Abstract—Robotic Assisted Feeding (RAF) addresses the fundamental need for individuals with mobility impairments to regain autonomy in feeding themselves. The goal of RAF is to use a robot arm to acquire and transfer food to individuals from the table. Existing RAF methods primarily focus on solid foods, leaving a gap in manipulation strategies for semi-solid and deformable foods. This study introduces Long-horizon Visual Action (LAVA) based food acquisition of liquid, semisolid, and deformable foods. Long-horizon refers to the goal of “clearing the bowl” by sequentially acquiring the food from the bowl. LAVA employs a hierarchical policy for long-horizon food acquisition tasks. The framework uses high-level policy to determine primitives by leveraging ScoopNet. At the mid-level, LAVA finds parameters for primitives using vision. To carry out sequential plans in the real world, LAVA delegates action execution which is driven by Low-level policy that uses parameters received from mid-level policy and behavior cloning ensuring precise trajectory execution.

We validate our approach on complex real-world acquisition trials involving granular, liquid, semisolid, and deformable food types along with fruit chunks and soup acquisition. Across 46 bowls, LAVA acquires much more efficiently than baselines with a success rate of 89 ± 4%, and generalizes across realistic plate variations such as different positions, varieties, and amount of food in the bowl. Code, datasets, videos, and supplementary materials can be found on our website.

I. INTRODUCTION

For individuals with limited mobility or disabilities, the act of feeding themselves can pose a significant challenge. This challenge has motivated the development of Robotic Assisted Feeding (RAF) [1] aiming to restore independence and enhance the quality of life for those affected, while also alleviating the caregiver burden. A key component of such an assistive feeding system is bite acquisition, i.e., the act of a robotic arm picking up morsels of food from a plate to transfer the food to a person’s mouth.

Navigating the diverse array of foods—from granular cereals to semi-solid food such as yogurt and deformable food items such as tofu—without breakage or deformation presents significant challenges for RAF [2], [3]. Additionally, the dynamic positioning of food chunks within a fluid medium complicates the prediction of their exact location at the time of scooping, requiring sophisticated sensing and real-time adaptation capabilities. This underscores the need for RAF systems to exhibit not only dexterity but also an advanced understanding of the dynamic nature of various food types and operate over a long horizon to clear the bowl.

Traditional RAF methodologies have depended on hard-coded adaptation strategies and low-level vision-based action execution. This approach, while effective for singular, isolated actions, falls short in replicating the complex, sequential behaviors exhibited by humans during feeding. Humans adeptly combine various actions, such as scooping both solid chunks and liquid from a bowl in a single motion or rearranging food items for easier acquisition, demonstrating a nuanced understanding and strategy that spans the entire meal. This limitation underscores a gap in RAF technology and highlights the need for an advanced understanding and replication of human-like, long-horizon feeding strategies capable of managing both the rigidity of solid foods and the complexity of deformable items.

Recent advancements in skill-based reinforcement learning (RL) offer promising methodologies for modeling these complex, long-horizon manipulation sequences in a hierarchical manner. This entails first learning a high-level policy for composing skills [12], and then optionally inferring the parameters of low-level skills separately [13], [14]. Such approaches have shown potential, yet they face limitations when applied to the food domain, which demands high-fidelity models for food deformation, visual recognition, and utensil interaction not fully captured in current simulations. While VAPORS [7] demonstrates effective long-horizon planning for specific food items such as noodles, it relies heavily on simulation for learning plate dynamics and lacks adaptability and applicability in real-world scenarios.
A. Robotic-Assisted Feeding

Robotic Assisted Feeding (RAF) can be split into two stages: bite acquisition and bite transfer. Previous work in RAF focused on bite acquisition and transfer with the aid for broader categories of food types.

Thus, we seek to find an appropriate layer of abstraction for feeding, which can leverage the benefits of (1) hierarchical planning for long-horizon manipulation; (2) vision-based primitives for fine-grained control; and (3) flexible approach that can dynamically adapt to the wide variety of challenges presented by different food types.

Recognizing these challenges, this work introduces LAVA (Long-horizon Acquisition via Visual Action), as a hierarchical policy for sequential planning of food acquisition (see Figure 1). Our approach is decoupled into three level-policy: a high-level policy that identifies primitives based on visual inputs; a mid-level policy to refine these primitives and parameterize the actions of the lower-level policy and a low-level policy to use those parameters to rearrange and acquire food items and sequentially clear the bowl.

The key contributions of this paper are:

- We introduce a dataset of food items, showcasing different volumes and spatial arrangements within the bowl.
- We evaluate the learned scooping policies through real-world deployment system with UR5e and end-effector coupled with spoon attachment and D435i RealSense camera on the wrist.

The remaining of the paper is organized as follows.

In Section II, we review related work on robotic assisted feeding, imitation learning, and long-horizon planning. In Section III, we present the problem statement. In Section IV, we introduce the multi-level policy including high-level policy, mid-level policy, and low-level policy. In Section V, we present and discuss the real robot experimental results and compare them with the baseline. Finally, Section VI, concludes the paper with lessons learned and possible future work.

II. RELATED WORK

We build on prior works studying multisensory robot learning and long-horizon task planning both within and beyond the food domain. In this section, we will discuss related work in robot-assisted feeding, learning from demonstration, and more generally long-horizon planning and control.

A. Robotic-Assisted Feeding
of robotic arms and specialized tools such as spoons and forks [2], [4], [6], [8], [15]. The incorporation of computer vision has enabled these systems to adapt to various food types and user preferences, with models such as SPANet [5] demonstrating proficiency in mapping food images to actions. However, challenges remain in handling semi-solid and deformable foods, where generalizable strategies are scarce and bimanual scooping [2] techniques have shown limited success. Market-available devices [11] offer real-time assistance but are constrained by their reliance on teleoperation and the physical limitations of their design. In isolation, this does not capture many long-horizon real-world feeding scenarios with multiple utensils and strategies. While prior research has made strides in visual planning and manipulation for specific food items [7], a comprehensive approach that addresses the adaptability to a wide array of food types and real-world feeding scenarios is still needed.

B. Learning from Demonstration

Learning from Demonstration (LfD) is a methodology where robots learn new skills by observing expert demonstrations, which can be performed by humans or intelligent agents. This approach is particularly useful for tasks that are challenging to pre-program but can be easily demonstrated. LfD has been applied across various domains, including robotic assembly in manufacturing [16], path planning for complex tasks [17], assistive technologies in rehabilitation [18], and intricate picking and placing tasks [19]. The technique is divided into three main approaches: kinesthetic teaching [20], where a human physically guides the robot; teleoperation [21], where the robot is remotely controlled; and passive observation [22], where the robot learns by watching. Our research primarily utilizes kinesthetic teaching to instruct a UR5e robot arm in scooping tasks. Within LfD’s learning objectives, our focus is on developing policies for handling semi-solid and deformable food items, and optimizing their scooping trajectories.

C. Long-Horizon Planning and Control

Recent works in long-horizon manipulation frameworks have explored separating high-level strategic decision-making from detailed motion planning. Traditional task-and-motion planning approaches rely on extensive domain knowledge and fixed task sequences [23]–[25], but falter due to the unpredictable dynamics of food on a plate and the complexity of state estimation. Model-based planning has shown promise in tasks such as dough manipulation by using environment dynamics learned from visual inputs to plan action sequences [12], [26]. However, these methods struggle with the high-dimensional action spaces typical in food acquisition. Hierarchical reinforcement Learning offers a solution by dividing decision-making into high-level strategic planning and execution by discrete, parameterized low-level primitives [27]. While promising in simulation for tasks such as tabletop manipulation [13], [14], these methods have limitations in real-world application and handling the diverse and complex manipulation tasks required for effective feeding. Our work aims to address these gaps by focusing on the adaptation to real-world scenarios and the development of specialized primitives for a wide variety of food items, challenging the scalability of current approaches and introducing the necessity for innovative solutions in robotic feeding.

III. Problem Statement

In this work, we tackle the challenge of sequential bite acquisition to maximize the success rate and efficiency of long-horizon food acquisition to ensure efficient bowl clearance. The focus is on a variety of food types, ranging from granular items such as cereals to semi-solid foods such as yogurt, and deformable substances such as tofu, all within a bowl fixed in position and assumed to be scoopable with a spoon.

We assume access to bowl image observations \( o \in \mathbb{R}_{W \times H \times C} = \mathcal{O} \) of unknown bowl states \( S \). Here, \( W, H \), and \( C \) denote the image dimensions. The image is sourced from a camera attached to the wrist of the robotic arm with a custom spoon attachment as an end-effector. We have access to expert demonstration data for robot proprioceptive information (joint positions). Our goal is to learn a policy \( \pi(\phi_i|o_i) \) that takes RGB images as input \( o_i \) and returns output as joint angles \( \theta_i \) of the arm for efficient long-horizon food acquisition. In this context, long-horizon refers to a series of sequential actions aimed at complete bowl clearance.

IV. Proposed Approach

We formalize the long-horizon food acquisition setting as a hierarchical policy \( \pi \). To do so we decouple \( \pi \) into separate high, mid, and low-level sub-policies. We assume access to \( K \) discrete manipulation primitives \( P_k^H \), \( k \in 1, ..., K \), and learn a high-level policy \( \pi_H \) which selects amongst these primitives based on visual input \( o_i \). The mid-level policy \( \pi_M \) further refines this selection, parameterizing the low-level policy \( \pi_L \) based on both the chosen primitive and additional visual inputs.

This low-level policy then executes a sequence of actions \( \theta_i^k \), aimed at achieving precise food acquisition. This hierarchical arrangement is formalized as follows:

- **High-level policy**: \( \pi_H(P_k^H|o_i) \) focuses on selecting the manipulation primitive suitable for the current visual scene.
- **Mid-level policy**: \( \pi_M(P_k^M, \psi^k_i|o_i, P_k^H) \) refines this choice by parameterizing actions tailored to the specific food item's characteristics.
- **Low-level policy**: \( \pi_L(\theta_i^k|P_M^k, \psi_i^k_M, P_k^H) \) executes the action sequence, utilizing parameters and primitives from the mid-level policy.

We consider low-level actions \( \theta_i \), parameterized by the position of the tip of a spoon \((x, y)\) and spoon roll and pitch \((\gamma, \beta)\) in the wrist frame of reference. As shown in Figure 2 detailing the LAVA setup, we describe each module in LAVA in further detail.
A. High-level Policy

At the highest level of our hierarchical model, the high-level policy \( \pi_H(P^k_H|o_t) \) uses visual cues to select the most suitable scooping primitive—Wide Primitive \( (P^w_H) \) and Deep Primitive \( (P^d_H) \), based on the food type present.

1) Wide Primitive \( (P^w_H) \): Wide Primitive, is a strategy developed for handling foods that lack cohesion or are deformable, such as tofu or certain types of jelly. This method involves using the bowl’s wall as a guide and support mechanism for the scooping action. By gently pressing the food against the wall of the bowl, it creates a pseudo-cohesive mass that can be scooped more easily. This technique is especially useful for foods that tend to scatter or break apart, as the wall provides the necessary containment to gather and scoop the food effectively. Instance scooping requires sophisticated control over the spoon’s movement, including adjusting the angle applied against the food and the bowl wall, to achieve the desired outcome without compromising the integrity of the food or missing the target. It requires identifying the target instance to not collide with other instances or break them in the process. The wide primitive is implemented with the other two mid-level primitives align and wall-guided scooping described in Section IV-B.1.

2) Deep Primitive \( (P^d_H) \): Deep Primitive, on the other hand, is a straightforward approach designed for foods that possess enough cohesion to be picked up directly by a spoon without requiring additional support or manipulation. This method is particularly effective for liquid and semi-solid foods such as yogurt or porridge, where the food’s natural consistency allows it to adhere to the spoon when scooped directly from the top or side. The key to successful direct scooping lies in the precise control of the spoon’s trajectory and depth of penetration into the food, ensuring that a sufficient quantity is acquired without disturbing the remaining contents of the bowl excessively. The deep primitive is implemented with the direct scooping mid-level primitive described in Section IV-B.2.

3) ScoopNet \( (\pi_H) \): ScoopNet is a network designed to select between the two high-level primitives based on the type of food, utilizing the MobileNetV2 architecture [28] as the base. We train on a dataset of 5316 images from a custom collection and additional sources, targeting binary classification of high-level primitives \( P^w_H, P^d_H \). We used data augmentation (including rotations, zooms, and flips) to increase robustness against food image variations. Our dataset is available online.

The network is initially trained on the ImageNet dataset, with a customized final layer for specific task adaptation. This configuration, along with a Global Average Pooling layer and two dense layers ending in a sigmoid activation, uses the Adam optimizer and binary cross-entropy loss for accurate classification. The detailed architecture, ScoopNet, is depicted in Figure 3. The output of ScoopNet is softmax probabilities over high-level primitives.

B. Mid-level Policy

The Mid-level Policy \( \pi_M(P^K_M|o_t, P^k_H) \) serves as the intermediary layer that refines and parameterizes the chosen primitive for execution. This refinement is crucial for bridging the gap between high-level strategy selection and low-level action execution.

1) TargetNet \( (\pi_M^1) \) for Wide Primitive: We have designed TargetNet, shown in Figure 4, that uses Mask R-CNN, tailored for the task of identifying and segmenting target food items such as tofu in a bowl, crucial for executing wide primitives. This model precisely segments food items, enabling the selection of appropriate mid-level primitives: wall-guided scooping and center align (described later in this section).

We use a custom dataset annotated for bowl, tofu, and target scooping areas, TargetNet employs transfer learning to accurately segment food items against diverse backgrounds, increasing its generalizability. In Section V-D.3, we report zero-shot generalization results for other types of food items. The model’s training includes a COCOEvaluator to ensure segmentation accuracy meets COCO dataset [29] standards.

Fig. 3: ScoopNet outputs the softmax probabilities over the high-level primitive depending on the type of food items present in the image.

Fig. 4: TargetNet finds the next “target” item for the wide high-level primitive and the mid-level primitive that decides whether to scoop the target item or to align it first.

Post-training, TargetNet creates a binary mask for pixels that are “occupied” by instances of food items. We divide the surrounding region of interest into sub-regions. If a sub-region intersects the bowl boundary, it is considered to be “occupied.” Otherwise, it is “unoccupied.” A food item is classified as “R1” if it is rightmost and closest to the wall, “R2” if the food item is at the center of the bowl, and “R3” otherwise. The subsequent visualization and centroid calculation steps of detected instances help with determining its location in subregions of the bowl and its location with respect to the center of the bowl, selecting between mid-level
primitives—Wall-guided Scooping ($P_{M1}^1$, $P_{M2}^1$) or Align($P_{M1}^2$, $P_{M2}^2$) and predicting parameters for low-level policy.

Wall-guided Scooping ($P_{M1}^1$, $P_{M2}^1$) The Wall-guided Scooping strategy, parameterized by $\delta$—the centroid distance of the target instance from the bowl’s center—adapts its approach based on the target’s proximity to the bowl’s wall and the sub-region. For food items in subregion R1, the strategy uses the wall’s structural support for a scooping action. Conversely, items in central subregion R2 require a pre-scooping alignment, tactically moving the food towards the wall to simplify the scooping motion.

Align ($P_{M1}^2$, $P_{M2}^2$) The alignment step is essential for orienting the spoon to the target instance and guiding its movement toward the bowl’s center. This procedure takes into consideration the centroid coordinates of the tofu ($x_t, y_t$) and the bowl’s center ($x_b, y_b$) as well as the spoon’s roll ($\gamma$) and pitch ($\beta$). Two key parameters are computed:

- **Spoon Orientation Angle:** Calculated as $\gamma = \arctan\left(\frac{y_t - y_b}{x_t - x_b}\right)$, this angle determines the necessary rotation of the spoon to align with the target instance, ensuring the spoon is positioned for optimal interaction and is utilized for planar push ($\beta = 0^\circ$).
- **Instance Push distance:** Determined by ($x_t$, $y_t$, $x_b$, $y_b$), the instance is pushed from its current position towards the bowl’s center, optimizing the positioning for the scooping action.

2) DepthNet ($\pi_M^2$) for Deep Primitive: DepthNet is designed for depth detection of food in a bowl based on visual input and high-level primitive received from high-level policy. The architecture of DepthNet is outlined in Figure 5.

![Figure 5: DepthNet detects the depth ($h$) of the food in the bowl.](image)

**V. EXPERIMENTS**

For the experiments, we begin by describing the experimental setup. Following this, we discuss the data collection procedure and the baseline for comparison. Subsequently, we present and analyze the experimental results.

A. Experimental Setup

The setup comprises a UR5e robot arm, a custom spoon attachment, a RealSense camera, and a fixed-position bowl, depicted in Figure 1. The spoon is affixed to the arm, with a length measuring 10.0 cm. The RealSense camera is attached to the wrist of the arm.
During experiments, we explore varied configurations across the amount, size, position, and depth of food as well as different food types including granular cereals, liquid water, and semi-solid yogurt in the bowl. Food position configurations encompass multiple numbers of tofu and fruit chunks placed in different instance positions across the bowl. The varied amount and food depth included cereals, water, yogurt, and jelly filled at different depth levels inside the bowl. Additionally, we conduct tests with tofu chunks inside soup as shown in Figure 8. For each food type, and depth, we conduct 10 trials of long-horizon food scooping attempts and for each position configuration in case of multiple tofu and fruit chunks, we conduct 5 trials of long-horizon food scooping attempts.

<table>
<thead>
<tr>
<th>Time (s)</th>
<th>0.0</th>
<th>0.4</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breakage</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>Spillage</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>Success</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**Fig. 6:** Breakdown of experimental performance comparison between LAVA, LAVA-low, and Fixed Trajectory Scooping (FTS). * represents zero-shot experiments.

### D. Experimental Results

In this section, we present and analyze the experimental results. We first present the success rate of LAVA’s networks. Then, following the training of the hierarchical policy, we evaluate its performance on the robot and compare it with the baseline methods. We test across varied food items and varied food configurations, including granular food cereals, liquid food water, semi-solid yogurt, deformable tofu, and multi-medium soup with tofu chunks. To assess performance, we employ the criteria of success rate, which indicates the successful scooping of food items from a bowl without spillage and breakage and successful long-horizon food acquisition by clearing the bowl efficiently. Instances where some spillage occurs are considered partial success.

1) **LAVA’s Network Success rates:** ScoopNet achieved 100% accuracy in choosing correct high-level primitives across 46 bowls, TargetNet accurately predicted bite targets at 87.9% over 83 instances, and DepthNet successfully determined correct spoon depths for bite sizes at 85.7% across 175 instances, demonstrating the LAVA networks’ effectiveness in robotic-assisted feeding.

2) **Comparison with Baselines:** In our experimental analysis, we evaluate the success rates of LAVA against two baseline models, LAVA-low and FTS, across a variety of food types and scooping dynamics, as shown in Figures 6 and 7. Our evaluation focused on several key metrics: efficiency (total time taken to clear the bowl), scoop size, and spillage for granular, semisolid, and liquid foods. For deformable foods and fruit chunks, we recorded configuration, number of scoop attempts, and instances of food breakage. In particular, for complex scenarios such as soup with tofu chunks, our assessment averaged efficiency, spillage, and breakage metrics.

**How do all the methods handle the challenge of scooping liquids, such as water and soup, which are prone to spillage?** The analysis, particularly visible in Figure 6c, reveals that both baseline models struggle with the fluidity of water and soup, leading to significant spillage.
The FTS model, with its fixed end-effector orientation and height, is not equipped to adjust to the varying dynamics of liquid scooping, resulting in spillage and ineffective scooping. LAVA-low initially copes but struggles as water levels decrease, showing inefficiency in maintaining adequate scoop sizes. In contrast, LAVA adeptly adjusts to real-time changes in food depth, achieving optimal scoop sizes and minimizing spillage for efficient bowl clearance.

What about the acquisition of more solid, yet deformable food types, such as tofu? Our findings, demonstrated in Figures 7a, 7b, and 6d, indicate that Both baselines encounter issues with deformable foods such as tofu, often resulting in food breakage. The FTS model’s rigid scooping motion damages the food, while Lava-low, despite managing to scoop, causes tofu to accumulate and break as shown by instances of food breakage in Figure 6d due to lack of strategic food prioritization based on subregions. LAVA, however, prioritizes tofu chunks based on their subregion, aligning them for easier access and significantly reducing breakage, mimicking human scooping strategies.

How does each method fare in preventing spillage and ensuring efficient scoop attempts with solid foods such as fruit chunks? The evaluation as visible in Figure 6c and 6a reveals that the baselines are less adept with solid, irregularly shaped foods such as fruit chunks, prone to rolling or falling off the spoon. This issue is exacerbated for fruits with curved surfaces. LAVA, employing an align-then-scoop strategy, ensures better alignment and significantly less spillage by adjusting to the fruit’s shape for secure scooping.

We see that LAVA consistently outperforms the baselines, achieving higher success rates and more effective plate clearance. It surpasses FTS and Lava-low by adapting its strategy for efficient, minimal-breakage scooping across all tested food types, demonstrating the benefits of its hierarchical policy framework. As expected, FTS and Lava-low, limited by their static approaches, fail to optimize for future scooping advantages, leading to increased breakage and inefficiency, especially without considering food prioritization and arrangement strategies.

LAVA’s comprehensive strategy ensures efficient, adaptive, and precise food acquisition, significantly improving upon the limitations of existing models.

3) Zero-shot Generalization: As detailed in Section V-B, our data collection process exclusively involved the transparent glass bowl containing granular cereals and tofu. However, we evaluated our approach to soup with tofu chunks and different food types such as liquid water and semi-solid yogurt, and solid apple chunks during testing. Remarkably, our approach demonstrates robust performance across these varied configurations, as depicted in Figure 6 and 8.

Especially with soup and tofu chunks, scooping up both the solid pieces and the liquid at the same time is tricky. Our system, LAVA, is designed to adjust to these challenges. Despite the tofu chunks’ tendency to float away from the desired central scooping area, LAVA’s adaptive strategy realigns and reorients to scoop the tofu effectively. Following the tofu acquisition, LAVA continues to adapt and clear the remaining soup, showcasing its capability to handle various food textures and types within the same meal, leading to efficient bowl clearance.

![Zero-shot long-horizon food acquisition with tofu chunks in soup.](image-url)

Fig. 8: (Zero-shot) long-horizon food acquisition with tofu chunks in soup. The top sequence of images shows the spoon aligning the target tofu towards the bowl’s center, which then drifts away during the scooping attempt due to the soup’s fluidity. The bottom sequence shows the system’s subsequent attempt to realign the tofu to the center, followed by a successful scooping action.
VI. CONCLUSION, LIMITATION AND FUTURE WORK

In this work, we have developed and presented a hierarchical policy framework designed to enhance robotic systems' capability in the acquisition of diverse food types, ranging from liquids to solids and deformable items. Through integrating DepthNet, TargetNet, and ScoopNet, our approach leverages representation learning, alongside sophisticated planning and execution strategies, to address the challenges associated with the variability in food textures, sizes, and positions within the bowl.

Our experimental analysis demonstrates the framework’s better performance in achieving better efficiency, minimum spillage, and breakage as well as adaptive food scooping compared to baseline models. Specifically, it showcases improvements in success rates across various food configurations. Despite the promising result towards generalization, limitations exist, particularly in handling thin, flat, or irregular foods needing specialized strategies. Future efforts will focus on broadening the action space for diverse food types and exploring efficient data acquisition methods, including leveraging internet video resources for complex food handling strategies in real-world scenarios.

REFERENCES


