notably, few-shot MonoDuo boosts success rates on *box lifting* from 30% to 100%, *backpack packing* from 25% to 90%, and *jacket zipping* from 5% to 75%, highliting our approach's ability to complement limited real-world bimanual data.

Comparison with Baselines. We evaluate MonoDuo against 4 baselines: (1) *Robot-Only Policies*: Two independent single-arm policies, each conditioned on the cross-painted global observation and its respective robot state, trained to predict the action for a single arm. (2) *Pure Human Videos*: A policy trained solely on bimanual human-only demonstrations, using cross-painting to simulate robot embodiment and pose estimation of both hands. (3) *No Cross-Paint*: An ablation that removes the visual domain alignment step, training instead on raw images while still leveraging both human and robot action supervision. (4) *No Retargeted Actions:* Similar to MonoDuo in using a unified policy architecture, but excludes human retargeted actions during training, relying only on robot action supervision. Quantitative results in Table I show that all three core components of MonoDuo —joint human-robot demonstrations, robot-robot cross-painting, and human-robot cross-painting—are essential for learning effective bimanual coordination policies. Figure 4 show some qualitative examples, and we analyze key insights from ablation studies below.

Importance of Weight-Sharing. Our results indicate that using two disjoint robot-only policies, even when paired with cross-painted visual input, fails to produce reliable coordination. Each arm tends to execute its part of the task independently, leading to asynchronous behavior. While some tasks may succeed occasionally, the overall quality is poor—for example, lifting a box unevenly—and no success is observed in tasks demanding precise temporal synchronization.

Importance of Cross-Painting. Removing cross-painting results in a 20–30% drop in success across all tasks, demonstrating the critical role of visual domain alignment in enabling effective transfer. Cross-painting helps mitigate the embodiment gap and enables the model to generalize better.

Value of Human Retargeted Actions. Training solely on human video data yields near-zero success in zero-shot settings, primarily due to noisy hand pose estimations. To enable a fair comparison, we also include results of MonoDuo trained without wrist cameras in Table II. MonoDuo addresses this limitation through structured augmentation: it preserves real robot actions for high-fidelity supervision while incorporating human retargeted actions to balance the training distribution. This reduces reliance on perfect human pose estimation, enhances policy robustness, and enables fair cross-embodiment transfer. Empirically, combining human retargeted actions with robot actions improves performance across tasks, even when hand pose estimation is imperfect, demonstrating that leveraging both sources produces more reliable bimanual policies than using human data alone.

VI. Conclusion

We present MonoDuo, a framework for learning bimanual robot policies using only single-arm robot demonstrations paired with a human. By alternating roles and applying vision-based augmentation, MonoDuo generates synthetic bimanual demonstrations for a target robot, enabling zero-shot generalization to unseen configurations and improved sample efficiency in low-data regimes. We validate MonoDuo on five challenging bimanual tasks, showing superior performance over baselines, highlighting its potential as a scalable solution for bimanual robot learning.

VII. Limitations and Future Work

While MonoDuo provides a scalable framework for learning bimanual policies from a single-arm robot, several limitations remain. First, the approach assumes fixed and known camera calibration across domains, simplifying rendering and cross-painting but limiting use in uncalibrated settings. Second, it requires depth sensing for 3D hand pose refinement and segmentation, necessitating an RGB-D camera.

MonoDuo's mix of human-retargeted and real robot actions mitigates some challenges, but our augmentation pipeline is sensitive to noise in hand-pose estimation and inpainting. Under occlusion or poor lighting, this can degrade policy quality. However, our pipeline's modular design enables integration of newer, improved models for hand-pose estimation, segmentation, and inpainting, which can improve the quality of human retargeted actions and overall policy performance. Including such improvements in future work could further enhance robustness and reliability.

References

[1] S. Stepputtis, M. Bandari, S. Schaal, and H. B. Amor, "A system for imitation learning of contact-rich bimanual manipulation policies," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2022, pp. 11 810–11 817.

[2] J. Grannen, Y. Wu, B. Vu, and D. Sadigh, "Stabilize to act: Learning to coordinate for bimanual manipulation," in *Conference on Robot Learning*, PMLR, 2023, pp. 563–576.

[3] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *RSS*, 2023.

[4] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu, "Low-cost exoskeletons for learning whole-arm manipulation in the wild," in *ICRA*, 2023.

[5] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: Teleoperation with immersive active visual feedback," *arXiv preprint arXiv:2407.01512*, 2024.

[6] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2024, pp. 12 156–12 163.

[7] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto, "Open teach: A versatile teleoperation system for robotic manipulation," *arXiv preprint arXiv:2403.07870*, 2024.

[8] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, "Learning visuotactile skills with two multifingered hands," *arXiv:2404.16823*, 2024.

[9] B. Huang, Y. Chen, T. Wang, Y. Qin, Y. Yang, N. Atanasov, and X. Wang, "Dynamic handover: Throw and catch with bimanual hands," *arXiv preprint arXiv:2309.05655*, 2023.

[10] T. Lin, Z.-H. Yin, H. Qi, P. Abbeel, and J. Malik, "Twisting lids off with two hands," *arXiv:2403.02338*, 2024.

[11] T. Lin, K. Sachdev, L. Fan, J. Malik, and Y. Zhu, "Sim-to-real reinforcement learning for vision-based dexterous manipulation on humanoids," *arXiv:2502.20396*, 2025.

[12] H. Xiong, Q. Li, Y.-C. Chen, H. Bharadhwaj, S. Sinha, and A. Garg, "Learning by watching: Physical imitation of manipulation skills from human videos," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2021, pp. 7827–7834.

[13] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild," *arXiv preprint arXiv:2207.09450*, 2022.

[14] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar, "Mimicplay: Long-horizon imitation learning by watching human play," *arXiv preprint arXiv:2302.12422*, 2023.

[15] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu, "Okami: Teaching humanoid robots manipulation skills through single video imitation," in *8th Annual Conference on Robot Learning*, 2024.

[16] A. Bahety, P. Mandikal, B. Abbatematteo, and R. Martín-Martín, "Screwmimic: Bimanual imitation from human videos with screw space projection," *arXiv preprint arXiv:2405.03666*, 2024.

[17] Y. Zhu, A. Lim, P. Stone, and Y. Zhu, "Vision-based manipulation from single human video with open-world object graphs," *arXiv preprint arXiv:2405.20321*, 2024.

[18] H. Zhou, R. Wang, Y. Tai, Y. Deng, G. Liu, and K. Jia, "You only teach once: Learn one-shot bimanual robotic manipulation from video demonstrations," *arXiv preprint arXiv:2501.14208*, 2025.

[19] Y. Chen, C. Wang, Y. Yang, and C. K. Liu, "Object-centric dexterous manipulation from human motion data," *arXiv preprint arXiv:2411.04005*, 2024.

[20] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "Dexcap: Scalable and portable mocap data collection system for dexterous manipulation," *arXiv preprint arXiv:2403.07788*, 2024.

[21] M. Zare, P. M. Kebria, A. Khosravi, and S. Nahavandi, "A survey of imitation learning: Algorithms, recent developments, and challenges," *IEEE Transactions on Cybernetics*, 2024.

[22] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson, "Implicit behavioral cloning," in *Conference on robot learning*, PMLR, 2022, pp. 158–168.

[23] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Learning agile robotic locomotion skills by imitating animals*, 2023.

[24] O. X.-E. Collaboration *et al.*, *Open X-Embodiment: Robotic learning datasets and RT-X models*, IEEE International Conference on Robotics and Automation, 2024.

[25] C. Yu, W. Zhang, H. Lai, Z. Tian, L. Kneip, and J. Wang, "Multi-embodiment legged robot control as a sequence modeling problem," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023, pp. 7250–7257.

[26] T. Chen, A. Murali, and A. Gupta, "Hardware conditioned policies for multi-robot transfer learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[27] L. Shao, F. Ferreira, M. Jorda, V. Nambiar, J. Luo, E. Solowjow, J. A. Ojea, O. Khatib, and J. Bohg, "Unigrasp: Learning a unified model to grasp with multifingered robotic hands," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2286–2293, 2020.

[28] Z. Xu, B. Qi, S. Agrawal, and S. Song, "Adagrasp: Learning an adaptive gripper-aware grasping policy," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021, pp. 4620–4626.

[29] T. Wang, R. Liao, J. Ba, and S. Fidler, "Nervenet: Learning structured policy with graph neural networks," in *International conference on learning representations*, 2018.

[30] A. Sanchez-Gonzalez, N. Heess, J. T. Springenberg, J. Merel, M. Riedmiller, R. Hadsell, and P. Battaglia, "Graph networks as learnable physics engines for inference and control," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, Oct. 2018, pp. 4470–4479. [Online]. Available: https://proceedings.mlr.press/v80/sanchez-gonzalez18a.html.

[31] D. Pathak, C. Lu, T. Darrell, P. Isola, and A. A. Efros, "Learning to control self-assembling morphologies: A study of generalization via modularity," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[32] W. Huang, I. Mordatch, and D. Pathak, "One policy to control them all: Shared modular policies for agent-agnostic control," in *International Conference on Machine Learning*, PMLR, 2020, pp. 4455–4464.

[33] 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*, OpenReview, 2021.

[34] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 3511–3516.

[35] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, *et al.*, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Conference on robot learning*, PMLR, 2018, pp. 651–673.

[36] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International journal of robotics research*, vol. 37, no. 4-5, pp. 421–436, 2018.

[37] C. Eppner, A. Mousavian, and D. Fox, "ACRONYM: A large-scale grasp dataset based on simulation," in *2021 IEEE Int. Conf. on Robotics and Automation, ICRA*, 2020.

[38] N. M. M. Shafiullah, A. Rai, H. Etukuru, Y. Liu, I. Misra, S. Chintala, and L. Pinto, *On bringing robots home*, 2023. arXiv: 2311.16098 [cs.RO].

[39] H.-S. Fang, H. Fang, Z. Tang, J. Liu, J. Wang, H. Zhu, and C. Lu, "RH20T: A robotic dataset for learning diverse skills in one-shot," in *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[40] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine, "Bridge data: Boosting generalization of robotic skills with cross-domain datasets," in *Robotics: Science and Systems (RSS) XVIII*, 2022.

[41] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning*, PMLR, 2023, pp. 1723–1736.

[42] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning*, PMLR, 2022, pp. 991–1002.

[43] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "RT-1: Robotics transformer for real-world control at scale," *Robotics: Science and Systems (RSS)*, 2023.

[44] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*, PMLR, 2023, pp. 2165–2183.

[45] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "VIMA: General robot manipulation with multimodal prompts," *International Conference on Machine Learning (ICML)*, 2023.

[46] D. Shah, A. Sridhar, A. Bhorkar, N. Hirose, and S. Levine, "GNM: A general navigation model to drive any robot," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023, pp. 7226–7233.

[47] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine, "ViNT: A Foundation Model for Visual Navigation," in *7th Annual Conference on Robot Learning (CoRL)*, 2023.

[48] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence, "Interactive language: Talking to robots in real time," *IEEE Robotics and Automation Letters*, 2023.

[49] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on Robot Learning*, PMLR, 2022, pp. 894–906.

[50] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, S. Kirmani, B. Zitkovich, F. Xia, *et al.*, "Open-world object manipulation using pre-trained vision-language models," in *Conference on Robot Learning*, PMLR, 2023, pp. 3397–3417.

[51] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

[52] S. Reed *et al.*, "A generalist agent," *Transactions on Machine Learning Research*, 2022, ISSN: 2835-8856.

[53] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *Conference on Robot Learning*, 2022.

[54] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar, "Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2024, pp. 4788–4795.

[55] X. Chen *et al.*, *Pali-x: On scaling up a multilingual vision and language model*, 2023. arXiv: 2305.18565 [cs.CV].

[56] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model," in *International Conference on Machine Learning*, PMLR, 2023, pp. 8469–8488.

[57] L. Y. Chen, C. Xu, K. Dharmarajan, M. Z. Irshad, R. Cheng, K. Keutzer, M. Tomizuka, Q. Vuong, and K. Goldberg, "Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning," in *Conference on Robot Learning (CoRL)*, Munich, Germany, 2024.

[58] L. Y. Chen, K. Hari, K. Dharmarajan, C. Xu, Q. Vuong, and K. Goldberg, "Mirage: Cross-embodiment zero-shot policy transfer with cross-painting," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[59] M. Lepert, R. Doshi, and J. Bohg, "Shadow: Leveraging segmentation masks for zero-shot cross-embodiment policy transfer," in *Conference on Robot Learning (CoRL)*, Munich, Germany, 2024.

[60] M. Lepert, J. Fang, and J. Bohg, *Phantom: Training robots without robots using only human videos*, 2025. arXiv: 2503.00779 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2503.00779.

[61] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, *Egomimic: Scaling imitation learning via egocentric video*, 2024. arXiv: 2410.24221 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2410.24221.

[62] M. Lepert, J. Fang, and J. Bohg, *Masquerade: Learning from in-the-wild human videos using data-editing*, 2025. arXiv: 2508.09976 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2508.09976.

[63] V. Jain, M. Attarian, N. J. Joshi, A. Wahid, D. Driess, Q. Vuong, P. R. Sanketi, P. Sermanet, S. Welker, C. Chan, *et al.*, "Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers," *arXiv preprint arXiv:2403.12943*, 2024.

[64] K. Kedia, P. Dan, A. Chao, M. A. Pace, and S. Choudhury, *One-shot imitation under mismatched execution*, 2024. arXiv: 2409.06615 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2409.06615.

[65] R. Doshi, H. Walke, O. Mees, S. Dasari, and S. Levine, "Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation," *arXiv preprint arXiv:2408.11812*, 2024.

[66] Z. Jiang, Y. Xie, K. Lin, Z. Xu, W. Wan, A. Mandlekar, L. Fan, and Y. Zhu, "Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[67] G. Lu, T. Yu, H. Deng, S. S. Chen, Y. Tang, and Z. Wang, "Anybimanual: Transferring unimanual policy for general bimanual manipulation," *arXiv preprint arXiv:2412.06779*, 2024.

[68] M. Kobayashi, J. Yamada, M. Hamaya, and K. Tanaka, "Lfdt: Learning dual-arm manipulation from demonstration translated from a human and robotic arm," in *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, 2023, pp. 1–8. DOI: 10.1109/Humanoids57100.2023.10375192.

[69] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[70] J. H. Yang, D. Sadigh, and C. Finn, "Polybot: Training one policy across robots while embracing variability," in *Conference on Robot Learning*, PMLR, 2023, pp. 2955–2974.

[71] J. Yang, C. Glossop, A. Bhorkar, D. Shah, Q. Vuong, C. Finn, D. Sadigh, and S. Levine, "Pushing the limits of cross-embodiment learning for manipulation and navigation," 2024.

[72] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, "Reconstructing hands in 3D with transformers," in *CVPR*, 2024.

[73] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, Spie, vol. 1611, 1992, pp. 586–606.

[74] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," *Image and vision computing*, vol. 10, no. 3, pp. 145–155, 1992.

[75] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 245:1–245:17, vol. 36, no. 6, Nov. 2017.

[76] N. Ravi *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024. [Online]. Available: https://arxiv.org/abs/2408.00714.

[77] Z. Li, C.-Z. Lu, J. Qin, C.-L. Guo, and M.-M. Cheng, "Towards an end-to-end framework for flow-guided video inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.