53rd SME North American Manufacturing Research Conference (NAMRC 53, 2025)

Learning Force-Conditioned Visuomotor Diffusion Policy from Human Demonstrations for Complex Robotic Assembly Tasks

Rishabh Shukla[a], Samrudh Moode[a], Raj Talan[a], Satyandra K. Gupta[a,*]

[a] *Center for Advanced Manufacturing, University of Southern California, Los Angeles, CA, USA*

Abstract

Assembly operations in manufacturing, especially those involving precise alignment and force control, pose significant challenges for automation. Tasks like fitting a battery cover onto a housing require careful manipulation to ensure proper alignment and insertion without causing damage. We propose leveraging imitation learning by collecting demonstrations through hand-guided manipulation, capturing both vision and force/torque data from sensors mounted on the robot's end-effector. Although hand-guided manipulation may introduce minor imprecisions, our approach compensates by integrating high-fidelity force/torque sensing and run-time visual feedback, along with post-processing filters, to ensure the precision required for complex robotic assembly. These demonstrations are used to train a bimanual robotic system where one arm holds the battery housing securely while the other inserts the top cover. To enable this, we extend the diffusion policy framework by incorporating run-time force feedback and visual observations. Additionally, we introduce data segmentation and augmentation methods to reduce the number of required demonstrations, enhancing the policy's robustness to task failures. Our findings demonstrate that our approach, despite being trained on a limited dataset, improves success rate and efficiency over conventional diffusion techniques. In addition, we present a case study in which our bimanual robotic system performs precise alignment and insertion of the battery cover, highlighting its potential for complex assembly tasks in manufacturing settings. However, the approach remains sensitive to sensor drift and has not yet been tested on highly deformable or ultra-tight-tolerance assemblies, highlighting opportunities for future improvement.

Keywords: Learning from Demonstration; Bimanual Manipulation; Robotic Assembly

1. Introduction

Many modern products require complex assembly operations. For example, consider the task of aligning two mating features with tight tolerances and then performing motion to insert the component in place as shown in fig. 1. Such assembly tasks require relative motion among parts while maintaining contacts. Human operators excel at such tasks by using multimodal sensory feedback, primarily vision and force sensing, to make fine adjustments in run-time. Force needs to be carefully controlled in such operations. Application of excessive forces can damage parts. On the other hand, the use of insufficient force leads to unsuccessful task completion. With increasing labor shortages, there is a growing demand to automate these complex assembly operations [1]. Automating these processes allows robots to take on tedious, physically demanding work, freeing humans to focus on higher-value tasks [2].

There has been good progress in deploying robots to automate simple assembly operations that require peg-in-hole insertion for simple geometries [3]. Off-line programming alone is insufficient for high-contact assembly tasks, since run-time force/visual adjustments are critical to success. Robots need to fine tune motions based on the force and vision data to efficiently and safely perform assembly operations. Robots need to utilize controllers that select actions based on the current state of the assembly. It is difficult for humans to design controllers for tasks that utilize complex multi-modal sensing [4]. Recent advances in machine learning have been exploited by robots to learn policies to execute complex tasks.

* Corresponding author.
E-mail address: guptask@usc.edu (Satyandra K. Gupta).

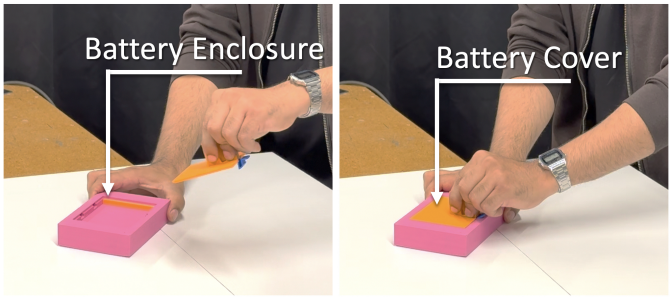

Figure 1: A human operator demonstrating the assembly task involving aligning two mating features and performing the necessary motion to insert the component in place. The figure illustrates key stages of the task: (A) precise positioning of the components, and (B) applying coordinated forces and adjustments to ensure correct insertion and alignment, emphasizing the importance of multimodal sensory feedback for successful assembly.

When good simulation models are available, reinforcement learning works well for robots to learn new policies. Unfortunately, simulation models do not work well in contact-rich complex assembly tasks. Therefore, we cannot effectively utilize reinforcement learning. Another popular approach is for robots to learn a task execution policy using learning from demonstrations (LfD) [5, 6]. For many simple tasks, behavior cloning, a common imitation learning method, has been utilized to learn a policy from human demonstrations (see Section 2 for details). However, traditional behavior cloning techniques suffer from two major limitations when applied to complex tasks: (1) they are prone to compounding errors over long execution horizons; and (2) they inadequately capture the multimodal nature of human demonstrations, failing to model multiple valid strategies for performing a task. In contrast, diffusion policies (DPs) have emerged as a promising alternative for learning complex tasks from human demonstrations. They reduce error accumulation and capture a diverse range of strategies via the diffusion process. Most prior work on diffusion-based policy learning relies only on vision-based inputs—often called visuomotor diffusion policies [7, 8, 9, 10]. In contrast, our approach enables the robot to adjust its actions using both real-time visual and force feedback.

We collect demonstration data using hand guidance, allowing human operators to naturally perform the assembly task while the robot records the necessary data. We integrate vision and force feedback by utilizing a wrist-mounted camera and a force/torque sensor on the robot's end-effector. This setup captures rich sensory information, including image and force measurements in the tool frame, providing the diffusion policy with comprehensive observations for decision-making.

Experimental results show that our method outperforms baseline diffusion policy (see Section 5 for details) in assembling constrained components, even with a limited dataset. The robot successfully performs precise alignment and insertion of the battery cover, adjusting its actions in run-time based on visual and force feedback. These findings highlight the potential of our approach for automating complex assembly tasks in manufacturing environments.

Our main contributions in this paper are as follows. First, we extend the diffusion policy method to incorporate both vision and force feedback, enabling the robot to adjust its actions based on multi-modal sensory information. Second, by segmenting task-level demonstrations into sub-tasks—such as positioning the cover, aligning it with the housing, and applying the necessary force for insertion—we reduce the number of required demonstrations. Third, we develop a data augmentation approach to make the training process robust to failures, enhancing the policy's ability to handle variability in the assembly process. Finally, we design a new testbed specifically for bimanual robotic assembly of constrained components, demonstrating practical applications in battery assembly tasks. A video demonstrating our work is available at `https://www.youtube.com/watch?v=GPplJa_7xAw`.

2. Related Work

Bimanual Manipulation is essential for complex assembly tasks in manufacturing [11]. Early methods relied on classical control and model-based planning [12, 13], achieving success in tasks like precise component insertion and alignment [14, 15]. However, these approaches often require detailed models and are time-consuming to develop, especially for tasks involving contact-rich interactions and variability in part geometries. Recent advancements in learning-based methods, including reinforcement learning [16, 17], low-level primitives [18, 19], and imitation learning [20, 21], have enabled robots to perform more dexterous assembly tasks like knot-tying [22, 23] and tool use [20]. Bimanual systems now leverage large datasets of human demonstrations to master tasks such as battery insertion and shoelace tying [24]. Building on these systems, our bimanual setup utilizes force feedback and visual data to coordinate arms for precise assembly tasks like fitting a battery cover, which requires accurate alignment and controlled force application.

Imitation Learning is a common technique within the Learning from Demonstration (LfD) paradigm [5, 25]. Previous approaches employed motion primitives to encode complex motions [26, 27]. With the advent of deep learning, methods like behavioral cloning emerged to learn an observed state-to-action mapping that replicates expert behavior [28, 29]. While effective, behavioral cloning often struggles with generalizing to unseen scenarios and is prone to compounding errors—issues that are critical in manufacturing settings where variability is common.

State-of-the-art methods have demonstrated that robots can learn visuomotor policies for complex tasks. For instance, Diffusion Policies [7, 30, 31] can handle multimodal action distributions, Action Chunking Transformers (ACT) [20] ensure coherent sequential decision-making, ALOHA [32] combines imitation learning with co-training to learn contact-rich manipulation. Recent advances, including self-supervised learning [33] and differentiable trajectory optimization [9], have improved these approaches. Complementary formulations like Generative Skill Chaining [34] and Consistency Policy [10] address inference speed challenges. Our work builds upon diffusion policy [8], which uses a denoising process to transform noise into