RAMPS: Robot-Assisted Bite Acquisition Through Integrated Multimodal Visuo-Haptic Perception and Common Sense

Zhanxin Wu
zw754@cornell.edu
Cornell University
Ithaca, NY, USA

Tapomayukh Bhattacharjee
tapomayukh@cornell.edu
Cornell University
Ithaca, NY, USA

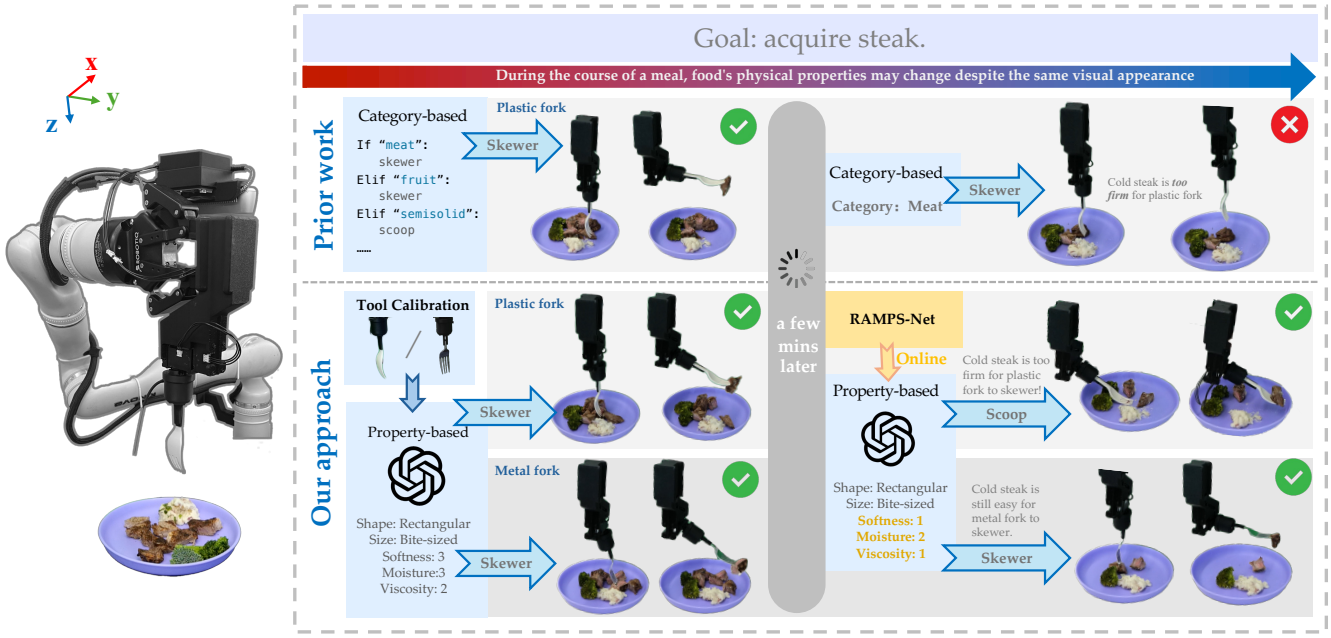

Fig. 1: We propose **RAMPS**, a method that combines commonsense knowledge from a visually-conditioned language model and multimodal visuo-haptic perception to dynamically estimate food properties. Based on food properties, RAMPS select the appropriate manipulation skill to enable robust bite acquisition.

***Abstract*—Robot-assisted feeding is a critical yet challenging task, particularly due to the diverse physical properties of food. Even the same food item can vary significantly—for example, fruit ripens day-to-day, and steak becomes firm when cooled. Bite acquisition, a key step in robot-assisted feeding, involves selecting appropriate skills, such as skewering or scooping, to pick up food. Existing approaches rely solely on visual information to do category-based skill selection (e.g., always skewering fruits). This leads to unreliable performance when important physical properties cannot be estimated from vision alone. To overcome these limitations, we propose RAMPS, a method that combines commonsense knowledge from a visually-conditioned language model and multimodal visuo-haptic perception to dynamically estimate food properties including shape, size, softness, moisture, and viscosity during interaction. RAMPS trains a RAMPS-Net offline and uses it to online predict food properties from vision and haptic inputs and performs calibration to generalize across utensils. Our key insight is that these visuo-haptic-derived food properties enable much finer-grained skill selection than the category-based state-of-the-art. Through extensive experiments on 20 single food items and 10 in-the-wild dishes, we show that our approach achieves a 10% average improvement in bite acquisition success rates. Our findings confirm that multimodality is key to generalization in bite acquisition and beyond.**

I. Introduction

Eating is a fundamental activity central to human life, yet millions of individuals face significant challenges feeding themselves independently due to mobility limitations [39]. A robot-assisted feeding system has the potential to empower individuals by helping them regain independence and dignity while ensuring their needs are met reliably and consistently [28]. Bite acquisition, the process of picking up a food item from a plate or bowl, is a critical step in robot-assisted feeding. This task presents several key challenges: (i) Food items are diverse and have different physical properties. Even within the same food type, significant variations can occur—for instance, different ripeness levels of a fruit, or food that is raw, fully cooked, or cooked to varying degrees. (ii) Physical

properties can change even during the course of a meal while maintaining a similar visual appearance. For example, meat and rice may become firm as they cool, while bananas may turn sticky when exposed to air. (iii) Some food items, such as tofu, are fragile, and without careful manipulation, they can break, making them increasingly difficult to pick up again. (iv) Physical interaction with food items may vary depending on what utensil is used. Because of all these challenges, identifying and applying an appropriate manipulation skill for a given food item is crucial to ensuring successful bite acquisition and advancing robot-assisted feeding systems.

Successful bite acquisition depends on robust utensil-food interaction. Our key insight is that *a better understanding of utensil and food physical properties and their interaction can enhance bite acquisition success across a wide range of food items*. Imagine having a meal with a variety of food items. Given a utensil, how do humans decide which manipulation skill to execute to pick up a food item? As humans, we possess a common-sense understanding of tool affordance. For example, when using a plastic fork, we are more likely to scoop rather than skewer a steak, recognizing its inability to penetrate the steak. Moreover, we effortlessly form initial assumptions about the properties of each food item based on visual cues. When observing a piece of steak, we might assume it is moderately soft and decide that skewering is a suitable skill. However, our common-sense understanding maybe not always be accurate. Consider a well-done steak: when we attempt to skewer it with a plastic fork, we may instantly realize that it is firmer than expected with our visual and haptic feedback. This sensory feedback allows us to refine our understanding of the physical properties of food items and adapt our strategies accordingly. This process highlights the importance of (i) understanding tool capabilities for effective skill selection, (ii) leveraging prior knowledge of food item physical properties for initial action selection, and (iii) updating physical property estimation using online visuo-haptic perception.

Using these insights, we present RAMPS (Figure 1), a method that combines commonsense knowledge from a visually-conditioned language model and online visuo-haptic perception for bite acquisition. RAMPS trains a RAMPS-Net offline and utilizes it for online prediction of food properties from vision and haptic inputs. In RAMPS, food items are represented with explicit physical properties including shape, size, softness, moisture, and viscosity. Most existing work on food manipulation represents food items as latent features [10, 35, 15, 21, 26]. Unlike prior work, our physical property representation is human-interpretable and could be useful when a human is in-the-loop. RAMPS has two phases: (i) Before deployment, we conduct tool calibration by evaluating various skills, such as skewering, scooping, and pushing on diverse food items with the given tool. This process forms an implicit understanding of the capability of the current tool embodiment. (ii) During deployment, we first leverage a visually-conditioned language model for commonsense understanding of the physical properties of food items based on visual cues. We use this as our initial estimate, which we adapt online using time-series visuo-haptic physical property perception for bite acquisition. We leverage these online physical property estimates coupled with tool embodiment capability estimates to select bite acquisition actions.

Overall, our contributions include:

- An **algorithm** that learns a food representation using physical properties through visuo-haptic perception. This representation is informed by common sense from foundation models, and is adaptable through visuo-haptic feedback during interaction.
- A **method** to better understand food-tool interaction by calibrating a tool embodiment on diverse food items.
- **Evaluation** of our framework on 20 single food items and 10 diverse in-the-wild dishes, showing improvement in bite acquisition success rates over category-based state-of-the-art approaches.
- A **dataset** of bite acquisition trials, utilizing different skills to manipulate food with varying physical properties.

II. Related Work

Our RAMPS framework combines the commonsense knowledge in foundation models and visuo-haptic perception to estimate food properties for bite acquisition. We briefly review prior literature in bite acquisition, visuo-haptic perception, and foundation models below.

Food Manipulation for Meal Assistance. Existing approaches for food manipulation for meal assistance explore various tasks, including cutting [42, 16, 45, 21], peeling [44, 7], and feeding [19, 15, 5, 18, 29]. Recent works in bite acquisition make significant progress in developing various skills, such as skewering [35, 10, 11, 12, 9], twirling [36], and scooping [38, 13]. To combine these skills, VAPORS [36] focus on noodle dishes and employs physics-based simulations for decision-making between twirling and grouping noodles. The closest relevant work, FLAIR [19] integrates multiple food manipulation strategies into a skill library and selects strategies based on food categories inferred from vision information. However, it overlooks intricate physical properties, which leads to failures in food items of varying softness. To address this limitation, we propose RAMPS that selects manipulation skills based on food properties instead of their categories. RAMPS enables more reliable bite acquisition, improving bite acquisition success rate across diverse food items.

Integrating Vision and Haptics in Robotics. Vision-only manipulation has demonstrated effectiveness in robotic domains such as semantic grasping and deformable object manipulation [32, 6, 10]. However, contact-rich tasks, such as in-hand manipulation [30, 37] and dense packing [3], have been shown to benefit significantly from the integration of vision and haptic information. Some work explores the benefits of multi-modality using paired visual, tactile, and auditory information [45, 35]. Previous studies [11, 4, 43] have highlighted the importance of haptic feedback in food manipulation. More recently, Sundaresan et al. [35] and Gordon et al. [11] combine a single post-contact image with time-series

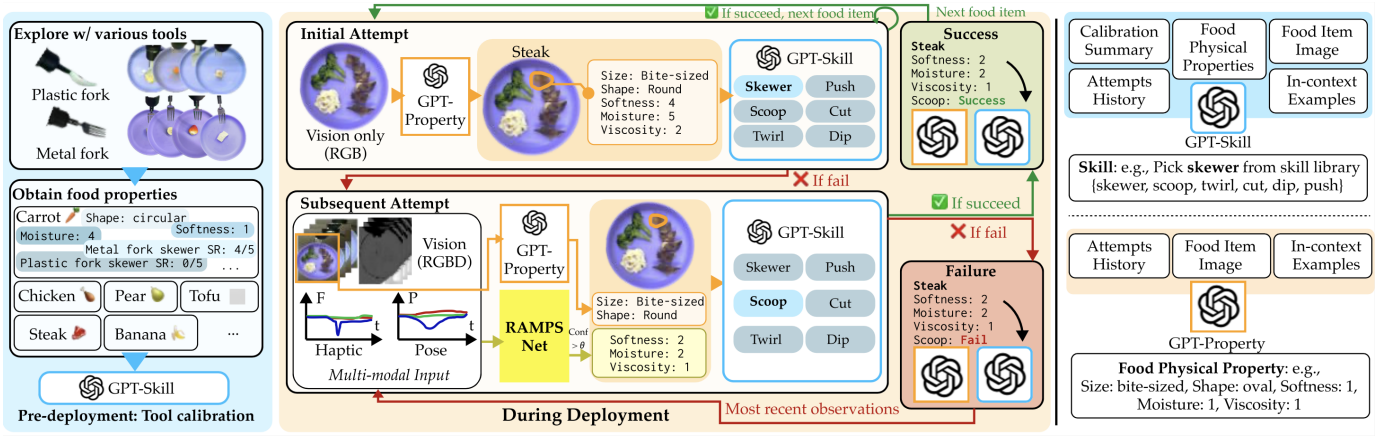

Fig. 2: **RAMPS Framework.** Before deployment, we perform an offline tool calibration phase to understand the affordances of different bite acquisition skills, such as skewering, scooping, and pushing. This calibration involves executing these skills on diverse food items using the corresponding utensils. During deployment, we first use a visually-conditioned language model to estimate the physical properties of food items from visual cues, providing an initial belief. As the robot interacts with the food, we re-estimate online using time-series visual and haptic observations. These updated physical property estimates, combined with tool calibration results, guide the selection of bite acquisition actions to ensure effective and stable bite acquisition.

haptic readings to learn optimal skewering strategies. However, these studies either rely solely on haptic data or utilize only a single image during interaction with the food. In contrast, RAMPS focuses on learning food physical properties from history observations of both vision and haptics, allowing us to capture dynamic changes, such as deformation, surface texture variations, and moisture accumulation. These dynamic changes provide clues for predicting physical properties and enable more effective manipulation skill selection for bite acquisition.

Foundation Models in Robot Manipulation. Foundation models are widely used in high-level planning for robotic manipulation. Most works prompt foundation models with context, including available skills and agent states, to generate action sequences [2, 17, 40, 41, 33, 8, 34, 25]. To provide richer context for better planning, many works provide sensor feedback such as vision and audio information as natural language to foundation models [24, 14]. Building on this, RAMPS extracts semantic information relevant to bite acquisition from visuo-haptic perception and provides this information to foundation models, enabling in-context reasoning.

III. Problem Formulation

We consider the problem of robot-assisted bite acquisition. A robot with a utensil tool on its end-effector is presented with a plate containing multiple bite-size food items. A user specifies the order in which they would like to receive the food items (e.g., through natural language [19]). The robot executes one or more actions to acquire each food item with its tool. If a bite acquisition attempt is successful, we assume that the food is subsequently transferred to the user and removed from the tool. The objective is to maximize the number of food items acquired within a limited number of attempts.

Formally, we frame the problem as a Partially Observable Markov Decision Process (POMDP) [20]. The POMDP is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}_0, \mathcal{O}, \mathcal{T}, \mathcal{Z}, R, L)$ where

- Each **state** $s \in \mathcal{S}$ is characterized by the robot end-effector pose, the target food item, and the positions and physical properties of all the food items on the plate.
- Each **action** $a \in \mathcal{A}$ is a discrete skill from a predefined library: $\{\texttt{push}, \texttt{cut}, \texttt{skewer}, \texttt{dip}, \texttt{scoop}, \texttt{twirl}\}$ [19].
- Each **initial observation** $o_0 \in \mathcal{O}_0$ comprises an RGB-D image $I_0 \in \mathbb{R}^{W \times H \times 4}$ of the plate from a camera mounted on the robot's wrist.
- Each subsequent **observation** $o \in \mathcal{O}$ includes a *time series* of RGB-D images $I \in \mathbb{R}^{T \times W \times H \times 4}$ from the wrist-mounted camera, a corresponding time series of force and torque readings $F \in \mathbb{R}^{T \times 6}$ from an F/T sensor, and a corresponding time series of robot end-effector poses $P \in \mathbb{R}^{T \times 6}$. The length of the time series T varies depending on the execution of the previous skill. The target food item and the positions of the food items on the plate are also directly observed.
- The **transition model** $\mathcal{T}$ and **observation model** $\mathcal{Z}$ are unknown. The **reward function** $R : \mathcal{S} \to \mathbb{R}$ is 1 when the target food item is successfully acquired and 0 otherwise. The **time horizon** $L \in \mathbb{Z}$ is finite.

Importantly, physical food properties are not directly observed, but we hypothesize that the visuo-haptic observations I and F can be used to infer these properties towards completing an estimate of the hidden state s. We consider five physical properties: shape, size, softness, moisture, and viscosity. Shape and size are represented in natural language (e.g., "bite-sized" and "round"). Softness, moisture, and viscosity are each numerical quantities represented on a scale from 1 to 5. We assume that shape and size can be inferred from vision alone, but softness, moisture, and viscosity require integrating vision and haptics. The robot should use its estimates of these

latent physical properties to choose actions that maximize the probability that each target food item is successfully acquired.

IV. RAMPS: Integrating Visuo-Haptic Perception and Common Sense for Bite Acquisition

A. Overview

For robust bite acquisition, we need to understand both the physical properties of food and the general affordances of the utensil tool on the robot's end effector to determine what skills to execute. To understand tool affordances, we start with an offline tool calibration stage where the robot evaluates different skills on diverse food items (Section IV-B). The calibration dataset collected during this stage itself serves as an implicit representation of tool affordances. Then, while still offline, we train RAMPS-Net, a neural network that predicts physical food properties from raw visuo-haptic observations (Section IV-C). Once online, when presented with a new plate of food, we initialize our estimate of food properties with a visually-conditioned language model (VLM) and refine this estimate with RAMPS-Net as interaction data is collected (Section IV-D). To select actions, we prompt a VLM-based planner (Section IV-E) with the tool calibration dataset and the estimated food properties. We detail these steps below.

B. Pre-Deployment: Tool Calibration

The goal of tool calibration is to give the robot a physical understanding of the utensil tool on its end effector. This understanding is important for later determining which skills should be applied to different food items. One potential approach is to learn a complete skill-level transition model, but this is often challenging due to the complexity and variability of interactions [22]. Instead, we capture the transition model implicitly by collecting a small offline calibration dataset with the tool. This dataset consists of randomly-sampled skill executions on diverse food items with manual annotations for the food item, physical properties, and execution outcomes. For example:

```
The robot arm interacts with various food items using a
plastic fork. We summarize the history as follows:

Food Item: Nuts
Properties: Shape: Oval, Size: Bite-sized, Softness: 1,
Moisture: 1, Viscosity: 1
Action with Success Rate: Skewer 0/5, Scoop 3/5, Cut 0/5,
Push 5/5, Dip 5/5

Food Item: Cheese
Properties: Shape: Block, Size: Small, Softness: 5,
Moisture: 2, Viscosity: 5
Action with Success Rate: Skewer 5/5, Scoop 3/5, Cut 5/5,
Push 5/5, Dip 5/5
```

This calibration dataset serves as our implicit representation of tool affordances. We will later include these data in the natural-language format shown above as part of the prompt for VLM-based planning (Section IV-E).

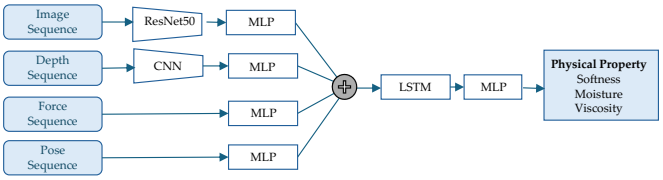

Fig. 3: **RAMPS-Net model architecture.** The network takes images, depths, haptics, and robot end-effector poses as input, and predicts discrete physical property classes.

C. Pre-Deployment: Training RAMPS-Net

While still offline, we next train RAMPS-Net, a neural network that predicts latent physical food properties—softness, moisture, and viscosity—from visuo-haptic observations. We discretize food property values into C classes (e.g., $\texttt{Viscosity} \in \{1, 2, \ldots, 5\}$) so that the problem of predicting food properties is one of multiclass classification. In practice, we set C as 5 for each physical property. The input to RAMPS-Net at time t includes all three time-series observations (I_t, F_t, P_t) and the output is $\psi_t \in \mathbb{R}^{3\times C}$, a vector of log probabilities for each of the 3 predicted food properties: softness, moisture, and viscosity. RAMPS-Net uses separate encoders for each of the time series and further splits the RGB-D inputs into RGB and depth for separate encoding. The encoder for RGB images is a pre-trained ResNet50 followed by a two-layer MLP. The encoder for depth images is a 4-layer convolutional neural network, followed by a two-layer MLP, where each convolutional layer has a 3×3 kernel and is followed by Leaky ReLU activation. The encoder for haptics F_t is a two-layer MLP and the encoder for end-effector poses P_t is a two-layer MLP. Each encoder outputs a vector in $\mathbb{R}^{128}$. The four vectors are concatenated into a unified multimodal representation and then passed to an LSTM with 2 layers and a hidden size of 512. A three-layer MLP takes output from LSTM and produces the final output ψ_t. We show the structure of the model in Figure 3.

To train RAMPS-Net, we first train on an existing dataset [27] and then fine-tune on our own dataset. The existing dataset contains 400 examples of a human skewering a diverse range of food items with varying softness, moisture, and viscosity, such as bagels, mashed potatoes, jelly, and lettuce. Since this dataset only involves skewering and does not involve a robot, we also collect and fine-tune with our own dataset, which contains 300 examples of the robot skewering, pushing, twirling, cutting, dipping, and scooping diverse food items. The latter dataset was collected over a span of three hours. See Appendix B2 for additional training details.

D. During Deployment: State Estimation

With a tool calibration dataset and a trained RAMPS-Net, we are ready to acquire bites online with dishes and food items not seen during training (Figure 2). At the first time step, the robot must rely on visual observations only to initialize a state estimate. As it takes actions towards acquiring a food item (e.g., first `cutting` and then `skewering` a banana), it

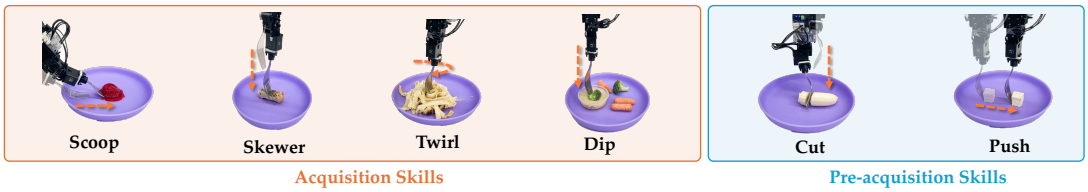

Fig. 4: **Skill library for bite acquisition.** Our framework includes a set of skills categorized into two types: *acquisition skills* (skewer, twirl, scoop, dip), which directly facilitate food pickup, and *pre-acquisition skills* (cut, push), which modify food placement or shape to improve acquisition success. These skills lay the foundation for versatile bite acquisition systems.

gains visuo-haptic information about that food item, which it inputs to RAMPS-Net to generate a refined state estimate when the acquisition attempt fails. After the food item is acquired, a new target is specified. The robot does not yet have haptic information about that food item, but it can nonetheless leverage the full history of interactions towards improved state estimation (e.g., banana slices on the same plate likely have similar properties). We now describe each of these state estimation steps in more detail.

Initializing State Estimates. Given an initial RGB-D observation I_0, we prompt a VLM (GPT-4V [1]) to extract semantic labels of food items (e.g., `['potato', 'chicken', 'onion']`). We then use an open-vocabulary object detector (GroundingDINO [23]) to obtain segmentation masks for all food items. Finally, the VLM is prompted again with these segmentation masks and in-context examples (Appendix C) to generate estimated physical properties for each of the food items. These estimates, together with the directly observed robot end-effector pose and target food item, comprise our initial state estimate $\hat{s}_0$.

Refining State Estimates with RAMPS-Net. We use RAMPS-Net to refine the state estimate for the target food item for each timestep $t > 0$ such that the target food item in o_t is unchanged from o_{t-1}. The time series (I_t, F_t, P_t) in o_t are input to RAMPS-Net, which produces log probability outputs ψ_t. For each predicted property, if the respective log probability is less than a threshold θ_{th}, the prediction is ignored and the property estimate remains unchanged in the state. Otherwise, the prediction is accepted and the property is directly overwritten. The output of this process is a refined state estimate $\hat{s}_t$ where only the target food item properties are potentially modified from $\hat{s}_{t-1}$.

Generating State Estimates for New Targets. For time steps $t > 0$ where the target food item *has* changed, we do not invoke RAMPS-Net because we do not yet have haptic information for the new target. However, we do have an interaction history that we can leverage to generalize between different food items on the plate. We start by following the same detection and segmentation procedure as in initialization to generate segmented masks for the remaining food items. We then build a prompt (Appendix C) for a VLM that includes (i) the segmented food item images; (ii) a natural-language history of acquisition attempts and whether they were successful; (iii) the most recent state estimates $\hat{s}_{t-1}$; and (iv) in-context examples. The VLM directly generates a new state estimate $\hat{s}_t$.

E. During Deployment: Planning

Given the tool calibration dataset and the estimated food state $\hat{s}_t$, we query a VLM (GPT-4V [1]) to select an action from the skill library. Our prompt for the VLM includes (i) the calibration dataset; (ii) a brief description of each skill in the library; (iii) a natural-language history of acquisition attempts; (iv) the segmented food item image; and (v) the estimated physical properties of the target food item. Below, we present an example input prompt (shown in gray) and the corresponding next bite action, generated by the VLM. The full prompting strategy is detailed in Appendix C.

```
< Calibration Summary >
< Skill descriptions >
< History of acquisition attempts >
This is a food item: Mashed Potatoes. <Image>
The robot uses a plastic fork to try picking up the food.
The food physical properties, which range from 1 to 5,
are as follows:
Shape: Circular
Size: bite-sized
Softness: 4
Moisture: 3
Viscosity: 3
Please select an action from ['skewer', 'scoop', 'twirl']
to pick up the food item. Notice that if all acquisition
skills are not immediately feasible, please select an
action from ['cut', 'dip', 'push'] to rearrange
or manipulate items to facilitate subsequent
acquisition. Always follow the format:
Reasoning: <your reason>. Answer: <your answer>.
```

```
Reasoning: The food is soft and moist, making it suitable
to scoop rather than skewer or cut. The viscosity
indicates it will adhere moderately to the fork.
Answer: scoop
```

F. Control

Similar to FLAIR [19], we categorize manipulation skills into two types: *acquisition* and *pre-acquisition* skills. *Acquisition skills* are those that directly pick up food, such as skewering a food item. However, when these actions are not immediately feasible, we employ auxiliary strategies, which we refer to as *pre-acquisition skills*. These actions do not

Fig. 5: **Experimental plate setup.** We prepare 10 plates containing a diverse set of 20 food items, each exhibiting different physical properties such as texture, firmness, and viscosity. These plates are used to evaluate the effectiveness of our bite acquisition framework across a variety of real-world food interactions.

directly acquire food but instead rearrange or manipulate items to facilitate subsequent acquisition.

We follow the skill library from FLAIR [19], which consists of six parameterized manipulation skills, as shown in Figure 4. This includes four acquisition skills (`skewer`, `twirl`, `scoop`, `dip`) and two pre-acquisition skills (`push`, `cut`). For further details on these skill definitions and their execution, please refer to [19].

V. Experiments

In this section, we seek to answer the following questions:

Q1. How does tool calibration affect bite acquisition action selection?

Q2. How much do vision and haptic feedback contribute to estimating the physical properties of food items?

Q3. How effective is action selection based on estimated physical properties?

Q4. How well does RAMPS generalize to unseen food items?

A. Setup

Evaluation Scenarios. We evaluate our approach on a diverse set of food items, including 20 individual food items and 10 in-the-wild dishes (Figure 5). The dishes comprise three categories: (i) fruits and appetizers, (ii) grocery store and restaurant meals, and (iii) homemade dishes.

Baselines. We compare our approach against five methods and baselines, including ablated versions of our approach.

- **RAMPS w/o calibration**. In this baseline, we do not provide calibration information to the VLM planner.
- **Vision-only RAMPS**: This approach differs from ours in that it has no haptic perception.
- **Haptics-only RAMPS**: This approach differs from ours in that it has no visual perception.
- **FLAIR [19]**: This approach selects bite acquisition action based on food category.
- **VLM w/o history**: This method queries a VLM to predict the physical properties of the food from a single post-contact RGB image and subsequently selects a bite acquisition action based on the estimated properties.

Evaluation Metrics. We evaluate each method using two key metrics: *Success Rate* and *Plate Clearance Rate*. A trial is considered successful if the utensil picks up the food item and retains it for at least 5 seconds, after which a human operator removes it. 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 acquisitions relative to total attempts, defined as Success Rate $= \frac{\#\text{Items Acquired}}{\#\text{Total Attempts}}$, capturing acquisition reliability across repeated trials. Plate Clearance Rate measures the proportion of distinct items successfully acquired relative to the total items present on the plate, given by Plate Clearance Rate $= \frac{\#\text{Items Acquired}}{\#\text{Total Items in the Plate}}$, reflecting how effectively a method clears the plate. Together, these metrics provide a comprehensive evaluation of both acquisition efficiency and overall effectiveness.

Hardware. RAMPS is implemented on the Kinova Gen3 robot arm equipped with a motorized feeding utensil mounted at the end-effector [19] (Figure 6). The utensil contains a

TABLE I: **Quantitative results on bite acquisition.** We evaluate our approaches on 10 different dishes and report success rate and coverage. Asterisks (*) denotes unseen food items.

Plate Type				Success Rate				
Plate	**Food Items**	**Visual**	**Haptics**	**RAMPS**	**Haptic-only**	**Vision-only**	**FLAIR**	**VLM**
1	strawberries*, watermelon, carrots	Similar	Diverse	**10/15**	9/17	6/26	9/18	9/19
2	tomatoes*, broccoli, carrots	Similar	Diverse	**7/13**	4/17	4/18	3/15	4/17
3	avocado*, banana, sauce	Diverse	Diverse	**7/11**	7/11	6/14	5/11	7/11
4	muffin*, cake*, jello	Diverse	Diverse	**6/12**	6/13	6/13	5/15	5/14
5	cookie*, bread*, cheese	Diverse	Diverse	**6/13**	6/15	5/21	5/14	7/12
6	roasted turkey, chicken nuggets*, mashed potatoes, green beans*	Diverse	Diverse	**10/18**	8/20	7/29	8/23	8/23
7	salmon, mushrooms	Diverse	Similar	**5/17**	6/17	4/19	5/17	6/14
8	chicken breast, tofu, mushrooms	Diverse	Diverse	**5/9**	4/12	4/16	4/12	5/10
9	chicken, broccoli*, noodles*	Diverse	Diverse	**7/10**	6/12	5/17	7/11	7/10
10	steak*, broccoli*, mashed potatoes	Diverse	Diverse	**6/16**	5/18	5/19	7/16	6/15
Aggregated Success Rate (%)				**51.5 ± 12.7**	40.1 ± 12.2	27.1 ± 8.8	38.2,± 12.2	44.1 ± 14.3
Aggregated Plate Clearance Rate (%)				**87.3 ± 10.0**	77.2 ± 13.0	65.8 ± 11.2	73.4 ± 15.2	81.0 ± 13.1

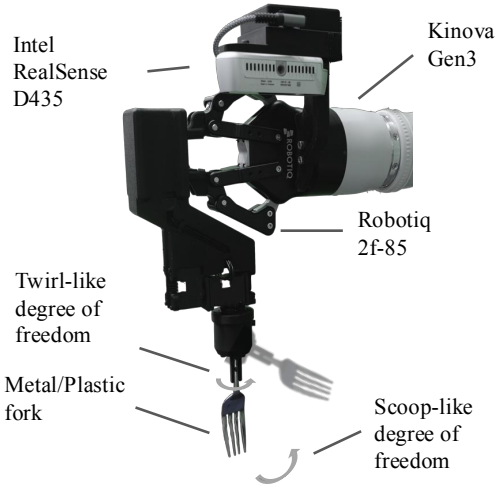

Fig. 6: **System setup for bite acquisition.** Our robotic system consists of a Kinova Gen3 robotic arm equipped with a Robotiq 2F-85 gripper and an Intel RealSense D435 depth camera for visual perception.

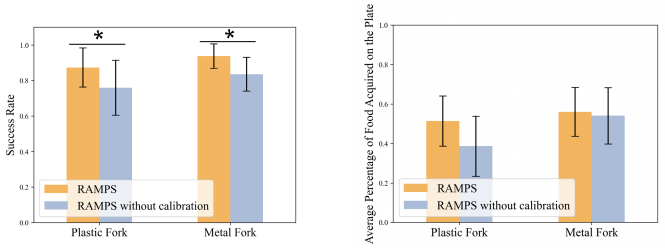

Fig. 7: **System performance using plastic and metal forks.** We report (a) success rate and (b) plate clearance rate of bite acquisition for 10 in-the-wild dishes.

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 for visual perception and an F/T sensor for haptic perception.

B. Evaluating Tool Calibration

We investigate the role of tool calibration in bite acquisition action selection through controlled experiments using two utensils: a plastic fork and a metal fork. As shown in Figure 7, RAMPS without calibration exhibits a significantly lower success rate, acquiring less food items in a plate. We perform chi-square significance tests on the success rates for each utensil. The result suggests RAMPS performs significantly better than RAMPS without calibration ($p < 0.05$). Qualitative observations reveal that RAMPS-Net correctly classifies steak as firm, assigning it a softness score of 2. However, without calibration, RAMPS still selects the skewer action for the plastic fork, erroneously assuming it can pierce through the steak. This performance degradation stems from an inadequate understanding of the utensil's physical capabilities.

For the metal fork, calibration primarily enhances bite acquisition success for food items with high softness and moisture content. Without tool calibration, RAMPS tends to default to skewering most food items, even when inappropriate. In particular, the metal fork struggles with soft, high-moisture foods like tofu, which frequently slip during skewering. With tool calibration, our method better captures the interaction between the utensil and the food, leading to more effective

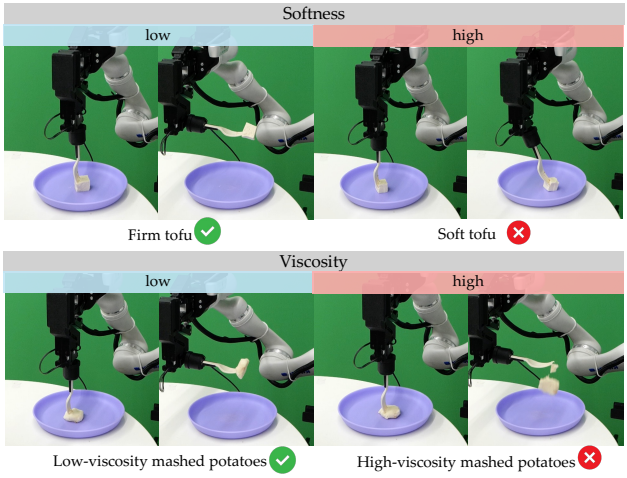

Fig. 8: **Effect of food physical properties on utensil interactions.** The robot skewers food items of varying *softness* (top) and *viscosity* (bottom). Soft tofu and low-viscosity mashed potatoes are successfully acquired, while firm tofu and high-viscosity mashed potatoes lead to failure, illustrating the challenges of bite acquisition.

bite acquisition action selection.

C. Evaluating State Estimation

To examine the role of visual and haptic perception in our framework, we compare RAMPS with its vision-only and haptics-only variants using plastic fork. As shown in Table I, the vision-only method significantly underperforms compared to RAMPS across different plates. This is primarily due to its inability to capture fine-grained local interactions between the utensil and food, as well as its difficulty in distinguishing visually similar food items with different physical properties.

For example, on Plate 1, strawberries, watermelon, and baby carrots share a similar color but exhibit vastly different physical properties. Strawberries, being softer, can be skewered, whereas carrots are too firm—attempting to skewer them may cause them to break apart, making scooping a more suitable action. The vision-only variant frequently confuses these food items and misestimates their physical properties, sometimes predicting unrealistically high softness for carrots. Additionally, we observe that the vision-only model is highly sensitive to the presence of other food items on the plate, likely due to overfitting to visual cues in the training set.

The haptics-only model demonstrates stronger performance than the vision-only variant, achieving a success rate of up to 40.1% but still underperforming with respect to RAMPS. We examine its failure modes and find that most failures occur with extremely firm or soft food items, such as extra soft tofu, nuts, raw carrots, and cookies. This is because haptic readings and pose features—are highly similar across these cases. Specifically, in such cases, the forces increase rapidly with slight changes in the end-effector pose. For extremely soft food items, the utensil often skewers through the food and applies force directly to the plate. In contrast, for firm food items, the utensil applies force directly to the surface of the food. However, due to the plastic utensil's properties, applying significant force deforms the utensil, which slightly alters the end-effector pose. As a result, haptic features appear similar for these extreme cases. To effectively distinguish between them, visual information becomes crucial for observing the deformation of both the food item and the utensil. Overall, we find that RAMPS, which integrates both vision and haptics, achieves the highest success rate of 51.5% and successfully picks up an average of 87.3% of food items on the plate.

D. Evaluating Action Selection Policy

In this section, we evaluate the effectiveness of our physical property-based action selection policy by comparing RAMPS with FLAIR and VLM. Our objective is to assess whether explicitly estimating food properties improves bite acquisition success over category-based or purely vision-based methods.

Both FLAIR and VLM rely on a single post-contact RGB image to select a bite acquisition skill. While FLAIR makes decisions based on predefined food categories, VLM instead infers food properties directly from visual input. As shown in Table I, VLM achieves a 44.1% success rate—slightly outperforming FLAIR—and successfully picks up 81.0% of food items, demonstrating more fine-grained skill selection than the category-based approach. For instance, when handling baby carrots and cookies, FLAIR rigidly applies the skewer action based on categorical reasoning, whereas VLM recognizes their firmness and selects scooping when appropriate. Similarly, for watermelon, VLM occasionally detects its high moisture content and opts for scooping to minimize juice spillage.

However, relying solely on vision introduces biases and inconsistencies. VLM frequently misclassifies avocado as extremely soft and defaults to scooping. While scooping is still a viable action, it is less robust than skewering and leads to additional failed attempts. This limitation highlights the necessity of real-time adaptation based on both visual and haptic feedback. Figure 8 further illustrates how food properties influence bite acquisition. For example, soft tofu easily slips from the utensil, whereas firm tofu adheres when skewered. Similarly, mashed potatoes with high viscosity fail to stick to the utensil.

Unlike the baselines, RAMPS continuously refines its physical property estimates using visuo-haptic perception. As shown in Table I, it achieves the highest overall success rate. Its advantage is particularly evident on Plate 6, where chicken nuggets pose a unique challenge due to their temperature-dependent firmness (Figure 9). Initially hot, the nuggets gradually cool and become firmer over time while maintaining the same visual appearance. RAMPS quickly adapts by detecting these changes and adjusting its strategy, switching from skewering to scooping as needed. Furthermore, since our VLM retains historical context, it generalizes this observation to other chicken nuggets on the plate, anticipating their firmness. In contrast, the baselines fail to recognize these physical changes, persist with skewering, and often struggle to pick up the nuggets—sometimes even causing the plastic fork to break.

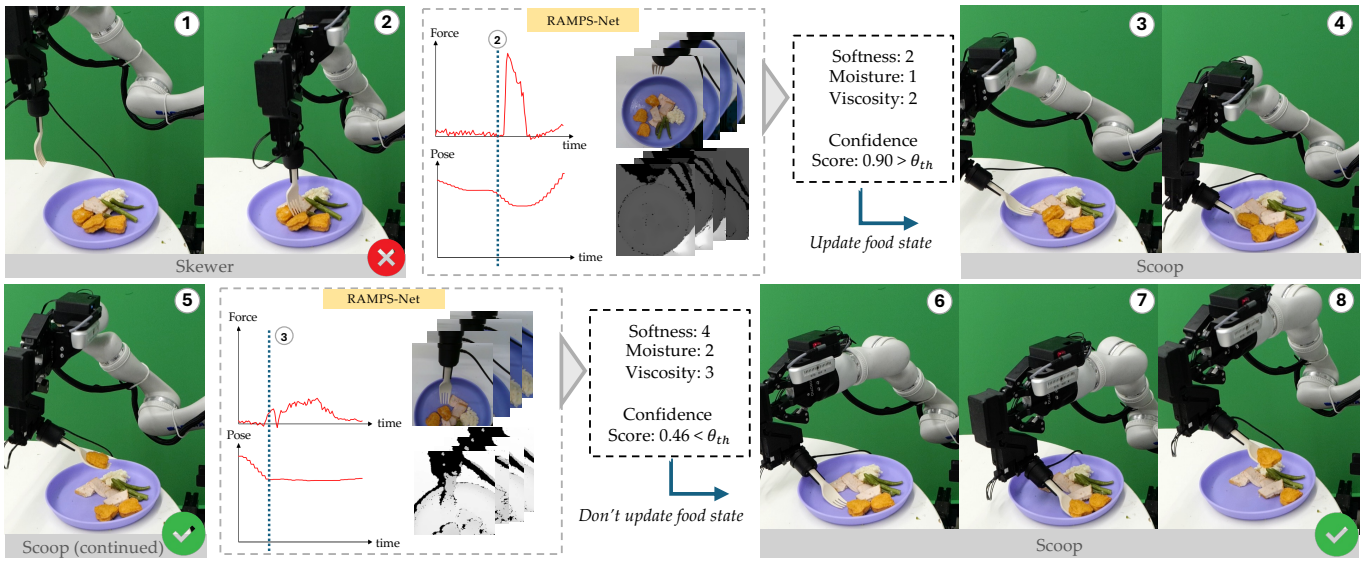

Fig. 9: **Qualitative results on bite acquisition.** The robot first executes the skewer action based on the initial food physical property estimation but fails to pick up the food item in step 2. The vision, haptic, and depth data from this attempt are processed by RAMPS-Net, which provides an updated physical property estimation with a high confidence score. If the confidence score exceeds a threshold, the food state is updated accordingly. A VLM planner then selects the scoop skill based on the latest estimated state in step 3.

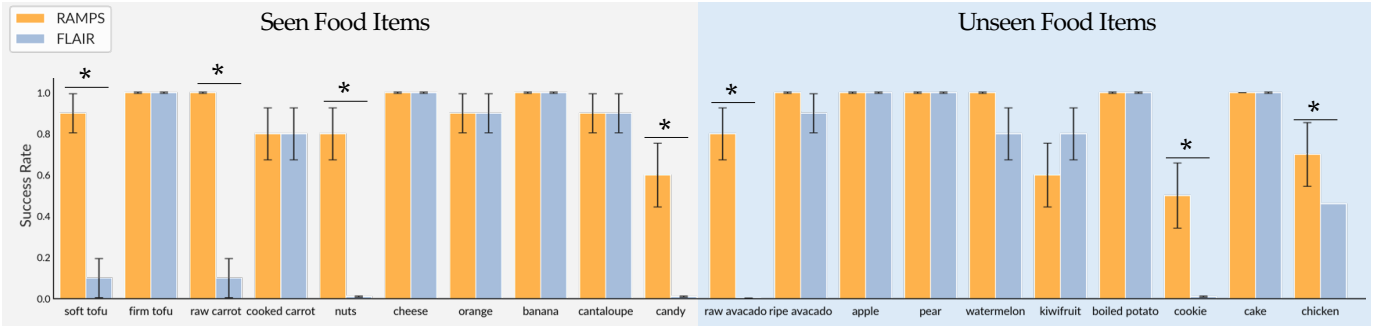

Fig. 10: **Generalization performance on seen and unseen food items.** We compare RAMPS and FLAIR across 20 food items, evaluating success rates on both seen (left) and unseen (right) categories. RAMPS demonstrates strong performance on unseen items, maintaining comparable success rates to seen items. Asterisks (∗) indicate statistically significant differences.

These results demonstrate that explicit, online estimation of physical properties is essential for robust and adaptive bite acquisition, outperforming both category-based and purely vision-driven approaches.

E. Evaluating the Generalizability of RAMPS

To assess the generalizability of RAMPS, we test it on 20 single food items, 10 of which are previously unseen and in-the-wild dishes. As shown in Figure 10, RAMPS achieves an approximately 80% success rate on these novel food items, demonstrating its ability to generalize beyond the training set.

Compared to FLAIR, RAMPS shows notable improvement on foods such as cookies and raw avocados. RAMPS-Net effectively differentiates avocados with varying ripeness levels and selects appropriate bite acquisition actions. However, RAMPS slightly underperforms FLAIR on kiwifruit, as the thin, soft pieces lead to rapid force increases when skewered. This abrupt change misleads RAMPS-Net to classify kiwifruit as firm and favor scooping instead of skewering.

Generalization errors primarily arise from slippage, which introduces noise into haptic feedback and disrupts physical property estimation. In Plate 7, for example, the oily surfaces of salmon and mushrooms cause them to slip upon contact, leading to inconsistent physical property predictions and, consequently, unreliable skill selection. These results highlight the challenge of handling out-of-distribution variations in surface properties and suggest that improving robustness to slippage could further enhance RAMPS's generalization capabilities.

VI. Limitations

Our approach effectively leverages multi-modal sensing to estimate food physical properties, enabling adaptive bite

acquisition. However, several areas remain for improvement. First, RAMPS-Net estimates properties only for the object in direct contact with the utensil, whereas interactions with one object could provide information about others. Extending our method with an explicit 3D scene representation could enhance multi-object reasoning [3]. Second, slippage can introduce noise in haptic signals, leading to inaccurate physical property estimates, which could be mitigated by incorporating contact feedback for refinement. Lastly, while our system already performs well with open-loop execution, it could further benefit from a closed-loop low-level policy to enhance real-time adaptability. This would complement our contribution in food state estimation and adaptive skill selection. These directions present promising avenues for future work.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Christopher Agia, Toki Migimatsu, Jiajun Wu, and Jeannette Bohg. Stap: Sequencing task-agnostic policies. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7951–7958, 2023. doi: 10.1109/ICRA48891.2023.10160220.

[3] Bo Ai, Stephen Tian, Haochen Shi, Yixuan Wang, Cheston Tan, Yunzhu Li, and Jiajun Wu. Robopack: Learning tactile-informed dynamics models for dense packing. *Robotics: Science and Systems (RSS)*, 2024. URL https://arxiv.org/abs/2407.01418.

[4] Tapomayukh Bhattacharjee, Gilwoo Lee, Hanjun Song, and Siddhartha S. Srinivasa. Towards robotic feeding: Role of haptics in fork-based food manipulation. *IEEE Robotics and Automation Letters*, 4(2):1485–1492, 2019. doi: 10.1109/LRA.2019.2894592.

[5] Tapomayukh Bhattacharjee, Gilwoo Lee, Hanjun Song, and Siddhartha S Srinivasa. Towards robotic feeding: Role of haptics in fork-based food manipulation. *IEEE Robotics and Automation Letters*, 4(2):1485–1492, 2019.

[6] Tadhg Brosnan and Da-Wen Sun. Improving quality inspection of food products by computer vision—a review. *Journal of Food Engineering*, 61(1):3–16, 2004. ISSN 0260-8774. doi: https://doi.org/10.1016/S0260-8774(03)00183-3. URL https://www.sciencedirect.com/science/article/pii/S0260877403001833. Applications of computer vision in the food industry.

[7] Chenyu Dong, Liangliang Yu, Masaru Takizawa, Shunsuke Kudoh, and Takashi Suehiro. Food peeling method for dual-arm cooking robot. pages 801–806, 01 2021. doi: 10.1109/IEEECONF49454.2021.9382700.

[8] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.

[9] Ryan Feng, Youngsun Kim, Gilwoo Lee, Ethan K Gordon, Matt Schmittle, Shivaum Kumar, Tapomayukh Bhattacharjee, and Siddhartha S Srinivasa. Robot-assisted feeding: Generalizing skewering strategies across food items on a plate. In *The International Symposium of Robotics Research*, pages 427–442. Springer, 2019.

[10] Ethan K Gordon, Xiang Meng, Tapomayukh Bhattacharjee, Matt Barnes, and Siddhartha S Srinivasa. Adaptive robot-assisted feeding: An online learning framework for acquiring previously unseen food items. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9659–9666. IEEE, 2020.

[11] Ethan K Gordon, Sumegh Roychowdhury, Tapomayukh Bhattacharjee, Kevin Jamieson, and Siddhartha S Srinivasa. Leveraging post hoc context for faster learning in bandit settings with applications in robot-assisted feeding. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10528–10535. IEEE, 2021.

[12] Ethan Kroll Gordon, Amal Nanavati, Ramya Challa, Bernie Hao Zhu, Taylor Annette Kessler Faulkner, and Siddhartha Srinivasa. Towards general single-utensil food acquisition with human-informed actions. In *Conference on Robot Learning*, pages 2414–2428. PMLR, 2023.

[13] Jennifer Grannen, Yilin Wu, Suneel Belkhale, and Dorsa Sadigh. Learning bimanual scooping policies for food acquisition. In *6th Annual Conference on Robot Learning*, 2022. URL https://openreview.net/forum?id=qDtbMK67PJG.

[14] Yanjiang Guo, Yen-Jen Wang, Lihan Zha, and Jianyu Chen. Doremi: Grounding language model by detecting and recovering from plan-execution misalignment. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12124–12131. IEEE, 2024.

[15] Nayoung Ha, Ruolin Ye, Ziang Liu, Shubhangi Sinha, and Tapomayukh Bhattacharjee. Repeat: A real2sim2real approach for pre-acquisition of soft food items in robot-assisted feeding. 2024.

[16] Eric Heiden, Miles Macklin, Yashraj S Narang, Dieter Fox, Animesh Garg, and Fabio Ramos. DiSECt: A Differentiable Simulation Engine for Autonomous Robotic Cutting. In *Proceedings of Robotics: Science and Systems*, Virtual, July 2021. doi: 10.15607/RSS.2021.XVII.067.

[17] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In *arXiv preprint arXiv:2207.05608*, 2022.

[18] Rajat Kumar Jenamani, Daniel Stabile, Ziang Liu, Abrar Anwar, Katherine Dimitropoulou, and Tapomayukh Bhattacharjee. Feel the bite: Robot-assisted inside-mouth bite transfer using robust mouth perception and physical interaction-aware control. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 313–322, 2024.

[19] Rajat Kumar Jenamani, Priya Sundaresan, Maram Sakr, Tapomayukh Bhattacharjee, and Dorsa Sadigh. Flair: Feeding via long-horizon acquisition of realistic dishes. *arXiv preprint arXiv:2407.07561*, 2024.

[20] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artif. Intell.*, 101(1-2): 99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X. URL https://doi.org/10.1016/S0004-3702(98)00023-X.

[21] Ian Lenz, Ross A Knepper, and Ashutosh Saxena. Deepmpc: Learning deep latent features for model predictive control. In *Robotics: Science and Systems*, volume 10, page 25. Rome, Italy, 2015.

[22] Sizhe Li, Zhiao Huang, Tao Chen, Tao Du, Hao Su, Joshua B. Tenenbaum, and Chuang Gan. Dexdeform: Dexterous deformable object manipulation with human demonstrations and differentiable physics. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=LIV7-_7pYPl.

[23] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[24] Zeyi Liu, Arpit Bahety, and Shuran Song. Reflect: Summarizing robot experiences for failure explanation and correction. *arXiv preprint arXiv:2306.15724*, 2023.

[25] Joel Loo, Zhanxin Wu, and David Hsu. Open scene graphs for open world object-goal navigation. *arXiv preprint arXiv:2407.02473*, 2024.

[26] Silvia B Matiacevich, Domingo Mery, and Franco Pedreschi. Prediction of mechanical properties of corn and tortilla chips by using computer vision. *Food and Bioprocess Technology*, 5:2025–2030, 2012.

[27] Amal Nanavati, Ramya Challa, Ethan K. Gordon, and Siddhartha S. Srinivasa. A Dataset of Food Manipulation Strategies for Diverse Foods, 2022. URL https://doi.org/10.7910/DVN/C8SI1D.

[28] Amal Nanavati, Patricia Alves-Oliveira, Tyler Schrenk, Ethan K. Gordon, Maya Cakmak, and Siddhartha S. Srinivasa. Design principles for robot-assisted feeding in social contexts. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, page 24–33, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399647. doi: 10.1145/3568162.3576988. URL https://doi.org/10.1145/3568162.3576988.

[29] Amal Nanavati, Ethan K Gordon, Taylor A Kessler Faulkner, Yuxin (Ray) Song, Johnathan Ko, Tyler Schrenk, Vy Nguyen, Bernie Hao Zhu, Haya Bolotski, Atharva Kashyap, Sriram Kutty, Raida Karim, Liander Rainbolt, Rosario Scalise, Hanjun Song, Ramon Qu, Maya Cakmak, and Siddhartha S Srinivasa. Lessons learned from designing and evaluating a robot-assisted feeding system for out-of-lab use. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*, 2025.

[30] Haozhi Qi, Brent Yi, Sudharshan Suresh, Mike Lambeta, Yi Ma, Roberto Calandra, and Jitendra Malik. General In-Hand Object Rotation with Vision and Touch. In *Conference on Robot Learning (CoRL)*, 2023.

[31] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[32] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, pages 894–906. PMLR, 2022.

[33] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530, 2023. doi: 10.1109/ICRA48891.2023.10161317.

[34] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023.

[35] Priya Sundaresan, Suneel Belkhale, and Dorsa Sadigh. Learning visuo-haptic skewering strategies for robot-assisted feeding. In *6th Annual Conference on Robot Learning*, 2022. URL https://openreview.net/forum?id=ILq09gVoaTE.

[36] Priya Sundaresan, Jiajun Wu, and Dorsa Sadigh. Learning sequential acquisition policies for robot-assisted feeding. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 1282–1299. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/sundaresan23b.html.

[37] Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, Joseph Ortiz, and Mustafa Mukadam. Neural feels with neural fields: Visuo-tactile perception for in-hand manipulation. *Science Robotics*, page adl0628, 2024.

[38] Yen-Ling Tai, Yu Chien Chiu, Yu-Wei Chao, and Yi-

Ting Chen. Scone: A food scooping robot learning framework with active perception. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 849–865. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/tai23a.html.

[39] World Health Organization. *Global report on health equity for persons with disabilities*. World Health Organization, 2022. ISBN 9789240063600. URL https://www.who.int/publications/i/item/9789240063600.

[40] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 2023.

[41] Zhanxin Wu, Bo Ai, and David Hsu. Integrating common sense and planning with large language models for room tidying. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[42] Zhenjia Xu, Zhou Xian, Xingyu Lin, Cheng Chi, Zhiao Huang, Chuang Gan, and Shuran Song. Roboninja: Learning an adaptive cutting policy for multi-material objects. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[43] Akihiko Yamaguchi and Christopher G. Atkeson. Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pages 1045–1051, 2016. doi: 10.1109/HUMANOIDS.2016.7803400.

[44] Ruolin Ye, Yifei Hu, Yuhan Anjelica Bian, Luke Kulm, and Tapomayukh 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)*, pages 9540–9547, 2024. doi: 10.1109/ICRA57147.2024.10610050.

[45] Kevin Zhang, Mohit Sharma, Manuela Veloso, and Oliver Kroemer. Leveraging multimodal haptic sensory data for robust cutting. In *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*, pages 409–416. IEEE, 2019.