

# Learning Generalizable Tool-use Skills through Trajectory Generation

Carl Qi<sup>\*1</sup> Yilin Wu<sup>\*2</sup> Lifan Yu<sup>2</sup> Haoyue Liu<sup>2</sup> Bowen Jiang<sup>2</sup> Xingyu Lin<sup>\*\*3</sup> David Held<sup>\*\*2</sup>

**Abstract**—Autonomous systems that efficiently utilize tools can assist humans in completing many common tasks such as cooking and cleaning. However, current systems fall short of matching human-level of intelligence in terms of adapting to novel tools. Prior works based on affordance often make strong assumptions about the environments and cannot scale to more complex, contact-rich tasks. In this work, we tackle this challenge and explore how agents can learn to use previously unseen tools to manipulate deformable objects. We propose to learn a generative model of the tool-use trajectories as a sequence of tool point clouds, which generalizes to different tool shapes. Given any novel tool, we first generate a tool-use trajectory and then optimize the sequence of tool poses to align with the generated trajectory. We train a *single model* on four different challenging deformable object manipulation tasks, using demonstration data from only one tool per task. The model generalizes to various novel tools, significantly outperforming baselines. We further test our trained policy in the real world with unseen tools, where it achieves the performance comparable to human. Additional materials can be found on our project website.<sup>1</sup>

## I. INTRODUCTION

Building autonomous systems that leverage tools can greatly enhance efficiency and assist humans in completing many common tasks in everyday life [1], [2], [3], [4], [5], [6], [7]. As humans, we possess an innate ability to adapt quickly to use novel tools. However, replicating such adaptability in autonomous systems presents a significant challenge. To solve this task of novel tool manipulation, prior work has explored different representations for tools. A good tool representation should contain a rich visual understanding of the object and be useful for downstream physical interactions. Prior work [2] uses data-driven approaches to learn the latent representations for tools but such representation cannot generalize because of the lack of compositionality and interpretability. Another line of the work studies keypoints as a representation for tool which works only for rigid object manipulation including hammering, pushing and reaching.

In this work, we explore how agents can learn to use novel tools to manipulate deformable objects. Beyond the challenges of representing novel tools, manipulating deformable objects with tools adds considerable difficulties. For one, manipulating deformable objects often results in rich, continuous contact between the tool and the object; the contacts between a roller tool and dough, for example, are

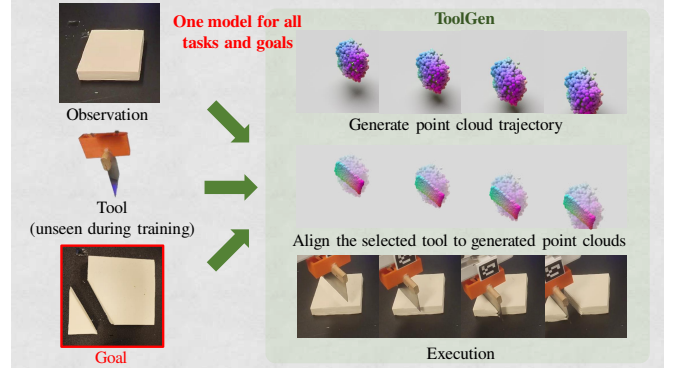


Fig. 1: Our method ToolGen can solve deformable object manipulation with diverse tasks and goals. It does so by first generating a point cloud trajectory of the desired tool and then aligning the actual tool to the generated point clouds for execution. We train a single model for four different challenging deformable object manipulation tasks. Our model is trained with demonstration data from just a single tool for each task and is able to generalize to various unseen tools.

continuous and cannot be easily discretized, which makes specifying discrete affordance labels to describe such interactions difficult. Further, defining rewards or keypoints (as is sometimes used for tool and environment representations [3], [4]) for deformable objects is also challenging. Therefore, operating novel tools to solve diverse tasks calls for an approach that makes few assumptions about the task and the environment. Our goal is to train a policy to solve various manipulation tasks with multiple tools, including tools that were not seen during training. We propose a novel approach, ToolGen, which learns tool-use skills via trajectory generation and sequential pose optimization. Given the scene, the goal, and a tool, ToolGen first generates a point cloud of a tool in the desired initial pose, and it subsequently predicts how this generated tool would move to perform the task. Finally, we sequentially align the actual tool to the generated tool to extract the actions for the agent to execute. Fig 1 offers an overview of our task setting and ToolGen’s outputs. We evaluate ToolGen against several baselines in deformable object manipulation with diverse tasks, goals, and tools. Impressively, with just a single model trained across all tasks and tools, ToolGen significantly outperforms the baselines and generalizes to many novel tools. Further, ToolGen achieves this despite being trained on demonstrations from just one tool for each task.

To summarize our contribution, we propose ToolGen,

<sup>1</sup>University of Texas at Austin, United States

<sup>2</sup>Carnegie Mellon University, United States

<sup>3</sup>University of California, Berkeley, United States

\* equal contribution

\*\* equal advising

<sup>1</sup><https://sites.google.com/view/toolgen>

which represents tool use via trajectory generation. We have shown that generating a point cloud trajectory of the tool can effectively capture the essence of tool use, i.e. how the tool should be placed in relation to the dough and how it should move over time, which allows us to generalize to a variety of unseen tools and goals. Furthermore, we transfer the policy to the real world without any finetuning to demonstrate our method’s effectiveness on three real world manipulation tasks with unseen novel tools and different goals.

## II. RELATED WORK

**Learning Generalizable Tool-use Skills:** Prior works have explored training robots to perform manipulation tasks with tools. To enable generalization, some approaches predict intermediate “affordances” and then generate actions based on these affordances [2], [8], [9], [10]. For example, affordances like grasping points or functional points and be represented as key points [2], [3], [4], [9], [10]. Similarly, concepts like contacts and forces [11], [12] can also be used. However, obtaining labels for these affordances can be difficult, and such affordance labels do not easily extend to deformable object manipulation, since the contacts with deformable objects (e.g. rolling a piece of dough) are complex and cannot be modeled by a few keypoints. Comparing to these methods, our method is capable of learning from unlabeled interaction data, as it implicitly learns affordances from the point clouds of the tool and the dough. This data-driven approach is similar to prior work [13], but we do not explicitly specify the structure of the shape embedding space, leaving more flexibility in tool shapes.

Another approach is to discover affordance regions in a self-supervised way by running parameterized motion primitives [2] or affordance-conditioned policies [3], [4] in simulation. In the image space, prior works have explored training an action-conditioned video prediction model [1] for planning actions for different tools. However, the video prediction model lacks 3D structure and has difficulty representing fine-grained action trajectories. Another research direction for generalizable tool use is to utilize the pretrained Large Language Models (LLMs) for long horizon reasoning. Prior work [14] designs four consecutive modules to prompt LLMs to directly generate code for robotic tool use. However, they make an assumption that we have state information of the tools and objects and directly include them into the prompt for LLMs. For deformable objects, state estimation is very challenging so this approach doesn’t generalize to deformable object manipulation with tool use.

**Deformable Object Manipulation with Tools:** Prior works with deformable objects often consider using a fixed set of tools. For example, Some approaches [5], [7] aim to solve the task of dough manipulation with a differentiable simulator but their tool sets are fixed with rolling pin and spatula. Other work [15], [16] use a fixed tool set of knives for cutting. These works do not consider generalization to novel tools, which is the focus of this work.

## III. PROBLEM STATEMENT AND ASSUMPTIONS

Consider a set of point clouds  $(P^o, P^g, P^{tool})$ , where  $P^o$  represents the initial observation of the scene,  $P^g$  stands for the goal, and  $P^{tool}$  for a tool to use for execution. Our task is to predict an actions sequence of horizon  $H$ , where the tool transforms the initial pose into a predicted target pose. The actions is represented by a transformations sequence  $T_{0:H}$ . Here, all the point cloud positions as well as the objects’ orientations are relative to a reference frame located at the dough center. This design allows us to perform manipulation that is agnostic to the location of the dough on the table.

In the training stage, we use demonstrations of tools from a training set  $\{P^{train\ tool_i}\}_{i=1:K_{train}}$ , where  $K_{train}$  is the number of training tools we have. The demonstration data fed into the model are of the form:  $(P^o, P^g, P^{train\ tool_i}, T_{0:H})$ . The initial transformation  $T_0$  in the sequence brings the tool to a “reset pose”. The remaining terms  $T_{1:H}$  are the relative transformations from the previous timestep, which we call “delta poses.” For each task, we manually specify distributions of the initial and goal configurations. We then run trajectory optimization using a differentiable simulator to generate these demonstrations following prior works [17]. Human demonstrations could serve as an alternative source of the training data described above.

## IV. METHOD

We propose the following approach to obtain an trajectory executable for robots with any given tools:

- We first generate a point cloud of a reconstructed tool  $P^{gen}$  at a starting pose based on the given tool  $P^{tool}$  (Sec. IV-A).
- Next we generate a sequence of tool actions of how this generated tool would achieve the task (Sec. IV-A) based on policy learned with Behavior Cloning.
- We then align the actual tool to each of the point clouds in the generated trajectory (Sec. IV-B).

Below, we describe this approach in detail, and experiments in Sec. V demonstrate the remarkable improvements of this approach compared to other approaches.

### A. Representing tool-use through point cloud trajectory generation

In this section, we describe our approach for trajectory generation. A straightforward method for trajectory generation would be to directly predict the motion of the tool. However, directly regressing into the tool’s pose, particularly the orientation, is proved to be challenging, as indicated by prior studies [18], [19], [20]. To alleviate this challenge, we employ a generative module  $G_{traj}$  to produce a point cloud trajectory  $P_{0:H}^{gen}$  to complete the task with reconstructed tool. Our trajectory generation model consists of two parts. In the first part, a initial point cloud generator  $G_{reset}$  is utilized to reconstruct a tool point cloud at “reset pose”. In the second part, a path generator  $G_{path}$  is adopted for producing trajectory of  $P_{0:H}^{gen}$  based on the reconstructed tool. This generated trajectory will later be used to determine the actions of the actual tool.

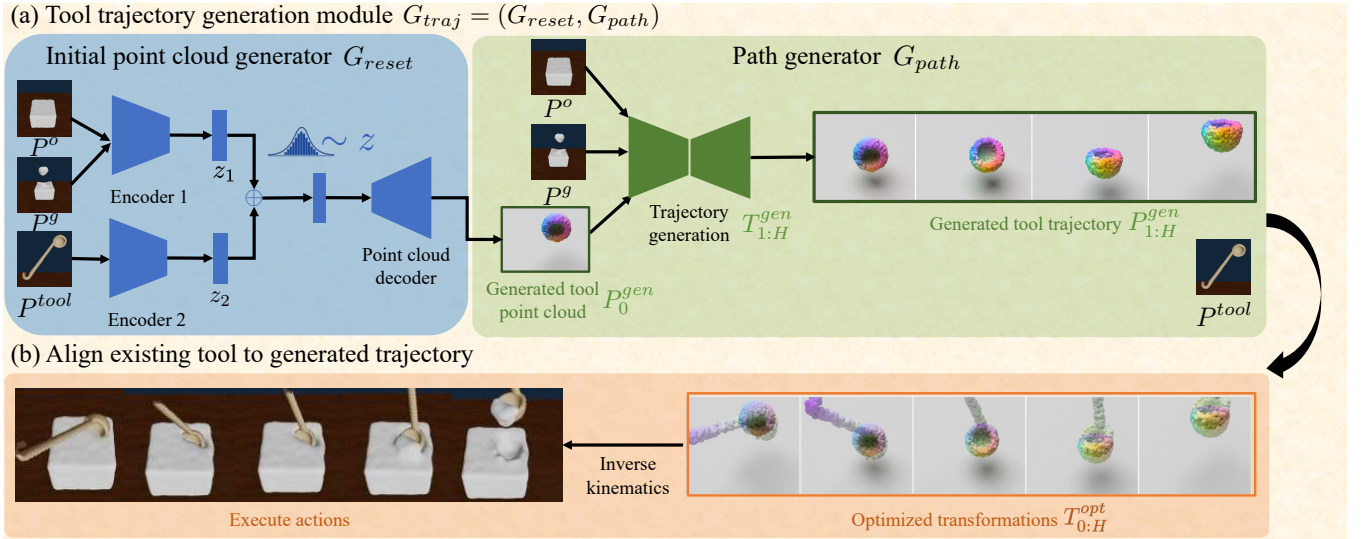


Fig. 2: Overview of our method: (a) Given an initial observation of the scene  $P^o$ , the goal  $P^g$ , and a tool  $P^{tool}$ , we first leverage the trajectory generation module  $G_{traj}$  to generate an ideal tool trajectory accomplishing the task  $P_{0:H}^{gen}$ . It encompasses two submodules: Initial point cloud generator  $G_{reset}$  generating reset pose  $P_0^{gen}$  of reconstructed tool and Path generator  $G_{path}$  generating  $P_{1:H}^{gen}$  (b) We then align the existing tool with the reconstructed tool via sequential pose optimization to extract the pose of the existing tool  $T_{0:H}^{opt}$ , and we subsequently use inverse kinematics to obtain the actions for the agent to execute.

$G_{reset}$  is a PointFlow-based [21] encode-decoder generation model. It conditions on the point cloud of the existing tool  $P^{tool}$ , the initial scene observation  $P^o$ , and the goal  $P^g$ , to reconstruct the tool at “reset pose”,  $P_0^{gen}$ . The architecture of our PointFlow-based [21] generator  $G_{reset}$  is shown in Fig. 2 (a) (top). It encodes the tool points and the concatenation of initial and target dough points to two sets of latent features with separate PointNet++ [22] encoders. These latent features are concatenated and inputted through an MLP to produce an estimation of Gaussian distribution. We then take a sample from this estimated distribution as the input of a PointFlow [21] decoder, which outputs the reconstruction point cloud of the given tool at the reset pose  $P_0^{gen}$ .

The second part  $G_{path}$  works on predicting a sequence of transformations of how this generated tool would move to achieve the task. The architecture of the path generator is shown in Fig. 2 (a) (top right). We follow the design in ToolFlowNet [23] to train a policy model through Behavior Cloning, which optimizes a combined loss of point content loss and consistency loss. The  $P_0^{gen}$  from  $G_{reset}$  is concatenated together with the initial scene observation  $P^o$ , and the goal state  $P^g$  and passed into the model. Transformations of  $H - 1$  time-steps,  $T_{1:H}^{gen}$ , are generated for the tool. Details for  $G_{path}$  can be found in Appendix B on the website.

Together, our generative module  $G_{traj} = (G_{reset}, G_{path})$  predicts a trajectory of point clouds  $P_{0:H}^{gen}$ , which shows the movement of a reconstructed tool accomplishing the manipulation task. Training details are described in Section IV-C.

### B. Execution via sequential pose optimization

In Section IV-A, the generated point cloud trajectory of the tool  $P_{0:H}^{gen}$  is built upon the reconstructed tool and is not guaranteed to be executable for the actual tool. In this

section, we describe the optimization procedure for aligning the actual tool with the generated tool reconstruction in order to extract reasonable actions for actual execution (visualized in Fig. 2 (b) and listed in detail in Algorithm 1).

The initial transformation at time-step 0 exerts a decisive influence on the overall trajectory. We therefore subdivide the optimized transformations  $T_{0:H}^{opt}$  into the reset transformation  $T_0^{opt}$  and delta pose optimization  $T_{1:H}^{opt}$ . To align the actual tool  $P^{tool}$  to the reconstructed tool in the first timestep  $P_0^{gen}$ , we consider the following terms: 1) the similarity between the predicted reset pose and actual tool pose, 2) the collision between tool and the initial scene observation. The loss function is given by:

$$J_{reset}(T) = \text{Chamfer}(T \circ P^{tool}, P_0^{gen}) - \lambda_c \cdot \text{Chamfer}(T \circ P^{tool}, P^o), \quad (1)$$

The first term is the Chamfer distance between the actual tool  $P^{tool}$  transformed by  $T$  and the reconstructed tool  $P_0^{gen}$  at reset pose. The second term is a penalty term computed as the Chamfer distance between the existing tool  $P^{tool}$  transformed by  $T$  and the observation of the dough  $P^o$ .  $\lambda_c$  is a hyper-parameter balancing the two terms. The aim of the penalty term is to avoid undesirable collisions between the tool in reset pose and the environment, while collisions will be allowed for subsequent time-steps.

For optimization, we use Projected Gradient Descent, detailed in Sec. IV-C, for different initializations of  $T$  and learn to start from the one that minimizes the objective described in Eq. 1.

Next we work on the optimization of the delta poses  $T_{1:H}^{opt}$ . Similar to that for reset pose, we evaluate the distance between the actual tool  $P^{tool}$  and the reconstructed tool

**Algorithm 1** Sequential pose optimization

- 1: **Input:** The current observation of the dough  $P^o$ , the existing tool  $P^{tool}$ , and the point cloud trajectory for the generated tool  $P_{0:H}^{gen}$
- 2: **// Optimize for the reset transformation**
- 3: Initialize random transformations  $T_0^1, \dots, T_0^N$  in  $SE(3)$ ;
- 4: Optimize  $T_0^1, \dots, T_0^N$  according to Eq. 1 to obtain costs  $J_{reset}^1 \dots J_{reset}^N$ ;
- 5: Choose the transformation that minimizes the costs, denoted as  $T_0^{opt}$ ;
- 6: **// Optimize for delta poses**
- 7: Initialize the delta poses as identities, i.e.,  $T_{1:H} = I$ ;
- 8: Optimize the delta poses according to Eq. 2 and obtain the final transformations  $T_{1:H}^{opt}$ ;
- 9: **Output:** Optimized transformations for the existing tool:  $T_{0:H}^{opt}$

at each time-step  $P_t^{gen}$ , with an additional penalty term to encourage small motions. The loss function for the delta poses is given by:

$$J_\delta(T_{1:H}) = \sum_{t=1:H} \text{Chamfer}(T_t \circ X_{t-1} \circ P^{tool}, P_t^{gen}) + \lambda_r \cdot \|T_t\|$$

where  $X_{t-1} = T_{t-1} \circ T_{t-2} \circ \dots \circ T_0^{opt}$  (2)

The first term is the Chamfer distance between the reconstructed tool points  $P^{tool}$  transformed by  $T_t \circ X_{t-1}$  and the generated tool points  $P_t^{gen}$  at time-step  $t$ ,  $\|\cdot\|$  is a regularization function to moderate the magnitude of the translation and rotation defined by the delta poses (see Sec. IV-C for details).  $\lambda_r$  is a hyper-parameter balancing the two terms.

Finally, we apply these objectives in an optimization routine, as outlined in Algorithm 1, to align the reconstructed tool with the generated one and produce the final trajectory  $T_{0:H}^{opt}$  for the reconstructed tool. Subsequently, we can utilize inverse kinematics to determine the required actions for our agent to execute the task. In our case, these actions comprise the translation and angular velocities of the tool.

### C. Implementation details

Before inputting the tool, the dough, and the goal point clouds into the PointNet++ networks, we use a one-hot encoding to differentiate points that belong to different objects. Therefore, the input features per point will be  $[x, y, z, \text{one-hot}]$ .

The two modules of trajectory generation,  $G_{reset}$  and  $G_{path}$ , are trained separately.  $G_{reset}$  learns by optimizing the evidence lower bound (ELBO) given the training tools and their reset poses  $T_0 \circ P^{train tool_i}$  from the demonstration dataset described in Sec. III. The trajectories of the training tools  $T_{1:H}$  from the demonstration dataset described in Sec. III are then used as labels for the training  $G_{path}$ .

We train a single set of modules ( $G_{reset}, G_{path}$ ) across a compact demonstration dataset comprised of multiple tasks rather than training separate networks for each task. To achieve this, we introduce a scoring module  $D_{score}$  to evaluate and select tool for each task. Details of  $D_{score}$

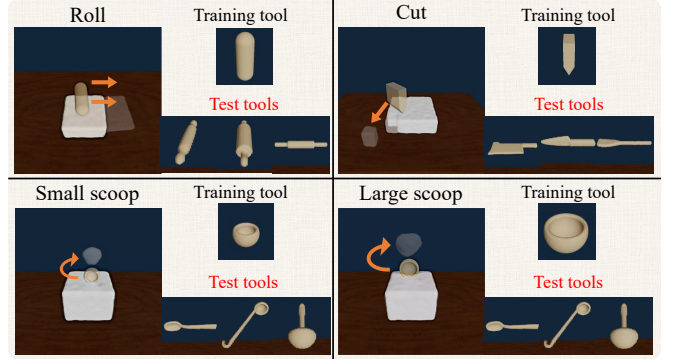


Fig. 3: We consider 4 tasks: Roll, Cut, Small scoop, and Large scoop. On the left side of each task, we illustrate how the training tool is used to achieve the goal, overlaying the goal on the initial observation. On the right side, we visualize the initial configurations of the training tool and test tools for each task, highlighting the ability of our method to generalize to novel tools.

module is shown in Appendix A on the website. In training data, we collect 200 demonstration trajectories for each task performed with just one training tool. Despite the limited training data, our model is demonstrated to be capable of generalizing to various unseen tools in both simulation and real world. See Appendix C on the website for more information on our demonstration dataset.

In trajectory optimization, we use the quaternion representation for the orientation of the transformation, and project the values onto a unit ball after each gradient update. Here, we use a step size of  $10^{-2}$ , and  $\lambda_c = 0.1$ . For optimizing the delta poses, we use the 3-DoF Euler angles representation with a step size of  $10^{-3}$ , a regularization factor of  $\lambda_r = 0.1$ , and we use the euclidean norm to regularize the translation as well as the rotation. We use a greedy IK solver [24] to obtain the robot actions from the simulation and the real world, and we find it to work well in our tasks. One could also use IK as the main objective in the sequential pose optimization step to produce better poses for IK to solve.

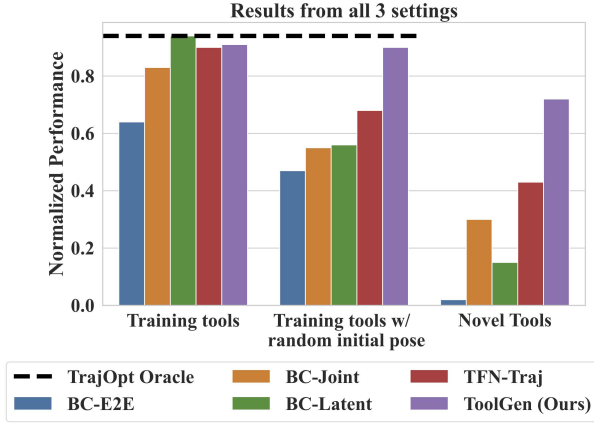
## V. RESULTS

As shown below, we demonstrate that ToolGen is able to perform well on a variety of manipulation tasks with novel tools using just a *single* model trained across multiple tasks and tools. Notably, we train with demonstrations from only one training tool per task and we test on several unseen tools, demonstrating our method’s generalization abilities. We additionally evaluate ToolGen on real world observations and use a Franka Panda robot to execute the predicted trajectory. For real world experiments, we include both the qualitative results and quantitative results to highlight our policy’s effectiveness when transferred to the real world.

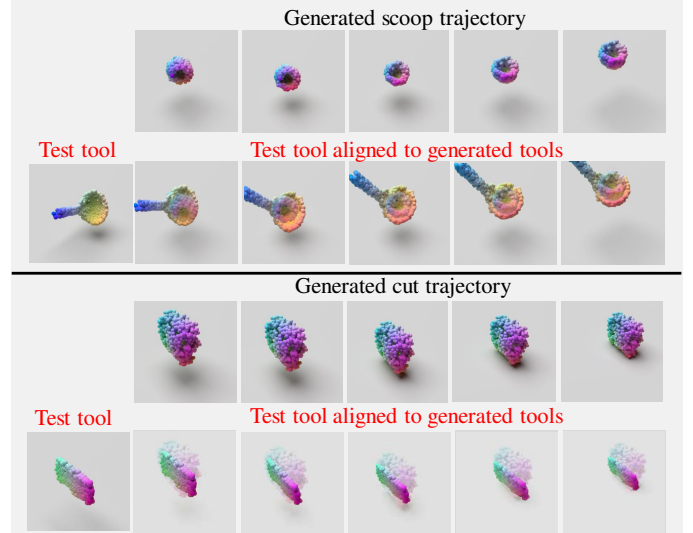
### A. Tasks and baselines

**Tasks:** We evaluate our method against several baselines in a soft body simulator, PlasticineLab [17]. We consider four tasks: “Roll”, “Cut”, “Small scoop” and “Large scoop”. Example configurations and their training and test tools for





(a) Performance of all the methods across 3 different settings. We evaluate 10 trajectories per task per tool and then aggregate the performance across all the tasks.



(b) Examples of generated tool trajectories for scoop (top) and cut (bottom), as well as the trajectories of the test tools aligned to these generated trajectories.

Fig. 4: Fig. 4a: Performance of all the methods across 3 settings. Fig. 4b: Examples of generated tool trajectories and test tool alignments.

these tasks are depicted in Fig. 3. In our setup, all of the tools are placed far from the dough at the start of each task, as would be the case in a normal tool-use scenario.

**Metric:** We specify goals as 3D point clouds of different geometric shapes. We report the normalized decrease in the Chamfer Distance between the observation and the goal, computed as  $s(t) = \frac{s_0 - s_H}{s_0}$ , where  $s_0, s_H$  are the initial and final Chamfer Distances to the goal respectively. To compute the performance of each method, we evaluate 10 trajectories per task per tool and then aggregate the performance across all the tasks.

**Baselines:** We evaluate the following baselines with different action representations. All of the baselines regress to reset transformations and delta poses, except for BC-E2E which predicts delta poses directly from the initial configuration without a reset transformation. Details on the architectures of the baselines are described in Appendix D on the website.

- **TrajOpt Oracle.** Differentiable trajectory optimization with ground truth dynamics from the simulator.
- **BC-E2E.** End-to-end behavioral cloning that outputs a  $H' \times 6, (H' > H)$  vector representing the delta poses of the tool relative to the initial tool pose. Unlike the other baselines, this baseline does not output a reset transformation.
- **BC-Joint.** Behavioral cloning that jointly regresses to the reset transformation and subsequent delta poses from the initial tool configuration.
- **BC-Latent.** Behavioral cloning that regresses to the reset transformation, moves the tool to the predicted reset pose, and then predict subsequent delta poses from a latent encoding of the scene with the tool in the reset pose.
- **TFN-Traj.** Behavioral cloning that regresses to the reset transformation, moves the tool to the predicted

reset pose, and then uses the updated scene to predict subsequent delta poses with the ToolFlowNet-based [23] trajectory model described in Appendix B on the website.

We examine three settings, each presenting a greater level of difficulty, detailed in Sec. V-B, Sec. V-C, and Sec. V-D, respectively. We demonstrate that ToolGen is robust to these generalization challenges and maintains superior performance over the baselines. We additionally conduct ablation studies by removing the path generator of ToolGen, detailed in Appendix E on the website.

### B. Leveraging training tools at test time

We first test the methods on a set of held out configurations using training tools. To successfully perform the manipulation, the methods need to output the appropriate poses for the training tools to complete the tasks. Fig. 4a shows the performance of all the methods. We see that most methods achieve reasonable performance. This shows that all these methods generalize reasonably well to different goal configurations given the same training tools. In contrast, BC-E2E achieves suboptimal performance on even this simple version of the task, showing the limitations of methods that do not predict a reset transformation.

### C. Generalization to unseen initial tool poses

To simulate the fact that a tool might be in any initial configuration in the real world, we randomize the initial poses of the training tools in  $SE(3)$  and rerun evaluations. From Fig. 4a, we observe that ToolGen is the only method that is robust to this perturbation. Despite the fact that the baselines are trained with the same tools, they fail to generalize to unseen initial poses of the tool. On the other

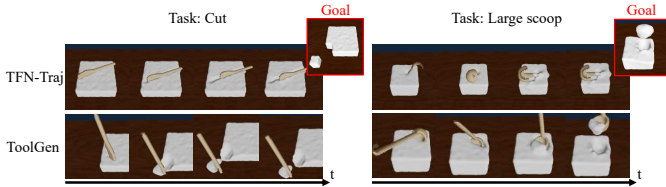


Fig. 5: Example rollouts of ToolGen (ours) compared to the baseline TFN-Traj. The goal configuration of each task is shown on the top right. ToolGen can effectively use the new tool while the baseline struggles.

Method	Roll	Cut	Small scoop	Large Scoop	Average
BC-E2E	$0.49 \pm 0.30$	$-0.22 \pm 0.44$	$-0.27 \pm 0.20$	$0.07 \pm 0.08$	$0.02 \pm 0.26$
BC-Joint	$0.64 \pm 0.26$	$0.00 \pm 0.09$	$-0.05 \pm 0.10$	$-0.01 \pm 0.07$	$0.30 \pm 0.41$
BC-Latent	$0.70 \pm 0.15$	$0.37 \pm 0.10$	$-0.15 \pm 0.30$	$0.34 \pm 0.41$	$0.15 \pm 0.33$
TFN-Traj	$0.70 \pm 0.19$	$0.29 \pm 0.19$	$0.40 \pm 0.44$	$0.35 \pm 0.40$	$0.43 \pm 0.36$
ToolGen (Ours)	<b><math>0.75 \pm 0.15</math></b>	<b><math>0.82 \pm 0.08</math></b>	<b><math>0.50 \pm 0.40</math></b>	<b><math>0.80 \pm 0.19</math></b>	<b><math>0.72 \pm 0.27</math></b>

TABLE I: Quantitative performance for different methods when using novel tools. Each value in the table represents the normalized decrease of Chamfer Distance for a specific task, measured across the use of 3 novel tools in 10 different goal configurations. The final column denotes the average performance of each method across all tasks.

hand, ToolGen is robust to the initial configuration of the tool and receives no performance loss.

#### D. Generalization to unseen tools

Finally, we evaluate the methods on a far more challenging scenario, in which our agents are given unseen tools. We evaluate each novel tool on 10 held out goals for each task and average their performances. See Fig. 3 for a visualization of the novel tools we consider. Since the novel tools are also in arbitrary initial poses, this scenario requires the method to be robust to tool shapes as well as initial poses of the tool. Fig. 4a and Table I shows the quantitative results of all the methods, and Fig. 5 show examples of rollouts by ToolGen (ours) and the baseline TFN-Traj. All of the baselines fail to obtain a high performance, especially in the more challenging task of scooping (see Table I). In contrast, ToolGen can leverage completely unseen tools in meaningful ways. This is because ToolGen leverages trajectory generation to alleviate the issues of distribution shift. It further uses a non-learned optimization procedure (gradient descent with multiple random initializations), which also does not suffer from a distribution shift. For more analysis, please see our Appendix E on the website.

We show examples of the tools generated by ToolGen (top row) as well as the test tools aligned to these generated tools (bottom row) in Fig. 4b. Overall, ToolGen achieves superior performance over the baselines in this challenging scenario of using novel tools. Remarkably, we train just a single ToolGen model across all tasks and tools, using merely one training tool per task. Despite this, ToolGen demonstrates the capacity to solve all tasks effectively when presented with novel tools.

#### E. Inference on real world observations

For our real world experiments, we select three representative tasks, *Cut*, *Roll* and *Scoop(Large)* to test our trained policy. In each task, we select two real world tools and attach each tool to a mount so that it can stay on the tool hanger for the robot to pick up. Details of our environment setup can be seen in Figure 7 and Appendix I on the website. Our ToolGen model is trained entirely with simulation data. To demonstrate the robustness of ToolGen, we record point clouds of tool and dough from the real world and use ToolGen to predict the trajectory of the real world tool. To obtain the point clouds from the real world, we use three Azure Kinect cameras to record the initial dough and the tool point clouds. We then manually manipulate the dough to a desired shape and record the final point cloud as the goal point cloud. We record the point cloud of the real dough at its initial and goal states and concatenate them with the tool point cloud. The initial pose of the tool is entailed in the point cloud input. The model then output a trajectory of horizon  $H = 50$ . In the execution stage, we use a mold to restore the dough to the recorded initial state. The robot picks up the tool and move to the recorded initial pose and executes the trajectory, where the tool first transforms from the initial pose to reset pose, and then moves through to execute the produced trajectory on the real dough. Fig. 6 includes qualitative results of the robot executing the policy’s rollouts. For each of the three different tasks, we show an example of the robot using one of the test tools to reshape the dough. Quantitatively, despite the gap in point cloud observations between sim and real, our method can effectively generalize to unseen tool in the real world with an average normalized decrease of Chamfer Distance of 0.77 as shown in Table II. This metric shows how close the final state gets to the goal compared to the initial state of the dough. The larger normalized decrease of Chamfer Distance indicates better performance. Compared to the baseline method *BC-Latent*, our method *ToolGen* outperforms it on all three tasks by a large margin. *BC-Latent* fails to generalize to some unseen real world tools and generates transformations that oscillate without further movements for those tools. In contrast, our method successfully generalizes to all tasks with different goals and tools. Hence, our method demonstrates smaller performance variance. To prove that our policy is comparable to human performance, two volunteers are asked to perform the same manipulation tasks as that for robots. From the table we can see that our policy’s task performance is very close to humans with the largest difference of 0.08 in normalized decrease of Chamfer Distance. We also notice that *Scoop* generally has worst performance compared to other two tasks because the real world dough we are using is so sticky that both the human and our trained policy struggle with detaching the scooped piece from the whole dough.

## VI. CONCLUSION AND LIMITATIONS

In this paper, we introduce ToolGen, a novel framework for learning generalizable tool-use skills. ToolGen uses a point cloud trajectory generation approach to represent tool

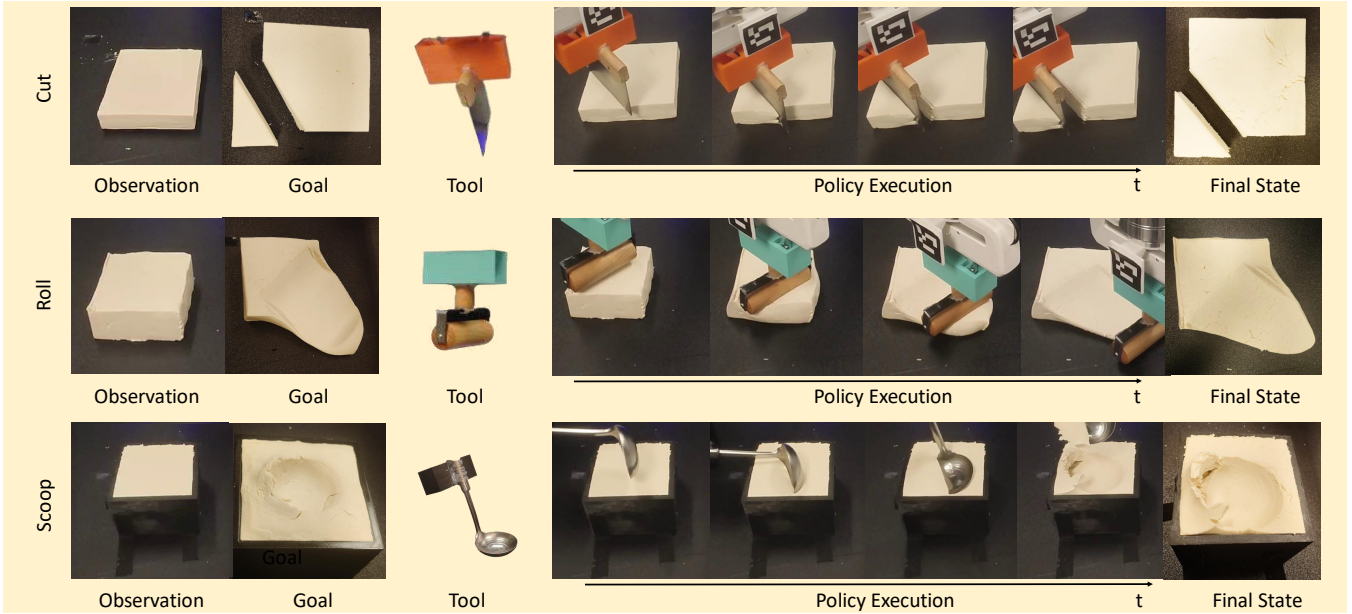


Fig. 6: Results of ToolGen on real world observations for Cut (top) and Roll (middle) and Scoop (bottom). For each task, we visualize initial dough observation, the goal, the real world tool, the policy rollout of our trained policy and the final state of the policy. As a result, ToolGen can effectively generate manipulation trajectories from real world observations even though the model is trained entirely in simulation.

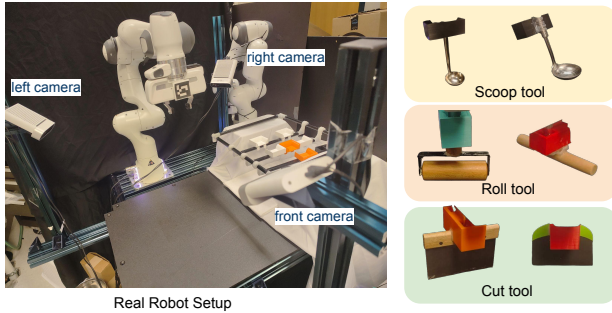


Fig. 7: On the left, it is the real world experiment setup with three cameras, one Franka Panda robot and a tool hanger. On the right, they are six tools we use for three different tasks.

use and then applies sequential pose optimization for execution. This representation circumvents the issues associated with using affordances to represent tool use, and it demonstrates superior generalization capabilities, especially when evaluating on unseen test tools, given only one tool per task for training. We applied a single ToolGen model to the manipulation of deformable objects, tackling diverse tasks, goals, and tools, and we found that ToolGen significantly outperforms the baselines and generalizes effectively to many novel tools. It is our hope that ToolGen will inspire more innovative approaches for tool use representation that enable broad ranges of generalization in the future.

**Limitations:** Our method has several limitations: First, our method’s execution time is considerably longer compared to that of a trained policy, due to the time needed for generating point clouds and optimizing the current tool’s poses. Quantitative results are shown in Appendix F on the website. We anticipate that the use of faster techniques for

Method	Roll $\uparrow$	Cut $\uparrow$	Scoop $\uparrow$	Average $\uparrow$
BC-Latent	$0.73 \pm 0.21$	$0.60 \pm 0.46$	$0.55 \pm 0.12$	$0.57 \pm 0.27$
ToolGen (Ours)	<b><math>0.83 \pm 0.09</math></b>	<b><math>0.86 \pm 0.16</math></b>	<b><math>0.63 \pm 0.14</math></b>	<b><math>0.77 \pm 0.16</math></b>
Human (Oracle)	$0.91 \pm 0.03$	$0.90 \pm 0.11$	$0.69 \pm 0.14$	$0.83 \pm 0.14$

TABLE II: Quantitative results for different methods when using real world novel tools. *Human Oracle* is not an automated method and serves as an upper bound for the performance of the dough manipulation tasks. Each value in the table represents the average normalized decrease of the Chamfer Distance and the standard deviation for a specific task, measured across 2 different goal configurations. Each goal configuration is tested with two different initial pose of tools. The final column denotes the average performance of each method across all tasks. The metric is computed the same way as in Table I

sequential pose optimization, such as second-order methods, could speed up our method. Secondly, as our point cloud generator is trained on limited tools, it is sometimes unable to generate accurate point clouds for novel tools and thus the alignment process could fail. A promising direction is to train on more variations of the tool to improve the generation process and make alignment easier. Further details on these failure cases are shown in Appendix G on the website.

## VII. ACKNOWLEDGEMENT

This work was supported by the National Science Foundation under Grant No. IIS-2046491, and the National Institute of Standards and Technology under Grant No. 70NANB23H178. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or the National Institute of Standards and Technology.

## REFERENCES

- [1] A. Xie, F. Ebert, S. Levine, and C. Finn, “Improvisation through physical understanding: Using novel objects as tools with visual foresight,” *Robotics: Science and Systems (RSS)*, 2019.
- [2] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, and S. Savarese, “Learning task-oriented grasping for tool manipulation from simulated self-supervision,” *The International Journal of Robotics Research (IJRR)*, 2020.
- [3] Z. Qin, K. Fang, Y. Zhu, L. Fei-Fei, and S. Savarese, “Keto: Learning keypoint representations for tool manipulation,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.
- [4] D. Turpin, L. Wang, S. Tsogkas, S. Dickinson, and A. Garg, “Gift: Generalizable interaction-aware functional tool affordances without labels,” in *Robotics: Science and Systems (RSS)*, 2021.
- [5] X. Lin, Z. Huang, Y. Li, J. B. Tenenbaum, D. Held, and C. Gan, “Diffskill: Skill abstraction from differentiable physics for deformable object manipulations with tools,” *International Conference on Learning Representations (ICLR)*, 2022.
- [6] C. Qi, X. Lin, and D. Held, “Learning closed-loop dough manipulation using a differentiable reset module,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9857–9864, 2022.
- [7] X. Lin, C. Qi, Y. Zhang, Z. Huang, K. Fragkiadaki, Y. Li, C. Gan, and D. Held, “Planning with spatial-temporal abstraction from point clouds for deformable object manipulation,” in *Conference on Robot Learning (CoRL)*, 2022.
- [8] Y. Zhu, Y. Zhao, and S. Chun Zhu, “Understanding tools: Task-oriented object modeling, learning and recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [9] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, “kPAM:KeyPoint affordances for Category-Level robotic manipulation,” *International Symposium of Robotics Research (ISRR)*, 2019.
- [10] W. Gao and R. Tedrake, “kpam 2.0: Feedback control for category-level robotic manipulation,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2962–2969, 2021.
- [11] Z. Zhang, Z. Jiao, W. Wang, Y. Zhu, S.-C. Zhu, and H. Liu, “Understanding physical effects for effective tool-use,” *IEEE Robotics and Automation Letters (R-AL)*, 2022.
- [12] Y. Wi, A. Zeng, P. Florence, and N. Fazeli, “Virdo++: Real-world, visuo-tactile dynamics and perception of deformable objects,” *arXiv preprint arXiv:2210.03701*, 2022.
- [13] S. Thompson, L. P. Kaelbling, and T. Lozano-Perez, “Shape-based transfer of generic skills,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5996–6002.
- [14] M. Xu, P. Huang, W. Yu, S. Liu, X. Zhang, Y. Niu, T. Zhang, F. Xia, J. Tan, and D. Zhao, “Creative robot tool use with large language models,” 2023.
- [15] E. Heiden, M. Macklin, Y. Narang, D. Fox, A. Garg, and F. Ramos, “Disect: A differentiable simulation engine for autonomous robotic cutting,” *arXiv preprint arXiv:2105.12244*, 2021.
- [16] Z. Xu, Z. Xian, X. Lin, C. Chi, Z. Huang, C. Gan, and S. Song, “Roboninja: Learning an adaptive cutting policy for multi-material objects,” *Robotics: Science and Systems (RSS)*, 2023.
- [17] Z. Huang, Y. Hu, T. Du, S. Zhou, H. Su, J. B. Tenenbaum, and C. Gan, “Plasticinellab: A soft-body manipulation benchmark with differentiable physics,” in *International Conference on Learning Representations*, 2021.
- [18] V. Peretroukhin, M. Giamou, D. M. Rosen, W. N. Greene, N. Roy, and J. Kelly, “A smooth representation of belief over so (3) for deep rotation learning with uncertainty,” *arXiv preprint arXiv:2006.01031*, 2020.
- [19] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, “On the continuity of rotation representations in neural networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5745–5753.
- [20] J. Chen, Y. Yin, T. Birdal, B. Chen, L. J. Guibas, and H. Wang, “Projective manifold gradient layer for deep rotation regression,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6646–6655.
- [21] G. Yang, X. Huang, Z. Hao, M.-Y. Liu, S. Belongie, and B. Hariharan, “Pointflow: 3d point cloud generation with continuous normalizing flows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4541–4550.
- [22] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *Advances in neural information processing systems*, vol. 30, 2017.
- [23] D. Seita, Y. Wang, S. J. Shetty, E. Y. Li, Z. Erickson, and D. Held, “Toolflownet: Robotic manipulation with tools via predicting tool flow from point clouds,” in *Conference on Robot Learning (CoRL)*, 2022.
- [24] K. Zhang, M. Sharma, J. Liang, and O. Kroemer, “A modular robotic arm control stack for research: Franka-interface and frankapy,” *arXiv preprint arXiv:2011.02398*, 2020.
- [25] O. Sorkine-Hornung and M. Rabinovich, “Least-squares rigid motion using svd,” *Computing*, vol. 1, no. 1, pp. 1–5, 2017.
- [26] J. Levinson, C. Esteves, K. Chen, N. Snively, A. Kanazawa, A. Ros-tamizadeh, and A. Makadia, “An analysis of svd for deep rotation estimation,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 554–22 565, 2020.



## APPENDIX

### A. Implementation of tool scoring module

Given a set of  $K$  training tools, represented as a set of point clouds,  $\{P^{train_{tool}_i}\}_{i=1:K}$ , we train a tool scoring module  $D_{score}$ , which takes in a tool point cloud  $P^{tool}$ , the initial observation  $P^o$ , and the goal  $P^g$ , and it predicts a score  $s$  for the tool indicating how suitable the tool is for the task. The architecture for the tool scoring module is shown in Fig. 2 (a). The module first encodes the tool points to a latent feature using a PointNet++ [22] encoder. It then encodes the concatenation of observation points and goal points to another latent feature using a separate PointNet++ encoder. These latent features are concatenated and inputted through a multi-layer perceptron (MLP) to output a score. We train the module with binary cross-entropy loss, in which the tool used in the demonstration to achieve the goal point cloud  $P^g$  is considered as a positive example, and randomly selected tools from the training set are considered as negative examples.

### B. Details on the path generator

The path generator  $G_{path}$  starts by encoding the concatenated point clouds into a latent vector using a PointNet++ [22] encoder. This vector is then input into a ToolFlowNet [23]-based trajectory model. The trajectory model is set to a flow dimension of  $(H - 1) \times 3$ . The resulting output is interpreted as the tool’s flow at each time step, thereby producing  $H - 1$  delta poses  $T_{1:H}^{gen}$  via singular value decomposition [25], [26]. Finally, by utilizing this path generator with the generated tool in the reset pose  $P_0^{gen}$ , we create a point cloud trajectory  $P_{1:H}^{gen}$ . We train the path generator using the delta poses of the training tools  $T_{1:H}$  as labels (from the demonstration dataset). At each timestep, we apply the ToolFlowNet [23] loss between the trajectory produced by  $G_{path}$  and the actual trajectory of the training tool.

### C. Details on tasks and demonstration data

	Per task	Overall
# of initial configurations	200	800
# of target configurations	200	800
# of training trajectories	180	720
# of testing trajectories	20	80
# of total trajectories	200	800
# of total transitions	$10^4$	$4 \times 10^4$

TABLE III: Summary of training/testing data

We inherit the data generation procedure from Diff-Skill [5]: first, we randomly generate initial and target configurations. The variations in these configurations include the location, shape, and size of the dough and the reset pose of the tool. We then sample a specific initial configuration and a target configuration and perform gradient-based trajectory optimization to obtain demonstration data. For each task, the demonstration data consists of all the transitions from

executing the actions outputted by the trajectory optimizer, and we use a task horizon of  $H = 50$ . For each task, we perform a train/test split on the dataset and select 10 configurations in the test split for evaluating the performance for all the methods. More information about training and testing data can be found in Table III.

### D. Details on baselines

We provide additional details on each baseline below:

- **BC-E2E.** End-to-end behavioral cloning that outputs a  $H' \times 6, (H' > H)$  vector representing the delta poses of the tool relative to the initial tool pose. Unlike the other baselines, this baseline does not output a reset transformation. Here, we set  $H' = 60$  and use delta poses in the entire trajectory (i.e. the delta poses from interpolating the initial pose and the reset pose, as well as the subsequent delta poses during manipulation) as the label to regress on. As for the architecture, it first encodes the tool points to a latent feature using a PointNet++ [22] encoder. It then encodes the concatenation of observation points and goal points to another latent feature using a separate PointNet++ encoder. These latent features are concatenated and inputted through an MLP to produce the delta poses (represented as a  $H' \times 6$  vector).
- **BC-Joint.** Behavioral cloning that jointly regresses to the reset transformation and subsequent delta poses from the initial tool configuration. As for the architecture, it first encodes the tool points to a latent feature using a PointNet++ [22] encoder. It then encodes the concatenation of observation points and goal points to another latent feature using a separate PointNet++ encoder. These latent features are concatenated and inputted through an MLP to produce the reset transformation as well as delta poses.
- **BC-Latent.** Behavioral cloning that regresses to the reset transformation, moves the tool to the predicted reset pose, and then predict subsequent delta poses from a latent encoding of the restrictions scene with the tool in the reset pose. As for the architecture, it first encodes the tool points to a latent feature using a PointNet++ [22] encoder. It then encodes the concatenation of observation points and goal points to another latent feature using a separate PointNet++ encoder. These latent features are concatenated and inputted through an MLP to produce the reset transformation. For the delta poses, we encode the concatenated point clouds of the scene (observation, goal, and tool in the reset pose) into a latent vector using a PointNet++ encoder and then pass the latent feature through an MLP to produce the delta poses (represented as a  $(H - 1) \times 6$  vector).
- **TFN-Traj.** Behavioral cloning that regresses to the reset transformation, moves the tool to the predicted reset pose, and then uses the updated scene to predict subsequent delta poses for the tool with the ToolFlowNet-based [23] trajectory model described in Appendix B on the website.

Ablation Method	Training tools	Random initial pose	Novel tools
ToolGen Reset w/ BC-Latent	<b>0.94 ± 0.05</b>	<b>0.93 ± 0.05</b>	0.30 ± 0.55
ToolGen Reset w/ TFN-Traj	0.86 ± 0.14	0.85 ± 0.07	0.36 ± 0.60
ToolGen (Ours)	<b>0.91 ± 0.05</b>	<b>0.90 ± 0.06</b>	<b>0.72 ± 0.27</b>

TABLE IV: Ablation results across 3 scenarios. Each value in the table represents the normalized performance across all tasks.

Method	Average Inference Time	Average Execution Time
TFN-Traj	0.2s	23.0s
ToolGen (Ours)	22.7s	19.1s

TABLE V: Execution times averaged for all simulation tasks.

### E. Ablation studies

We conduct an ablation study on ToolGen by modifying its point cloud generator: we only generate the initial point cloud using ToolGen and align the current tool with this point cloud to determine the current tool’s reset pose. Following this, we input the current tool at its reset pose into the delta pose predictors of BC-Latent and TFN-Traj to obtain the subsequent delta poses. This ablation provides a clear comparison between the process of directly regressing to the delta poses and the approach of using ToolGen to output delta poses. The performance gap between these two methods when using novel tools is displayed in Table IV, which underscores the significance of generalization occurring in trajectory prediction. Specifically, since these two ablations regress onto the delta poses of the training tools, they tend to overfit to the training tools, causing them to produce inaccurate trajectories when faced with out-of-distribution test tools. In contrast, ToolGen inputs the generated tool into the trajectory predictors during the generation process. The generated tool minimizes the distribution shift for the path generator and thus significantly enhances the accuracy of the resulting trajectory predictions.

### F. Execution time

We further compare the average inference time of ToolGen and a baseline in Table V. Due to the sequential pose optimization step, ToolGen requires significantly more time during inference compared to its baseline, TFN-Traj, which only requires a single forward pass in the networks. We leave improving the time efficiency of ToolGen’s trajectory generation for future work.

### G. Failure cases

In Figure 8, we present two typical failure scenarios that occur when trying to align a novel tool with the generated tool. The first scenario, displayed on the left of Figure 8, occurs when there is a substantial disparity between the generated tool (top row) and the test tool (bottom row). In this case, the optimization process fails to meaningfully align the test tool with the generated shape in the later timesteps of the trajectory. However, this issue can be alleviated by training on more diverse tool shapes, which will create a richer shape distribution for the point cloud generator to generate.

$\lambda_c \backslash \lambda_r$	0.01	0.1	0.5
0.01	0.45 ± 0.30	0.65 ± 0.23	0.60 ± 0.15
0.1	0.50 ± 0.33	<b>0.72 ± 0.27</b>	0.68 ± 0.10
0.5	0.48 ± 0.36	0.49 ± 0.29	0.53 ± 0.18

TABLE VI: The effects of hyperparameters in sequential pose estimation. Each entry shows the performance cross all tasks with a particular combination of hyperparameters.

The second type of failure results from the optimization of delta poses. The hyper-parameter  $\lambda_r$  regulates the balance between the actual alignment and the regularization of the rotation amount in delta poses, and can be sensitive to the task at hand. During our experiments, we found that a single  $\lambda_r$  value generally performs well across all tasks. However, in the “Roll” task, minor problems occurred - occasionally the tool would rotate itself when aligning with the generated tool (as shown on the right of Figure 8). This issue can be remedied by fine-tuning the optimization’s objective function and hyper-parameters for each task. For instance, by increasing the regularization parameter  $\lambda_r$ , we can prevent large rotations during the alignment of delta poses.

### H. Effects of hyperparameters

To investigate how sensitive ToolGen is to hyperparameters during sequential pose estimation, we vary  $\lambda_c$  and  $\lambda_r$  when optimizing for the transformations. To eliminate any stochasticity from the generation process, we only generate the trajectories once and use the same set of generated trajectories for optimization. Table VI shows the performances of executing the trajectories that are optimized with different hyperparameters. In summary,  $\lambda_c$ , which penalizes collisions between the tool in reset pose and the environment, seems to have a greater effect than  $\lambda_r$ . This is because the alignment at the rest pose is being used to optimize the subsequent poses, and the error in optimizing the reset pose might cascade to the later process. We also observe that choosing a larger  $\lambda_r$  will decrease the variance in performance. This is because a larger  $\lambda_r$  will penalize large motions and encourage smaller and safer motions. It is worth noting that even with a large variation of these hyperparameters, ToolGen still almost always outperforms the baselines.

### I. Real World Experiment Details

#### 1) Environment setup:

**Robot and workspace.** The robot used for real-world execution is a 7-axis Franka Panda robot arm with a two-finger Franka Hand. The robot is fixed on a table with a  $0.55m \times 0.55m$  space for execution. The tool rack is placed on the left side of the execution space.

**Dough.** In real world experiments, we use *modeling dough*<sup>2</sup> for our manipulation tasks. For each of the three tasks, we create a mold of the same size as the dough in

<sup>2</sup>Hygloss Products 48308 Dazzlin’ Dough 3lb. White, bought from Amazon: <https://www.amazon.com/Hygloss-Products-48308-Dazzlin-Dough/dp/B07SNX6BPK>

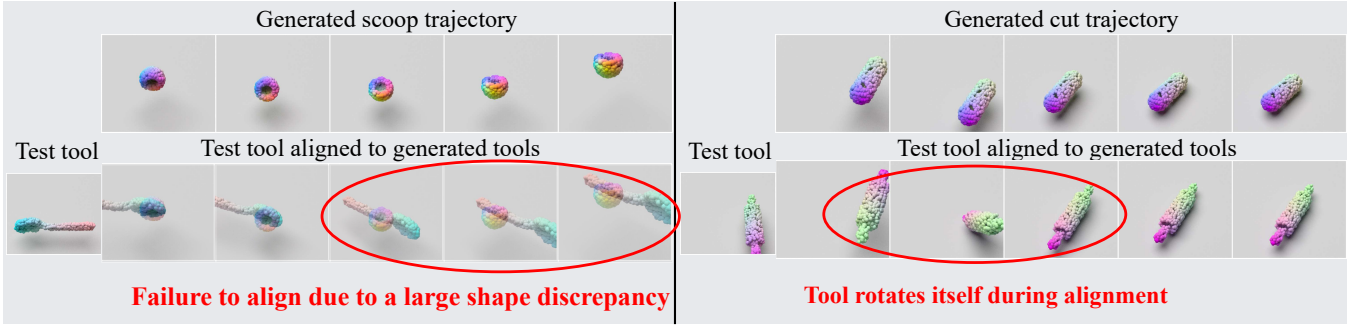


Fig. 8: Example failure cases of ToolGen (ours) when trying to align the test (actual) tool with the generated tool. Left: the alignment fails due to large difference in shapes of the generated tool and the test tool. Right: occasionally during the alignment process the aligned tool would have unexpected motions.

Taichi simulation. The real-world experiment goal states are created by human volunteers using real tools. The dough is reshaped and placed at a fixed center point before every experiment run.

**Multi-camera setup.** Here we set up multiple cameras to record dough state point clouds. Three Azure Kinect cameras are arranged around the workspace with equal distances from each other, i.e., placed in an equilateral triangle configuration, in front of the robot and on both sides of the robot, all pointing towards the geometric center of the workspace. The cameras are calibrated to form point clouds with re-projection errors less than  $0.01m$ . To synthesize a comprehensive view of the object, point clouds are further aligned using an Iterative Closest Point algorithm.

**Point cloud processing.** The collected tool and dough point clouds are hollow. We interpolate them by identifying cross-sections along the  $x$ ,  $y$ , or  $z$  axis and filling them with points. Then we downsample the interpolated point clouds using the same voxel size of  $0.002m$ . This produces a uniform distribution of points, and thus allows more accurate metric calculations for the dough’s target and goal point clouds.

2) *Robot execution details:* We use *Frankapy*<sup>3</sup> as the robot controller. The delta transformations from model output is under the tool frame. To execute the trajectory with robot arm, we calculate and apply the transformation from the recorded tool frame to robot end-effector. Each target pose  $(x, y, z, r, p, y)$  under robot frame is passed to Frankapy as a goal pose. Frankapy then calculates the inverse kinematics for the given goal state and execute each goal within 0.5 seconds.

<sup>3</sup><https://github.com/iamlab-cmu/frankapy>