RoboChef: Making a Sandwich with a Single Robot Arm

Karan Baijal, Zhanxin Wu

***Abstract*— Robot-assisted meal preparation is a critical yet challenging task, primarily due to the diverse physical properties of food and the long-horizon nature of the task. This project presents RoboChef, a system designed to autonomously prepare sandwiches using a single arm that can pick up a knife equipped with a T/F sensor. We separate low-level skill learning from high-level task planning, where the robot learns primitive skills such as picking, placing, and cutting from demonstrations and then uses these learned skills to execute a complete task sequence. We demonstrate a meal preparation system using a single-arm robot that can manipulate raw ingredients to assemble a sandwich.**

I. Introduction

Eating is a fundamental aspect of daily life, yet millions of individuals worldwide face significant challenges in independently preparing meals and feeding themselves due to mobility limitations [1]. Meal preparation is one of the most essential—and most challenging—instrumental activities of daily living (IADLs). Despite significant advancements in robotics, automating meal preparation, particularly tasks involving food manipulation such as sandwich-making, remains a difficult and largely unsolved problem. Key challenges stem from the inherent variability and complexity of food items, each with distinct physical properties such as shape, texture, and compliance, which demand precise and adaptable manipulation skills. Furthermore, successful meal preparation typically involves executing long-horizon task plans composed of multiple sequential subtasks, where errors in earlier steps can propagate and hinder overall task completion.

Most robotic systems rely on bimanual setups to replicate human-like dexterity in meal preparation. However, these configurations introduce practical challenges in household environments, such as higher costs, space constraints, and safety concerns. In contrast, single-arm robotic systems offer a more cost-effective and space-efficient alternative. However, it is challenging to achieve comparable effectiveness in tasks typically performed with two hands using a single arm due to factors such as object stabilization and sequential execution constraints, especially in tasks like securely cutting vegetables or assembling sandwiches. To address these challenges, we propose RoboChef, a robotic system that uses a single-arm robot to autonomously manipulate raw ingredients and prepare sandwiches.

II. Related Work

Recent research in meal preparation has explored a wide range of tasks, from peeling and cutting to more complex actions such as sandwich assembly. While most studies focus on individual skills—such as pushing [2], peeling [3][4], or cutting [5]—few have attempted to integrate these into a complete system. Early work [6] designs an automated system for sandwich assembly and packaging using modular workstations to handle compliant food items. Recently, Yi et al. [7] study task scheduling to optimize cooking sequences. Most studies focus on the perception of food items, including rigid [8], semi-fluid [9], and soft items [4]. Some works explore learning control policies for cooking tasks such as stir-frying [10]. Shi et al. [11] demonstrate learning to use diverse tools to make dumplings. While these works focus on individual components and often assume preprocessed food items, our project aims to develop a sandwich-making system that operates on raw ingredients.

III. Problem Formulation

We consider the problem of robot-assisted sandwich making. A robot with a utensil tool on its end-effector is presented with raw food materials and a cutting board. A user specifies the sandwich type they would like through natural language instructions. At timestep t, the system takes observation $o \in \mathcal{O}$, which comprises of an RGB-D image $I_t^{wrist} \in \mathbb{R}^{W \times H \times 4}$ of the plate from a camera mounted on the robot's wrist, an RGB-D image $I_t^{third} \in \mathbb{R}^{W \times H \times 4}$ of the plate from a third-person view camera, force and torque readings $F_t \in \mathbb{R}^6$ from an F/T sensor, and corresponding robot end-effector poses $P_t \in \mathbb{R}^6$. Based on observations, our system selects a skill $k \in \mathcal{K}$ that is a discrete skill from a predefined library: $\{\texttt{pickup}, \texttt{place}, \texttt{cut}, \texttt{spread sauce}\}$. Each skill takes a *time series* of observations o_T as input and outputs a sequence of action $a \in \mathbb{R}^7$ including end effector pose and the width of the gripper.

We assume we have a closed set of food items to work with and no meat cutting is required. The considered food items include:

- Meats: chicken breast, turkey breast, ham
- Vegetables: cucumber, tomato, lettuce
- Others: brown bread, cheese

IV. Method

We propose a robotic system for autonomous sandwich preparation, as shown in Figure 1. At the low level, our system learns primitive skills—including picking, placing, and cutting—from demonstrations. At the high level, the system combines these learned skills to prepare a sandwich by generating a sequence of actions derived from natural language instructions.

Perception. Given I_t^{wrist}, we utilize Grounding DINO and the Segment Anything Model (SAM) to detect a list of semantic labels l_t (e.g., bread, tomato, chicken) and

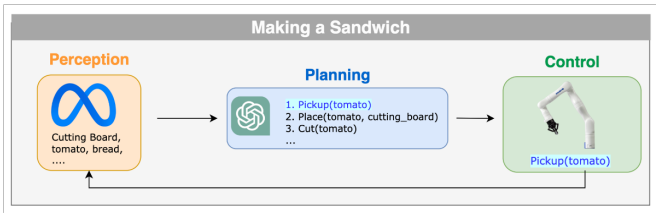

Fig. 1: Overview

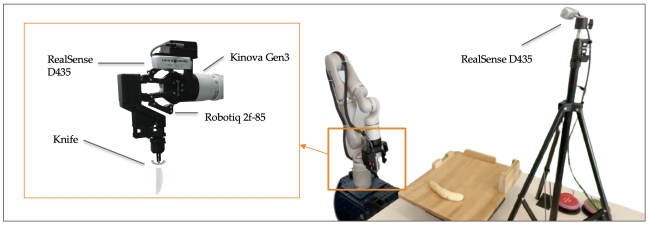

Fig. 2: Hardware

corresponding segmentation masks ${m^i}_t$ for each food item i on the table.

Planning. Given a natural language instruction (e.g., "I want a chicken breast cucumber sandwich"), the available food items l_t, and the skill library $\mathcal{K}$, we use a Large Language Model (LLM) as a planner to generate a skill sequence. For each skill, the LLM planner also incorporates situated state feedback from the environment by asserting preconditions of our plan, such as cutting the tomatoes before placing them on the bread, and responding to failed assertions with recovery actions [12].

Control. Before deployment, we collect demonstrations to train an imitation learning policy for each skill in the library $\mathcal{K}$. The policy takes as input a segmentation mask m^i for the item of interest, along with o_T, and outputs an action sequence. During deployment, given a skill sequence, each skill assumes access to the corresponding segmentation mask ${m^i}_t$, as well as o_T, and executes its associated action sequence to complete the task.

V. Evaluation

Hardware. RoboChef is implemented on the Kinova Gen3 robot arm equipped with a motorized utensil mounted at the end-effector (Figure 2). The utensil contains a fork attachment and has two degrees of freedom corresponding to the orientation of the fork tines and the tilt angle. This setup enables direct control of the utensil for dynamic movements such as scooping and cutting, while the robot executes waypoint-based navigation within the workspace using Cartesian position control. Additionally, we utilize an Intel RealSense D435 camera mounted on the wrist of the Kinova arm and another Intel RealSense D435 for a third-person view, both for visual perception, along with an F/T sensor for haptic perception.

Evaluation Scenarios. We evaluate our approach on three sandwiches: (i) chicken breast cucumber sandwich, (ii) turkey breast tomato sandwich and (iii) ham lettuce sandwich.

Metrics. We evaluate each method using the Success Rate (SR). A trial is considered to be successful if it completes a sandwich that contains all the food items required by the user. Failed attempts leave the item on the plate, allowing up to three re-attempts before it is manually removed and recorded as a failure. Success Rate quantifies the proportion of successful attempts relative to total attempts, defined as Success Rate (SR) = $\frac{\#\text{successful attempts}}{\#\text{total attempts}}$. We measure SR for the entire system as well as SR for each individual skill.

References

[1] World Health Organization, *Global report on health equity for persons with disabilities*. World Health Organization, 2022.

[2] T. B. Gilwoo Lee and S. S. Srinivasa, "Bite acquisition of soft food items via reconfiguration," *RSS Workshop on Task-Informed Grasping (TIG-II) - From Perception to Physical Interaction*, 2019.

[3] R. Ye, Y. Hu, Y. A. Bian, L. Kulm, and T. Bhattacharjee, "Morpheus: a multimodal one-armed robot-assisted peeling system with human users in-the-loop," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9540–9547, 2024.

[4] N. Ha, R. Ye, Z. Liu, S. Sinha, and T. Bhattacharjee, "Repeat: A real2sim2real approach for pre-acquisition of soft food items in robot-assisted feeding," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7048–7055, 2024.

[5] K. Zhang, M. Sharma, M. Veloso, and O. Kroemer, "Leveraging multimodal haptic sensory data for robust cutting," *IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*, pp. 409–416, 2019.

[6] S. Davis, M. King, J. Casson, J. Gray, and D. G. Caldwell, "Automated handling, assembly and packaging of highly variable compliant food products - making a sandwich," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pp. 1213–1218, 2007.

[7] J.-s. Yi, M. S. Ahn, H. Chae, H. Nam, D. Noh, D. Hong, and H. Moon, "Task planning with mixed-integer programming for multiple cooking task using dual-arm robot," pp. 29–35, 06 2020.

[8] A. Dikshit, A. Bartsch, A. George, and A. B. Farimani, "Robochop: Autonomous framework for fruit and vegetable chopping leveraging foundational models," 2023.

[9] X. Luo, S. Jin, H.-J. Huang, and W. Yuan, "An intelligent robotic system for perceptive pancake batter stirring and precise pouring," 2024.

[10] J. Liu, Y. Chen, Z. Dong, S. Wang, S. Calinon, M. Li, and F. Chen, "Robot cooking with stir-fry: Bimanual non-prehensile manipulation of semi-fluid objects," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5159–5166, 2022.

[11] H. Shi, H. Xu, S. Clarke, Y. Li, and J. Wu, "Robocook: Long-horizon elasto-plastic object manipulation with diverse tools," *arXiv preprint arXiv:2306.14447*, 2023.

[12] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11523–11530, 2023.