DemoGen: Synthetic Demonstration Generation for Data-Efficient Visuomotor Policy Learning

Zhengrong Xue[123*] **Shuying Deng**[1*] **Zhenyang Chen**[2] **Yixuan Wang**[1] **Zhecheng Yuan**[123] **Huazhe Xu**[123]
[1]Tsinghua University, [2]Shanghai Qi Zhi Institute, [3]Shanghai AI Lab, [*]Equal contribution

demo-generation.github.io

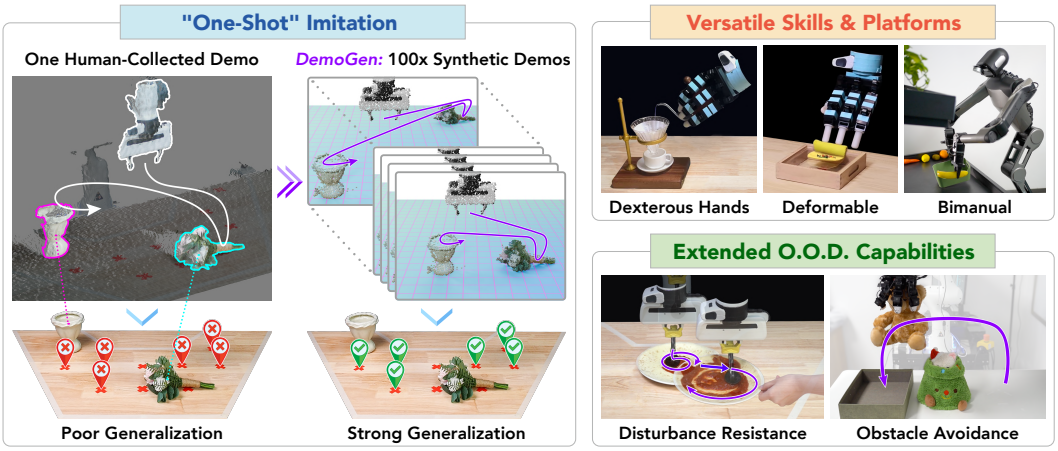

Figure 1: ***DemoGen*** is a fully synthetic approach for automatic demonstration generation. *DemoGen* promotes the spatial generalization ability of visuomotor policies and can facilitate one-shot imitation by adapting one human-collected demonstration into novel object configurations.

Abstract

Visuomotor policies have shown great promise in robotic manipulation but often require substantial human-collected data for effective performance. A key factor driving the high data demands is their limited spatial generalization capability, which necessitates extensive data collection across different object configurations. In this work, we present *DemoGen*, a low-cost, fully synthetic approach for automatic demonstration generation. Using only one human-collected demonstration per task, *DemoGen* generates spatially augmented demonstrations by adapting the demonstrated action trajectory to novel object configurations. Visual observations are synthesized by leveraging 3D point clouds as the modality and rearranging the subjects in the scene via 3D editing. Empirically, *DemoGen* significantly enhances policy performance across a diverse range of real-world manipulation tasks, showing its applicability even in challenging scenarios involving deformable objects, dexterous hand end-effectors, and bimanual platforms. Furthermore, *DemoGen* can be extended to enable additional out-of-distribution capabilities, including disturbance resistance and obstacle avoidance.

1 Introduction

Visuomotor policy learning has demonstrated remarkable competence for robotic manipulation tasks Chi et al. (2023); Zhao et al. (2023); Fu et al. (2024); Ze et al. (2024b), yet it typically demands large volumes of human-collected data. State-of-the-art approaches often require tens to hundreds of demonstrations to achieve moderate success on complex tasks, such as spreading sauce on pizza Chi et al. (2023) or making rollups with a dexterous hand Ze et al. (2024b). More intricate, long-horizon tasks may necessitate thousands of demonstrations Zhao et al. (2024).

One key factor contributing to the data-intensive nature of these methods is their limited **spatial generalization** Saxena et al. (2024); Tan et al. (2024) ability. Our empirical study in Sec. A suggests that visuomotor policies Chi et al. (2023), even when coupled with pre-trained or 3D visual

encoders Radford et al. (2021); Oquab et al. (2023); Ze et al. (2024b), exhibit limited spatial capacity, typically confined to regions adjacent to the demonstrated object configurations. Such limitation necessitates repeated data collection with repositioned objects until the demonstrated configurations sufficiently cover the full tabletop workspace. This creates a paradox: while the critical actions enabling dexterous manipulation are concentrated in a small subset of contact-rich segments, a substantial portion of human effort is spent teaching robots to approach objects in free space.

A potential solution to reduce redundant human effort is to replace the tedious relocate-and-recollect procedure with automatic demonstration generation. Recent advances such as MimicGen Mandlekar et al. (2023b) and its subsequent extensions Hoque et al. (2024); Garrett et al. (2024); Jiang et al. (2024) have proposed to generate demonstrations by segmenting the demonstrated trajectories based on object interactions. These object-centric segments are then transformed and interpolated into execution plans that fit desired spatially augmented object configurations. The resulting plans are then executed through open-loop rollouts on the robot, termed *on-robot rollouts*, to verify their correctness and simultaneously capture the visual observations needed for policy training. Despite their success in simulation, applying MimicGen-style strategies to real-world environments is hindered by the high costs of on-robot rollouts, which are nearly as expensive as collecting raw demonstrations. An alternative is to deploy via sim-to-real transfer Peng et al. (2018); Torne et al. (2024); Yuan et al. (2024), though bridging the sim-to-real gap remains a significant challenge.

In this work, we introduce *DemoGen*, a data generation system that can be seamlessly plugged into the policy learning workflow in both simulated and physical worlds. Recognizing the high cost of on-robot rollouts represents a major barrier to practical deployment, *DemoGen* adopts a **fully synthetic** pipeline that efficiently concretizes the generated plans into spatially augmented demonstrations.

For action generation, *DemoGen* develops the MimicGen strategy by incorporating techniques from Task and Motion Planning (TAMP) Dalal et al. (2023); Cheng et al. (2023); Mandlekar et al. (2023a), similar to the practice in the recently released SkillMimicGen Garrett et al. (2024). Specifically, we decompose the source trajectory into *motion segments* moving in free space and *skill segments* involving on-object manipulation through contact. During generation, the skill segments will be transformed as a whole according to the augmented object configuration, and the motion segments will be replanned via motion planning to connect the neighboring skill segments after transformation. With the processed actions in hand, a core challenge is obtaining spatially augmented visual observations without relying on costly on-robot rollouts. *DemoGen* employs a straightforward strategy: it selects point clouds as the observation modality and synthesizes the augmented visual observations through 3D editing. The key insight is that point clouds, which inherently live in the 3D space, can be easily manipulated to reflect the desired spatial augmentations. Generating augmented point cloud observations is reduced to identifying clusters of points corresponding to the interested subjects and then applying the same spatial transformations used in the generated action plans.

We manifest the effectiveness of *DemoGen* by evaluating the performance of visuomotor policies trained on *DemoGen*-generated datasets from **only one** human collected demonstration per task. To assess spatial generalization, we adhere to a rigorous evaluation protocol in which the objects are placed across the entire tabletop workspace within the end-effectors' reach. We conduct extensive real-world experiments, showing that *DemoGen* can be successfully deployed on both single-arm and bi-manual platforms, using parallel-gripper and dexterous-hand end-effectors, from both third-person and egocentric observation viewpoints, and with a range of rigid-body and deformable/fluid objects. Meanwhile, the cost of generating one demonstration trajectory with *DemoGen* is merely $\mathbf{0.01}$ seconds of computation. Empirically, *DemoGen* significantly enhances policy performance, generalizing to un-demonstrated configurations and achieving an average of $\mathbf{74.6}\%$ across $\mathbf{8}$ real-world tasks. Additionally, we demonstrate that simple extensions under the *DemoGen* framework can further equip imitation learning with acquired out-of-distribution generalization capabilities such as disturbance resistance and obstacle avoidance. The code and datasets will be open-sourced. ***Please refer to the website for robot videos.***

2 Related Works

Visuomotor policy learning. Represented by Diffusion Policy Chi et al. (2023) and its extensions Ze et al. (2024b); Ke et al. (2024); Prasad et al. (2024); Wang et al. (2024b;a), visuomotor policy learning refers to the imitation learning methods that learn to predict actions directly from vi-

sual observations in an end-to-end fashion Levine et al. (2016). The end-to-end learning objective is a two-edged sword. Its flexibility enables visuomotor policies to learn dexterous skills from human demonstrations, extending beyond rigid-body pick-and-place. However, the absence of structured skill primitives makes such policies intrinsically data-intensive. The conflicts between the huge data demands and the great expense of robotic data collection have driven recent data-centric research, including data collection systems Chi et al. (2024b); Cheng et al. (2024); Li et al. (2024), collaborative gathering of large-scale datasets O'Neill et al. (2024); Khazatsky et al. (2024), and empirical studies on data scaling Zhao et al. (2024); Lin et al. (2024). Instead of scaling up via pure human labor, *DemoGen* aims to show that synthetic data generation can help save much of the human effort.

Data-efficient imitation learning. Attempting to develop manipulation policies from only a handful of demonstrations, data-efficient imitation learning methods often build on the principles of Task and Motion Planning (TAMP), while incorporating imitation learning to replace some components in the TAMP pipeline. A common approach is to learn the end-effector poses for picking and placing Zeng et al. (2021); Simeonov et al. (2022); Wen et al. (2022); Xue et al. (2023); Gao et al. (2024). The whole trajectories are generated using motion planning toolkits Kuffner & LaValle (2000) and then executed in an open-loop manner. Some methods extend this idea to more complex scenarios by learning to estimate the states of manipulated objects in the environment and replaying demonstrated trajectory segments centered around the target objects Johns (2021); Valassakis et al. (2022); Di Palo & Johns (2022; 2024). While these approaches are effective for simpler, Markovian-style tasks Vosylius & Johns (2024), their reliance on open-loop execution limits their application to more dexterous tasks requiring closed-loop retrying and re-planning. In contrast, *DemoGen* leverages the TAMP principles for synthetic data generation to train closed-loop visuomotor policies, thus effectively combining the merits of both approaches.

Data generation for robotic manipulation. Automated demonstration generation offers the opportunity to breed capable visuomotor policies with significantly reduced human efforts. A branch of recent works attempts to generate demonstrations by leveraging LLM for task decomposition and then using planning or reinforcement learning for subtask resolution Wang et al. (2023a); Hua et al. (2024); Wang et al. (2023b). An alternative line of research is exemplified by MimicGen Mandlekar et al. (2023b) and its extensions Hoque et al. (2024); Garrett et al. (2024); Jiang et al. (2024). Unlike generating demonstrations from the void, MimicGen adapts some human-collected source demonstrations to novel object configurations by synthesizing corresponding execution plans. However, execution plans produced by the MimicGen framework are not ready-to-use demonstrations in the form of observation-action pairs. To bridge this gap, the MimicGen family Mandlekar et al. (2023b); Hoque et al. (2024); Garrett et al. (2024); Jiang et al. (2024) relies on costly on-robot rollouts, which poses significant challenges for the deployment on physical robots. Building upon MimicGen and its extensions, *DemoGen* incorporates their strategies for generating execution plans, but replaces the expensive on-robot rollouts with an efficient, fully synthetic generation process. This enables *DemoGen* to generate real-world demonstrations cost-effectively.

3 *DemoGen* METHODS

3.1 PROBLEM FORMULATION

A visuomotor policy $\pi : \mathcal{O} \mapsto \mathcal{A}$ directly maps the visual observations $o \in \mathcal{O}$ to the predicted actions $a \in \mathcal{A}$. To train such a policy, a dataset $\mathcal{D}$ of demonstrations must be prepared. We define a source demonstration $D_{s_0} \subseteq \mathcal{D}$ as a trajectory of paired observations and actions conditioned on an initial object configuration: $D_{s_0} = (d_0, d_1, \ldots, d_{L-1}|s_0)$, where each $d_t = (o_t, a_t)$ represents an observation-action pair, s_0 denotes the initial configuration, and L is the trajectory length. *DemoGen* is designed to augment a human-collected source demonstration by generating a new demonstration conditioned on a different initial object configuration: $\hat{D}_{s'_0} = (\hat{d}_0, \hat{d}_1, \ldots, \hat{d}_{L-1}|s'_0)$.

Specifically, assuming the task involves the sequential manipulation of K objects $\{O_1, O_2, \ldots, O_K\}$, the initial object configuration s_0 is defined as the set of initial poses of these objects: $s_0 = \{\mathbf{T}_0^{O_1}, \mathbf{T}_0^{O_2}, \ldots, \mathbf{T}_0^{O_K}\}$, where $\mathbf{T}_t^O$ denotes the $\mathrm{SE}(3)$ transformation from the world frame to an object O at time step t. The action a_t consists of the robot arm and robot hand commands, represented as $a_t = (a_t^{\mathrm{arm}}, a_t^{\mathrm{hand}})$, where $a_t^{\mathrm{arm}} \triangleq \mathbf{A}_t^{\mathrm{EE}}$ is the target $\mathrm{SE}(3)$ end-effector pose in the world frame, and a_t^{hand} can either be a binary signal for a parallel gripper's open/close

action or a higher-dimensional vector for controlling the joints of a dexterous hand. The observation o_t includes both the point cloud data and the proprioceptive feedback from the robot: $o_t = (o_t^{\text{pcd}}, o_t^{\text{arm}}, o_t^{\text{hand}})$, where o_t^{arm} and o_t^{hand} reflect the current state of the end-effector, with the same dimensionality as the corresponding actions.

3.2 Pre-processing Source Demonstration

Segmented point cloud observations. To improve the practical applicability in real-world scenarios, we utilize a single-view RGBD camera for point cloud acquisition. The raw point cloud observations are first preprocessed by cropping the redundant points from the background and table surface. We assume the retained points are associated with either the manipulated object(s) or the robot's end-effector. A clustering operation Ester et al. (1996) is then applied to filter out the outlier points in noisy real-world observations. Subsequently, the point cloud is downsampled to a fixed number of points (e.g., 512 or 1024) using farthest point sampling to facilitate policy learning Qi et al. (2017). For the first frame of the trajectory, we employ Grounded SAM Ren et al. (2024) to obtain the segmentation masks for the manipulated objects from the RGB image. These masks are then applied to the pixel-aligned depth image and projected onto the 3D point cloud, as shown in Fig. 2.

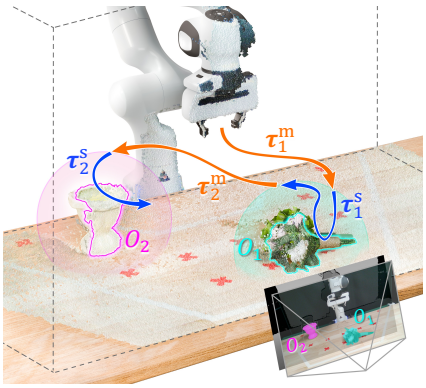

Figure 2: **Pre-processing the source demonstration.** The raw point cloud observations are processed by cropping, clustering, and down-sampling. The source action trajectory is parsed into motion and skill segments by referring to the object semantic masks.

Parsing the source trajectory. Following previous work Mandlekar et al. (2023b); Garrett et al. (2024), we assume that the execution trajectory can be parsed into a sequence of object-centric segments. Since the robot must initially approach the object in free space before engaging in on-object manipulation through contact, each object-centric segment can be further subdivided into two stages: motion and skill. For example, the trajectory in Fig. 2 is divided into four stages: 1) *move* to the flower, 2) *pick* up the flower, 3) *transfer* the flower to the vase, 4) *insert* the flower into the vase.

We can easily identify the skill segments associated with a given object by checking whether the distance between the geometric center of the object's point cloud and the robot's end-effector falls within a predefined threshold, illustrated by the spheres in Fig. 2. The intermediate trajectories between two skill segments are classified as motion segments. Formally, we represent an interval of time stamps as: $\boldsymbol{\tau} = (t_{\text{start}},\ t_{\text{start}}+1,\ \ldots,\ t_{\text{end}}-1,\ t_{\text{end}}) \subseteq (0, 1, \ldots, L-1)$, which can be used as an *index sequence* for the extraction of the corresponding segments from a sequence of demonstrations, actions, or observations. For instance, $d[\boldsymbol{\tau}] = (d_{t_{\text{start}}}, d_{t_{\text{start}}+1}, \ldots, d_{t_{\text{end}}-1}, d_{t_{\text{end}}})$ represents the extracted subset of source demonstration indexed by $\boldsymbol{\tau}$. Using this notation, we parse the source demonstration into alternating motion and skill segments according to the index sequence $(\boldsymbol{\tau}_1^{\text{m}}, \boldsymbol{\tau}_1^{\text{s}}, \ldots, \boldsymbol{\tau}_K^{\text{m}}, \boldsymbol{\tau}_K^{\text{s}})$: $D_{s_0} = (d[\boldsymbol{\tau}_1^{\text{m}}], d[\boldsymbol{\tau}_1^{\text{s}}], \ldots, d[\boldsymbol{\tau}_K^{\text{m}}], d[\boldsymbol{\tau}_K^{\text{s}}] | s_0)$.

3.3 TAMP-based Action Generation

The generation process begins by selecting a target initial configuration $s_0' = \{\mathbf{T}_0^{O_1\prime}, \mathbf{T}_0^{O_2\prime}, ..., \mathbf{T}_0^{O_K\prime}\}$. We can compute the spatial transformation under the homogeneous matrix representation by: $\Delta s_0 = \{(\mathbf{T}_0^{O_1})^{-1} \cdot \mathbf{T}_0^{O_1\prime}, \ldots, (\mathbf{T}_0^{O_K})^{-1} \cdot \mathbf{T}_0^{O_K\prime}\}$. Recall that the actions consist of both robot arm and robot hand commands. The robot hand commands define the interactive actions *on* the object, e.g., holding the flower with the gripper, or rolling up the dough with the dexterous hand. Since they are *invariant* of the spatial transformation, a_t^{hand} should remain unchanged regardless of the object configuration: $\hat{a}_t^{\text{hand}} = a_t^{\text{hand}},\ \forall\, t, s_o, s_0'$.

In contrast, the robot arm commands should be spatially *equivariant* to the object movements in order to adjust the trajectory according to the altered configuration. Specifically, for the motion and skill segments involving the k-th object, we adapt the robot arm commands $\mathbf{A}^{\text{EE}}[\boldsymbol{\tau}_k^{\text{m}}]$, $\mathbf{A}^{\text{EE}}[\boldsymbol{\tau}_k^{\text{s}}]$ following a TAMP-based procedure illustrated in Fig. 3.

For the skill segments with dexterous on-object behaviors, the spatial relations between end-effectors and objects must remain relatively static. Thus, the entire skill segments are transformed following the corresponding objects: $\hat{\mathbf{A}}^{\mathrm{EE}}[\boldsymbol{\tau}_k^{\mathrm{s}}] = \mathbf{A}^{\mathrm{EE}}[\boldsymbol{\tau}_k^{\mathrm{s}}] \cdot (\mathbf{T}_0^{O_k})^{-1} \cdot \mathbf{T}_0^{O_k{}'}$. For the motion segments moving in free space, the goal of the generated actions is to chain the adjacent skill segments. Therefore, we plan the robot arm commands in the motion stage via motion planning: $\hat{\mathbf{A}}^{\mathrm{EE}}[\boldsymbol{\tau}_k^{\mathrm{m}}] = \texttt{MotionPlan}(\hat{\mathbf{A}}^{\mathrm{EE}}[\boldsymbol{\tau}_{k-1}^{\mathrm{s}}][-1],\ \hat{\mathbf{A}}^{\mathrm{EE}}[\boldsymbol{\tau}_k^{\mathrm{s}}][0])$, where the starting pose for motion planning is taken from the last frame of the previous skill segment, and the ending pose is from the first frame of the current skill segment. For simple uncluttered workspaces, linear interpolation suffices. For complex environments requiring obstacle avoidance, an off-the-shelf motion planning method Kuffner & LaValle (2000) is employed.

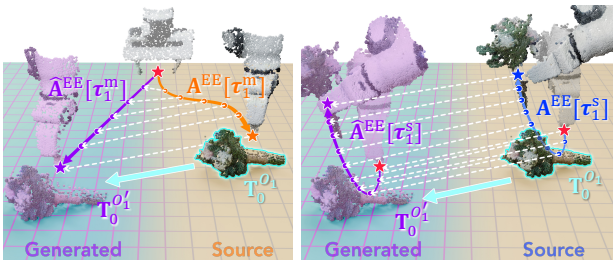

Figure 3: **Illustrations for action generation.** (Left) Actions in the motion stage are planned to connect the neighboring skill segments. (Right) Actions in the skill stage are transformed uniformly as a whole.

3.4 Fully Synthetic Observation Generation

Adapting proprioceptive states. The observations consist of point cloud data and proprioceptive states. Since the proprioceptive states share the same semantics with the actions, they should undergo the same transformation: $\hat{o}_t^{\mathrm{hand}} = o_t^{\mathrm{hand}},\ \forall\, t, s_o, s_0'$; $\hat{o}_t^{\mathrm{arm}} = o_t^{\mathrm{arm}} \cdot (\mathbf{A}_t^{\mathrm{EE}})^{-1} \cdot \hat{\mathbf{A}}_t^{\mathrm{EE}}$.

Synthesizing point cloud observations. To synthesize the spatially augmented point clouds for the robot and objects, we employ a simple segment-and-transform strategy. Apart from the target transformations, the only required information for synthesis is the segmentation masks for the K objects on the first frame of the source demonstration, obtained in Sec. 3.2.

For each object, we define 3 stages. In the *to-do* stage, the object is static and unaffected by the robot, and its point cloud is transformed according to the initial object configuration $(\mathbf{T}_0^{O_k})^{-1} \cdot \mathbf{T}_0^{O_k{}'}$. In the *doing* stage, the object is in contact with the robot, and its point cloud is merged with the end-effector's point cloud. In the *done* stage, the object remains in its final state. These stages are easily identified by referencing the trajectory-level motion and skill segments. For the robot's end-effector, its point cloud undergoes the same transformation as indicated by the proprioceptive states $(\mathbf{A}_t^{\mathrm{EE}})^{-1} \cdot \hat{\mathbf{A}}_t^{\mathrm{EE}}$. Given the assumption of a cropped workspace, the point clouds for the robot and the objects in the *doing* stage can be separated by subtracting the object point clouds in the *to-do* and *done* stages from the scene point cloud. A concrete example of this process is shown in Fig. 4. More examples of the synthetic trajectories in real-world experiments can be found in Fig. 21.

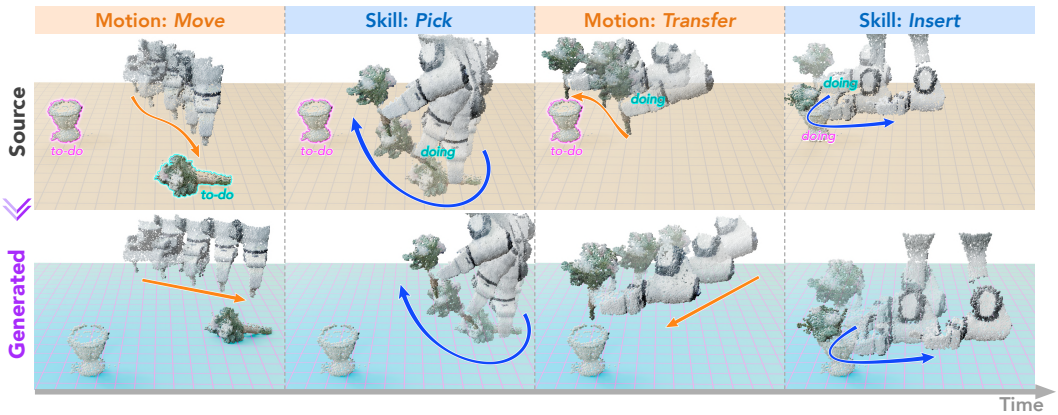

Figure 4: **Illustrations for synthetic visual observation generation.** Objects in the *to-do* stage are segmented and transformed by the target object configurations. Objects in the *doing* stage are merged with the end-effector and transformed according to the proprioceptive states.

4 EXPERIMENTS IN THE SIMULATOR

Policy. Both in the simulator and real world, we select DP3 Ze et al. (2024b) as the visuomotor policy, which predicts actions by consuming point cloud and proprioception observations. For a fair comparison, we fix the total training steps counted by observation-action pairs, resulting in an equal training cost regardless of the dataset size. The training details are listed in Appendix B.1.

Tasks. We design 8 tasks adapted from the MetaWorld Yu et al. (2020) benchmark, illustrated in Fig. 5. To strengthen the significance of spatial generalization, we modify these tasks to have enlarged object randomization ranges, as listed in Appendix D.1.

Generation and evaluation. We write scripted policies for these tasks and prepare only 1 source demonstration per task for demonstration generation. We also produce 10 and 25 source demonstrations per task using the scripted policy as a reference for human-collected datasets. Based on the one source demonstration, we leverage *DemoGen* to generate 100 spatially augmented demonstrations for the tasks containing the spatial randomization of one object. Since the tasks concerning two objects have a more diverse range of object configurations, 200 demonstrations are generated.

Results analysis. The evaluation results for the simulated tasks are presented in Tab. 1. *DemoGen* significantly enhances the policy performance compared with the source demonstration baseline. The policies trained on *DemoGen*-generated datasets also outperform those trained on 10 source demonstrations and get close to 25 source demonstrations. This indicates *DemoGen* has the potential to maintain the policy performance with over 20× reduced human effort for data collection. Additionally, we found a visual mismatch problem between the synthetic and real-captured observations, which poses a limitation for the effectiveness of *DemoGen*. Illustrations and the empirical consequence of the visual mismatch problem are provided in Appendix C.

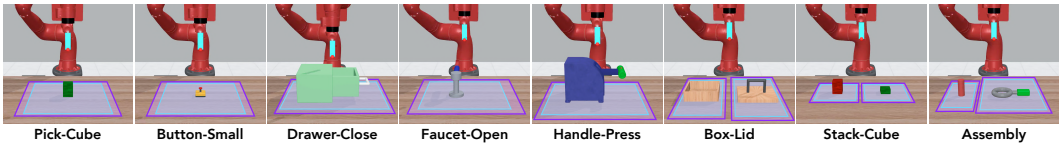

Figure 5: **Tasks for simulated evaluation on spatial generalization.** Purple and sky-blue rectangles mark the workspaces for demonstration generation and evaluation, respectively.

Table 1: **Simulated evaluation of *DemoGen* for spatial generalization.** We report the maximum/averaged success rates over 3 seeds.

	Pick-Cube	Button-Small	Drawer-Close	Faucet-Open	Handle-Press	Box-Lid	Stack-Cube	Assembly	Averaged
1 Source	0/0	4/4	55/50	39/23	17/16	11/11	0/0	0/0	16/13
DemoGen	**76/73**	**92/84**	**100/100**	**95/92**	**100/100**	**100/95**	**79/77**	**86/83**	**91/88**
10 Source	29/29	54/52	100/100	90/89	100/99	94/89	44/38	47/45	70/68
25 Source	82/74	90/84	100/100	100/100	100/100	100/100	95/93	83/79	94/91

5 EXPERIMENTS IN THE REAL WORLD

5.1 SPATIAL GENERALIZATION ON SINGLE-ARM PLATFORMS

Tasks. On the Franka Panda single-arm platform, we design 3 tasks using the original Panda gripper and 4 tasks using an Allegro dexterous hand as the end-effector. A task summary is provided in Tab. 2. The motion and skill trajectories of these tasks are visualized in Fig. 7 and the task descriptions are provided in Appendix D.2. For all tasks, a single Intel Realsense L515 camera is adopted to capture point cloud observations, as depicted in Fig. 6(a).

Evaluation protocol. To evaluate spatial generalization, we define a large planar evaluation workspace, the size of which corresponds to the maximum reach of the robot arm. Illustrated in Fig. 6(b), We uniformly sample 12 points within this irregularly-shaped workspace as the coordinates for potential object configurations, with a 15cm spacing between the neighbors. To determine the actual evaluated configurations for each task, we perform manual trials using kinematic teaching to confirm the feasibility of each configuration.

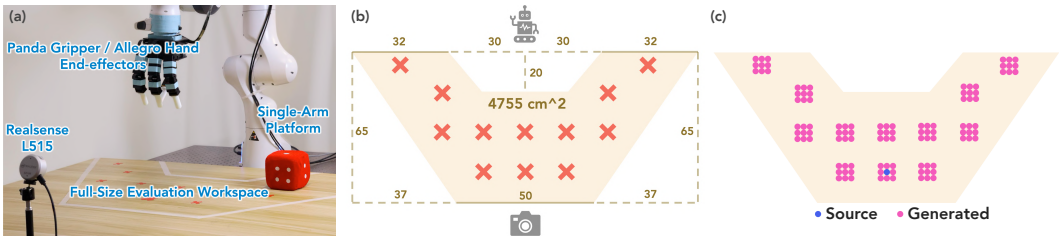

Figure 6: **Protocol for evaluating spatial generalization.** (a) Setups on the single-arm platform. (b) Illustration for the full-size evaluation workspace. (c) Illustration for the generation strategy targeting the evaluated configurations along with small-range perturbations.

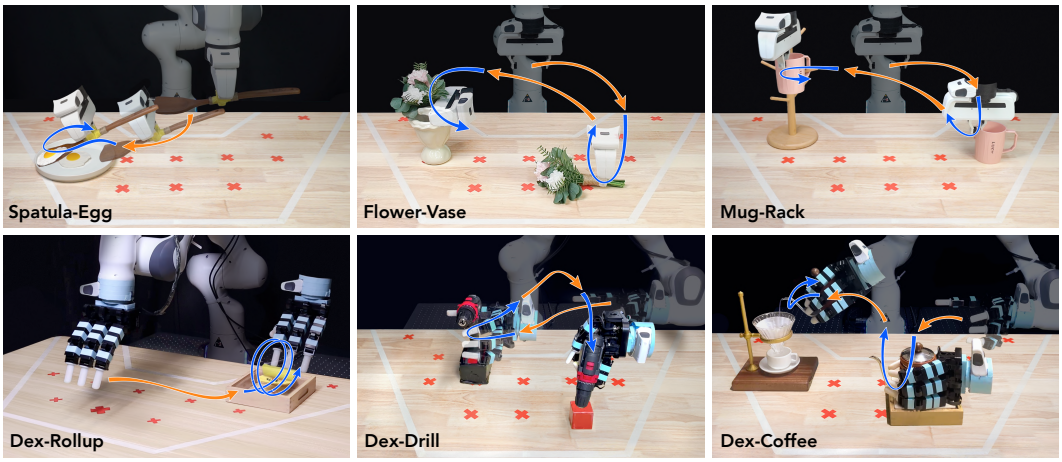

Figure 7: **Tasks for real-world evaluation on spatial generalization.** Spatula-Egg and Dex-Rollup are one-stage tasks involving contact-rich behaviors. Flower-Vase, Mug-Rack, Dex-Drill, and Dex-Coffee are two-stage tasks requiring precise manipulation.

Generation strategy. As in the simulated environments, we collect only one source demonstration for each task. However, real-world point cloud observations are often noisy, with issues such as flickering holes in the point clouds or projective smearing around object outlines. The imitation learning policy can overfit these irregularities if only one demonstration is provided. To mitigate this issue, we replay the source demonstration twice and capture the corresponding point cloud observations. The altogether 3 point cloud trajectories enrich the diversity in visual degradations and help alleviate the overfitting problem.

Table 2: **Real-world tasks for spatial generalization evaluation.** ActD: action dimension. #Obj: number of manipulated objects. #Eval: number of evaluated configurations. #GDemo: number of generated demonstrations.

Task	Platform	ActD	#Obj	#Eval	#GDemo
Spatula-Egg	Gripper	6	1	10	270
Flower-Vase	Gripper	7	2	4×4	432
Mug-Rack	Gripper	7	2	4×4	432
Dex-Cube	Dex. Hand	22	1	10	270
Dex-Rollup	Dex. Hand	22	1	12	324
Dex-Drill	Dex. Hand	22	2	3×3	243
Dex-Coffee	Dex. Hand	22	2	3×3	243
Fruit-Basket	Bimanual	14	2	4×6	72

We set the generated object configurations to correspond to the evaluated configurations. However, human operators cannot always place objects with perfect precision in the real world, yet we found visuomotor policies are sensitive to even small deviations. Thus, we further augment the generated object configurations by adding small-range perturbations. Specifically, for each target configuration, we generate 9 demonstrations with $(\pm 1.5\text{cm}) \times (\pm 1.5\text{cm})$ perturbation to mimic slight placement variations in the real world. The final generated configurations are shown in Fig. 6(c).

In summary, the total number of generated demonstrations is calculated as $3 \times (\#\text{Eval}) \times 9$, which represents the 3 source demonstrations, multiplied by the number of evaluated configurations, and further multiplied by the 9 perturbations. The detailed counts for each task are listed in Tab. 2.

Results analysis. The performance of visuomotor policies Ze et al. (2024b) trained on 3 source demonstrations and *DemoGen*-generated demonstrations are reported in Tab. 3. Agents trained solely on source demonstrations exhibit severe overfitting behaviors, blindly replicating the demonstrated trajectory. In Appendix D.3, we evaluate the policy performance trained on datasets contain-

Table 3: **Real-world evaluation of *DemoGen* for spatial generalization.** The success rates are averaged on 5 repetitions for each evaluated configuration. The evaluated configurations for each task are counted in Tab. 2, and visualized in Fig. 8.

	Spatula-Egg	Flower-Vase	Mug-Rack	Dex-Cube	Dex-Rollup	Dex-Drill	Dex-Coffee	Fruit-Basket	Averaged
Source	10.0	6.3	6.3	10.0	8.3	11.1	11.1	25.0	11.0
DemoGen	**88.0**	**82.5**	**85.0**	**78.0**	**76.7**	**55.6**	**40.0**	**90.8**	**74.6**

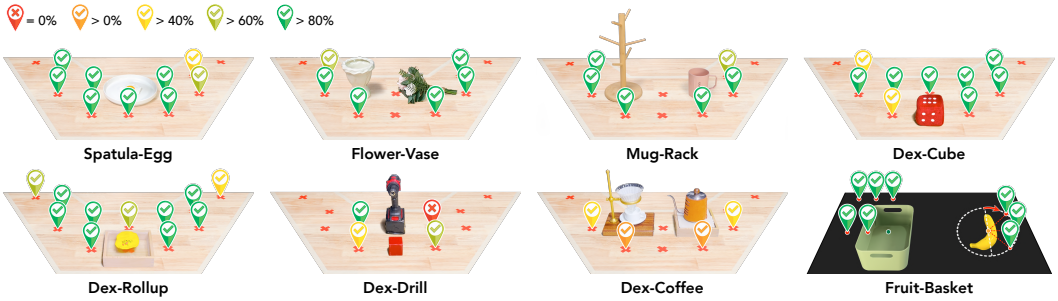

Figure 8: **Spatial heatmaps for real-world evaluation.** The success rate for each coordinate is calculated as the average across all relevant trials. For example, each coordinate of the vase in the Flower-Vase task is in combination with 4 coordinates of the flower, including the one appearing in the source demonstration. This results in a total of 20 trials, given 5 repetitions per combination.

ing additional human-collected demonstrations. We found the spatial effective range of the trained policies is upper-bounded by the demonstrations, aligned with the empirical study in Appendix A.

Similar to the effects of manually covering the workspace with human-collected demonstrations, *DemoGen*-generated datasets enable the agents to display a more adaptive response to diverse evaluated configurations, resulting in significantly higher success rates. *DemoGen* consistently enhances the performance across all the evaluated tasks. To further investigate the generalization capabilities enabled by *DemoGen*, we visualize the spatial heatmaps for the evaluated configurations in Fig. 8. The heatmaps reveal high success rates on configurations close to the demonstrated ones, while the performance diminishes as the distance from the demonstrated configuration increases. We attribute this decline to the visual mismatch problem, as discussed in Appendix C.

Generation cost. We compare the time cost of real-world demonstration generation between MimicGen Mandlekar et al. (2023b) and *DemoGen*. We estimate MimicGen's time cost by multiplying the duration of replaying a source trajectory by the number of generated demonstrations and adding an additional 20 seconds per trajectory for human operators to reset the object configurations. Note that MimicGen involves continuous human intervention, while the time cost of *DemoGen* is purely computational, without any human/robot involvement.

Table 4: **The time cost for generating real-world demonstrations.** The computational cost is measured on a single-process procedure.

	Single o-a Pair	A Trajectory	Whole Dataset
MimicGen	2.1 s	2.1 min	83.7 h
DemoGen	**0.00015** s	**0.010** s	**22.0** s

5.2 Spatial Generalization on Bimanual Humanoid

Task. In addition to the tasks on the single-arm platform, we also designed a Fruit-Basket task on a bimanual humanoid platform, illustrated in Fig. 9. The Fruit-Basket task is distinguished from the previous tasks by three key features: *1) Bimanual manipulation.* The robot simultaneously grasps the basket with one arm and the banana with the other. The right arm then places the basket in the center of the workspace, while the left arm places the banana into the basket. *2) Egocentric observation.* The camera is mounted on the robot's head Ze et al. (2024a). While the robot's base is immobilized in this task, the first-person view opens opportunities for future deployment in mobile manipulation scenarios. *3) Out-of-distribution orientations.* Still using a single human-collected demonstration, the banana is placed with orientational offsets (i.e., 45°, 90°, and 135°) relative to the original demonstration during evaluation, while the basket position is randomized within a translational 10 cm $\times$ 5 cm workspace.

Generation strategy. The generation procedure follows a similar approach as on the single-arm platform. Specifically, the human-collected demonstration is replayed twice, yielding 3 source demonstrations. *DemoGen* generates synthetic demonstrations by independently adapting the actions of both arms to the respective transformations of the objects. Small-range perturbations are omitted due to lower precision requirements. A challenge in synthesizing point cloud observations with orientational offsets lies in the limited view provided by the single camera. To address this issue, the humanoid robot adopts a stooping posture, enabling a near bird's-eye view perspective. This adjustment allows for more effective point cloud editing to simulate full-directional yaw rotations.

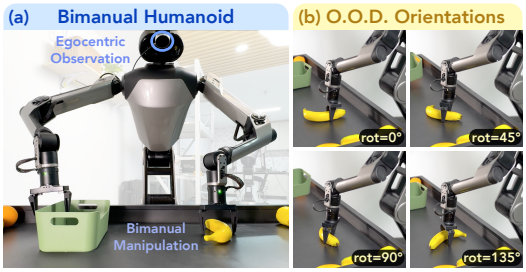

Figure 9: **Bimanual humanoid platform.** (a) Egocentric observations and bimanual manipulation. (b) The Fruit-Basket task involves the out-of-distribution orientations during evaluation.

Results analysis. The success rates for both the source and generated datasets are compared in Tab. 3, and the spatial heatmap is shown in Fig. 8. The high success rate of 90.8% demonstrates the effectiveness of *DemoGen* on bimanual humanoid platforms and its ability to help policies generalize to out-of-distribution orientations. A more detailed analysis is presented in Appendix D.4.

5.3 Disturbance Resistance

Task and evaluation protocol. We consider a Sauce-Spreading task (Fig. 10(a)) adapted from DP Chi et al. (2023). Initially, the pizza crust contains a small amount of sauce at its center (Fig. 10(b)). The gripper maneuvers the spoon in hand to approach the sauce center and periodically spread it to cover the pizza crust in a spiral pattern (Fig. 10(c)). During the sauce-spreading process, disturbances are introduced by shifting the pizza crust twice to the neighboring spots within the workspace. We consider 5 neighboring spots (Fig. 10(d)) and conduct 5 trials per spot, resulting in 25 trials. For quantitative evaluation, we measure the sauce coverage on the pizza crust. Additionally, we report a normalized sauce coverage score, where 0 represents no operation taken, and 100 corresponds to human expert performance. Detailed calculations are provided in Appendix D.5.

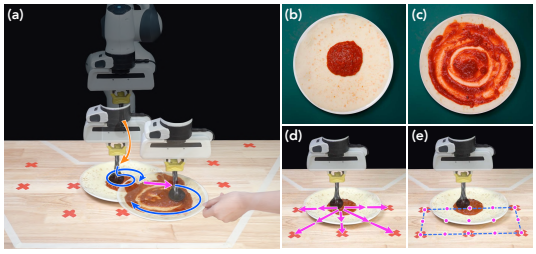

Figure 10: ***DemoGen* for disturbance resistance.** (a-c) Illustration, initial, and ending states of the Sauce-Spreading task. (d) Disturbance applied for evaluation. (e) Standard generation strategy.

Generation strategies. A standard generation strategy selects 15 intermediate spots (Fig. 10(e)) observed during the disturbance process as the initial object configurations for a standard *DemoGen* data generation procedure. To specifically enhance disturbance resistance, we propose a specialized strategy named Augmentation for Disturbance Resistance (ADR), illustrated in Fig. 11. In ADR, the pizza crust is artificially displaced to nearby positions at certain time steps to simulate the disturbance. The robot's end-effector, holding the spoon, initially remains static and subsequently interpolates its motion to re-approach the displaced crust before continuing the periodic spreading motion.

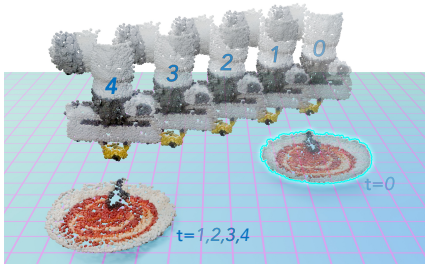

Figure 11: **Illustration for ADR.**

Results analysis. Tab. 5 presents the sauce coverage and normalized scores for both the standard *DemoGen* and the ADR-enhanced *DemoGen* strategies. Raw evaluation results and detailed definitions for the metrics are presented in Appendix D.5. We found the ADR strategy significantly outperforms the standard *DemoGen*, achieving performance comparable to human experts. In the video, we showcase the ADR-enhanced policy is still robust under up to 5 successive disturbances. These findings underscore the critical role

Table 5: **Evaluation results for disturbance resistance.**

	Coverage	Score
Standard	34.2	40.4
w/ ADR	**61.2**	**92.3**
Initial State	13.2	0
Human Expert	65.2	100

of the demonstration data in enabling policy capabilities. The ability to resist disturbances does not emerge naturally but is acquired through targeted disturbance-involved demonstrations.

5.4 Obstacle Avoidance

Task. Similar to the case of disturbance resistance, the visuomotor policy's ability to avoid obstacles is also imparted through demonstrations containing obstacle-avoidance behaviors. To investigate such capability, we introduce obstacles to a Teddy-Box task, where the dexterous hand grasps the teddy bear and transfers it into the box on the left (Fig. 12(a)). Trained on the source demonstrations without obstacles, the visuomotor policy fails to account for potential collisions, e.g., it might knock over the coffee cup placed in the middle (Fig. 12(b)).

Generation strategy. To generate obstacle-involved demonstrations, we augment the real-world point cloud observations by sampling points from simple geometries, such as boxes and cones, and fusing these points into the original scene (Fig. 12(c)). Obstacle-avoiding trajectories are generated by a motion planning tool Kuffner & LaValle (2000), which enables collision-free actions.

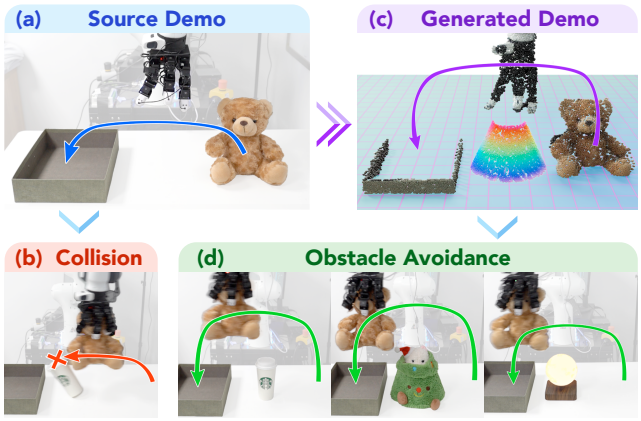

Figure 12: ***DemoGen* for obstacle avoidance.** (ab) Policy trained on the source demonstration collides with the unseen obstacle. (cd) Policy trained on the generated dataset could avoid diverse-shaped obstacles.

Evaluation and results analysis. For evaluation, we position 5 everyday objects with diverse shapes in the middle of the workspace (Fig. 12(d)) and conduct 5 trials per object, resulting in a total of 25 trials. The agent trained on the augmented dataset successfully bypasses obstacles in 22 out of 25 trials. Notably, in scenarios without obstacles, the agent follows the lower trajectory observed in the source demonstrations, indicating its responsiveness to environmental variations.

6 Conclusion

In this work, we introduced *DemoGen*, a fully synthetic data generation system designed to facilitate visuomotor policy learning by mitigating the need for large volumes of human-collected demonstrations. Through TAMP-based action adaption and 3D point cloud manipulation, *DemoGen* generates spatially augmented demonstrations with minimal cost, significantly improving visuomotor policy's spatial generalization capability across a wide range of real-world tasks and platforms. Furthermore, we extend *DemoGen* to generate demonstrations incorporating disturbance resistance and obstacle avoidance behaviors, endowing the trained policies with the corresponding capabilities.

Limitations. Although we have demonstrated the effectiveness of *DemoGen*, it has several limitations. First, *DemoGen* relies on the availability of segmented point clouds, which limits its applicability in highly cluttered or unstructured environments. Second, *DemoGen* is not suitable for tasks where spatial generalization is not required, such as in-hand reorientation Chen et al. (2022) or push-T Florence et al. (2022); Chi et al. (2023) with a fixed target pose. Third, the performance of *DemoGen* is affected by the visual mismatch problem caused by the constraint of single-view observation, which is discussed in detail in Appendix C.

Future works. Future works could explore mitigating the impact of visual mismatch, potentially by leveraging techniques such as contrastive learning or 3D generative models. Another avenue for future research is to use additional human-collected demonstrations as source data, aiming to identify the optimal balance between policy performance and the overall cost of data collection.

REFERENCES

Tao Chen, Jie Xu, and Pulkit Agrawal. A system for general in-hand object re-orientation. In *CoRL*, 2022.

Shuo Cheng, Caelan Reed Garrett, Ajay Mandlekar, and Danfei Xu. Nod-tamp: Multi-step manipulation planning with neural object descriptors. In *CoRL 2023 Workshop on Learning Effective Abstractions for Planning (LEAP)*, 2023.

Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *RSS*, 2023.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024a.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024b.

Murtaza Dalal, Ajay Mandlekar, Caelan Reed Garrett, Ankur Handa, Ruslan Salakhutdinov, and Dieter Fox. Imitating task and motion planning with visuomotor transformers. In *Conference on Robot Learning*, pp. 2565–2593. PMLR, 2023.

Norman Di Palo and Edward Johns. Learning multi-stage tasks with one demonstration via self-replay. In *Conference on Robot Learning*, pp. 1180–1189. PMLR, 2022.

Norman Di Palo and Edward Johns. Dinobot: Robot manipulation via retrieval and alignment with vision foundation models. *arXiv preprint arXiv:2402.13181*, 2024.

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 *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pp. 226–231, 1996.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *CoRL*, 2022.

Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *arXiv*, 2024.

Chongkai Gao, Zhengrong Xue, Shuying Deng, Tianhai Liang, Siqi Yang, Lin Shao, and Huazhe Xu. Riemann: Near real-time se (3)-equivariant robot manipulation without point cloud segmentation. *arXiv preprint arXiv:2403.19460*, 2024.

Caelan Garrett, Ajay Mandlekar, Bowen Wen, and Dieter Fox. Skillmimicgen: Automated demonstration generation for efficient skill learning and deployment. *arXiv preprint arXiv:2410.18907*, 2024.

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*, pp. 770–778, 2016.

Ryan Hoque, Ajay Mandlekar, Caelan Garrett, Ken Goldberg, and Dieter Fox. Intervengen: Interventional data generation for robust and data-efficient robot imitation learning. *arXiv preprint arXiv:2405.01472*, 2024.

Pu Hua, Minghuan Liu, Annabella Macaluso, Yunfeng Lin, Weinan Zhang, Huazhe Xu, and Lirui Wang. Gensim2: Scaling robot data generation with multi-modal and reasoning llms. *arXiv preprint arXiv:2410.03645*, 2024.

Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. *arXiv preprint arXiv:2410.24185*, 2024.

Edward Johns. Coarse-to-fine imitation learning: Robot manipulation from a single demonstration. In *2021 IEEE international conference on robotics and automation (ICRA)*, pp. 4613–4619. IEEE, 2021.

Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*, 2024.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

James J Kuffner and Steven M LaValle. Rrt-connect: An efficient approach to single-query path planning. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, volume 2, pp. 995–1001. IEEE, 2000.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

Jinhan Li, Yifeng Zhu, Yuqi Xie, Zhenyu Jiang, Mingyo Seo, Georgios Pavlakos, and Yuke Zhu. Okami: Teaching humanoid robots manipulation skills through single video imitation. In *8th Annual Conference on Robot Learning*, 2024.

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

I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ajay Mandlekar, Caelan Reed Garrett, Danfei Xu, and Dieter Fox. Human-in-the-loop task and motion planning for imitation learning. In *Conference on Robot Learning*, pp. 3030–3060. PMLR, 2023a.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*, 2023b.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning*, pp. 892–909. PMLR, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

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

Aaditya Prasad, Kevin Lin, Jimmy Wu, Linqi Zhou, and Jeannette Bohg. Consistency policy: Accelerated visuomotor policies via consistency distillation. In *Robotics: Science and Systems*, 2024.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Vaibhav Saxena, Matthew Bronars, Nadun Ranawaka Arachchige, Kuancheng Wang, Woo Chul Shin, Soroush Nasiriany, Ajay Mandlekar, and Danfei Xu. What matters in learning from large-scale datasets for robot manipulation. In *CoRL 2024 Workshop on Mastering Robot Manipulation in a World of Abundant Data*, 2024. URL `https://openreview.net/forum?id=oY0PQdLA6c`.

Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural descriptor fields: Se (3)-equivariant object representations for manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 6394–6400. IEEE, 2022.

Hengkai Tan, Xuezhou Xu, Chengyang Ying, Xinyi Mao, Songming Liu, Xingxing Zhang, Hang Su, and Jun Zhu. Manibox: Enhancing spatial grasping generalization via scalable simulation data generation. *arXiv preprint arXiv:2411.01850*, 2024.

Marcel Torne, Anthony Simeonov, Zechu Li, April Chan, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. *arXiv preprint arXiv:2403.03949*, 2024.

Eugene Valassakis, Georgios Papagiannis, Norman Di Palo, and Edward Johns. Demonstrate once, imitate immediately (dome): Learning visual servoing for one-shot imitation learning. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8614–8621. IEEE, 2022.

Vitalis Vosylius and Edward Johns. Instant policy: In-context imitation learning via graph diffusion. *arXiv preprint arXiv:2411.12633*, 2024.

Dian Wang, Stephen Hart, David Surovik, Tarik Kelestemur, Haojie Huang, Haibo Zhao, Mark Yeatman, Jiuguang Wang, Robin Walters, and Robert Platt. Equivariant diffusion policy. *arXiv preprint arXiv:2407.01812*, 2024a.

Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. Gensim: Generating robotic simulation tasks via large language models. *arXiv preprint arXiv:2310.01361*, 2023a.

Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. *arXiv preprint arXiv:2311.01455*, 2023b.

Zhendong Wang, Zhaoshuo Li, Ajay Mandlekar, Zhenjia Xu, Jiaojiao Fan, Yashraj Narang, Linxi Fan, Yuke Zhu, Yogesh Balaji, Mingyuan Zhou, et al. One-step diffusion policy: Fast visuomotor policies via diffusion distillation. *arXiv preprint arXiv:2410.21257*, 2024b.

Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. You only demonstrate once: Category-level manipulation from single visual demonstration. *Robotics: Science and Systems 2022*, 2022.

Zhengrong Xue, Zhecheng Yuan, Jiashun Wang, Xueqian Wang, Yang Gao, and Huazhe Xu. Useek: Unsupervised se (3)-equivariant 3d keypoints for generalizable manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1715–1722. IEEE, 2023.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *CoRL*, 2020.

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.

Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, Xue Bin Peng, and Jiajun Wu. Generalizable humanoid manipulation with improved 3d diffusion policies. *arXiv preprint arXiv:2410.10803*, 2024a.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy. *arXiv preprint arXiv:2403.03954*, 2024b.

Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, et al. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, pp. 726–747. PMLR, 2021.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

Tony Z Zhao, Jonathan Tompson, Danny Driess, Pete Florence, Kamyar Ghasemipour, Chelsea Finn, and Ayzaan Wahid. Aloha unleashed: A simple recipe for robot dexterity. *arXiv preprint arXiv:2410.13126*, 2024.

A EMPIRICAL STUDY: SPATIAL GENERALIZATION CAPABILITY OF VISUOMOTOR POLICIES

In this section, we present a brief empirical study examining the spatial generalization capability of visuomotor policies. We demonstrate how the lack of such generalization contributes to the data-intensive nature of learning visuomotor policies.

A.1 VISUALIZATION OF SPATIAL EFFECTIVE RANGE

Spatial generalization refers to the ability of a policy to perform tasks involving objects placed in configurations that were not seen during training. To gain an intuitive understanding of spatial generalization, we visualize the relationship between the spatial effective range of visuomotor policies and the spatial distribution of demonstration data.

Tasks. We evaluate a Button-Large task adapted from the MetaWorld Yu et al. (2020) benchmark, where the robot approaches a button and presses it down. The object randomization range is modified to a $30\,\text{cm} \times 40\,\text{cm} = 1200\,\text{cm}^2$ area on the tabletop workspace, covering most of the end-effector's reachable space. Noticing the large size of the button makes it pressed down even if the press motion does not precisely hit the center, we also examine a more precision-demanding variant, Button-Small, where the button size is reduced by a factor of 4.

Policy. We adopt 3D Diffusion Policy (DP3)Ze et al. (2024b) as the studied policy, as our benchmarking results indicate that 3D observations provide superior spatial generalization compared to 2D approaches. Training details are provided in Appendix B.1.

Evaluation. To visualize the spatial effective range, we uniformly sample 21 points along each axis within the workspace, resulting in a total of 441 distinct button placements. Demonstrations are generated using a scripted policy, with 4 different spatial distributions ranging from `single` to `full`. The performance of each configuration is evaluated on the 441 placements, enabling a comprehensive assessment of spatial generalization. The visualization result is presented in Fig. 13.

Key findings. Overall, the spatial effective range of visuomotor policies is closely tied to the distribution of object configurations seen in the demonstrations. Specifically, the effective range can be approximated by the union of the areas surrounding the demonstrated object placements. Thus, to train a policy that generalizes well across the entire object randomization range, demonstrations must cover the full workspace, resulting in substantial data collection costs. Furthermore, as task precision requirements increase, the effective range shrinks to more localized areas, necessitating a greater number of demonstrations to adequately cover the workspace.

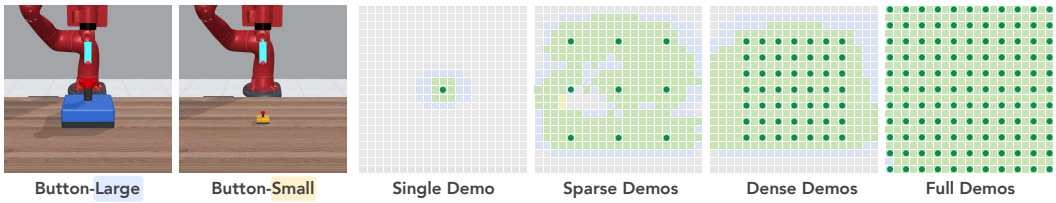

Figure 13: **Qualitative visualization of the spatial effective range.** The grid maps display discretized tabletop workspaces from a bird's-eye view under different demonstration configurations. Dark green spots mark the locations where buttons are placed during the demonstrations. Each grid cell corresponds to a policy rollout with the button placed at that location. Blue, yellow, green, and gray grids denote successful executions for the Button-Large, Button-Small, both tasks, and no tasks, respectively.

A.2 BENCHMARKING SPATIAL GENERALIZATION CAPABILITY

The practical manifestation of the spatial generalization is reflected in the number of demonstrations required for effective policy learning. In the following benchmarking, we explore the relationship between the number of demonstrations and policy performance to determine how many demonstrations are sufficient for effective training.

Tasks. To suppress the occurrence of inaccurate but successful policy rollouts, we design a Precise-Peg-Insertion task. We construct a T-shaped peg, whose upper end has a cross-section of $6\,\text{cm} \times 6\,\text{cm}$, and the bottom end has a cross-section of $3\,\text{cm} \times 3\,\text{cm}$. The hole in the green socket has a cross-section of $4\,\text{cm} \times 4\,\text{cm}$. This shape enforces a strict fault tolerance of $1\,\text{cm}$ during both the picking and insertion stages, asking for millimeter-level precision. Both objects are randomized in a $40\,\text{cm} \times 20\,\text{cm}$ workspace in the `full` setting. The randomization range is halved into $20\,\text{cm} \times 10\,\text{cm}$ in the `half` setting.

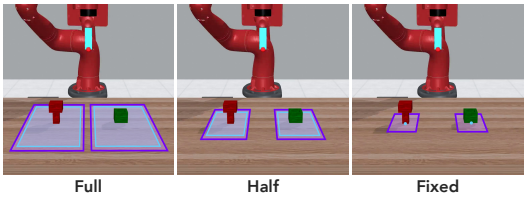

Figure 14: **The Precise-Peg-Insertion task.** A total of 3 workspace sizes is considered. Purple and sky-blue rectangles mark the workspaces for demonstration and evaluation, respectively.

Policies. In addition to Diffusion Policy (DP)Chi et al. (2023) and 3D Diffusion Policy (DP3)Ze et al. (2024b) trained from scratch, we explore the potential of pre-trained visual representations to enhance spatial generalization. Specifically, we replace the train-from-scratch ResNet He et al. (2016) encoder in DP with pre-trained encoders including R3M Nair et al. (2023), DINOv2 Oquab et al. (2023), and CLIP Radford et al. (2021). Detailed implementations are provided in Appendix B.2.

Demonstrations. We vary the number of demonstrations from 25 to 400. The object configurations are randomly sampled from a slightly larger range than the evaluation workspace to avoid performance degradation near workspace boundaries. A visualization is provided in Fig. 14.

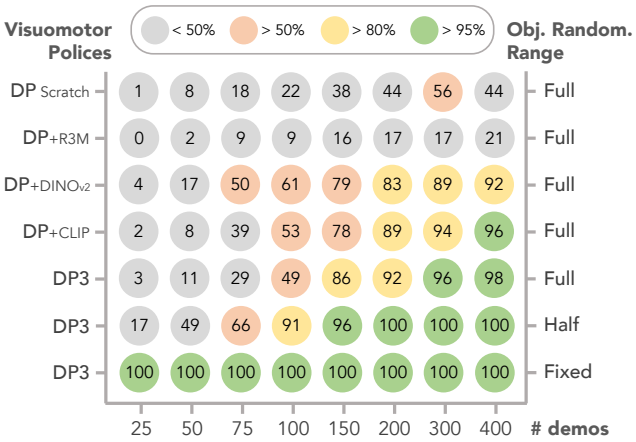

Figure 15: **Quantitative benchmarking on the spatial generalization capacity.** We report the relationship between the agent's performance in success rates and the number of demonstrations used for training when different visuomotor policies and object randomization ranges are adopted. The results are averaged over 3 seeds.

Evaluation. In the `full` workspace, both the peg and socket are placed on 45 uniformly sampled coordinates, resulting in 2025 distinct configurations for evaluation. For the `half` and `fixed` settings, the number of evaluated configurations is 225 and 1, respectively.

Key findings. The degree of object randomization significantly influences the required demonstrations. Therefore, an effective evaluation protocol for visuomotor policies must incorporate a sufficiently large workspace to provide enough object randomization. On the other hand, both 3D representations and pre-trained 2D visual encoders contribute to improved spatial generalization capabilities. However, none of these methods fundamentally resolve the spatial generalization problem. This indicates the agent's spatial capacity is not inherently derived from the policy itself but instead develops through extensive traversal of the workspace from the given demonstrations.

B Policy Training and Implementation Details

We select 3D Diffusion Policy (DP3) Ze et al. (2024b) as the visuomotor policy used for real-world and simulated experiments. We compare its performance against 2D Diffusion Policy (DP) Chi et al. (2023) in the empirical study in Sec. A. We list the training and implementation details as follows.

B.1 Details for Policy Training

For a fair comparison, we fix the total training steps counted by observation-action pairs to be $2\mathrm{M}$ for all evaluated settings, resulting in an equal training cost regardless of the dataset size. To stabilize the training process, we use AdamW Loshchilov (2017) optimizer and set the learning rate to be $1\mathrm{e}^{-4}$ with a 500 step warmup.

In real-world experiments, we use the DBSCAN Ester et al. (1996) clustering algorithm to discard the outlier points and downsample the number of points in the point cloud observations to 1024. In the simulator, we skip the clustering stage and downsample the point clouds to 512 points.

We follow the notation in the Diffusion Policy Chi et al. (2023) paper, where T_o denotes the observation horizon, T_p as the action prediction horizon, and T_a denotes the action execution horizon. In real-world experiments, we set $T_\mathrm{o} = 2$, $T_\mathrm{p} = 8$, $T_\mathrm{a} = 5$. We run the visuomotor policy at $10\mathrm{Hz}$. Since T_a indicates the steps of actions executed on the robot without re-planning, our horizon settings result in a closed-loop re-planning latency of 0.5 seconds, responsive enough for conducting dexterous retrying behaviors and disturbance resistance. In the simulator, since the tasks are simpler, we set $T_\mathrm{o} = 2$, $T_\mathrm{p} = 4$, $T_\mathrm{a} = 3$.

B.2 Pre-Trained Encoders for Diffusion Policies

To replace the train-from-scratch ResNet18 He et al. (2016) visual encoder in the original Diffusion Policy architecture, we consider 3 representative pre-trained encoders: R3M Nair et al. (2023), DINOv2 Oquab et al. (2023), and CLIP Radford et al. (2021). R3M utilizes a ResNet He et al. (2016) architecture and is pre-trained on robotics-specific tasks. DINOv2 and CLIP employ ViT Dosovitskiy et al. (2021) architectures and are pre-trained on open-world vision tasks. These encoders are widely used in previous works Chi et al. (2024a); Lin et al. (2024) to enhance policy performance.

C LIMITATION: THE VISUAL MISMATCH PROBLEM

Figure 16: **Illustration for the visual mismatch problem.** The appearance changes due to the perspective change.

While the one-shot imitation experiment verifies the effectiveness of *DemoGen*, it also reveals its limitation: synthetic demonstrations generated from one source demonstration are not as effective as the same number of human-collected demonstrations. We attribute the performance gap to the *visual mismatch* problem under the constraint of a single-view observation perspective. When objects move through 3D space, their appearance changes due to variations in perspective. An illustration is provided in Fig. 16. However, synthetic demonstrations consistently reflect a fixed side of the object's appearance seen in the source demonstration. This discrepancy causes a visual mismatch between the synthetic and real-captured data.

C.1 PERFORMANCE SATURATION.

A notable consequence of the visual mismatch problem is the phenomenon of performance saturation. An empirical analysis is conducted on the Pick-Cube task. In Fig. 17(a), we fix the spatial density of target object configurations in the synthetic demonstrations and increase their spatial coverage by adding more synthetic demonstrations. The curve indicates that the performance improvement plateaus once the spatial coverage exceeds a certain threshold. This saturation occurs because the visual mismatch intensifies as the distance between the source and synthetic object configurations increases, making additional synthetic demonstrations ineffective. In Fig.17(b), similar performance saturation is observed when we increase the density while fixing the spatial coverage. This indicates excessive demonstrations are unnecessary once they sufficiently cover the workspace.

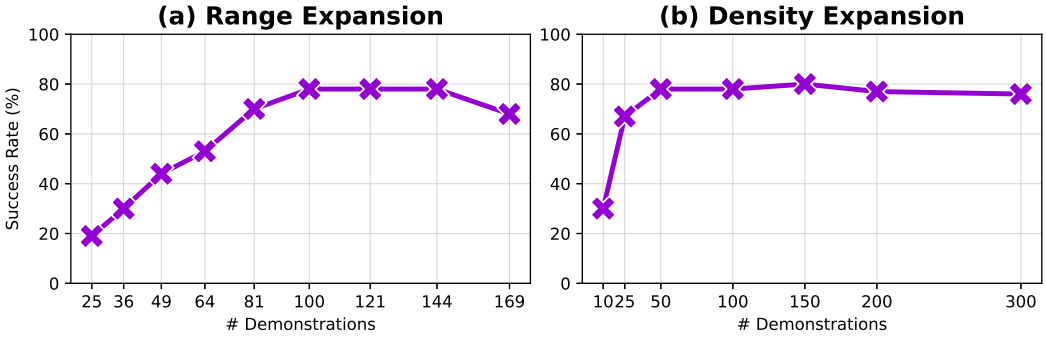

Figure 17: **Performance Saturation.** We report the policy performance boost w.r.t. the increase of synthetic demonstrations over 3 seeds.

D EXPERIMENTAL DETAILS

D.1 RANDOMIZATION RANGES FOR SIMULATED TASKS

In Fig. 5, we illustrated the simulated tasks for the evaluation on spatial generalization. To strengthen the significance of spatial generalization, we enlarge the original object randomization ranges in the MetaWorld Yu et al. (2020) tasks. For demonstration generation, we select a slightly larger range than the evaluation workspace to avoid performance degradation near the workspace boundaries. The detailed workspace sizes are listed in Tab. 6.

Table 6: **Object randomization ranges in simulated tasks.** All the reported sizes have the units in centimeters.

	Pick-Cube	Button-Small	Drawer-Close	Faucet-Open	Handle-Press	Box-Lid	Stack-Cube	Assembly
Object(s)	Cube	Button	Drawer	Faucet	Toaster	Box × Lid	Red × Green	Pillar × Hole
Evaluation	40×40	40×40	15×15	30×30	20×30	$(2.5\times30)^2$	$(15\times15)^2$	$(10\times30)^2$
DemoGen	48×48	48×48	20×20	40×40	25×40	$(7.5\times40)^2$	$(20\times20)^2$	$(15\times40)^2$

D.2 TASK DESCRIPTIONS FOR REAL-WORLD TASKS

In Fig. 7, we illustrated the real-world tasks for the evaluation on spatial generalization. We describe these tasks in the text as follows, where we mark the verbs for motion and skill actions in the corresponding colors.

1. **Spatula-Egg.** The gripper holds a spatula in hand. The robot maneuvers the spatula to first move toward the fried egg and then 1) slide beneath the egg, 2) lift the egg leveraging the contact with the plate's rim, 3) carry the egg and maintain stable suspension.
2. **Flower-Vase.** The gripper moves toward the flower, picks it up, reorients it in the air while transferring toward the vase, and finally inserts it into the vase.
3. **Mug-Rack.** The gripper moves toward the mug, picks it up, reorients it in the air while transferring toward the rack, and hangs it onto the rack.
4. **Dex-Cube.** The dexterous hand moves toward the cube and grasps up the cube.
5. **Dex-Rollup.** The dexterous hand moves toward a piece of plasticine and wraps it multiple times until it is fully coiled. The required times of the wrapping motion may vary due to the distinct plasticity of every hand-molded piece of plasticine.
6. **Dex-Drill.** The dexterous hand moves toward the drill, grasps it up, transfers it toward the cube, and finally touches the cube with the drill.
7. **Dex-Coffee.** The dexterous hand moves toward the kettle, grasps it up, transfers it toward the coffee filter, and finally pours water into the filter.

D.3 INCREASED HUMAN-COLLECTED DEMONSTRATIONS

In Tab. 3, we compare the *DemoGen*-generated dataset against 3 human-collected source demonstrations. In Fig. 18, we provide a reference on how the increase of source demonstrations leads to the enhancement of policy performance on the Dex-Cube task. To further understand the policy capacity enabled by human-collected demonstrations, we visualize the spatial heatmaps of human-collected datasets in Fig. 19. By comparing the demonstrated configurations and the spatial effective range of the resulting policies, we found the policy capacity is upper-bounded by the demonstrated configurations. This is in line with the findings in the empirical study.

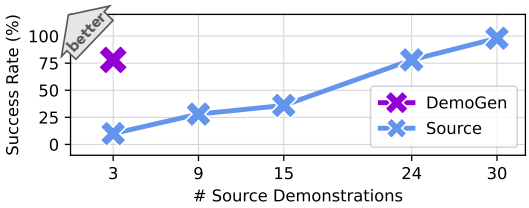

Figure 18: **Real-world comparison between *DemoGen*-generated and human-collected datasets.** The *DemoGen*-generated dataset is based on 3 source demonstrations.

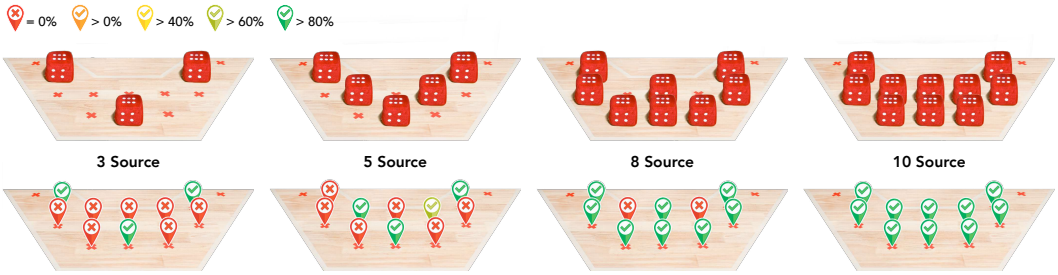

Figure 19: **Visualization of the policy performance trained on human-collected datasets.** (Upper row) The demonstrated configurations. (Bottom row) The spatial heatmaps with success rates averaged on 5 trials.

D.4 DETAILED ANALYSIS OF THE BIMANUAL HUMANOID EXPERIMENT

The orientational augmentations share the same visual mismatch problem as translational augmentation. The policy performs as expected when the generated orientations are close to the orientation in the source demonstration. As the orientational difference increases, we observed the policy might react to the orientation in the current visual observation with actions for mismatched orientations.

Additionally, we found the spatial generalization problem persists in mobile manipulation scenarios. This is mainly due to the physical constraints of real-world environments, such as kitchen countertops or fruit stands, as demonstrated in our experiments, where terrain limitations prevent the base from approaching objects at arbitrary distances. Consequently, the base typically moves to a fixed point at a specific distance from the object, after which the robot conducts a standard non-mobile manipulation process at the fixed base position.

D.5 Disturbance Resistance Experiments Details

D.5.1 Evaluation Metrics

The sauce coverage score is computed as follows. First, we distinguish between green background and red sauce in the HSV color space. The identified background is set to black, the sauce is set to red, and the rest which should be the uncovered crust is set to white. Second, due to the highlights on the sauce liquid, some small fragmented points of the sauce may be identified as the crust. To address this, we apply smoothing filtering followed by dilation and erosion, where the kernel size is 9×9. Finally, the coverage is calculated as the ratio of red areas (sauce) over non-black areas (sauce + uncovered crust).

D.5.2 Raw Evaluation Results

For quantitative evaluation, we perform 5 repetitions for each of the 5 disturbance directions, resulting in 25 trials for both strategies.

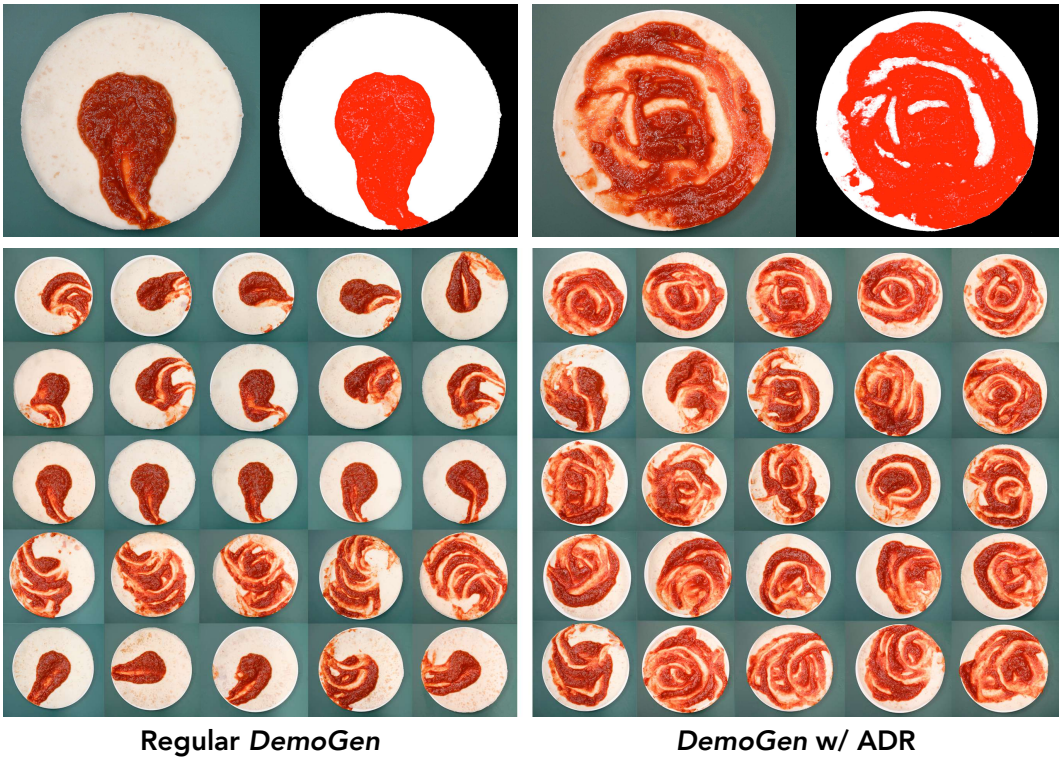

Figure 20: **Raw evaluation results in the Sauce-Spreading task.** (Top) Examples of the processing results for metric calculation. (Bottom) Compared with the regular *DemoGen*, the policy trained with the ADR strategy better spreads the sauce to cover the crust under external disturbance.

D.6 VISUALIZATION OF *DemoGen*-GENERATED TRAJECTORIES

In Fig. 4, we gave a concrete example of the trajectory of synthetic visual observations. We provide more examples in Fig. 21 by showcasing the key frames of source and generated demonstrations.

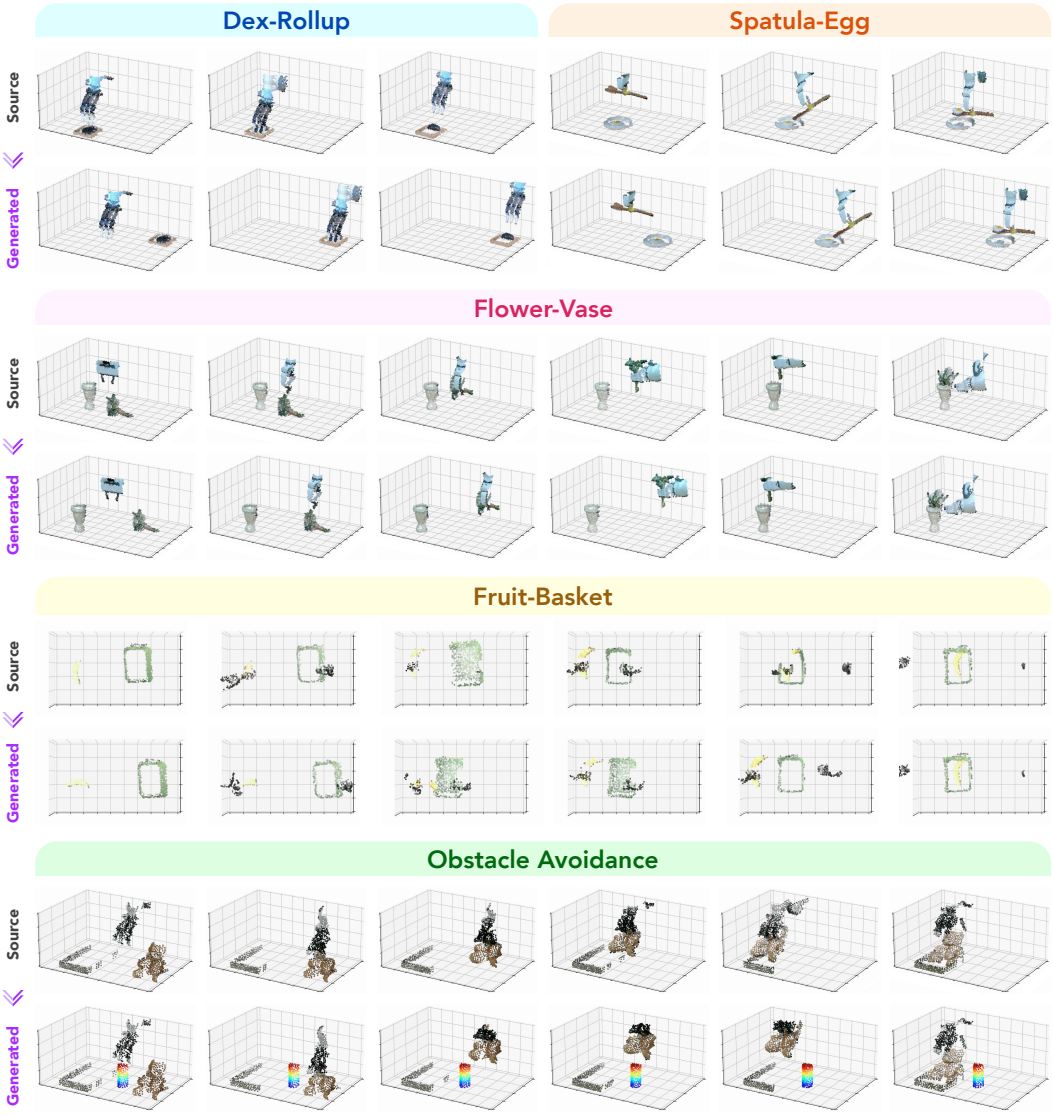

Figure 21: **More examples of the trajectories consisting of synthetic visual observations.**