H2R: A Human-to-Robot Data Augmentation for Robot Pre-training from Videos

Anonymous CVPR submission

Paper ID 00076

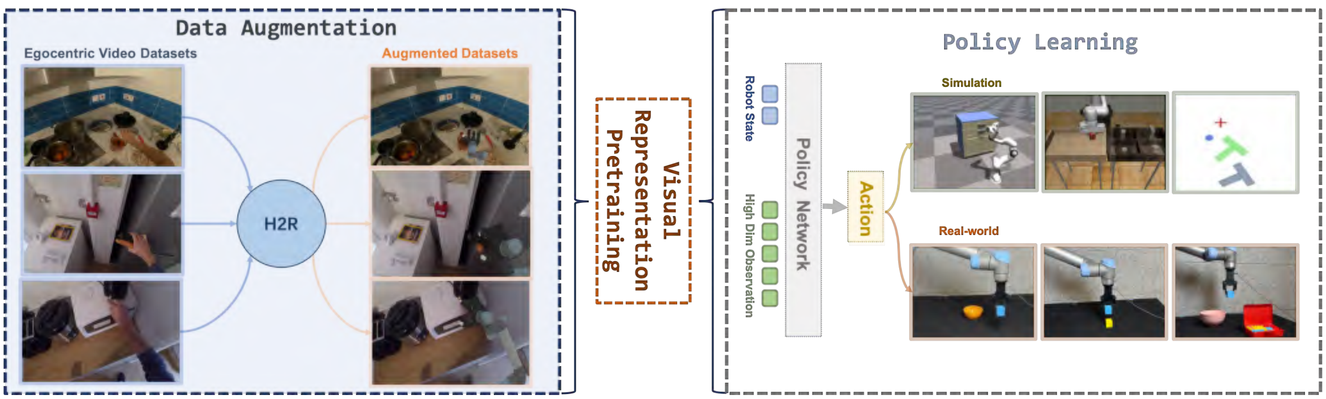

Figure 1. **H2R** is a data augmentation technique designed to enhance robot pre-training by converting first-person human hand operation videos into robot-centric visual data. By bridging the visual domain gap, H2R improves pre-trained visual encoders for downstream robot policies (imitation/reinforcement learning), validated across simulation benchmarks and real-world robotic tasks.

Abstract

Large-scale pre-training using videos has proven effective for robot learning, as it enables the model to acquire task knowledge from first-person human operation data that reveals how humans perform tasks and interact with their environment. However, the models pre-trained on such data can be suboptimal for robot learning due to the significant visual gap between human hands and those of different robots. To remedy this, we propose H2R, a simple data augmentation technique for robot pre-training from videos, which extracts the human hands from first-person videos and replaces them with those of different robots to generate new video data for pre-training. Specifically, we start by detecting the 3D position and key points of human hands, which serves as the basis for generating robots in the simulation environment that exhibit similar motion postures. Then, we calibrate the intrinsic parameters of the simulator camera to match the camera in the original video and render the images of generated robots. Finally, we overlay these images onto the original video to replace human hands. Such a procedure bridges the visual gap between the human hand and the robotic arm and produces an augmented dataset for pre-training. We conduct extensive experiments on a variety of robotic tasks, ranging from standard simulation benchmarks to robotic real-world tasks, with varying pre-training strategies, video datasets, and policy learning methods. The experimental results show that H2R can improve the representation capability of visual encoders pre-trained by various methods. In imitation learning, H2R consistently enhances the average success rate across different pre-training methods, with improvements ranging from 0.9% to 10.2%. The effect of this improvement is highly stable. In reinforcement learning, most pre-training methods show improvements. Our real-world evaluations across diverse manipulation tasks demonstrate that H2R-enhanced visual representations consistently outperform baseline models, achieving success rate improvements ranging from 6.7% to 15% across all model-task configurations.

1. Introduction

Pre-training of generalizable robotic features for object manipulation and motion navigation constitutes a crucial objective within the realm of robotics. Inspired by the remarkable accomplishments of large scale pre-training in computer vision [18, 28, 34, 46, 68] and natural language processing [1, 12, 41, 47, 65], many efforts have been devoted to harness large-scale data to construct generalizable representations in the robotics field [2, 3, 8]. Nevertheless, when it comes to robot manipulation, the process of collecting demonstrations is not only labor-intensive but also highly costly [3, 15, 16, 19, 29, 32, 33, 35, 37]; at the same time, there exist many large-scale egocentric video datasets showing how human perform manipulation and navigation, which can serve as a can serve as a cheap alternative of demonstrations for the pre-training of generalizable visual features for robotics.

Recent works [46, 63, 69] analyze such egocentric human video datasets such as Ego4D [23], SSV2 [20], and Epic Kitchens [9] with the aim of gleaning prior knowledge about object manipulation and enabling the acquisition of general and robust feature representations. However, the gap in visual representations between the human arm and the robotic arm remain largely unaddressed in prior work and can hinder the transferability of models trained on egocentric datasets to robotic systems. Specifically, when utilizing the robot expert data to fine-tune the pre-trained robotic representations for downstream robotic tasks, the model has to learn to bridge the visual gap between the first person human hand and the robots in addition to acquiring nuanced task-specific skills demonstrated in the robot expert data. This would result in increased complexity during the fine-tuning process and suboptimal performance.

To mitigate this issue, we propose H2R (as shown in Figure 1), a simple data augmentation method that converts videos of **H**uman hand operations into that of **R**obotic arm manipulation. H2R consists of two major procedures: the first part is to generate the robotic arm's movements to imitate the human hand movements in a video, followed by the second stage that overlays the robotic arm's movements onto the human hand's movements in the video. Specifically, in the first part, we employ state-of-the-art 3D hand reconstruction model HaMeR [50] to accurately detect the position and posture of the human hand in egocentric videos. Then, we simulate the same robot state in simulators to obtain the mask of robot hands. While in the second stage, we use the Segment Anything Model [36] to automatically separate human hand from background, and use the inpainting model LaMa [58] to fill the removed hand mask. After that, we align the camera intrinsic parameters of the images detected in HaMeR with those in the simulator, and then achieve pixel-level matching between the robotic arm images in the simulators and the human hand images in the egocentric video. Finally, we overlay the robotic arm images captured by the simulator's camera onto the areas where the human hands are removed. Through such a process, H2R explicitly reduces the gap between human and robot hands by creating realistic robotic arm movements that visually mimic human hand actions. It allows the model to learn the task-specific actions demonstrated by the human hand, but with robotic arm visual representations that are more suitable for robotic systems.

For pre-training, we used the SSV2 dataset with 62,500 videos, from which 16 keyframes were randomly sampled per video for MAE [28] and R3M [46]. Additionally, we extracted 117,624 action clips from 2,486 videos in the Ego4D dataset for MPI [68]. Specific settings are detailed in the experimental section.

We demonstrate the effectiveness of H2R by integrating it into a holistic policy learning framework. We trained standard MAE [28], R3M [46], and MPI [68] vision encoders on egocentric videos obtained by the proposed H2R. We then freeze the encoder model as a feature extractor and train both an Reinforcement Learning (RL) policy by employing mainstream RL learning methods such as PPO [56] and Imitation Learning (IL) policy with behavior clone and Diffusion Policy [6]. Finally, for the RL policies, we evaluate them on MVP [51], a closed-loop benchmark, and compare with results where the encoders are trained on the original egocentric video data. We observe a significant improvement of the training stability, which brings more effectiveness in the context of RL policy learning seeing the unstable nature of the bare RL training. For the IL policies, the BC policies are trained and tested on Robomimic [45], while the Diffusion Policy models are trained and tested on their own baseline. Both of the BC and Diffusion policies showed a significant improvement on the success rate and stability on the downstream tasks.

Through extensive real-world experiments, we validate the effectiveness of H2R in real-world robotic manipulation tasks. We employ Diffusion Policy [6] (DP) and Equivariant Diffusion Policy [64] (eq-dp) as policy frameworks for downstream training, integrating pre-trained visual representation models MAE and R3M into the policy networks. The results demonstrate that H2R significantly enhances the performance of both MAE and R3M-based policies.

Our paper provides three contributions:

- We propose a data-centric pipeline, H2R, to mitigate the gap between human and robot hands when utilizing large-scale egocentric video datasets to pre-train generalizable visual features for robots.
- We apply H2R to SSV2 and Ego4D datasets and train a visual encoder that is more suitable for robotic tasks. Built upon this, we yield a robust robotic manipulation policy through RL and IL training on robot expert data.
- We demonstrate the effectiveness of H2R through extensive experiments on closed-loop benchmarks.

CVPR #00076

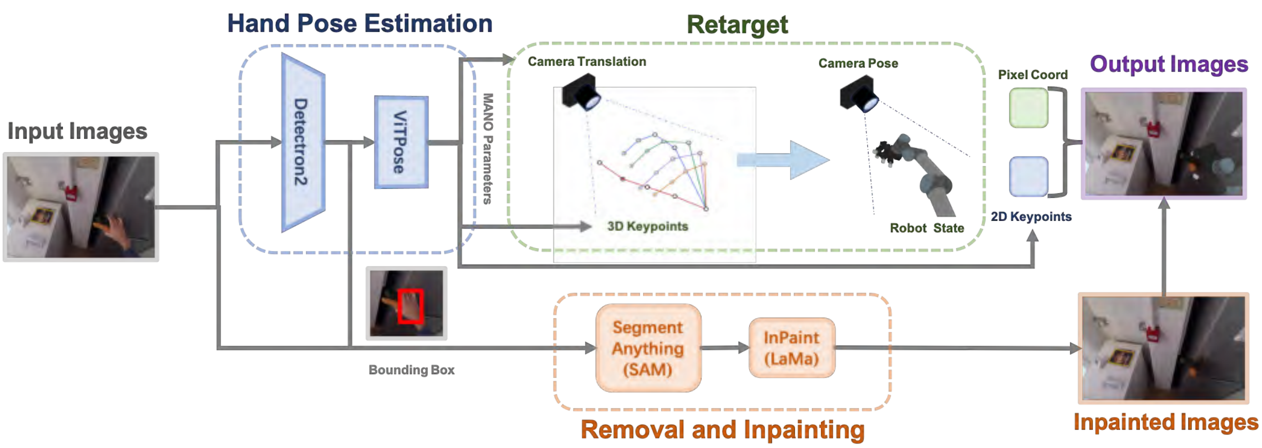

Figure 2. **H2R Pipeline.** H2R involves replacing human hands with robot arms by first using the HaMeR model to detect hand poses and camera parameters. The human hand is then removed using the SAM, and the inpainting model LaMa fills in the gap. A robot hand is constructed based on the detected pose and keypoints, with the camera perspective adjusted to match the original image. Finally, the robot hand is overlaid onto the image, ensuring accurate alignment with the human hand.

2. Related Work

2.1. Robot Policy Learning

Training robot policies [6, 8, 40, 43, 63, 64] in a data-driven manner have been adopted by the robotics community as well as the machine learning community. This serves as a paradigm to automatically yield models for performing robotic tasks including grasping, manipulation, locomotion, navigation, and other complex tasks. [39, 59]. Currently, policy learning methods can be classified into two types: imitation learning (IL)-based [6, 45, 67, 70] and reinforcement learning (RL)-based [24, 56].

IL-based methods [6, 45, 67, 70] train robot policies based on successful demonstration of task execution within the dataset. Supervised by behavior cloning [17, 60] objective along with other auxiliary objectives, the policy predicts a sequence of future actions based on current and past observations. To deal with the non-markovian transition of robot configuration under scenario such as stationary process, ACT [70] employes a temporal fusion of sequence predicted at multiple time steps and thus mitigates the related confounder problem. To deal with the multi-modality nature of robot motion, diffusion models are adopted [6, 67]. For IL-based methods, data diversity contributes largely to the generalizability of model.

On the other hand, RL-based methods [66] resort for the RL paradigm of learning an optimal policy by defining a reward function. These methods formulate robotic manipulation tasks as MDP processes and apply RL algorithms such as PPO [56], SAC [24], and more. Typically, RL for robotics tasks are realized by researchers via RL training in simulator, sim2real transfer, and policy deploying on real robots such as legged robots or aerial drones for locomotion [30, 38, 71], robot arms and dexterous hands for manipulation [62, 66], mobile robots for navigation [7].

For both IL and RL-based methods, a strong feature extractor backbone serves as a cornerstone for learning a robust policy. Therefore, Well-conceived data-centric pipeline is crucial and contributes to the backbone training.

2.2. Visual Encoder Pretraining for Robotics

Researchers investigated visual representation [48] under various perspectives such as model architecture [13, 25], training objective [26, 27], dataset [11, 42, 55, 57], and more. PVR-Control [49] demonstrates the effectiveness of visual representation which surpasses the state representation under the investigated scenarios. RPT [53] explores tokenized representation of transformer and trains the corresponding encoder through masked token-prediction.

Unsupervised training methods such as Masked Auto-Encoder (MAE) [27] and contrastive learning [4, 5] are employed by researchsers [46, 51] for training video encoder and enhancing generalizability. Specifically, MVP [51] introduced video representation for downstream RL tasks while R3M [46] combines time-contrastive learning and video-language alignment. To effectively perform language-guided robotic tasks, researchers of Voltron [34] utilize MAE [27] and contrastive learning [4, 5] for low-level control and high-level planning, respectively.

2.3. Data Quality to Learning Method

Yielding a universal visual representation through a data-centric fashion is crucial for the visual encoder along with the policy to generalize to in-domain and even out-of-domain scenarios [66]. Data-centric analysis indicates the importance of data regarding to the pretraining of visual

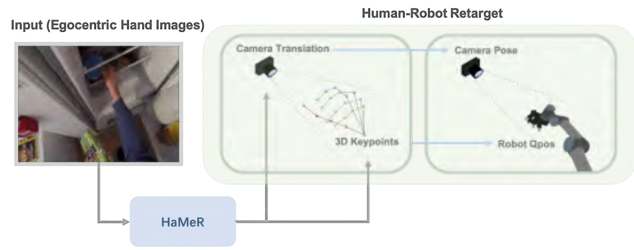

Figure 3. **Human-Robot retarget.** We adopt the HaMeR model to extract hand keypoints and obtain camera parameters, corresponding coordinate systems are constructed, and this information is used to adjust the robot hand's pose and camera perspective, enabling precise hand pose retargeting.

representation such as data distribution and inclusion of out-of-domain data [10], task and domain adaptation [44], etc. Mirage [52] realizes domain transfer of policy between different robots through a pre-processing process within the image space of training data. Our paradigm differs from Mirage from the perspective that we further investigate a more generalized form for robotic representation learning in a data-centric way which results in a more robust policy training.

3. H2R: Human-to-Robot Data Augmentation

In this section, we describe H2R, a data augmentation pipeline for robot pre-training from videos, our key insight is to remove the hands in every frame and replace them with robotic arms. Figure 2 shows H2R pipeline.

Our proposal is to replace the human hands in every frame with that of a robot, generating an augmented dataset D_{aug}. This approach aims to mitigate the visual gap between human hands and robots, facilitating the transfer of knowledge for easier adaptation of models trained on egocentric data to robotic tasks. In particular, we hope that the vision encoder trained on the augmented dataset D_{aug} would outperform that trained on the original dataset D in downstream robotic tasks.

3.1. Pipeline of H2R

3D Hand Pose Estimation. In order to overlay the human hands in the image with different robots, we firstly need an efficient and accurate model to detect the hand information. Recent HaMeR [50], a state-of-the-art 3D hand detection and reconstruction model. we detect the position of the hand and its key points, and the internal and external parameters of the rendering camera of the RGB image. Such position information of the identified hand is then used to remove the hands from the image.

Human Arm and Hand Remove. We leverage the Segment Anything Model (SAM) [36] to automatically separate the human hand from the background using hand pose information detected by HaMeR. Even though there is only hand information provided but no arm information Thus,SAM could detect both hand and arm as a single object and separate it from background, showing good robustness under the varying conditions of clothing and occlusion. Finally, a state-of-the-art inpainting model LaMa [58] is used to fill the removed hand mask. After this step, we obtain the RBG images with the human hand removed for the later stage of adding robot hands.

Robotic Arm and End Effector Construction. The final step involves constructing the robot arm and end effector, then overlaying it onto the generated images from the previous stage (as shown in Figure 3). For the robotic arm reconstruction, Since HaMeR does not provide information about the arm keypoints, we initially set the target robot to a neutral pose and then adjust the missing joint point information. For the robotic end effector reconstruction For the dexterous hand, the angles of each joint are determined by the angles formed by the corresponding three keypoints, while for the gripper, the degree of opening and closing is determined by the distance between the corresponding fingers.

Simulator Camera Position Alignment. The visual bias introduced by the camera perspective is significantly larger than the action retargeting itself; thus, we leverage the hand keypoints and camera parameters from HaMeR to adjust the camera pose in the simulator. Specifically, the two coordinate systems C_H and C_S can be uniquely determined by the human hand and the robot arm, and the camera's position in C_H can be used in C_S to ensure the same perspective of the camera. We build the coordinate system ${}^{W}\mathbf{I}_H$ based on the hand keypoints:

$$ {}^{W}\mathbf{I}_H = \{{}^{w}\mathbf{i}_{H,x}, {}^{w}\mathbf{i}_{H,y}, {}^{w}\mathbf{i}_{H,z}\} \tag{1} $$

Where ${}^{w}\mathbf{i}_{H,x}, {}^{w}\mathbf{i}_{H,y}, {}^{w}\mathbf{i}_{H,z}$ are unit vectors along the x-axes, y-axes and z-axes of the human hand coorinate system. With the keypoints get in HaMeR, we build the three axis of coordinates with the following functions:

$$ \begin{aligned} {}^{w}\mathbf{i}_{H,x} &= {}^{w}\mathbf{i}_{0,9} \\ {}^{w}\mathbf{i}_{H,y} &= {}^{w}\mathbf{i}_{0,9} \times {}^{w}\mathbf{i}_{0,13} \\ {}^{w}\mathbf{i}_{H,z} &= {}^{w}\mathbf{i}_{H,x} \times {}^{w}\mathbf{i}_{H,y} \end{aligned} \tag{2} $$

Where${}^{w}\mathbf{i}_{0,9}, \mathbf{i}_{0,13}$ are unit vectors along middle finger and ring finger. Similarly, To construct the mapping from hand pose to robot arms, we need to get another coordinate system ${}^{W}\mathbf{I}_S$ in the simulator:

$$ {}^{W}\mathbf{I}_S = \{{}^{w}\mathbf{i}_{S,x}, {}^{w}\mathbf{i}_{S,y}, {}^{w}\mathbf{i}_{S,z}\} \tag{3} $$

The method of determaining the axis of coordinates is the same:

$$ \begin{aligned} {}^{w}\mathbf{i}_{S,x} &= {}^{w}\mathbf{i}_{0,2} \\ {}^{w}\mathbf{i}_{S,y} &= {}^{w}\mathbf{i}_{0,2} \times {}^{w}\mathbf{i}_{0,3} \\ {}^{w}\mathbf{i}_{s,z} &= {}^{w}\mathbf{i}_{S,x} \times {}^{w}\mathbf{i}_{S,y} \end{aligned} \tag{4} $$

Where $\mathbf{i}_{0,2}, \mathbf{i}_{0,3}$ are unit vectors along robot fingers that correspond to human middle and ring fingers. We build the following two coordinate transformation matrix to construct the mapping:

$$
\begin{aligned}
{}^{W}_{H}\mathbf{R} &= \begin{pmatrix} {}^{W}\mathbf{I}_H & \mathbf{key}_0 \\ \mathbf{O} & 1 \end{pmatrix} \\
{}^{W}_{S}\mathbf{R} &= \begin{pmatrix} {}^{W}\mathbf{I}_S & \mathbf{ee\_pos} \\ \mathbf{O} & 1 \end{pmatrix}
\end{aligned} \tag{5}
$$

Where $\mathbf{key}_0, \mathbf{ee\_pos}$ are the positions of human wrist and robot wrist. After obtaining the two coordinate systems, we need to determine the position of the camera in the simulator (${}^{W}\mathbf{cam}_{sim}$) and the position of the camera in the real world (${}^{H}\mathbf{cam}_{Real}$), thus we can ensure we get the same pose of the human hand and robot arms

$$
\begin{aligned}
{}^{H}\mathbf{cam}_{Real} &= {}^{W}_{H}\mathbf{R}^{-1} \times {}^{W}\mathbf{cam}_{Real} \\
{}^{S}\mathbf{cam}_{sim} &= {}^{H}\mathbf{cam}_{Real} \\
{}^{W}\mathbf{cam}_{sim} &= {}^{W}_{S}\mathbf{R} \times {}^{W}_{H}\mathbf{R}^{-1} \times {}^{W}\mathbf{cam}_{Real}
\end{aligned} \tag{6}
$$

Robot Hand Rendering and Copy-paste. After setting the action, the segmentation mask of the robot arm is obtained by shooting with the camera. The result of the HaMeR model contains the pixel coordinates of the key points of the human hand. By calculating the pixel coordinates of the corresponding links in the robot hand, the robot hand can be copy-pasted to the original image based on the corresponding relationship, ensuring that it is pixel-level aligned with the original human hand in the image (see Figure 4).

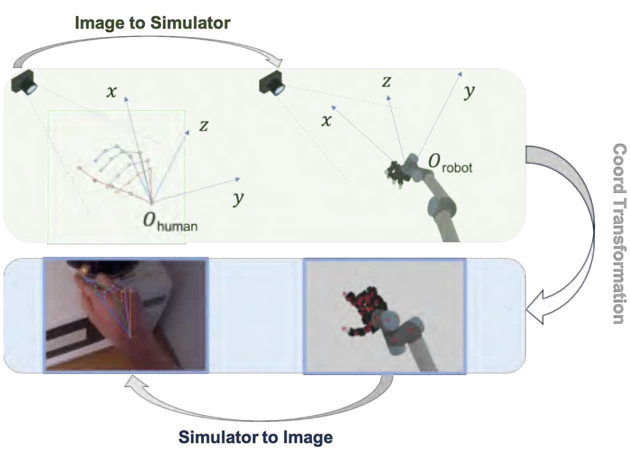

Figure 4. **Robot hand rendering and copy-paste.**The HaMeR model provides hand keypoints and camera parameters, which are used to align the simulator's camera pose with the original view. The robot arm is then rendered in the simulator, and by matching the pixel coordinates of the arm's links, it is overlaid onto the original image with pixel-level alignment to the human hand.

3.2. Model Training

Encoder Pre-training. We adopt the MAE [28, 63], R3M [46], and MPI [68] frameworks for pre-training, each employing a Vision Transformer (ViT) Base [14] model as the visual encoder. The SSv2 dataset [21] is used for MAE and R3M training, whereas the Ego4D dataset [22] is employed for MPI training. For the MAE and R3M pretraining methods, in addition to pre-training on the H2R data and raw data, we also applied a simple CutMix baseline to demonstrate the effectiveness of using the robotic arm to cover the human hand, which overlays a fixed set of specific images of robotic arms with grippers onto the original images, ensuring that the overlaid images cover the human hands as much as possible, without exceeding the detected bounding box. Our H2R is different from such baseline by employing robot hand construction to better match the pose of the hand and arm in the images. Based on the type of robotic arm used in CutMix, we categorize the augmented set into three types: CutMix1 represents the UR5 robotic arm, CutMix2 refers to the Franka robotic arm, and CutMix3 combines both the UR5 and Franka robotic arms.

Policy Training. Finally, we employ several existing policy training methods to fine-tune the pre-trained model for evaluations on downstream robotic tasks. We reuse their original implementations to ensure that any performance improvements are solely attributable to our data augmentation approach. For RL models, we evaluate downstream tasks using the PixMC [63] benchmark and employ PPO [56] for policy learning. Additionally, we utilize Robomimic [45] and Diffusion Policy [6] for evaluating IL models. The Robomimic baseline is primarily used for BC policies, and we test three tasks with the Robomimic datasets. For the Diffusion Policy, we specifically evaluate the push task to assess the robustness of our method across different approaches.

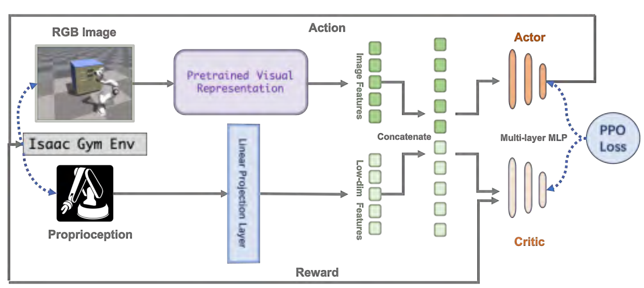

Figure 5. **RL-driven policy training pipeline.** We propose a training pipeline for RL-driven policy learning, designed to evaluate performance across various simulation benchmarks.

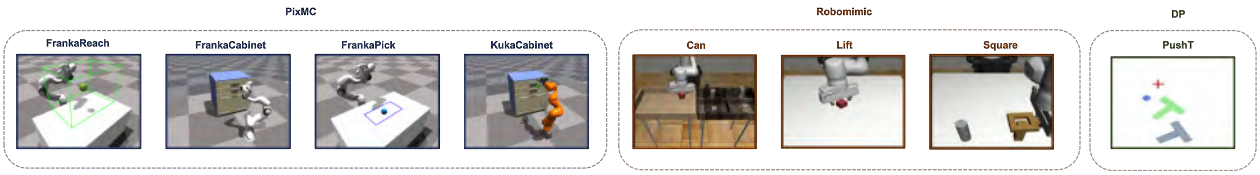

Figure 6. **Simulation benchmark.** We choose 4 tasks from the PixMC and 3 tasks from the Robomimic, covering a range of robotic manipulation skills. We also include the PushT task, designed for the Diffusion Policy framework, as an additional benchmark to evaluate performance in a different task setup.

4. Simulation Experiment

4.1. Experiment Setup

Simulation Benchmark. For evaluations in simulation, we select a total of 8 simulation benchmarks across different environments, which are PixMC [63], Robomimic [45], and Diffusion Policy [6]. In particular, for PixMC, we select **FrankaReach**, **FrankaCabinet**, **FrankaPick**, and **KukaCabinet** to assess the robot's ability to interact with objects. For Robomimic, we include tasks such as **MoveCan**, **Square**, and **Lift**, where the robot performs actions like moving or lifting objects. We also use the **PushT** task, designed for the Diffusion Policy framework, which evaluates a robot's ability to push an object to a target location. These simulation tasks, visualized in Figure 6, span a range of manipulation skills, providing a comprehensive evaluation of robot performance. For each pre-training method (MAE [28, 63], R3M [46], MPI [68]), we evaluate the performance of pre-trained encoders with H2R in reinforcement learning and imitation learning. For the PixMC, most tasks involve motion control of robotic arms, we primarily use reinforcement learning (RL) methods to validate the effectiveness of H2R. However, for more complex simulation tasks, such as Robomimic benchmark, experimental results tend to be more sensitive to the reward mechanisms in reinforcement learning. Therefore, to avoid the impact of reward on task success rates, we adopt imitation learning methods (IL) combined with a pre-trained visual encoder for testing. This approach will evaluate the effectiveness of H2R in bridging the gap between human hand and robotic arm visual perception.

Pre-training Dataset. We select SSV2 (Something-Something V2) [21] and Ego4D [22] as the primary datasets for our experiments. SSV2 contains 220,847 video clips of human actions with everyday objects, designed to help models understand fine-grained hand gestures. Ego4D is a large-scale egocentric dataset with 3,670 hours of video collected from 923 participants worldwide, aimed at advancing first-person visual perception. We use the SSV2 dataset in the MAE and R3M methods, and the Ego4D dataset in the MPI method. For MAE and R3M, we select 62,500 videos from the SSV2 dataset and randomly sample 16 keyframes from each video. For MPI, we extract 117,624 action clips from 2,486 videos in the Ego4D dataset, each clip consisting of three frames (start, middle, and end).

Evaluation Metric. We repeat each experiment three times with different seeds and report the averaged results. For tasks in PixMC, we train the models using reinforcement learning for 2,000 steps and report the final success rate. For tasks in Robomimic, we train for 200 steps and reporting the mean success rate. For tasks in RLBench, rlbench For the PushT task, we train the Diffusion Policy model for 200 epoches and report the success rate in the simulation environment. The training hyperparameters used in this work are identical to those described in the original paper.

4.2. Results.

Reinforcement Learning. From Table 1, we observe that the improvement brought by H2R in reinforcement learning shows more variation depending on the task. Some tasks see an improvement, while others experience a decline. However, on average, across all tasks, there is still an overall improvement. Additionally, the performance with CutMix data is particularly better with R3M, while the use of H2R data yields excellent results with MAE. For example, when using the MAE pretraining method, the use of our H2R data results in a 29.7% improvement in the average success rate of the tasks. On the other hand, encoders trained with CutMix data show improvements ranging from 18.0% to 21.4%. When using the R3M pretraining method, the improvement in average success rate with H2R data is smaller, but the performance boost with CutMix data is more pronounced. Finally, when using the MPI pretraining method, the use of H2R data results in a modest reduction in the average success rate.

Imitation Learning. From Table 2, we observe that the encoder trained on H2R data shows consistent improvements across various tasks compared to the encoder trained on the original data, with the average success rate improvement on all tasks ranging from 0.9% to 10.2%. Especially for the more challenging MoveCan task, it can improve the success rate by 25.5%. Additionally, while encoders trained on the relatively simple CutMix data show improvement on tasks in Robomimic, their performance in the PushT task remains slightly worse than the encoders trained on original data. These results demonstrate the effectiveness of using

CVPR #00076

	FrankaReach	FrankaCabinet	FrankaPick	KukaCabinet	Average
MAE	97.5	88	0	90	46.9
MAE+CutMix1	93.5 (-4.0%)	90.5 (+2.5%)	0.0 (0.0%)	80.5 (-9.5%)	66.1 (+19.2%)
MAE+CutMix2	96.5 (-1.0%)	100.0 (+12.0%)	0.0 (0.0%)	63.0 (-27.0%)	64.9 (+18.0%)
MAE+CutMix3	98.5 (+1.0%)	90.5 (+2.5%)	0.0 (0.0%)	84.0 (-6.0%)	68.3 (+21.4%)
MAE+H2R	96.0 (-1.5%)	92.0 (+4.0%)	31.5 (+31.5%)	87.0 (-3.0%)	76.6 (+29.7%)
R3M	63	0	0	0	15.8
R3M+CutMix1	95.5 (+32.5%)	99.5 (+99.5%)	0.0 (0.0%)	1.0 (+1.0%)	49.0 (+33.2%)
R3M+CutMix2	98.5 (+35.5%)	97.5 (+97.5%)	0.0 (0.0%)	0.0 (0.0%)	49.0 (+33.2%)
R3M+CutMix3	97.5 (+34.5%)	85.5 (+85.5%)	0.0 (0.0%)	0.0 (0.0%)	45.8 (+30.0%)
R3M+H2R	8.5 (-54.5%)	0.0 (0.0%)	0.0 (0.0%)	81.0 (+81.0%)	17.9 (+2.1%)
MPI	83.5	0	0	58	35.4
MPI+H2R	88.0 (+4.5%)	20.0 (+20.0%)	0.0 (0.0%)	0.0 (-58.0%)	27 (-8.4%)

Table 1. **Reinforcement learning experiment result.** We report the success rate (%) over RL-based tasks for MAE, R3M, and MPI.

	MoveCan	Square	Lift	Average	PushT
MAE	54	25.5	94.5	58	59.2
MAE+CutMix1	72.0 (+18.0%)	30.0 (+4.5%)	95.0 (+0.5%)	65.7 (+7.7%)	37.5 (-21.7%)
MAE+CutMix2	58.0 (+4.0%)	36.0 (+10.5%)	90.0 (-4.5%)	61.3 (+3.3%)	40.0 (-19.2%)
MAE+CutMix3	78.0 (+24.0%)	32.0 (+9.3%)	92.0 (-2.5%)	67.3 (+2.7%)	42.0 (-17.2%)
MAE+H2R	79.5 (+25.5%)	29.5 (+4.0%)	95.5 (+1.0%)	68.2 (+10.2%)	64.5 (+5.3%)
R3M	59.5	20.5	85	55	15
R3M+CutMix1	69.5 (+10.0%)	30.0 (+9.5%)	91.0 (+6.0%)	63.5 (+8.5%)	19.0 (+4.0%)
R3M+CutMix2	66.0 (+6.5%)	26.0 (+5.5%)	83.0 (-2.0%)	58.3 (+3.3%)	17.0 (+2.0%)
R3M+CutMix3	68.0 (+8.5%)	26.0 (+5.5%)	84.0 (-1.0%)	59.3 (+4.3%)	14.0 (-1.0%)
R3M+H2R	61.5 (+2.0%)	37.5 (+17.0%)	85.0 (0.0%)	61.3 (+6.3%)	22.0 (+7.0%)
MPI	58	21	96	58.3	62.7
MPI+H2R	62.5 (+4.5%)	24.5 (+3.5%)	94.5 (-1.5%)	60.5 (+2.2%)	63.8 (+0.9%)

Table 2. **Imitation learning experiment result.** We report the success rate (%) over IL-based tasks for MAE, R3M, and MPI.

the robotic arm to cover the human hand in video data, as well as the effectiveness of H2R in imitation learning.

5. Real World Experiment

5.1. Experiment Setup

Real-world Tasks. To validate the effectiveness of H2R in downstream manipulation tasks, we implement three real-world manipulation tasks using a UR5 [61] robotic arm with a Robotiq [54] Gripper integration. The single RealSense L515 [31] camera is used to obtain visual observations in the real world. Realsense L515 is set above and behind the robotic arm, which provides a similar viewpoint to the human video data used in the pretrained visual model. Our real-world setup are shown in Figure 7. We provide detailed descriptions of the three implemented manipulation tasks as follows.

1. **Pick and Place:** Grasp a cube and place it into a bowl.
2. **Stack Cubes:** Stack a blue cube atop a yellow cube.
3. **Pick from Box:** Retrieve a cube from a box, place it into a bowl, and then close the box lid.

All the tasks are visualized in Appendix. Pick and Place task is the simplest of three tasks but still requires precise cube recognition, grasping, and placement within a designated bowl area. By contrast, Stack Cubes task demands higher positional accuracy during placement,

requiring precise identification of the yellow cube's location for successful stacking. Pick from Box task combines grasping with articulated object manipulation, necessitating longer-horizon planning and higher precision. For instance, the robot must avoid the lid while retrieving the cube and select optimal contact points to close the lid post-placement, challenging its ability to learn from high-dimensional visual inputs.

Policy Model	Tasks	MAE	MAE+H2R	R3M	R3M+H2R
Diffusion Policy	Pick and Place	45	65(+20%)	40	50(+10%)
	Stack Cubes	50	55(+5%)	55	70(+15%)
	Pick from Box	55	50(-5%)	45	65(+20%)
	Average	50	56.7(+6.7%)	46.7	61.7(+15%)
Equivariant Diffusion Policy	Pick and Place	55	70(+15%)	60	70(+10%)
	Stack Cubes	50	50(0%)	65	75(+10%)
	Pick from Box	55	75(+25%)	50	70(+20%)
	Average	53.3	65(+11.7%)	58.3	75(+13.3%)

Table 3. **Real-World success rate.**We report the success rate (%) over real-world tasks for MAE, R3M. Percentage changes due to H2R are shown in parentheses, with blue indicating improvement and red indicating degradation.

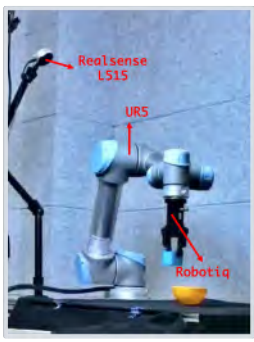

Figure 7. **Real-world experiment scene.**

Policy Training Process Details. For dataset collection, we collect expert demonstrations through human teleoperation, comprising 30 demonstrations per task. For downstream policy training, we select the Diffusion Policy(DP) [6] and Equivariant Diffusion Policy(eq-dp) [64] as policy frameworks. We apply upstream pre-trained MAE and R3M visual representation models to downstream policy learning, selecting four configurations for comparison: MAE, MAE+H2R, R3M, and R3M+H2R. Pretrained models are incorporated as frozen vision encoders in the policy network to evaluate their effectiveness. We use a single RGB camera image as the high-dimensional observation space and the robot proprioception as the low-dimensional observation space. Both are combined as input observations to the policy network. Policy is trained for 300 epochs using the collected data for each task.

5.2. Experiment Results

We evaluate the success rates of each model-task combination in real-world deployments. The results, as shown in Table 3, demonstrate that H2R significantly enhance the performance of visual encoders across diverse robotic tasks. H2R augmentation improves MAE-based policies in 6 out of 7 task configurations, with the largest gains in Pick from Box (+25% for Equivariant Diffusion) and Pick and Place (+20% for Diffusion Policy). Across all tasks, H2R augmentation consistently enhanced R3M-based policies, with the most notable improvements observed in geometrically complex scenarios such as Pick from Box, where R3M+H2R paired with Equivariant Diffusion Policy achieved a 20% success rate increase. These results highlight the potential of our approach to enhance visual encoders for real-world robotic applications, even in complex and dynamic environments.

6. Conclusion

We proposed H2R, a data augmentation technique that bridges the visual gap between human hand demonstrations and robotic arm manipulations by replacing human hands in first-person videos with robotic arm movements. Using 3D hand reconstruction and image inpainting models, H2R generates synthetic robotic arm manipilation sequences, making them more suitable for robot pre-training. Experiments across simulation benchmarks and real-world tasks demonstrate consistent improvements in success rates for encoders trained with various pre-training methods(MAE, R3M, MPI), highlighting its effectiveness and generalizability. H2R enables efficient transfer of task knowledge from human demonstrations to robotic systems, reducing reliance on costly robot-specific data collection.

References

[1] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023. 2

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. 2

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale, 2023. 2

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 3

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 3

[6] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023. 2, 3, 5, 6, 8

[7] Reinis Cimurs, Il Hong Suh, and Jin Han Lee. Goal-driven autonomous exploration through deep reinforcement learning. *IEEE Robotics and Automation Letters*, 7(2):730–737, 2022. 3

[8] Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, 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, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, 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, Jay Vakil, 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, Shubham Tulsiani, 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, Vikash Kumar, 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, Yongqiang Dou, 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, 2024. 2, 3

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 2

[10] Sudeep Dasari, Mohan Kumar Srirama, Unnat Jain, and Abhinav Gupta. An unbiased look at datasets for visuo-motor pre-training. In *7th Annual Conference on Robot Learning*, 2023. 4

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 5

[15] 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*, pages 2838–2848. PMLR, 2023. 2

[16] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot, 2023. 2

[17] Pete Florence, Corey Lynch, Andy Zeng, Oscar Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning, 2021. 3

[18] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[19] Jensen Gao, Annie Xie, Ted Xiao, Chelsea Finn, and Dorsa Sadigh. Efficient data collection for robotic manipulation via compositional generalization, 2024. 2

[20] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 2

[21] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense, 2017. 5, 6

[22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson

Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022. 5, 6

[23] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[24] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018. cite arxiv:1801.01290Comment: ICML 2018 Videos: sites.google.com/view/soft-actor-critic Code: github.com/haarnoja/sac. 3

[25] 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 (CVPR)*, 2016. 3

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 3

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 5, 6

[29] 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, 2023. 2

[30] Kevin Huang, Rwik Rana, Alexander Spitzer, Guanya Shi, and Byron Boots. DATT: Deep adaptive trajectory tracking for quadrotor control. In *7th Annual Conference on Robot Learning*, 2023. 3

[31] Intel Corporation. Intel®realsense™lidar camera l515. https://www.intelrealsense.com/lidar-camera-l515/, 2025. Accessed: 2025-02-01. 7

[32] 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, 2022. 2

[33] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale, 2021. 2

[34] Siddharth Karamcheti, Suraj Nair, Annie S. Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. In *Robotics: Science and Systems (RSS)*, 2023. 2, 3

[35] 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. 2

[36] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao,

Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 2, 4

[37] Alex X. Lee, Coline Devin, Yuxiang Zhou, Thomas Lampe, Konstantinos Bousmalis, Jost Tobias Springenberg, Arunkumar Byravan, Abbas Abdolmaleki, Nimrod Gileadi, David Khosid, Claudio Fantacci, Jose Enrique Chen, Akhil Raju, Rae Jeong, Michael Neunert, Antoine Laurens, Stefano Saliceti, Federico Casarini, Martin Riedmiller, Raia Hadsell, and Francesco Nori. Beyond pick-and-place: Tackling robotic stacking of diverse shapes, 2021. 2

[38] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47):eabc5986, 2020. 3

[39] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies, 2016. 3

[40] 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. 3

[41] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2024. 2

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 3

[43] Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning, 2024. 3

[44] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence? In *Advances in Neural Information Processing Systems*, pages 655–677. Curran Associates, Inc., 2023. 4

[45] 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, 2021. 2, 3, 5, 6

[46] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022. 2, 3, 5, 6

[47] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex

CVPR #00076

Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. 2

[48] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 3

[49] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. In *Proceedings of the 39th International Conference on Machine Learning*, pages 17359–17371. PMLR, 2022. 3

[50] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers, 2023. 2, 4

[51] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training, 2022. 2, 3

[52] Ilija Radosavovic, Baifeng Shi, Letian Fu, Ken Goldberg, Trevor Darrell, and Jitendra Malik. Robot learning with sensorimotor pre-training. In *7th Annual Conference on Robot Learning*, 2023. 4

[53] Ilija Radosavovic, Baifeng Shi, Letian Fu, Ken Goldberg, Trevor Darrell, and Jitendra Malik. Robot learning with sensorimotor pre-training. In *Conference on Robot Learning*, pages 683–693. PMLR, 2023. 3

[54] Robotiq Inc. Adaptive grippers. https://robotiq.com/products/adaptive-grippers, 2025. Accessed: 2025-02-01. 7

[55] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *ArXiv*, abs/2111.02114, 2021. 3

[56] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. 2, 3, 5

[57] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438, 2019. 3

[58] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions, 2021. 2, 4

[59] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world, 2017. 3

[60] Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018. 3

[61] Universal Robots A/S. Ur5 - collaborative robots. https://www.universal-robots.com/products/ur5-robot/, 2025. Accessed: 2025-02-01. 7

[62] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. *arXiv preprint arXiv:2304.00464*, 2023. 3

[63] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022. 2, 3, 5, 6

[64] Jingyun Yang, Zi ang Cao, Congyue Deng, Rika Antonova, Shuran Song, and Jeannette Bohg. Equibot: Sim(3)-equivariant diffusion policy for generalizable and data efficient learning, 2024. 2, 3, 8

[65] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19313–19322, 2022. 2

[66] Zhecheng Yuan, Tianming Wei, Shuiqi Cheng, Gu Zhang, Yuanpei Chen, and Huazhe Xu. Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning. *arXiv preprint arXiv:2407.15815*, 2024. 3

[67] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024. 3

[68] Jia Zeng, Qingwen Bu, Bangjun Wang, Wenke Xia, Li Chen, Hao Dong, Haoming Song, Dong Wang, Di Hu, Ping Luo,

Heming Cui, Bin Zhao, Xuelong Li, Yu Qiao, and Hongyang Li. Learning manipulation by predicting interaction, 2024. 2, 5, 6

[69] Jia Zeng, Qingwen Bu, Bangjun Wang, Wenke Xia, Li Chen, Hao Dong, Haoming Song, Dong Wang, Di Hu, Ping Luo, et al. Learning manipulation by predicting interaction. *arXiv preprint arXiv:2406.00439*, 2024. 2

[70] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, 2023. 3

[71] Ziwen Zhuang, Zipeng Fu, Jianren Wang, Christopher Atkeson, Sören Schwertfeger, Chelsea Finn, and Hang Zhao. Robot parkour learning. In *Conference on Robot Learning (CoRL)*, 2023. 3

CVPR #00076

Appendix

CLIP-based Evaluation of Augmentation Effectiveness. To quantitatively assess the effectiveness of H2R in bridging the visual gap between human and robot hands, we employ CLIP to measure the similarity between images and action descriptions before and after augmentation. Specifically, we generate two text prompts for an image:

- **Human-centric prompt**: "A human is *[action]*."
- **Robot-centric prompt**: "A robot is *[action]*."

where *[action]* describes the task being performed (e.g., "picking up a cube") in the image. We compute CLIP similarity scores between the image and their respective prompts, where higher robot-prompt scores indicate better visual alignment with robotic manipulation.

	Img1	Img2	Img3	Img4	Img5	Img6
ori	31.4	23.7	31.2	32.0	27.5	29.7
aug	28.4	28.8	32.7	29.6	28.7	27.2

Table 4. **CLIP similarity scores.** Higher values indicate better alignment between images and robot-centric descriptions.

Visual examples in Figure 11 demonstrate how our augmentation successfully adapts human motions to robotic kinematics. The CLIP similarity scores in Table 4 confirm that the augmented images maintain comparable alignment with robot-centric descriptions. CLIP scores occasionally decrease for some tasks, likely due to minor artifacts in hand-object interaction synthesis. However, as demonstrated in the following subsections, these discrepancies do not impede downstream policy performance, suggesting that H2R prioritizes functionally relevant visual features over pixel-perfect realism.

Failure Cases Analysis of the Real-world Experiment In addition to evaluating the success rate, we performed a detailed analysis of all failure cases by decomposing each task into distinct operational phases, as shown in Table 5. We divided three real-world tasks into multiple stages based on the complete motion sequence of a robotic arm. For each task, we classified the failure cases according to the furthest phase achieved, where later stages correspond to higher task completion levels.

In our real-world evaluation, we show the frequency distribution of task-specific failure cases in Figure 8.we find that regardless of the task or model used in the experiments, Case 1 constitutes a significant proportion of the failure cases. A major reason for this is the policy's inability to accurately locate the position to interact with the target object. This misalignment can be attributed to various factors such as camera noise, environmental lighting changes, object occlusions, or the model's limited adaptability to new environments. To delve deeper into these issues, we present detailed examples of typical failure cases in Figure 9. These factors can prevent the end-effector from correctly identifying and approaching the target location, leading to task failure. We also observe that H2R-augmented visual representation models not only improve overall success rates in real-world tasks but also significantly reduce the occurrence of Case 1 failures across most of the tasks, which indicates that even in failed attempts, the robot demonstrates higher partial-task completion.

tasks	Pick and Place	Stack Cubes	Pick from Box
case1	Picking failure	Picking failure	Picking failure
case2	Placing failure	Stacking failure	Placing failure
case3	Success	Success	Closing failure
case4	/	/	Success

Table 5. **Cases of each real-world task.** We list 3 cases for Pick and Place task, 3 cases for Stack Cubes task and 4 cases for Pick from Box task.

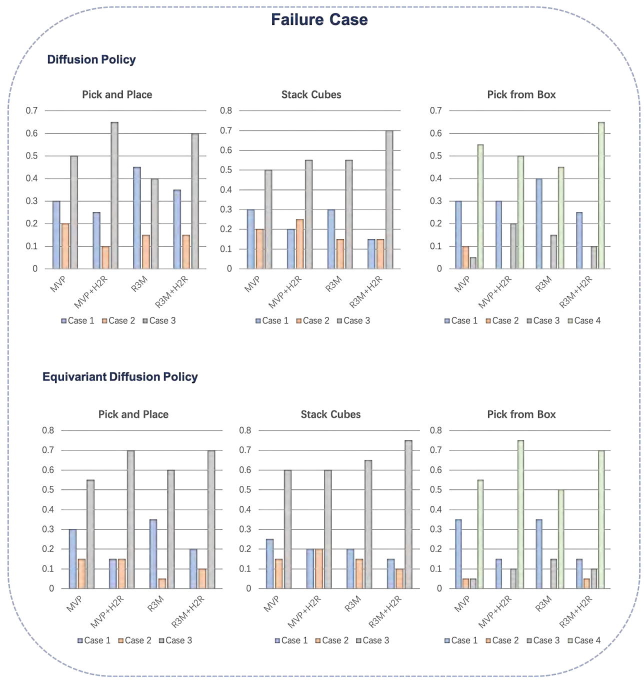

Figure 8. **Failure case analysis.** We divided each task into 3-4 cases to enable a detailed analysis of execution failure causes and finer-grained evaluation.

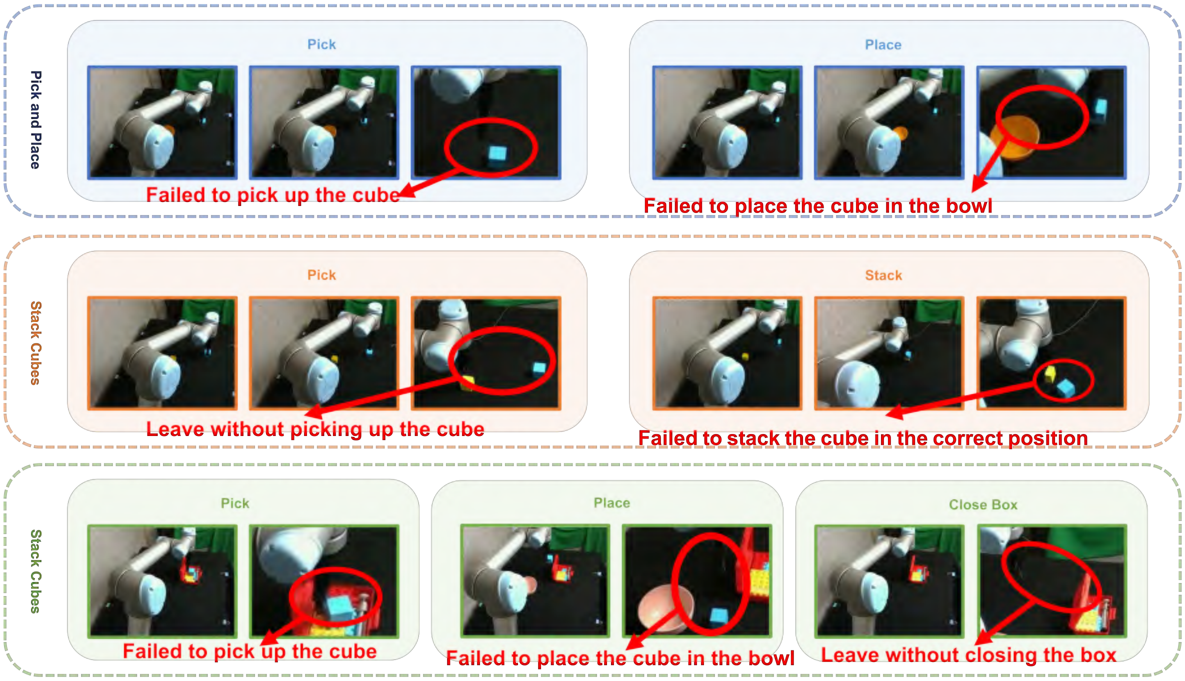

Figure 9. **Failure case visualization.** We provided first-person visualizations from the robot's viewpoint for every failure scenario.

Figure 10. **Real-world task.** Illustration of three real-world manipulation tasks ranging from simple to complex.

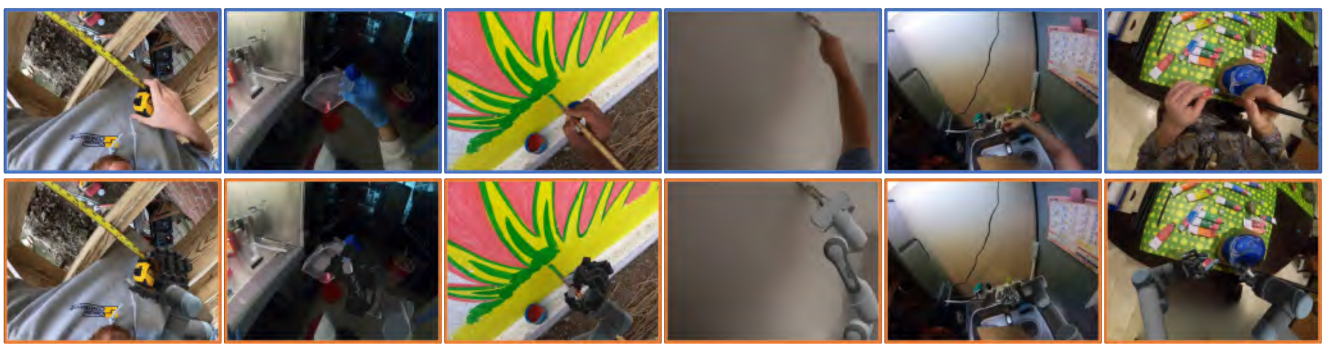

Figure 11. **H2R samples.** Visual comparison between original human data (top) and our augmented data (bottom).