AUTONOMOUS PLAY WITH CORRESPONDENCE-DRIVEN TRAJECTORY WARPING

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability to conduct and learn from self-directed interaction and experience is a central challenge in robotics, offering a scalable alternative to labor-intensive human demonstrations. However, realizing such "play" requires (1) a policy robust to diverse, potentially out-of-distribution environment states, and (2) a procedure that continuously produces useful, task-directed robot experience. To address these challenges, we introduce Tether, a method for autonomous play with two key contributions. First, we design a novel non-parametric policy that leverages strong visual priors for extreme generalization: given two-view images, it identifies semantic correspondences to warp demonstration trajectories into new scenes. We show that this design is robust to significant spatial and semantic variations of the environment, such as dramatic positional differences and unseen objects. We then deploy this policy for autonomous multi-task play in the real world via a continuous cycle of task selection, execution, evaluation, and improvement, guided by the visual understanding capabilities of vision-language models. This procedure generates diverse, high-quality datasets with minimal human intervention. In a household-like multi-object setup, our method is among the first to perform many hours of autonomous real-world play, producing a stream of data that consistently improves downstream policy performance over time. Ultimately, Tether yields over 1000 expert-level trajectories and trains policies competitive with those learned from human-collected demonstrations.

1 INTRODUCTION

Recent advances in robotic manipulation have been powered by imitation learning policies (Zhao et al., 2023; Chi et al., 2023; Zhao et al.; Black et al., 2024a; Kim et al., 2024; Brohan et al., 2022; 2023; Collaboration et al., 2024) trained on real-world teleoperated demonstrations. In all these cases, the human effort involved in teaching a skill is substantial: while there are continued efforts to simplify teleoperation interfaces (Zhao et al., 2023; Shafiullah et al., 2023; Chi et al., 2024; Cheng et al., 2024; Fu et al., 2024), such demo datasets can fundamentally only scale linearly with human time, and these data-hungry policy architectures need large spatially and semantically diverse datasets in order to generalize usefully (Lin et al., 2024). In this paper, we propose an alternative paradigm: autonomous play with robust policies where data scales primarily with robot time, minimizing the human effort bottleneck.

Our method, Tether, involves two key components. First, autonomous play requires an extremely robust policy that can recover from mistakes and out-of-distribution states. Without relying on massive datasets for training large neural policy architectures, we instead design a new non-parametric policy class that specifically supports generalization with few demonstrations. Specifically, our architecture exploits the remarkable leaps in semantic image keypoint correspondences: given a new scene with potentially new task-relevant object instances in new spatial layout with new distractors, it first computes keypoint correspondences with the demo images, selects the closest-matched demo, computes 3D transformations associated with each correspondence, and accordingly warps the robot trajectory to fit the new scene. Validated on 12 manipulation tasks in a household-like setting, we show that our policy surpasses the performance of alternative methods, including those that rely on foundation models or pretraining with large robotics or internet datasets.

Second, we run our policy within a cyclic multi-task play procedure that autonomously produces data for continuous downstream policy training. For a collection of tasks with few human demonstrations each, we query a vision-language model (VLM) to repeatedly plan and select tasks our policy should attempt. This is real-world task-directed play without any manual resets: the procedure naturally induces resets with a growing initial state distribution as the object configurations drift away from the beginning of play. Additionally, to evaluate the suboptimal play data, we query a VLM and compute correspondences to detect successful executions; these are then used downstream for filtered imitation learning. We show that this procedure can collect over 1000 new expert-level demonstrations across 26 hours with minimal human intervention (5 cases requiring a minute total of human time, 0.26% of executions). We also validate that this stream of newly generated demonstrations progressively improves the training of neural policies, which consistently improve with more play data and reach high success rates competitive with policies trained on human-collected demonstrations.

In summary, our contributions are:

1. A keypoint correspondence-driven trajectory warping policy that exhibits impressive generalization and robustness for diverse manipulation tasks.
2. A multi-task VLM-guided play procedure that generates increasingly diverse demonstrations over many hours, powering downstream neural policy training.

2 Related Work

Robustness in Imitation Learning. To build robust policies that excel in diverse environments, many prior efforts turn to scaling data and policy architecture, often with foundation models or large human-collected demo datasets (Khazatsky et al., 2024; Collaboration et al., 2024; AgiBot-World-Contributors et al., 2025). Prominent techniques include training representations for robotics on non-robot data (e.g., human videos) (Nair et al., 2022; Ma et al., 2023; Shi et al., 2025), querying vision-language models (VLMs) or large language models (LLMs) (Nasiriany et al., 2024; Goetting et al., 2024; Fang et al., 2024), and finetuning vision-language-action models (VLAs) (Black et al., 2024a; NVIDIA et al., 2025; Intelligence et al., 2025).

Another class of approaches design strong built-in priors for models trained on much fewer demos, with some executing actions open-loop. Some build action affordances or primitives (Kuang et al., 2024; Haldar & Pinto, 2025), retrieve from existing datasets (Du et al., 2023; Memmel et al., 2024; Xie et al., 2025), leverage pretrained representations and models (Pari et al., 2021; Burns et al., 2023; Shi et al., 2024), or exploit 3D scene geometry (Rashid et al., 2023; Goyal et al., 2024; Ke et al., 2024; Ze et al., 2024). Closest to our work are methods that operate on visual semantic keypoint correspondences. Like object-centric approaches (Shi et al., 2024; Qian et al., 2024), they benefit from recent advances in scene understanding and are naturally robust to distractors, yet provide higher spatial precision and avoid the rigid "objectness" assumptions that fail on deformable objects and granular particles. One class of approaches tracks keypoints through frames of human or robot videos and retargets the dense trajectory to the desired setting (Wen et al., 2023; Bharadhwaj et al., 2024; Ren et al., 2025). Another class instead uses keypoint correspondences as a compact trajectory representation. Among these, KAT (Di Palo & Johns, 2024) queries an LLM to generate open-loop actions based on keypoints, while P3-PO (Levy et al., 2024) and SKIL (Wang et al., 2025) input keypoints to point-conditioned policies. We too use keypoint correspondences, but we demonstrate the advantages of a more direct approach: using correspondences to select and warp a demonstrated trajectory to fit the new scene. We compare against KAT (Di Palo & Johns, 2024) in our experiments.

Autonomous Data Generation. To reduce the need for large, human-collected datasets, recent efforts in robotic manipulation have explored autonomously generating data for downstream policy learning. Previous works collect data in simulation by querying foundation models to propose and solve tasks (Ha et al., 2023; Wang et al., 2024) or leveraging privileged simulation state to adapt human-collected demos for new scene configurations (Mandlekar et al., 2023; Jiang et al., 2025; Lin et al., 2025a). However, simulation-based approaches struggle with sim-to-real transfer within cluttered, unstructured environments, which remains an open challenge especially for tasks involving complex contacts and vision-based policies trained on synthetic renders (Blanco-Mulero et al., 2024; Yu et al., 2024; Lin et al., 2025b). Alternatively, another class of works autonomously generate data directly in the real world. Some require initial policies that are trained on hundreds of human-

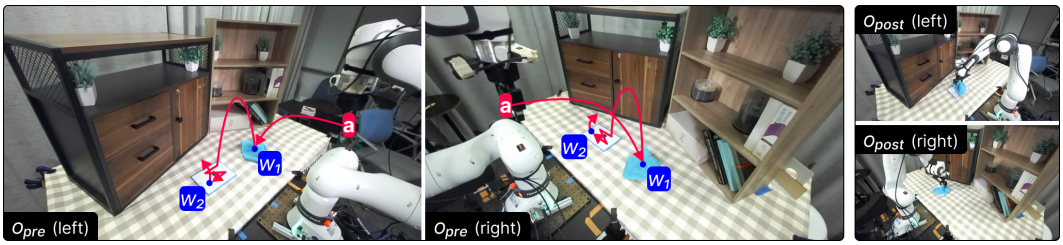

Figure 1: **Demonstration Summaries.** Tether summarizes demonstrations into the pre- and post-images, action sequence (red), waypoints (blue), and keypoints.

collected demos (Zhou et al., 2024; Mirchandani et al., 2024), which still pose a bottleneck when transferring to new tasks and environments. Most similar to our work are methods that initialize from few demos using specially-designed policies. Manipulate-Anything (Duan et al., 2024b) introduces a zero-shot foundation model policy and applies it to autonomous data collection. However, performing multiple rounds of foundation model inference significantly hinders its throughput, and it collects under 50 real world demos; additionally, environment reset during data collection is not considered. In contrast, we introduce a system that autonomously collects over 1000 demonstrations in the real world with minimal human intervention.

3 Robust Imitation And Autonomous Play

We formulate robot manipulation as a Controlled Markov Process (CMP). A CMP is represented as a tuple $M = (S, A, \mathcal{T})$, where S is the set of states, A is the set of actions, $\mathcal{T} : S \times A \to \Delta(S)$ is the transition dynamics function.

In the imitation learning problem, for a given CMP M and a desired task, we have a dataset of N expert demonstrations $\mathcal{D} = \{\tau_1, \tau_2 \ldots, \tau_N\}$ that perform that task, with trajectories $\tau_i = \{s_1, a_1, s_2, a_2, \ldots\}$. In practice, our robot does not have direct access to state s_t and instead receives visual observations $o_t = O(s_t)$, consisting of two third-person RGB camera views from the left and right sides, as shown in Figure 1. There is no significant partial observability. The actions a_t are the 6-DOF pose of robot gripper and the gripper's binary open/close command. Given this demonstration dataset $\mathcal{D}$, our goal is to learn a policy $\pi : O \to \Delta(A)$ for the task.

As motivated in the introduction, today's standard imitation learning approaches require extensive human demonstration collection that is difficult to scale. Our paper directly addresses this problem. First, we introduce a non-parametric trajectory warping-based policy that generates action plans from a few training demonstrations (Section 3.1), much fewer than standard neural policies. Next, to address the challenge of scaling data, we deploy our policy in a VLM-guided multi-task play procedure, which autonomously produces demos for the downstream training of parametric policies.

3.1 Trajectory Warping with Keypoint Correspondences

Towards robust imitation, Tether leverages semantic visual priors in the form of image correspondence matching algorithms to interpolate within and generalize beyond a few demonstrations.

At a high level, our policy works as follows. Every demonstration is represented by the pre-image, i.e. the camera observations at the beginning of the demo, the post-image from the end of the demo, and a sequence of 3-D waypoints for the gripper motion. These waypoints, projected onto the pre-image, identify important visual "keypoints" for that demo. To execute the policy starting from an initial observation of the scene, we identify the best-matched demo based on the quality of keypoint correspondences, backproject those keypoints to compute desired 3-D gripper waypoints, and warp the demonstrated trajectory based on those waypoints. We walk through these steps in detail below.

Demonstration Summaries: Pre-Image, Post-Image, Waypoints, and Keypoints. Our policy is non-parametric, and relies on accessing demonstrations at test time. For convenience, we first preprocess all demonstrations into concise summaries in preparation for execution. This is a "training time" operation, it needs to be done once for each demonstration, and once summarized, the original demonstration can be discarded. For each demonstrated trajectory $\tau_i \in \mathcal{D}$, we summarize the key information for Tether in a tuple $\kappa_i = (o_{pre}, o_{post}, W, K_{pre}, K_{post}, \mathbf{a})$, where o_{pre} and o_{post}

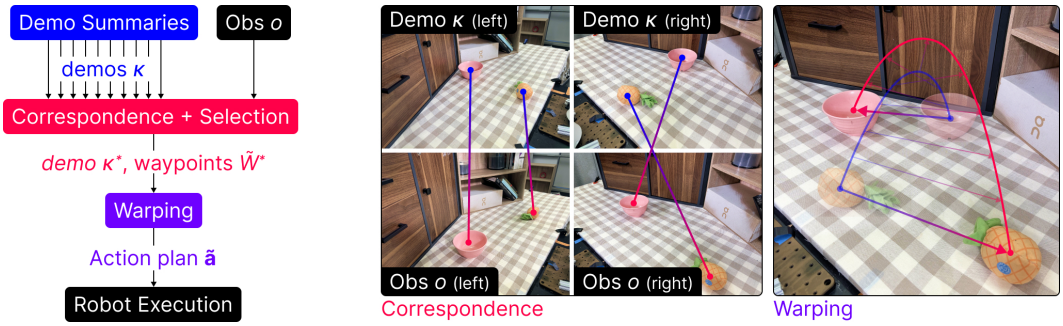

Figure 2: **Policy Inference**. During inference, Tether (left) computes correspondences (middle) and produces a warped trajectory action plan (right).

are respectively the pre- and post-images i.e., two-view camera observations at the beginning and end of the trajectory. W is a "waypoint" sequence $[w_1, w_2, ..., w_T]$ of task-critical 3-D gripper locations. In practice, we simply use the gripper locations from frames where the gripper open/close position toggles, following the convention for selecting critical frames from prior work (Johns, 2021; Mandlekar et al., 2023; Vecerik et al., 2024), and which is applicable to diverse manipulation tasks. Next, $\mathbf{a} = [a_0, ..., a_M]$ is the full sequence of robot actions during the trajectory i.e. 6-DOF gripper positions and gripper state. Finally, for each of the T waypoint locations, we project them onto o_{pre} and o_{post} to identify visually important keypoints $K_{pre} = [k_{pre,1}, k_{pre,2}, ..., k_{pre,T}]$ and K_{post} respectively. We depict this process in Figure 1.

This remaining subsection outlines executing our policy for an observation o, visualized in Figure 2.

Correspondence Matching and Source Demo Selection. To execute a Tether policy, we start by matching the current camera observations o to the demos in $\mathcal{D}$ to find a nearest-neighbor demo. In particular, for each demo summary κ_i, we search the current images o for correspondences for all keypoint locations $K_{pre,i}$ in the demo pre-images $o_{pre,i}$. We do this separately for the left-image and right-image to find a set of corresponding pixels in each image $\tilde{K}_i = [\tilde{K}_{i,left}, \tilde{K}_{i,right}]$. To find these 2D correspondences, we use a state-of-the-art model (Zhang et al., 2024) built on DINOv2 (Oquab et al., 2023) and Stable Diffusion (Rombach et al., 2022) features in our implementation.

We then backproject these images using calibrated camera extrinsics, to obtain a sequence of target 3-D waypoints $\tilde{W}_i$. If the backprojections fail to intersect, then the match is deemed to have failed i.e. demo i is an *infeasible* match for the current observation o. For the feasible matches, we rank the demos in order of dissimilarity to o by computing the Euclidean distance between the original and target waypoints, i.e. $score_i(o) = \|W_i - \tilde{W}_i(o)\|_2$. The closest demo is selected as the source demo κ^*, with its original gripper waypoints W^*, original robot action sequence $\mathbf{a}$ and the translated target waypoints for the current scene $\tilde{W}^*(o)$.

Warping the Source Demo Trajectory. The "target waypoints" $\tilde{W}^*(o)$ above provide a scaffold for how to warp the robot trajectory for the current scene. However, we still need to fill in the fine-grained robot actions in-between by warping the intermediate segments between waypoints.

Consider the segment $[w_t, w_{t+1}]$ between two waypoints from the selected source demo κ^*. The target waypoints are $[\tilde{w}_t, \tilde{w}_{t+1}]$. Denote the action sequence segment for this waypoint as $\mathbf{a}_t$. We first compute the 3-D displacements of the two waypoints that mark the beginning and end of the segment $d_t = \tilde{w}_t - w_t$ and $d_{t+1} = \tilde{w}_{t+1} - w_{t+1}$. We now perform linear interpolation between those displacements and add the resulting displacements to the original action sequence $\mathbf{a}_t$ to get the transformed action plan segment $\tilde{\mathbf{a}}_t$. Concatenating these segments produces the full action plan $\tilde{\mathbf{a}} = [\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, ...]$ to be executed in the new scene.

To prioritize preserving spatial relationships, we perform this linear interpolation in space rather than time. For waypoints w_t, w_{t+1}, we define a local 1-D coordinate frame mapping w_t to 0 and w_{t+1} to 1. Then, the interpolation coefficient α for each $a \in$ the source action segment $\mathbf{a}_t$ is simply its coordinate in this frame. Geometrically, this can be thought of as projecting the gripper position a onto the line spanning w_t and w_{t+1}, then computing the projection's relative distance to w_t and w_{t+1}. Then, the corresponding displacement that a must undergo when warped into the new scene is: $d_a = (1 - \alpha)d_t + \alpha d_{t+1}$. The new action is thus $a + d_a$.

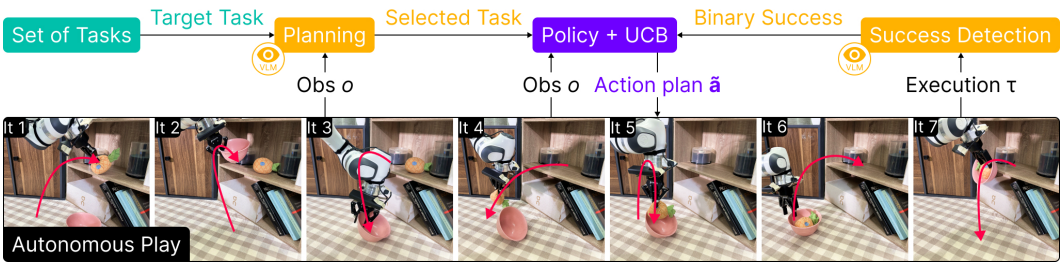

Figure 3: **Autonomous Play.** Our iterative procedure runs Tether for multiple tasks and uses VLMs for plan generation and success detection.

While simple, we show in Section 4.2 that our correspondence-driven trajectory warping is robust and performs remarkably well with as few as 10 demos, across challenging manipulation tasks—including those with out-of-distribution objects, millimeter-level precision, and complex contacts.

3.2 Autonomous Multi-Task Play with Vision-Language Models

Now, we address the issue of the scalability of human data collection by using our robust and efficient policy as a spark to set autonomous play in motion, so that the hundreds of new diverse trajectories can then be used to train larger, more flexible neural policy architectures.

To maximize the autonomy of our data generation, we apply our policy towards continuous task-directed play and design a set of tasks that facilitate natural resets and randomization: the end state of each task is a valid start state for another task (e.g., "place pineapple on table" leads into "place pineapple on shelf" or "place pineapple in bowl"), even in the event of failures. This approximately indefinitely composable structure is an extension of forward-backward tasks in prior reset-free learning (Eysenbach et al., 2017; Sharma et al., 2021; Mirchandani et al., 2024), and it allows previous tasks (and potential mistakes) to naturally randomize both relevant and background object locations and states for each task. This formulation is further illustrated by experiments in Section 4.3.

With these tasks, we run an iterative procedure where each step applies the policy to complete one task from our set. At a high level, each iteration proceeds as follows: we first query a VLM with an image of the scene and ask for an appropriate task to attempt. Then, we run the corresponding Tether policy, record its execution, and evaluate it with a two-stage process using VLMs and correspondences. These steps are visualized in Figure 3, and we describe them in detail below.

Task Selection And Planning. At each iteration, we select tasks based both on which demos we would like to add to our collection, and which tasks are actually executable. To prioritize demos to add, we maintain a running count of the number of successes for each task, and weight rare tasks higher. In particular, we sample the target task from a softmax over the negated success counts.

However, this rare target task might not be instantly executable. For instance, to attempt to “move object from shelf to table”, the object of interest must have been placed on the shelf in the first place. If that object could be in many other locations in the scene, it might only rarely reach the shelf in the course of undirected play. To overcome this, we query a VLM to provide a task plan, i.e., a sequence of executable tasks that culminates in the target task, of which we attempt the first task within the current iteration, similar to receding horizon control. Our prompt and examples are in the Appendix.

Success Evaluation. Having attempted a task, we determine the success of the resulting trajectory via two evaluations, employed in sequence to minimize false positives. First, from the final camera image o^*_{post} of the source demo κ^* that produced this trajectory, we look for correspondences of its final keypoint $k^*_{post,T}$ in the Tether execution’s post-image o_{post}. If one (or both) of the rays projected by the best-matched point from either camera view is far from the executed gripper position (beyond a preset threshold of 10 centimeters), we mark the trajectory a failure. If this test is passed, we then further query a VLM with the image frames from the trajectory to double-check for task success. This two-stage procedure alleviates the false positive problem with naively querying off-the-shelf VLMs as success detectors observed in prior works (Duan et al., 2024a).

Improving For and Through Play. There are a few additional considerations when deploying Tether. First, play benefits from injecting some stochasticity to help generate exploratory data to search for

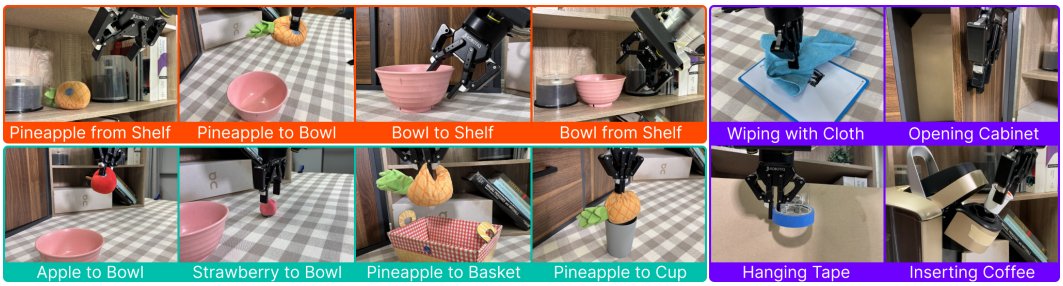

Figure 4: **Evaluation Tasks.** Our tasks involving moving fruits and containers with in-distribution (orange) and out-of-distribution (green) objects, as well as challenging manipulation skills (purple).

improved policies. This also expands the set of potential actions, decreasing the chance of our policy being trapped in states where all actions fail to induce environment transitions. Thus, rather than providing our policy with the full set of demos given for play, we first sub-select k demos and then run our policy to warp the closest one amongst them.

While we can select these k demos randomly, there is a second consideration: human demos could be of varying quality, in which case, play using various source demos could help identify *consistently better demos* to warp from, avoiding, for example, demos that involve non-robust fingertip grasps. Thus, we select the top k demos by formulating a multi-arm bandit problem: arms are demos, and the reward for picking a demo is the binary success of the executed trajectory warped from that demo. For this problem, we use upper confidence bounds (Garivier & Moulines, 2011), which balances exploring relatively less-tested source demonstrations with exploiting high-success ones. We provide pseudocode in Appendix.

4 Experiments

We conduct experiments evaluating our policy design and autonomous play procedure. Specifically, we study (1) policy robustness, particularly for out-of-distribution settings and challenging tasks, and (2) the effectiveness of autonomous play in generating a stream of data for policy training.

4.1 Experimental Setup

We run all experiments on the 7 DOF Franka Emika Panda arm running at 15 Hz and record two RGB views from calibrated ZED cameras. Across all experiments, we provide 10 demonstrations for each task. We run our semantic correspondence on 1 A6000 GPU. In autonomous play, we use Gemini-2.5 Flash for task selection and GPT-4.1 for success evaluation. Additional details are in Appendix.

Baselines And Ablations. We compare our Tether policy with recent imitation learning methods. These baselines are representative of the state-of-the-art across various levels of data-efficiency. First, π_0 (Black et al., 2024a) (with FAST (Pertsch et al., 2025)) is an open-source VLA intended to operate entirely zero-shot in new scenes. Second, Keypoint Action Tokens (KAT) (Di Palo & Johns, 2024) queries a LLM for in-context action sequence generation, given a few demos (10 in the original work). Third, Diffusion Policy (DP) (Chi et al., 2023) is a general imitation learning algorithm typically trained with a few tens up to a few hundreds of demos, depending on task complexity. Here, we evaluate π_0 zero-shot and provide DP and KAT with 10 demos each, same as our approach. Additionally, we ablate the number of demos given to our method, evaluating with 1, 5, and 10.

Tasks. We visualize our 12 tasks in Figure 4. First, we have 4 tasks that involve moving fruits and containers on a table and shelf. We instantiate these tasks with a soft pineapple toy and a curved rigid bowl. The main challenges are the bowl, which requires careful orientation of the gripper to prevent slipping, and the shelf, which requires a horizontal orientation approach to avoid collision.

We start with in-distribution objects (the same pineapple and bowl from demos) to specifically evaluate spatial generalization, since our few demos do not fully cover the entire distribution of object positions. Afterwards, we test with out-of-distribution objects to also assess semantic generalization: focusing on the "Pineapple to Bowl" task, we change the pineapple to an apple (color change) or strawberry (size change) and the bowl to a basket (appearance change) or cup (geometry change).

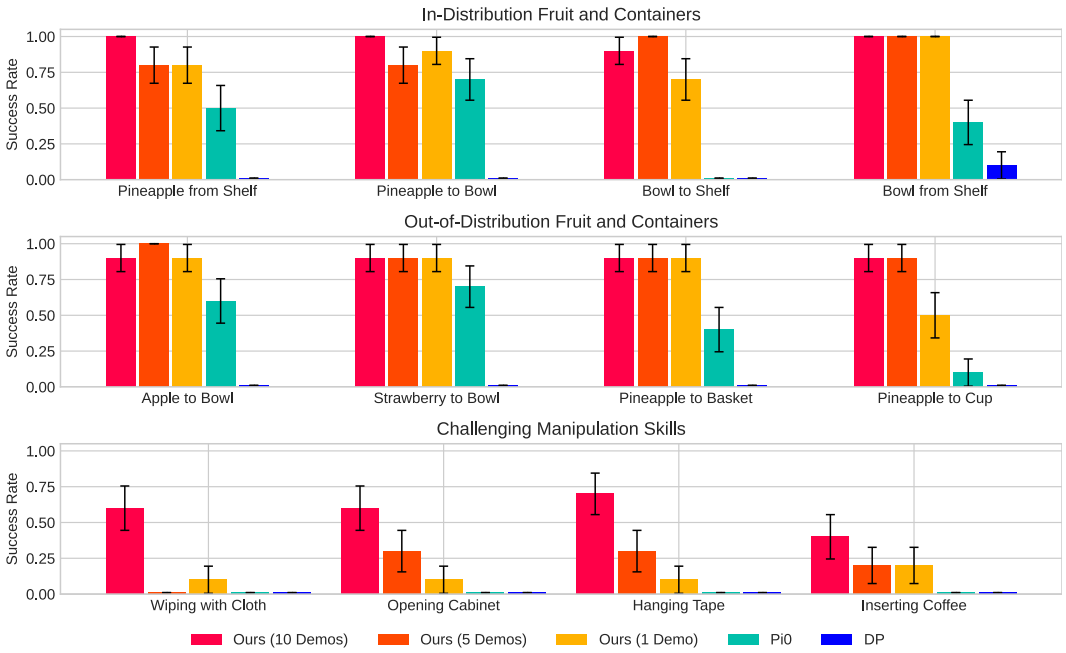

Figure 5: **Main Policy Comparison.** We compare the Tether policy with baselines across 12 tasks.

Next, we probe the limits of our policy design across 4 challenging tasks involving deformation, sustained contacts, articulation, and precision. First, the robot picks up a soft cloth from the table and maintains consistent contact to wipe marks off a whiteboard. Second, it grasps a cabinet doorknob only 0.5 centimeters thick (1/4 of gripper width) and applies stable contacts to fully open the cabinet's tight hinge. Third, it places a roll of tape on a small silver hook, 3 centimeters deep and visible in only a few pixels. Fourth, it inserts a K-cup pod into a coffee machine, which demands precision with an error margin of 8 millimeters.

We report success rates over 10 trials. For each trial, we randomize object positions and use the same two camera views that capture the entire scene (depicted in Figure 1). This naturally includes irrelevant objects and locations in the majority of the image, requiring methods to focus on task-relevant regions of the scene. The "Inserting Coffee" task is the sole exception: since the 8-millimeter margin of error projects onto the camera views as 2 to 3 pixels, we move the cameras closer to zoom in on the coffee pod and machine compartment.

4.2 Robust Imitation

Tether policy outperforms alternative imitation learning approaches. In Figure 5, we find that given few demos, our policy surpasses baselines across all tasks. Diffusion Policy, being an end-to-end model trained from scratch without built-in priors, fails to generalize from just 10 demos. Meanwhile, π_0 performs well on standard tabletop pick-and-place, likely benefiting from its pretraining on datasets containing similar behaviors, but fails with more complex tasks due to (1) incomplete command understanding (e.g., grasping the cloth but failing to locate the whiteboard) and (2) imprecise manipulation (e.g., approaching the coffee pod but missing the grasp).

Our few-shot learning baseline, KAT, did not achieve any successes on our tasks. KAT extracts visual tokens without considering task relevance, and the task-irrelevant features in our cluttered scene significantly outnumbers the task-relevant objects. We tried to fix this by manually annotating task-relevant scene regions, as well as running multiple language models (Gemini-2.5 Flash and GPT-4.1, and GPT-4 Turbo from the original work) and subsampling frequencies (between 3 and 15 Hz). We believe KAT's failures are due to the LLMs failing to handle complex multi-dimensional numerical patterns present in our tasks, caused by orientation changes and non-linear velocities.

Tether policy excels at both spatial and semantic generalization. First, our policy excels in tasks involving spatial robustness with in-distribution objects, including the bowl, which requires accurate orientation and position to avoid slipping. Second, our policy performs well even with

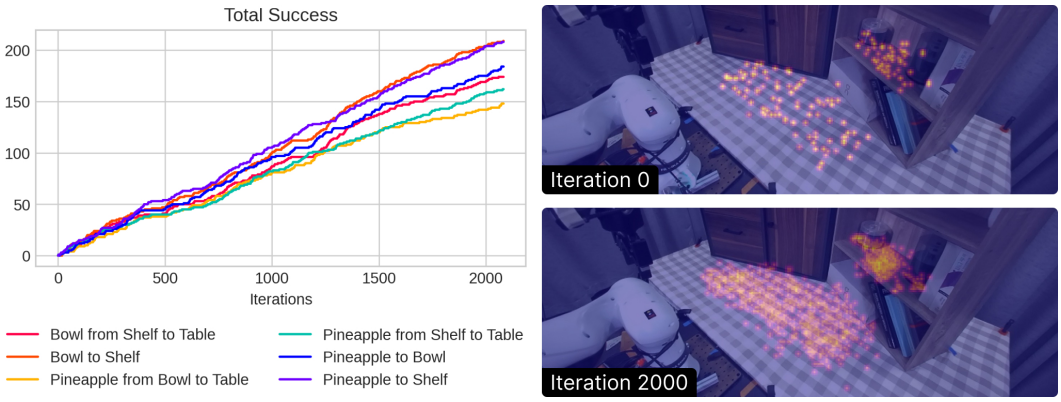

Figure 6: **Autonomous Play Statistics.** In around 26 hours of play, our method produces over 1000 diverse trajectories across 6 tasks.

out-of-distribution objects, attesting to the strong semantic generalization inherited from visual correspondence: without the demo object being present at test-time, correspondence finds the most semantically similar object and pinpoints the relative region of the keypoint (e.g., center of fruit or rim of container). This capability is significant especially for grasping the strawberry, which is vastly different in appearance and 1/4th the size of the demonstrated pineapple, and cup, which has 1/2 the diameter of the bowl and just big enough to fit the pineapple, thus requiring precision along with generalization. Visualizations are in Appendix.

Tether policy succeeds even on challenging manipulation tasks. Our policy is effective at more difficult manipulation tasks despite their challenges with deformation, sustained contacts, articulation, and precision. Successes with small features of the scene like the cabinet knob, hook, and coffee machine compartment demonstrate the high accuracy of our semantic correspondences. Most notably, our method achieves non-trivial success for the coffee insertion task without using the wrist camera; this task requires significant accuracy during grasping, since the deformable cylindrical pod crumbles if held just 1 centimeter off, and insertion, since the difference in pod and compartment diameter is a mere 8 millimeters. Additionally, fully opening the cabinet and wiping marks off the whiteboard show that trajectory warping can maintain consistent contacts with the scene, even without closed-loop adjustments. Finally, accurate grasps of the soft cloth demonstrate our method's flexibility with deformable objects, in contrast with prior work (Wen et al., 2022; Mandlekar et al., 2023; Zhu et al., 2024) that rely on rigid objects for pose estimation.

4.3 Autonomous Play

Given our policy's strong performance, we now deploy it on a subset of evaluation tasks for autonomous play. Specifically, we use the demos from Section 4.2 (10 per task) for moving the pineapple between the table, shelf, and bowl, as well as moving the bowl between the table and shelf. As described in Section 3.2, this task set is chosen so that there is a high certainty of at least one viable subsequent task that can be executed after previous tasks, even in their most common failure cases (e.g. object drops to the table).

Tether runs across multiple hours without resets. We run autonomous play for 4 sessions totaling around 26 hours. Success statistics for each task are shown in Figure 6 (left). Our policy generates 1085 successes from 1946 attempts across our 6 tasks, averaging around 1 success every 86 seconds and 1 attempt every 48 seconds, with a cumulative success rate of 55.8%. We intervene a total of 5 times (due to the bowl flipping), amounting to 0.26% of the attempts and an average of once every 5.2 hours. We provide a time-lapse in the Appendix.

We observe that these success rates are lower than those in Section 4.2 due to the uncontrolled nature of play. For instance, the bowl is sometimes tilted on its side due to imprecise placements or accidental pushes; these make grasps significantly harder, though our policy is still able to recover and fix the bowl. On rare occasions, the bowl is flipped completely upside-down after dropping from the shelf. With only one arm, this state is generally irrecoverable and requires interventions. However, on two separate instances, the robot accidentally recovers by squeezing it against the shelf and forcing

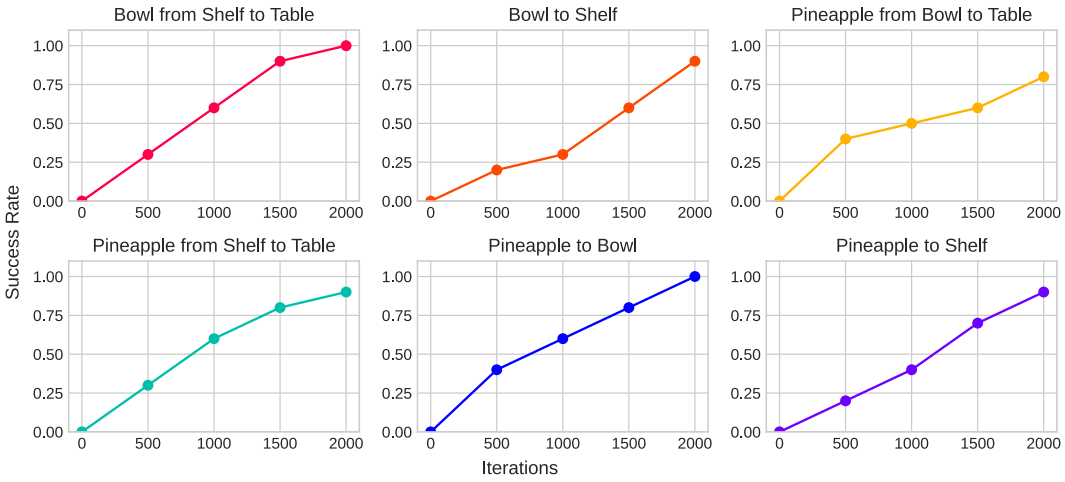

Figure 7: **Downstream Policy Learning Results.** The stream of data generated by autonomous play consistently improves policy performance over time.

it back upright. While such recoveries are not intended and occur purely by luck, they highlight the interesting nature of play: that at scale, coincidences may result in unexpected novel behaviors.

Tether produces diverse trajectories. In Figure 6 (right), we visualize the diversity of keypoints from demonstrations and successful play trajectories. We see that while our demos sparsely cover the table and shelf, using them to seed play allows us to not only interpolate between the demos but also expand on the edges of the distribution, such as the area around the cabinet.

Tether produces a stream of data that trains effective policies. Next, we train parametric policies on the generated data. While there are numerous algorithms for learning from suboptimal data, we adopt filtered behavioral cloning as a straightforward and effective approach. Integrating alternative methods that make fuller use of suboptimal trajectories remains a key direction for future work.

After every 500 iterations of play, we train Diffusion Policies on the cumulative successful trajectories for each task. We track their performance in Figure 7. Across all tasks, these policies progressively improve over time, with most eventually reaching near-perfect success rates. Thus, the data generated by Tether is consistently high-quality and effective for downstream policy learning. Diving deeper into the nature of this improvement, we see that as our ever-expanding datasets scale with more play, they increase in diversity and naturally build a stronger coverage of the object pose distribution. Thus, policies trained on this data improve primarily on their spatial robustness to different object positions: earlier policies succeed only when the object is in some specific locations, whereas our final policies are generally effective for any random object placement. Note that this robustness also extends to distractor objects that were involved in other tasks during play: for instance, policies that interact only with the bowl perform well irrespective of the pineapple position, and vice versa.

Finally, we compare our results with policies trained on an equivalent number of human-collected demos. For the "Pineapple from Shelf," "Pineapple to Bowl," and "Bowl to Shelf" tasks, these baselines achieve 80%, 100%, and 70% success rates, compared with 90%, 100%, and 90% from Figure 7. Thus, we confirm that Tether-generated data is competitive with human-collected data, while requiring minimal human effort and scaling simply with robot time.

5 Conclusion

We have presented Tether, a system for autonomous play with robust policies that scales data primarily with robot time. We introduce a novel policy design with semantic correspondence and trajectory warping that excels across a diverse set of tasks, and we deploy it within a VLM-guided multi-task play procedure that successfully produces over 1000 trajectories in 26 hours with minimal human intervention. This generated data, funneled downstream for filtered imitation learning, consistently improves the performance of neural policies, which ultimately reach near-perfect success rates. We believe that Tether demonstrates the potential for an alternative path in robot learning: one driven by scalable methods that perform and learn from autonomous interaction and experience.

REFERENCES

AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, Shu Jiang, Yuxin Jiang, Cheng Jing, Hongyang Li, Jialu Li, Chiming Liu, Yi Liu, Yuxiang Lu, Jianlan Luo, Ping Luo, Yao Mu, Yuehan Niu, Yixuan Pan, Jiangmiao Pang, Yu Qiao, Guanghui Ren, Cheng Ruan, Jiaqi Shan, Yongjian Shen, Chengshi Shi, Mingkang Shi, Modi Shi, Chonghao Sima, Jianheng Song, Huijie Wang, Wenhao Wang, Dafeng Wei, Chengen Xie, Guo Xu, Junchi Yan, Cunbiao Yang, Lei Yang, Shukai Yang, Maoqing Yao, Jia Zeng, Chi Zhang, Qinglin Zhang, Bin Zhao, Chengyue Zhao, Jiaqi Zhao, and Jianchao Zhu. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems, 2025. URL https://arxiv.org/abs/2503.06669.

Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision (ECCV)*, 2024.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024a. URL https://arxiv.org/abs/2410.24164.

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

David Blanco-Mulero, Oriol Barbany, Gokhan Alcan, Adrià Colomé, Carme Torras, and Ville Kyrki. Benchmarking the sim-to-real gap in cloth manipulation. *IEEE Robotics and Automation Letters*, 9(3):2981–2988, 2024. doi: 10.1109/LRA.2024.3360814.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

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

Kaylee Burns, Zach Witzel, Jubayer Ibn Hamid, Tianhe Yu, Chelsea Finn, and Karol Hausman. What makes pre-trained visual representations successful for robust manipulation? *arXiv preprint arXiv:2312.12444*, 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, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, pp. 02783649241273668, 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*, 2024.

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. URL https://arxiv.org/abs/2310.08864.

Norman Di Palo and Edward Johns. Keypoint action tokens enable in-context imitation learning in robotics. *arXiv preprint arXiv:2403.19578*, 2024.

Maximilian Du, Suraj Nair, Dorsa Sadigh, and Chelsea Finn. Behavior retrieval: Few-shot imitation learning by querying unlabeled datasets. *arXiv preprint arXiv:2304.08742*, 2023.

Jiafei Duan, Wilbert Pumacay, Nishanth Kumar, Yi Ru Wang, Shulin Tian, Wentao Yuan, Ranjay Krishna, Dieter Fox, Ajay Mandlekar, and Yijie Guo. Aha: A vision-language-model for detecting and reasoning over failures in robotic manipulation. *arXiv preprint arXiv:2410.00371*, 2024a.

Jiafei Duan, Wentao Yuan, Wilbert Pumacay, Yi Ru Wang, Kiana Ehsani, Dieter Fox, and Ranjay Krishna. Manipulate-anything: Automating real-world robots using vision-language models. In *8th Annual Conference on Robot Learning*, 2024b. URL https://openreview.net/forum?id=2SYFDG4WRA.

Benjamin Eysenbach, Shixiang Gu, Julian Ibarz, and Sergey Levine. Leave no trace: Learning to reset for safe and autonomous reinforcement learning. *arXiv preprint arXiv:1711.06782*, 2017.

Kuan Fang, Fangchen Liu, Pieter Abbeel, and Sergey Levine. Moka: Open-world robotic manipulation through mark-based visual prompting. *Robotics: Science and Systems (RSS)*, 2024.

Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. *arXiv preprint arXiv:2406.10454*, 2024.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International conference on algorithmic learning theory*, pp. 174–188. Springer, 2011.

Dylan Goetting, Himanshu Gaurav Singh, and Antonio Loquercio. End-to-end navigation with vlms: Transforming spatial reasoning into question-answering. In *Workshop on Language and Robot Learning: Language as an Interface*, 2024.

Ankit Goyal, Valts Blukis, Jie Xu, Yijie Guo, Yu-Wei Chao, and Dieter Fox. Rvt-2: Learning precise manipulation from few demonstrations. *arXiv preprint arXiv:2406.08545*, 2024.

Huy Ha, Pete Florence, and Shuran Song. Scaling up and distilling down: Language-guided robot skill acquisition, 2023. URL https://arxiv.org/abs/2307.14535.

Siddhant Haldar and Lerrel Pinto. Point policy: Unifying observations and actions with key points for robot manipulation. *arXiv preprint arXiv:2502.20391*, 2025.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054.

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. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

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.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Yuxuan Kuang, Junjie Ye, Haoran Geng, Jiageng Mao, Congyue Deng, Leonidas Guibas, He Wang, and Yue Wang. Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation. *arXiv preprint arXiv:2407.04689*, 2024.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pp. 71–91. Springer, 2024.

Mara Levy, Siddhant Haldar, Lerrel Pinto, and Abhinav Shirivastava. P3-po: Prescriptive point priors for visuo-spatial generalization of robot policies, 2024. URL https://arxiv.org/abs/2412.06784.

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.

Kevin Lin, Varun Ragunath, Andrew McAlinden, Aaditya Prasad, Jimmy Wu, Yuke Zhu, and Jeannette Bohg. Constraint-preserving data generation for visuomotor policy learning, 2025a. URL https://arxiv.org/abs/2508.03944.

Toru Lin, Kartik Sachdev, Linxi Fan, Jitendra Malik, and Yuke Zhu. Sim-to-real reinforcement learning for vision-based dexterous manipulation on humanoids. *arXiv preprint arXiv:2502.20396*, 2025b.

Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. Liv: Language-image representations and rewards for robotic control. In *International Conference on Machine Learning*, pp. 23301–23320. PMLR, 2023.

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

Marius Memmel, Jacob Berg, Bingqing Chen, Abhishek Gupta, and Jonathan Francis. Strap: Robot sub-trajectory retrieval for augmented policy learning. *arXiv preprint arXiv:2412.15182*, 2024.

Suvir Mirchandani, Suneel Belkhale, Joey Hejna, Evelyn Choi, Md Sazzad Islam, and Dorsa Sadigh. So you think you can scale up autonomous robot data collection? In *Conference on Robot Learning*, 2024.

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.

Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, Quan Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Heess, Chelsea Finn, Sergey Levine, and Brian Ichter. Pivot: Iterative visual prompting elicits actionable knowledge for vlms. 2024.

NVIDIA, :, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots, 2025. URL https://arxiv.org/abs/2503.14734.

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.

Jyothish Pari, Nur Muhammad Shafiullah, Sridhar Pandian Arunachalam, and Lerrel Pinto. The surprising effectiveness of representation learning for visual imitation. *arXiv preprint arXiv:2112.01511*, 2021.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models, 2025. URL https://arxiv.org/abs/2501.09747.

Jianing Qian, Yunshuang Li, Bernadette Bucher, and Dinesh Jayaraman. Task-oriented hierarchical object decomposition for visuomotor control. *arXiv preprint arXiv:2411.01284*, 2024.

Adam Rashid, Satvik Sharma, Chung Min Kim, Justin Kerr, Lawrence Chen, Angjoo Kanazawa, and Ken Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping, 2023. URL https://arxiv.org/abs/2309.07970.

Juntao Ren, Priya Sundaresan, Dorsa Sadigh, Sanjiban Choudhury, and Jeannette Bohg. Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning. *arXiv preprint arXiv:2501.06994*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.

Archit Sharma, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Autonomous reinforcement learning via subgoal curricula. *Advances in Neural Information Processing Systems*, 34:18474–18486, 2021.

Junyao Shi, Jianing Qian, Yecheng Jason Ma, and Dinesh Jayaraman. Composing pre-trained object-centric representations for robotics from" what" and" where" foundation models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15424–15432. IEEE, 2024.

Junyao Shi, Zhuolun Zhao, Tianyou Wang, Ian Pedroza, Amy Luo, Jie Wang, Jason Ma, and Dinesh Jayaraman. Zeromimic: Distilling robotic manipulation skills from web videos, 2025. URL https://arxiv.org/abs/2503.23877.

Mel Vecerik, Carl Doersch, Yi Yang, Todor Davchev, Yusuf Aytar, Guangyao Zhou, Raia Hadsell, Lourdes Agapito, and Jon Scholz. Robotap: Tracking arbitrary points for few-shot visual imitation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5397–5403. IEEE, 2024.

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, 2024. URL https://arxiv.org/abs/2310.01361.

Shengjie Wang, Jiacheng You, Yihang Hu, Jiongye Li, and Yang Gao. Skil: Semantic keypoint imitation learning for generalizable data-efficient manipulation. *arXiv preprint arXiv:2501.14400*, 2025.

Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. You only demonstrate once: Category-level manipulation from single visual demonstration, 2022. URL https://arxiv.org/abs/2201.12716.

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

Amber Xie, Rahul Chand, Dorsa Sadigh, and Joey Hejna. Data retrieval with importance weights for few-shot imitation learning. In *9th Annual Conference on Robot Learning*, 2025. URL https://arxiv.org/abs/2509.01657.

Alan Yu, Ge Yang, Ran Choi, Yajvan Ravan, John Leonard, and Phillip Isola. Learning visual parkour from generated images. In *8th Annual Conference on Robot Learning*, 2024.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3076–3085, June 2024.

Tony Z Zhao, Jonathan Tompson, Danny Driess, Pete Florence, Seyed Kamyar Seyed Ghasemipour, Chelsea Finn, and Ayzaan Wahid. Aloha unleashed: A simple recipe for robot dexterity. In *8th Annual Conference on Robot Learning*.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware, 2023. URL https://arxiv.org/abs/2304.13705.

Zhiyuan Zhou, Pranav Atreya, Abraham Lee, Homer Walke, Oier Mees, and Sergey Levine. Autonomous improvement of instruction following skills via foundation models, 2024. URL https://arxiv.org/abs/2407.20635.

Junzhe Zhu, Yuanchen Ju, Junyi Zhang, Muhan Wang, Zhecheng Yuan, Kaizhe Hu, and Huazhe Xu. Densematcher: Learning 3d semantic correspondence for category-level manipulation from a single demo. *International Conference on Learning Representations (ICLR) 2025*, 2024.