SAVOR: Skill Affordance Learning from Visuo-Haptic Perception for Robot-Assisted Bite Acquisition

Zhanxin Wu[1] **Bo Ai**[2] **Tom Silver**[1] **Tapomayukh Bhattacharjee**[1]
[1]Cornell University [2]UC San Diego

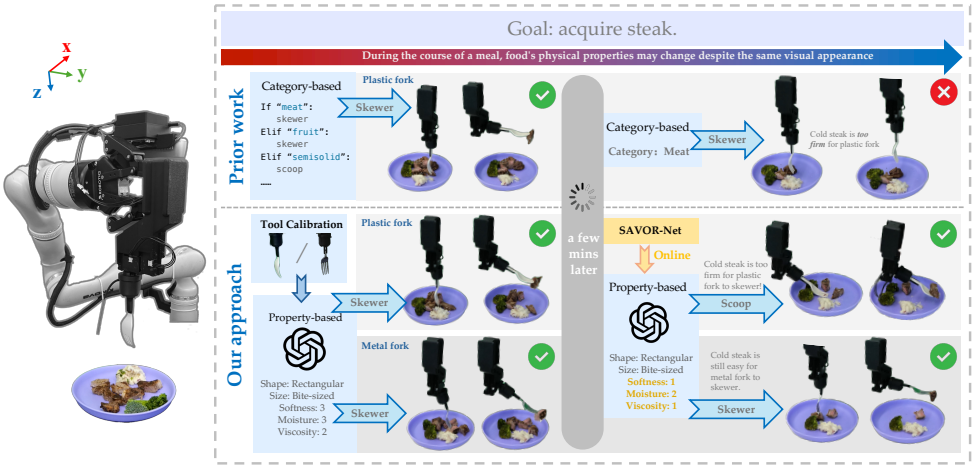

Figure 1: We propose **SAVOR**, a method that combines tool affordances and food affordances to select the appropriate manipulation skill for robust bite acquisition.

Abstract: Robot-assisted feeding requires reliable bite acquisition, a challenging task due to the complex interactions between utensils and food with diverse physical properties. These interactions are further complicated by the temporal variability of food properties—for example, steak becomes firm as it cools even during a meal. To address this, we propose SAVOR, a novel approach for learning skill affordances for bite acquisition—how suitable a manipulation skill (e.g., skewering, scooping) is for a given utensil-food interaction. In our formulation, skill affordances arise from the combination of tool affordances (what a utensil can do) and food affordances (what the food allows). Tool affordances are learned offline through calibration, where different utensils interact with a variety of foods to model their functional capabilities. Food affordances are characterized by physical properties such as softness, moisture, and viscosity, initially inferred through commonsense reasoning using a visually-conditioned language model and then dynamically refined through online visuo-haptic perception using SAVOR-Net during interaction. Our method integrates these offline and online estimates to predict skill affordances in real time, enabling the robot to select the most appropriate skill for each food item. Evaluated on 20 single food items and 10 in-the-wild meals, our approach improves bite acquisition success rate by 13% over state-of-the-art (SOTA) category-based methods (e.g. use skewer for fruits). These results highlight the importance of modeling interaction-driven skill affordances for generalizable and effective robot-assisted bite acquisition. Website: https://emprise.cs.cornell.edu/savor

Keywords: Assistive Robotics, Visuo-haptic Perception, Affordance Learning

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

1 Introduction

Eating is a fundamental human activity, yet millions struggle to feed themselves due to mobility limitations [1]. A robot-assisted feeding system has the potential to help them regain independence and dignity while ensuring their needs are met reliably [2]. Bite acquisition, the process of picking up a food item from a plate or bowl, is a critical step in robot-assisted feeding, but it is highly challenging: (i) Food items exhibit diverse and temporally variable physical properties (e.g., rice becomes firm when it cools down, and tofu is fragile and can break without careful manipulation). (ii) Physical interaction with food items varies significantly for different utensils. To address these challenges, a robot must reason about three types of affordances. First, **food affordances** describe what the food allows, such as whether the food can be skewered or scooped. Second, **tool affordances** characterize what a utensil can do, given its functionality. **Skill affordances** arise from reasoning jointly over food and tool affordances and capture whether a manipulation skill is appropriate, given the food's physical properties and the tool's capabilities.

To this end, our key insight is that *a better understanding of skill affordances—achieved by combining calibrated tool affordances and commonsense food affordances that are updated using online information gathering—enhances bite acquisition across diverse food items*. Take humans as an example. We begin with an understanding of tool affordance, e.g., knowing that a plastic fork may not penetrate firm foods. Upon seeing a food item, we form initial beliefs about its food affordances from visual cues, such as assuming a piece of steak is moderately soft and suitable for skewering. However, these assumptions are not always accurate. During interaction, visual and haptic feedback can reveal unexpected firmness (e.g., a well-done steak), prompting an adjustment of the chosen skill. This process highlights the importance of (i) understanding tool capabilities, (ii) estimating food properties visually, and (iii) refining estimates through visuo-haptic feedback for skill selection.

Building on these insights, we propose SAVOR, an approach for learning *skill affordances* from tool affordances and food affordances for bite acquisition (Figure 2). The system operates in two stages: (i) **Prior to deployment**, we learn *tool affordances* through offline calibration, where different utensils interact with various foods to evaluate manipulation skills to model tool capabilities. (ii) **During deployment**, we estimate *food affordances* by inferring food physical properties, including softness, moisture, and viscosity. We initially estimate food properties through commonsense reasoning using a visually-conditioned language model, and dynamically refine them through online visuo-haptic perception with our developed SAVOR-Net during the interaction. Together, combining offline tool affordances and online food affordances guides robust skill selection for bite acquisition.

Overall, our contributions are: (i) an **algorithm** that learns food affordances based on physical properties, initially informed by commonsense priors and dynamically refined through online visuo-haptic perception; (ii) a **method** that characterizes the affordances of manipulation skills across diverse food items and tools and uses the learned model for adaptive skill selection; (iii) a **comprehensive evaluation** on 20 single food items and 10 in-the-wild dishes, showing improved bite acquisition success over state-of-the-art approaches; and (iv) a **dataset** of bite acquisition trials, which is the first to involve varied manipulation skills applied to diverse foods, and provides synchronized visual and haptic signals throughout the trajectory.

2 Related Work

Food Manipulation for Robot-Assisted Feeding. Feeding [3–7] is an important activity of daily living (ADL). One crucial step in feeding is bite acquisition, which involves executing various skills, such as skewering [8–12], twirling [13], and scooping [14, 15]. To compose these skills, VAPORS [13] focuses on noodle dishes and employs physics-based simulations for decision-making between twirling and grouping noodles. The closest relevant work, FLAIR [3], proposes to select skills by food categories inferred from only visual observations. Such a category-based approach overlooks the intricate physical properties of food, often resulting in failures with items of varying softness. To address this, SAVOR selects manipulation skills based on estimated food affordances to enable more adaptive bite acquisition across diverse food items.

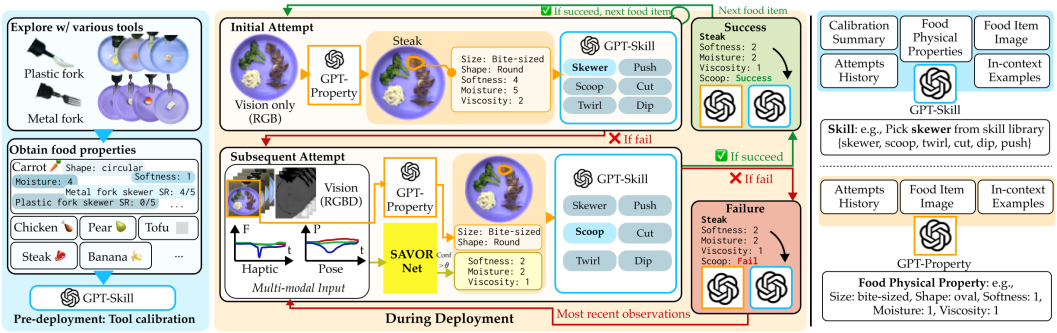

Figure 2: **SAVOR Framework.** Before deployment, we perform an offline tool calibration to understand tool affordances. During deployment, we first use a visually-conditioned language model to estimate food physical properties and then refine it through online visuo-haptic perception.

Integrating Vision and Haptics in Robotics. Vision-only manipulation has demonstrated effectiveness in robotic domains such as semantic grasping and deformable object manipulation [9, 16–19]. However, contact-rich tasks, such as in-hand manipulation [20, 21] and object packing [22], benefit significantly from the integration of vision and haptic information. Previous studies [5, 10, 23–31] have highlighted the importance of haptic feedback in food manipulation. More recently, Sundaresan et al. [8] and Gordon et al. [10] combine a single image with time-series haptics to learn skewering skills, but rely either on haptics alone or static visual input. In contrast, SAVOR focuses on learning food physical properties from time-series observations of both vision and haptics, allowing us to capture dynamic changes, such as deformation and surface texture variations, which enables a more comprehensive understanding of the food physical properties and their temporal variations.

Foundation Models in Robot Manipulation. Foundation models [32] are widely used for language-conditioned planning in robotics by prompting them with task context, available skills, and agent state information [33–44]. Similarly, SAVOR leverages a visually-conditioned language model to form initial beliefs about food physical properties from an image. These estimates are then dynamically refined through visuo-haptic perception during interaction.

3 Problem Formulation

We consider the problem of bite acquisition, where a robot equipped with a utensil needs to acquire food items from a plate in a specified sequence. The user specifies the desired food sequence (e.g., via natural language [3]), and the robot executes one or more actions per item to acquire it. Upon a successful attempt, the item is removed from the tool, and the robot proceeds to the next. The objective is to maximize the number of food items acquired within a limited number of attempts.

We formulate this as a Partially Observable Markov Decision Process (POMDP) [45], defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}_0, \mathcal{O}, \mathcal{T}, \mathcal{Z}, R, L)$. The state $s \in \mathcal{S}$ comprises the robot end-effector pose, the target food item, and the positions and physical properties of all food items on the plate. Actions $a \in \mathcal{A}$ are discrete skills selected from a predefined library: {`push`, `cut`, `skewer`, `dip`, `scoop`, `twirl`} [3]. The initial observation $o_0 \in \mathcal{O}_0$ consists of a single RGB-D image $I_0 \in \mathbb{R}^{W \times H \times 4}$ from a wrist-mounted camera. Subsequent observations $o \in \mathcal{O}$ include a time series of RGB-D images $I \in \mathbb{R}^{T \times W \times H \times 4}$, force-torque readings $F \in \mathbb{R}^{T \times 6}$, and end-effector poses $P \in \mathbb{R}^{T \times 6}$. The length of the time series T varies depending on the execution of the skill. The target food item and its position are assumed to be observable. The transition model $\mathcal{T}$ and observation model $\mathcal{Z}$ are unknown. The reward function $R : \mathcal{S} \to \mathbb{R}$ returns 1 for successful acquisition of the target item, and 0 otherwise. The time horizon $L \in \mathbb{Z}$ is finite.

Importantly, the physical properties of food 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 partially-observable state s. We consider five physical properties: shape, size, softness, moisture, and viscosity. We assume that shape and size can be estimated from vision alone, while softness, moisture, and viscosity require integrating vision and haptic feedback. Shape and size are represented as categorical descriptors in natural language (e.g., "bite-sized," "round"), and the re-

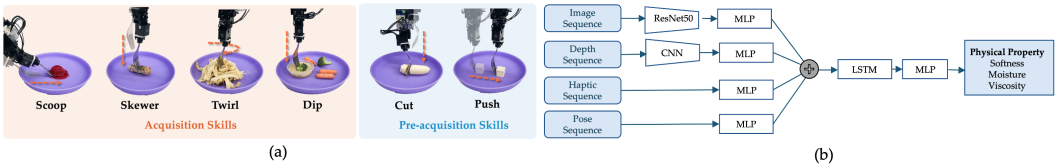

Figure 3: **(a) Skill library for bite acquisition. (b) SAVOR-Net model architecture.**

maining properties are scalar values from 1 to 5, similar to a 5-point Likert scale. For example, a softness score of 1 indicates very hard, while a score of 5 indicates very soft. We hypothesize that the visuo-haptic observations (I, F) provide sufficient information to infer these physical properties. The robot should use its estimates of these latent physical properties to guide skill selection and maximize bite acquisition success.

4 SAVOR

Robust bite acquisition requires reasoning over both food affordances and tool affordances. To estimate tool affordances, we begin with an offline calibration phase, where the robot executes various skills on diverse food items (Section 4.1). The resulting calibration dataset serves as an implicit representation of the tool's affordances. In parallel, we offline train SAVOR-Net, a neural network that predicts physical food properties from raw visuo-haptic observations (Section 4.2). At test time, given a new plate of food, we first initialize food physical property estimates using a visually-conditioned language model (VLM), and refine them online with SAVOR-Net (Section 4.3). For action selection, we prompt a VLM-based planner with the estimated food properties and the tool calibration dataset to generate a skill sequence (Section 4.4). We detail each component below.

4.1 Pre-Deployment: Offline Tool Calibration

The goal of tool calibration is to provide the robot with contextual knowledge of the utensil's affordances. We capture tool affordances implicitly by collecting a small offline calibration dataset consisting of randomly sampled skill executions on diverse food items, annotated by humans with utensil type, food type, 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, Shape: Oval, Size: Bite-sized, Softness: 1,  Moisture: 1, Viscosity: 1
Skill with Success Rate: Skewer 0/5, Scoop 3/5, Cut 0/5, Push 5/5, Dip 5/5
```

This calibration dataset serves as an implicit representation of tool affordances, particularly the success rates of various skills with a specific tool, expressed in natural language and later used as input for VLM-based planning (Section 4.4).

4.2 Pre-Deployment: Offline Training SAVOR-Net

While still offline, we next train SAVOR-Net, a neural network that predicts physical food properties (i.e., softness, moisture, and viscosity) from visuo-haptic observations. We discretize food property values into $C = 5$ levels (e.g., `Viscosity` $\in \{1, 2, \ldots, 5\}$). The input to SAVOR-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. We select these three properties based on prior work [4, 46] and insights obtained by querying a VLM about which food physical properties are relevant for the bite acquisition task. SAVOR-Net uses separate encoders for each of the time series and further splits the RGB-D inputs into RGB and depth for separate encoding. 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 the LSTM and produces the final output ψ_t. We present the structure of the model in Figure 3.

We pretrain SAVOR-Net on an existing dataset [47], which contains 400 examples of human skewering various food items, and then fine-tune it on our own dataset, consisting of 300 examples of a

robot performing skewering, pushing, twirling, cutting, dipping, and scooping across 20 food items. More details about the model and the training process are available in Appendix B.3.

4.3 During Deployment: State Estimation

Given a tool calibration dataset and a trained SAVOR-Net, the robot is ready to acquire bites from novel dishes and food items (Figure 2). At the first time step, it initializes a state estimate using only visual observations. As the robot attempts to acquire a target item (e.g., `cutting` then `skewering` a banana), it collects visuo-haptic data for SAVOR-Net to refine the estimate, which is especially important when the attempt fails. Once the item is acquired, a new target is specified. Although the robot initially lacks physical knowledge of the new item, it can leverage prior interaction history for improved estimation (e.g., banana slices on the same plate likely share physical properties). We now describe the state estimation process in more detail.

Initializing State Estimates. Given an initial RGB-D observation I_0, we prompt a VLM (GPT-4V [48]) to extract semantic labels of food items (e.g., `['potato', 'chicken']`). We then use GroundingDINO [49] to obtain segmentation masks for them. Finally, we prompt VLM again with the masks and in-context examples (Appendix B.4) to estimate physical properties for each item. These estimates are used by a VLM planner to select an appropriate skill (Section 4.4).

Refining State Estimates with SAVOR-Net. If the skill (Section 4.4) selected based on the initial estimate of food physical properties fails, we use SAVOR-Net to refine the state estimate for the target food item for each timestep $t > 0$. The time series (I_t, F_t, P_t) in o_t are input to SAVOR-Net, which produces log probability outputs ψ_t. For each predicted property, if its log probability is less than a threshold θ_{th}, the prediction is ignored. Otherwise, the estimate is updated. This process yields a refined $\hat{s}_t$, where only the target item's properties are potentially modified from $\hat{s}_{t-1}$.

Generating State Estimates for New Targets. When the target item changes at timestep $t > 0$ and no interaction data is available, we initialize the estimate using prior outcomes on similar items. We run the same detection and segmentation pipeline, then prompt a VLM (Appendix B.4) with: (i) segmented images, (ii) past attempt summaries, (iii) the previous estimate $\hat{s}_{t-1}$, and (iv) in-context examples. The VLM predicts $\hat{s}_t$ by matching the new item's visual features to similar examples. This estimate is later refined through interaction, using SAVOR-Net as mentioned above.

4.4 During Deployment: Planning

Given the tool calibration dataset and the estimated food state $\hat{s}_t$, we query a VLM (GPT-4V [48]) to select a skill from the skill library. Our prompt for the VLM includes (i) the calibration dataset; (ii) a brief description of each skill; (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 prompt. The full prompting strategy is detailed in Appendix B.4.

```
[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', 'dip'] to pick up the food item or
select one action from ['cut', 'push'] to manipulate items to facilitate subsequent
acquisition. Always follow the format: Reasoning: <your reason>. Answer: <your answer>.
```

For the low-level skills, we use the skill library from FLAIR [3], which includes four acquisition skills (`skewer`, `twirl`, `scoop`, `dip`) and two pre-acquisition skills (`cut`, `push`), illustrated in Figure 3. Given these low-level skills, SAVOR enables effective skill selection for each food item, despite variations in physical properties and the partial observability present during deployment.

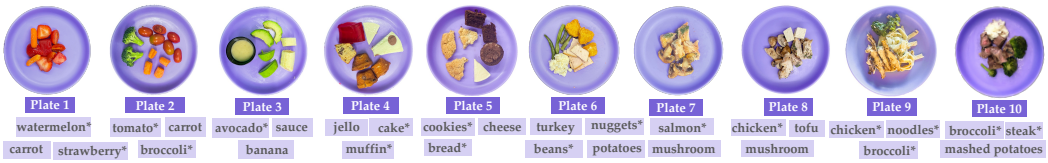

Figure 4: **Experimental setup: 10 in-the-wild dishes.** * denotes food items unseen during training.

Table 1: **Quantitative results on bite acquisition.** We evaluate our approach on 10 different dishes (Figure 4) and report both Attempt Efficiency (AE) and Success Rate (SR), averaged across plates.

Plate	#Items Acquired / #Total Attempts						
	SAVOR	Haptic-only	Vision-only	FLAIR	VLM	SayCan	End2End
1	**10/15**	9/17	6/26	9/18	9/19	9/22	9/15
2	**7/13**	4/17	4/18	3/15	4/17	5/13	4/13
3	**7/11**	**7/11**	6/14	5/11	**7/11**	5/15	6/11
4	6/12	6/13	6/13	5/15	5/14	**7/10**	6/12
5	6/13	6/15	5/21	5/14	**7/12**	5/14	5/17
6	**10/18**	8/20	7/29	8/23	8/23	8/23	8/16
7	5/17	6/17	4/19	5/17	6/14	4/15	**6/12**
8	5/9	4/12	4/16	4/12	5/10	5/12	**6/9**
9	**7/10**	6/12	5/17	7/11	**7/10**	5/18	4/17
10	6/16	5/18	5/19	**7/16**	6/16	5/18	5/16
Average AE (%)	**51.5 ± 12.7**	40.1 ± 12.2	27.1 ± 8.8	38.2 ± 12.2	43.8 ± 14.3	36.2 ± 12.6	42.7 ± 14.7
SR3 (%)	**87.3 ± 10.0**	77.2 ± 13.0	65.8 ± 11.2	73.4 ± 15.2	81.0 ± 13.1	73.4 ± 12.4	81.8 ± 14.7

5 Experiments

In this section, we aim to answer the following questions: **Q1.** Does tool calibration contribute to effective skill selection? **Q2.** How does visuo-haptic feedback help estimate the physical properties of food items? **Q3.** How effective is our overall approach compared to prior methods? **Q4.** How well does SAVOR generalize to unseen food items?

5.1 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 4). The dishes comprise three categories: (i) fruits and appetizers, (ii) grocery store and restaurant meals, and (iii) homemade dishes.

Baselines. We evaluate three ablations of our approach and compare against four existing methods. The ablations are: (i) **SAVOR w/o calibration**: In this ablation, we do not provide calibration information to the VLM planner. (ii) **Vision-only SAVOR** and (iii) **Haptic-only SAVOR**: These two ablations differ from ours in that it has no haptic/visual perception respectively. We also compare against the following baselines: (i) **FLAIR** [3]: This state-of-the-art approach selects bite acquisition action based on food category. (ii) **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 skill based on the estimated properties. (iii) **SayCan** [40]: This approach relies on a pre-contact image to estimate instruction relevance and success likelihood using a value function, which we train on the same SAVOR dataset. (iv) **End2End**: This method takes the same input as SAVOR-Net but directly predicts bite acquisition actions for execution. More details in Appendix B.1.

Evaluation Metrics. We evaluate each method using *Success Rate*. A trial is considered successful if the utensil picks up the food item and retains it for at least 5 seconds. Failed attempts leave the item on the plate, allowing up to three re-attempts before it is manually removed and recorded as a failure. To capture the efficiency of acquisition, we define Average Attempt Efficiency (average AE) $= \frac{\#\text{Items Acquired}}{\#\text{Total Attempts}}$ to quantify the proportion of successful acquisitions relative to total attempts across all plates. To further quantify overall performance, we define **SR1** as the proportion of items acquired within the first attempt in a meal, i.e., SR1 $= \frac{\#\text{Items Acquired}}{\#\text{Total Items}}$ within one attempt. Similarly, **SR2** and **SR3** represent the proportion of items acquired within two and three attempts, respectively.

Hardware. SAVOR is implemented on a Kinova Gen3 robot arm equipped with a motorized feeding utensil at the end-effector [3] (Figure 5a). The utensil includes a metal/plastic fork attachment with two degrees of freedom, fork orientation and tilt angle. We use an Intel RealSense D435 camera mounted on the arm wrist for visual perception and a Nano25 F/T sensor for haptic perception.

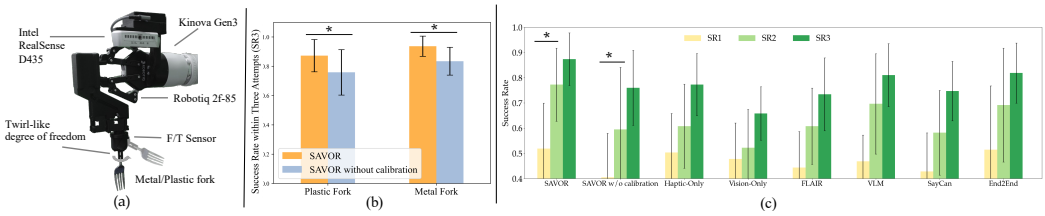

Figure 5: (a) System setup. (b) SR3 using a plastic/metal fork for 10 in-the-wild dishes. (c) SR{1,2,3} for 10 dishes. $*$ indicates statistically significant differences ($p < 0.05$).

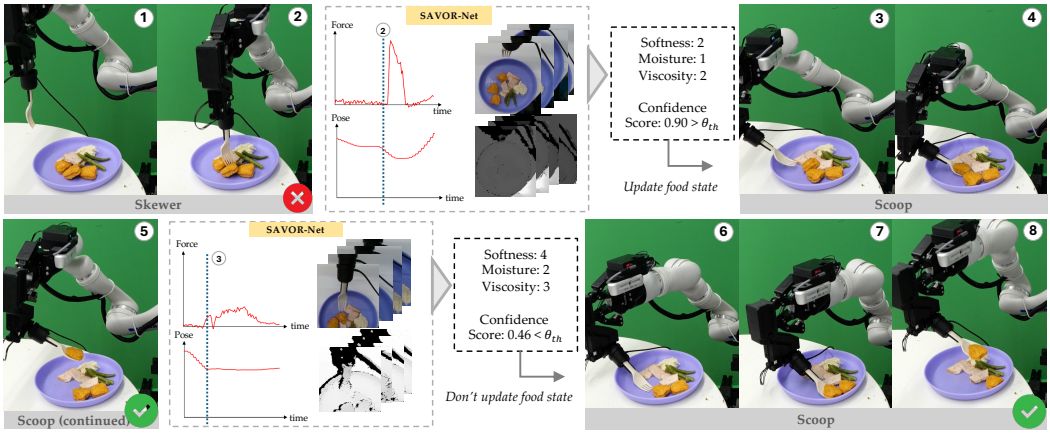

Figure 6: **Qualitative results on bite acquisition.** The robot first attempts to skewer the food based on its initial property estimate but fails (step 2). Vision and haptic data from this attempt are processed by SAVOR-Net, refining the estimate with high confidence. A VLM planner then selects the scoop skill based on this update (step 3). After executing the scoop (step 4), SAVOR-Net outputs a low-confidence estimate, so the properties are not updated in step 5, and the system proceeds directly to skill selection and execution (steps 6–8).

5.2 Results and Analysis

Evaluating the Contribution of Tool Calibration. We evaluate the effectiveness of offline tool calibration using two utensils: a plastic fork and a metal fork. As shown in Figure 5b and Figure 5c, tool calibration significantly improves success rates ($p < 0.05$), leading to an 18% increase in SR2 and a 13% increase in SR3. Without tool calibration, the system selects inappropriate skills, such as skewering a firm steak with a plastic fork. Similarly, without calibration with a metal fork, the system tries to skewer soft, high-moisture foods such as tofu instead of scooping them. Without calibration, SAVOR overuses skewering regardless of the context, leading to frequent failures. Tool calibration helps the system align skill selection with the physical interaction between the utensil and the food, resulting in more reliable bite acquisition.

Evaluating the Contribution of Visuo-Haptic Perception. We next evaluate how perception modalities affect performance by comparing SAVOR with its vision-only and haptics-only variants (Table 1). SAVOR achieves the highest success rate, confirming that combining vision and haptics improves food property estimation and significantly improves performance from SR1 to SR2 in Figure 5c ($p < 0.05$). The vision-only variant performs the worst (27.1% average AE), struggling with fine-grained utensil–food interactions and often misassigning visually similar items with inaccurate physical properties. For example, strawberries, watermelon, and carrots on Plate 1 appear similar in color, but differ significantly in physical properties. The model frequently overestimates softness for firm items like carrots, leading to repeated skill selection failures. The haptics-only variant performs better (40.1% average AE) but still falls short of SAVOR. The haptics-only variant often fails when food items share similar force signatures, such as tofu and nuts. In these cases, small changes in end-effector pose lead to rapid force increases—either due to skewering through soft items and contacting the plate, or due to deformation of the utensil against firm items. Without visual input, these ambiguities lead to misinterpretation of food properties. By integrating vision and haptics, SAVOR

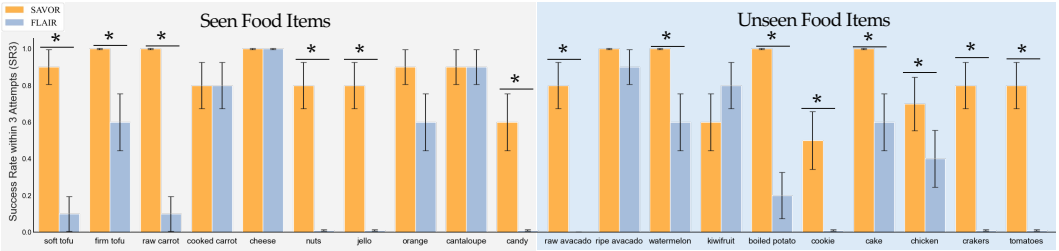

Figure 7: **Generalization performance on seen and unseen food items.** We compare SAVOR and SOTA FLAIR across 20 food items. We show SR3 on 10 episodes per food item, where each episode allows up to 3 attempts. Asterisks ($*$) indicate statistically significant differences ($p < 0.05$).

resolves ambiguities, achieving a 51.5% average AE and 87.3% SR3. Figure 5c demonstrates that SAVOR effectively learns from interaction during the initial attempt and adapts its skill selection dynamically, resulting in the greatest improvement in subsequent attempts. These results highlight the critical role of time-series visuo-haptic perception in estimating food affordances.

Evaluating the Overall Framework. To evaluate our full framework, we compare SAVOR with the SOTA method (FLAIR), a vision-only variant using a VLM, SayCan and the end2end method. SAVOR achieves the highest average AE (51.5%) and 87.3% SR3 across 10 diverse plates (Table 1).

FLAIR selects skills based on predefined food categories, achieving 38.2% average AE and 73.4% SR3. It performs well on stereotypical items (e.g., banana), but fails on atypical foods such as raw avocado, where category-based reasoning leads to inappropriate actions (e.g., skewering instead of scooping). VLM improves upon FLAIR, achieving 43.8% average AE and 81.0% SR3 by inferring food properties from visual input. However, it struggles with visually ambiguous or dynamically changing items, such as misclassifying avocado or failing to detect firmness changes in cooling chicken nuggets. Similarly, SayCan achieves only 36.2% average AE and 73.4% SR3, often oscillating between skewer and scoop skills. Without compact physical property representations, the end-to-end policy generalizes poorly, reaching 42.7% average AE and 81.8% SR3.

While all baselines show improvements from SR1 to SR3 in Figure 5c, these gains largely result from occasional success through repeated trials rather than informed adaptation. Among all methods, only SAVOR and its variant without calibration achieve significant improvement from SR1 to SR2 ($p < 0.05$). As seen in Figure 5c, SAVOR acquires 53.2% food items with its initial attempt (SR1). By inferring food properties during the initial attempt, the system adapts its skill selection and successfully picks up 77.2% of the food items with two attempts (SR2), further achieving 87.3% SR3. It dynamically adjusts skills (e.g., switching from skewering to scooping as nuggets firm up; Figure 6) and generalizes across similar items on the same plate. These results demonstrate the effectiveness of SAVOR for adaptive bite acquisition.

Evaluating Generalization to Unseen Food Items. We evaluate SAVOR on 20 single food items, including 10 that are unseen during training. As shown in Figure 7, SAVOR achieves higher SR3 ($p < 0.05$) than the SOTA baseline FLAIR in 70% of food items, including raw avocado, cookie, and chicken. These results highlight the advantage of explicitly estimating physical properties over relying on fixed food categories. SAVOR achieves the highest overall SR3 and maintains comparable performance on both seen and unseen items, demonstrating strong generalization.

Discussion. Our work highlights the importance of grounding skill affordance learning in physical interaction by combining tool and food affordances for effective bite acquisition. This integration allows the system to move beyond state-of-the-art category-based methods and instead make context-sensitive decisions based on how food interacts with a given tool. A key insight from our study is that affordances are not static properties, but rather emerge dynamically through interaction—reflecting how the tool and food jointly shape which skills are applicable. Moreover, we find that visuo-haptic sensing plays a critical role in understanding temporal variations of food physical properties and resolving ambiguity, particularly in cases where visual or haptic cues alone may be misleading. These findings underscore the importance of selecting manipulation skills that are physically grounded in food–utensil interaction for effective bite acquisition.

Limitations

Our approach effectively leverages visuo-haptic sensing to estimate food physical properties and tool calibration to assess tool capabilities for adaptive bite acquisition. SAVOR currently treats each food item as a whole when estimating its properties. However, a single food item may contain regions with differing physical properties. For instance, foods like broccoli or cauliflower can vary in softness across different parts due to their fibrous stems. While treating each food item as a whole has proven effective, incorporating a more fine-grained food representation could further improve bite acquisition success. Second, while our system already performs well with open-loop skill execution, it could further benefit from a closed-loop low-level policy to enhance real-time adaptability. Finally, slippage during manipulation can introduce noise in haptic signals, leading to inaccurate physical property estimates, which could be mitigated by detecting contact onset, loss, or lateral motion. These directions present promising avenues for future work.

Acknowledgments

This work was partly funded by National Science Foundation IIS #2132846, and CAREER #2238792. We thank Ruolin Ye and other members of the EmPRISE Lab for their help with the figures in the manuscript.

References

[1] 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.

[2] A. Nanavati, P. Alves-Oliveira, T. Schrenk, E. K. Gordon, M. Cakmak, and S. 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.

[3] R. K. Jenamani, P. Sundaresan, M. Sakr, T. Bhattacharjee, and D. Sadigh. Flair: Feeding via long-horizon acquisition of realistic dishes. *arXiv preprint arXiv:2407.07561*, 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. 2024.

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

[6] R. K. Jenamani, D. Stabile, Z. Liu, A. Anwar, K. Dimitropoulou, and T. 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.

[7] A. Nanavati, E. K. Gordon, T. A. Kessler Faulkner, Y. R. Song, J. Ko, T. Schrenk, V. Nguyen, B. H. Zhu, H. Bolotski, A. Kashyap, S. Kutty, R. Karim, L. Rainbolt, R. Scalise, H. Song, R. Qu, M. Cakmak, and S. 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.

[8] P. Sundaresan, S. Belkhale, and D. Sadigh. Learning visuo-haptic skewering strategies for robot-assisted feeding. In *6th Annual Conference on Robot Learning*, 2022. URL https://openreview.net/forum?id=lLq09gVoaTE.

[9] E. K. Gordon, X. Meng, T. Bhattacharjee, M. Barnes, and S. 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.

[10] E. K. Gordon, S. Roychowdhury, T. Bhattacharjee, K. Jamieson, and S. 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.

[11] E. K. Gordon, A. Nanavati, R. Challa, B. H. Zhu, T. A. K. Faulkner, and S. Srinivasa. Towards general single-utensil food acquisition with human-informed actions. In *Conference on Robot Learning*, pages 2414–2428. PMLR, 2023.

[12] R. Feng, Y. Kim, G. Lee, E. K. Gordon, M. Schmittle, S. Kumar, T. Bhattacharjee, and S. 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.

[13] P. Sundaresan, J. Wu, and D. Sadigh. Learning sequential acquisition policies for robot-assisted feeding. In J. Tan, M. Toussaint, and K. 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.

[14] Y.-L. Tai, Y. C. Chiu, Y.-W. Chao, and Y.-T. Chen. Scone: A food scooping robot learning framework with active perception. In J. Tan, M. Toussaint, and K. 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.

[15] J. Grannen, Y. Wu, S. Belkhale, and D. Sadigh. Learning bimanual scooping policies for food acquisition. In *6th Annual Conference on Robot Learning*, 2022. URL https://openreview.net/forum?id=qDtbMK67PJG.

[16] M. Shridhar, L. Manuelli, and D. Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, pages 894–906. PMLR, 2022.

[17] T. Brosnan and D.-W. 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.

[18] Z. He, B. Ai, Y. Liu, W. Wan, H. I. Christensen, and H. Su. Learning dexterous deformable object manipulation through cross-embodiment dynamics learning. In *RSS Workshop on Dexterous Manipulation: Learning and Control with Diverse Data*, 2025.

[19] T. Tian, H. Li, B. Ai, X. Yuan, Z. Huang, and H. Su. Diffusion dynamics models with generative state estimation for cloth manipulation. *Conference on Robot Learning (CoRL)*, 2025.

[20] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik. General In-Hand Object Rotation with Vision and Touch. In *Conference on Robot Learning (CoRL)*, 2023.

[21] S. Suresh, H. Qi, T. Wu, T. Fan, L. Pineda, M. Lambeta, J. Malik, M. Kalakrishnan, R. Calandra, M. Kaess, J. Ortiz, and M. Mukadam. Neural feels with neural fields: Visuo-tactile perception for in-hand manipulation. *Science Robotics*, page adl0628, 2024.

[22] B. Ai, S. Tian, H. Shi, Y. Wang, C. Tan, Y. Li, and J. Wu. Robopack: Learning tactile-informed dynamics models for dense packing. *Robotics: Science and Systems (RSS)*, 2024. URL https://arxiv.org/abs/2407.01418.

[23] T. Bhattacharjee, G. Lee, H. Song, and S. 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.

[24] A. Yamaguchi and C. 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.

[25] Z. Xu, Z. Xian, X. Lin, C. Chi, Z. Huang, C. Gan, and S. Song. Roboninja: Learning an adaptive cutting policy for multi-material objects. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[26] E. Heiden, M. Macklin, Y. S. Narang, D. Fox, A. Garg, and F. 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.

[27] K. Zhang, M. Sharma, M. Veloso, and O. 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.

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

[29] 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)*, pages 9540–9547, 2024. doi:10.1109/ICRA57147.2024.10610050.

[30] C. Dong, L. Yu, M. Takizawa, S. Kudoh, and T. Suehiro. Food peeling method for dual-arm cooking robot. pages 801–806, 01 2021. doi:10.1109/IEEECONF49454.2021.9382700.

[31] X. Luo, S. Jin, H.-J. Huang, and W. Yuan. An intelligent robotic system for perceptive pancake batter stirring and precise pouring. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5970–5977, 2024. doi:10.1109/IROS58592.2024.10802841.

[32] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. A. Creel, J. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. F. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. F. Nyarko, G. Ogut, L. J. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. H. Roohani, C. Ruiz, J. Ryan, C. R'e, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. P. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. A. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021.

[33] C. Agia, T. Migimatsu, J. Wu, and J. 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.

[34] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, P. Sermanet, N. Brown, T. Jackson, L. Luu, S. Levine, K. Hausman, and B. Ichter. Inner monologue: Embodied reasoning through planning with language models. In *arXiv preprint arXiv:2207.05608*, 2022.

[35] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 2023.

[36] Z. Wu, B. Ai, and D. 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.

[37] 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)*, pages 11523–11530, 2023. doi:10.1109/ICRA48891.2023.10161317.

[38] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.

[39] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. 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.

[40] B. Ichter, A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian, D. Kalashnikov, S. Levine, Y. Lu, C. Parada, K. Rao, P. Sermanet, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, M. Yan, N. Brown, M. Ahn, O. Cortes, N. Sievers, C. Tan, S. Xu, D. Reyes, J. Rettinghouse, J. Quiambao, P. Pastor, L. Luu, K. Lee, Y. Kuang, S. Jesmonth, N. J. Joshi, K. Jeffrey, R. J. Ruano, J. Hsu, K. Gopalakrishnan, B. David, A. Zeng, and C. K. Fu. Do as I can, not as I say: Grounding language in robotic affordances. In K. Liu, D. Kulic, and J. Ichnowski, editors, *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pages 287–318. PMLR, 2022. URL https://proceedings.mlr.press/v205/ichter23a.html.

[41] Y. R. Wang, J. Duan, D. Fox, and S. S. Srinivasa. NEWTON: are large language models capable of physical reasoning? In *Findings of the Association for Computational Linguistics (EMNLP)*, 2023.

[42] Q. Gao, X. Pi, K. Liu, J. Chen, R. Yang, X. Huang, X. Fang, L. Sun, G. Kishore, B. Ai, S. Tao, M. Liu, J. Yang, C.-J. Lai, C. Jin, J. Xiang, B. Huang, D. Danks, H. Su, T. Shu, Z. Ma, L. Qin, and Z. Hu. Do vision-language models have internal world models? towards an atomic evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

[43] L. Shi, Y. Liu, L. Zeng, B. Ai, Z. Hong, and H. Su. Learning adaptive dexterous grasping from single demonstrations. *arXiv preprint arXiv:2503.20208*, 2025.

[44] J. Loo, Z. Wu, and D. Hsu. Open scene graphs for open-world object-goal navigation. *The International Journal of Robotics Research*, 2025.

[45] L. P. Kaelbling, M. L. Littman, and A. 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.

[46] M. C. Gemici and A. Saxena. Learning haptic representation for manipulating deformable food objects. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 638–645, 2014. doi:10.1109/IROS.2014.6942626.

[47] A. Nanavati, R. Challa, E. K. Gordon, and S. S. Srinivasa. A Dataset of Food Manipulation Strategies for Diverse Foods, 2022. URL https://doi.org/10.7910/DVN/C8SI1D.

[48] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[49] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[50] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Appendix

A Additional Results

A.1 Timing and Latency Analysis

We present a timing analysis in Table 2, highlighting the processes involving VLM queries. The query for initial state estimation takes about 3.7 seconds to predict the food item's physical properties from an image. This step occurs only once per item. In subsequent attempts, SAVOR-Net updates the physical properties in 0.2 seconds. As shown in Table 2, the major source of latency stems from querying the VLM, and we envision that ongoing work on efficient VLMs holds promise for reducing query timing.

Table 2: **Timing of each component in SAVOR.** $*$ indicates processes that include VLM queries.

	Perception	State Estimation		Planning	Control
	Object Detection$*$	Initial Attempt$*$	Subsequent Attempt	Skill Selection$*$	Skill Execution
Time (s)	2.59 ± 0.32	3.69 ± 0.82	0.21 ± 0.01	3.58 ± 0.74	8.54 ± 1.21

A.2 Ablation Study on Tool Calibration

We provide detailed results of the ablation study on the calibration process for the 10 in-the-wild dishes (Table 3). Compared to the uncalibrated baseline, tool calibration significantly improves performance for both the plastic and metal forks. Specifically, with calibration, the plastic fork (PF) achieves a 51.5% average attempt efficiency and 87.3% SR3, compared to only 38.7% and 75.9% without calibration (PF-wo). Similarly, the metal fork (MF) benefits from calibration, improving from 83.5% SR3 to 93.7%. These results demonstrate that understanding tool capabilities through calibration helps the planner avoid infeasible actions, such as skewering tofu with a metal fork (Plate 8) or skewering firm steak with a plastic fork (Plate 10), thereby improving skill selection and acquisition success.

Table 3: **Ablation study on calibration.** Success rates of bite acquisition across 10 in-the-wild dishes, comparing the impact of calibration. PF: Plastic fork; PF-wo: Plastic fork without calibration; MF: Metal fork; MF-wo: Metal fork without calibration. Asterisks ($*$) indicates unseen food items.

Plate	**#Items Acquired / #Total Attempts**			
	PF	**PF-wo**	**MF**	**MF-wo**
1	10/15	9/19	11/15	11/15
2	7/13	4/17	6/15	5/17
3	7/11	5/15	7/10	7/9
4	6/13	6/13	7/11	7/11
5	6/13	7/12	7/12	7/12
6	10/18	8/21	10/17	10/18
7	5/17	6/17	7/17	7/15
8	5/9	3/12	6/9	5/11
9	7/10	7/10	6/12	6/12
10	6/16	5/19	8/13	7/13
Average AE (%)	51.5%	38.7%	56.1%	54.1%
SR3 (%)	87.3%	75.9%	93.7%	83.5%

A.3 Crumbly and Soft Foods

We evaluate our approach on foods that are particularly soft or crumbly, such as tofu and crackers. For crackers, our system selects scooping in 80% of trials and attempts skewering in 20% in the initial attempt. After a skewering attempt, SAVOR identifies low softness, logs the failure in the attempt history, and switches to scooping. For tofu, we test variants from extra soft to super firm. The system initially selects skewering for all tofu types but switches to scooping for extra soft tofu. We achieve a 70% average attempt efficiency for crackers and a 95% average attempt efficiency for tofu. These results suggest that our method can adapt to diverse food items when informed by interaction feedback.

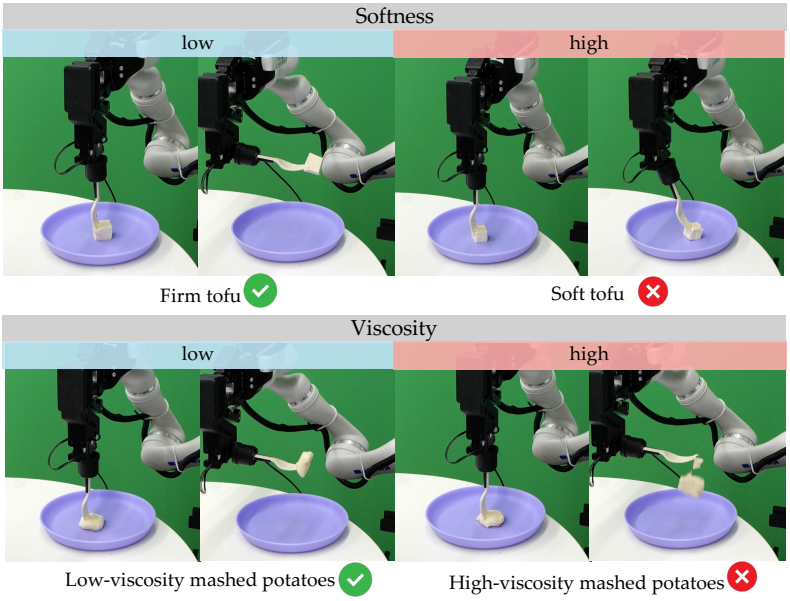

Figure 8: **Effect of food physical properties on utensil interactions.** The robot skewers food items of varying *softness* (top) and *viscosity* (bottom). Soft tofu and low-viscosity mashed potatoes are successfully acquired, while firm tofu and high-viscosity mashed potatoes lead to failure, illustrating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

A.4 VLMs Failures

We use a single VLM (GPT-4V) for both state estimation and skill planning, instantiated separately as GPT-Property and GPT-Skill. While the VLM itself is static, we augment its context after each interaction with the latest observations. After every interaction, GPT-Property can update food properties. GPT-Skill can also adjust skill choice because some of its inputs are from GPT-Property.

Most VLM-related failures stem from incorrect initial food property estimates by GPT-Property, when it infers physical properties from an RGB image before interaction. These vision-only priors can lead to misidentifications (e.g., confusing carrots with tomatoes) or overlook intra-category variations (e.g., assuming all avocados are soft). This also affects GPT-Skill as GPT-Skill uses inputs from GPT-Property. SR1=$\frac{\text{\#Items Acquired}}{\text{\#Total Items}}$ within 1 attempt ($\sim$50%) highlights the challenge of using vision priors alone to estimate food affordances, often leading to failures in first attempts. These failures lead to more attempts and thus impact the overall average attempt efficiency) in Table 1, though they are later corrected using online visuo-haptic updates (SR2: 77.2%, SR3: 87.3%).

A.5 Open-loop Nature of Skill Execution

As mentioned in our limitations section, though our system performs well with open-loop execution, we acknowledge that it could further benefit from a closed-loop policy, and we plan to address this in future work. However, we conduct further analysis of the experiments and find that the need for more closed-loop skills occurs in only specific cases such as slippage during picking up oily surfaces of salmon and mushrooms, which account for only 7.93% of trials. Note, despite slippage, re-attempts can potentially pick up the food item.

A.6 Multi-food Interactions on Cluttered Plates

Our study addresses multi-food interactions on cluttered plates. 8 out of 10 dishes contain overlapping food items, where interactions with one food item affect another food item. In such challenging

cases, SAVOR-Net often gives low confidence for push actions, but when the food is pushed toward a cluster of rigid items, it provides high confidence scores and meaningful property estimates.

B Implementation Details

B.1 Baselines

We provide implementation details for the baselines as follows:

(i) SayCan [40]: This method selects a skill by combining two scores: the skill's relevance to the instruction and its predicted likelihood of success. As the original work does not release the value function, we train a value function using our SAVOR dataset. The model takes a single RGB image as input and outputs success probabilities for each skill in our predefined skill library. For relevance estimation, we follow the original SayCan setup and use a vision-language model to compute instruction-skill alignment scores.

(ii) End2End: We train an end-to-end model for action selection in bite acquisition as a baseline. This model takes the same input as SAVOR-Net, which includes vision, haptics, and robot poses, and directly predicts one of the six manipulation skills: skewering, scooping, twirling, pushing, dipping, or cutting. The model is trained on the SAVOR dataset.

B.2 Data Collection

We collect data by applying each skill from a predefined skill library to food items. The library includes six manipulation skills: skewering, scooping, twirling, pushing, dipping, and cutting. For each skill, we perform 5 trials per food item, recording synchronized RGB-D images, haptic feedback, and pose data throughout each trajectory. The food items span a range of physical properties and include: bagel, nuts, mashed potatoes, broccoli, jello, carrot, tofu, pork, orange, cantaloupe, candy, lettuce, avocado, cheese, turkey, noodles, watermelon, banana, and tomatoes, along with variations in their cooking or ripeness levels.

B.3 SAVOR-Net

B.3.1 Model Architecture

SAVOR-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 the LSTM and produces the final output ψ_t.

B.3.2 Training

SAVOR-Net($\sim$14M parameters) is trained using cross-entropy loss with the hyperparameters listed in Table 4. Training is conducted on an NVIDIA RTX 4090 GPU and completes in approximately 40 minutes.

Table 4: **Training hyperparameters for SAVOR-Net**

Hyperparameter	Value
Epochs	200
Learning rate	1e-3
Optimizer	Adam
Batch size	16

B.3.3 Tool Calibration

Given the utensil and skill library, tool calibration is performed once and only needs to be repeated if the tool is modified. Before deployment, we conduct tool calibration by evaluating each skill five times using the current utensil on five food items with diverse physical properties. During this process, we record each item's physical properties and execution outcomes in natural language. The selected calibration items are raw carrot, cooked carrot, soft tofu, nuts, and cheese. The entire calibration process takes approximately 20 minutes.

B.4 Prompting Details

B.4.1 Perception

We prompt GPT-4V to generate a set of candidate labels, which are then used by open-set object detectors Grounded SAM [50] to generate masks for each food item. The prompt we use for this application is:

```
For the given image, please list the food items on the plate in a Python list
format.
Here are three examples:
Example Image 1; Answer: ['chicken','broccoli','sausage']
Example Image 2; Answer: ['steak','mushroom']
Example Image 3; Answer: ['carrot','watermelon','strawberries']
<Given Image>, please list down all the food items in the plate. Follow this
format: Answer: ['first_food', 'second_food', ..., 'last_food']
```

B.4.2 Calibration

We evaluate the utensil by executing different skills on a small set of diverse food items. During offline calibration, various utensils interact with a range of foods to assess their functional capabilities. We collect skill execution outcomes, annotated with food type and physical properties. The tool affordances are represented in natural language and later used as input to the VLM-based planner. An example of the calibration summary for the plastic fork is provided below:

```
The robot interacts with various food items using a plastic fork. We summarize
the history as follows:
Food Item: Nuts
Shape: Oval, Size: Bite-sized, Softness: 1, Moisture: 1, Viscosity: 2
Skill with Success Rate: Skewer 0/5, Scoop 3/5, Cut 0/5, Push 5/5, Dip 5/5

Food Item: Cheese
Shape: Block, Size: Bite-sized, Softness: 3, Moisture: 2, Viscosity: 4
Skill with Success Rate: Skewer 5/5, Scoop 3/5, Cut 5/5, Push 5/5, Dip 5/5

Food Item: Raw Carrot
Shape: Cylindrical, Size: Bite-sized, Softness: 2, Moisture: 2, Viscosity: 1
Skill with Success Rate: Skewer 0/5, Scoop 3/5, Cut 0/5, Push 5/5, Dip 5/5

Food Item: Cooked Carrot
Shape: Cylindrical, Size: Bite-sized, Softness: 2, Moisture: 3, Viscosity: 1
```

```
Skill with Success Rate: Skewer 5/5, Scoop 3/5, Cut 4/5, Push 5/5, Dip 5/5

Food Item: Soft Tofu
Shape: Cubic, Size: Large, Softness: 4, Moisture: 3, Viscosity: 2
Skill with Success Rate: Skewer 1/5, Scoop 4/5, Cut 5/5, Push 5/5, Dip 5/5
```

B.4.3 State Estimation

We prompt GPT-4V to estimate food physical properties based solely on visual cues. Specifically, we initialize the food property estimate using only an RGB image as input and ask the VLM to infer physical properties including shape, size, softness, moisture, and viscosity.

```
<Image on the target food item>
This is a plate of <food item>.
Please estimate the physical properties of the food item, including Shape, Size
, Softness, Moisture, and Viscosity, based on commonsense reasoning. For
Softness, Moisture, and Viscosity, provide a score ranging from 0 to 5, similar
 to a 5-point Likert scale (e.g., a softness score of 1 indicates very hard,
while 5 indicates very soft).

Always follow this format:
Answer: Shape: <shape> ; Size: <size>; Softness: <softness score>; Moisture: <
moisture score>; Viscosity: <viscosity score>
```

B.4.4 Skill Selection

We design prompting templates for skill selection. Each prompt includes a calibration summary, the history of past attempts, the available skills from the skill library, and the physical properties of the target food item. The prompt then asks the VLM to choose the most appropriate skill based on this context. We use a few-shot prompting setup with GPT-4V.

```
< Calibration Summary >
The robot is using a plastic fork to pick up the food. Please select an
appropriate skill by considering the food's category, shape, size, softness,
moisture, and viscosity.

We briefly describe the skills as follows:
< Skill description >

The attempt history is summarized as follows:
Steak:
shape: round
size: bite-sized
softness: 2
moisture:2
viscosity: 1
scoop: success

Example Prompt 1: <Image on sausage>
This is a food item: Sausage Slice.
The robot uses a plastic fork to try picking up the food.
The estimated food physical properties are as follows. The scores range from 0
to 5, similar to a 5-pt Likert scale. For example, a softness score of 1
indicates very hard, while a score of 5 indicates very soft.
Shape: cylinder
Size: bite-sized
Softness: 3
Moisture: 2
```

```
Viscosity: 1

Please select an action from ['skewer, 'scoop', 'twirl', 'dip'] to pick up the
food item or
select one action from ['cut', 'push'] to manipulate items to facilitate
subsequent
acquisition. Always follow the format: Reasoning: <your reason>. Answer: <your
answer>.

Example Answer 1:
Reasoning: Sausage slices are moderately firm to maintain their structure.
Skewering is suitable as the fork can easily pierce them without breaking them
apart.
Answer: skewer

Example Prompt 2: <Image on mashed potatoes>
This is a food item: Mashed Potatoes.
The robot uses a plastic fork to try picking up the food.
The estimated food physical properties are as follows. The scores range from 0
to 5, similar to a 5-pt Likert scale. For example, a softness score of 1
indicates very hard, while a score of 5 indicates very soft.
Shape: Amorphous
Size: bite-sized
Softness: 4
Moisture: 3
Viscosity: 2

Please select an action from ['skewer, 'scoop', 'twirl', 'dip'] to pick up the
food item or
select one action from ['cut', 'push'] to manipulate items to facilitate
subsequent
acquisition. Always follow the format: Reasoning: <your reason>. Answer: <your
answer>.

Example Answer 2:
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

This is a food item: Tofu. <image>
The robot uses a plastic fork to try picking up the food.
Food Item: Tofu
Shape: Cubic
Size: bite-sized
Softness: 4
Moisture: 3
Viscosity: 2


Please select an action from ['skewer, 'scoop', 'twirl', 'dip'] to pick up the
food item or
select one action from ['cut', 'push'] to manipulate items to facilitate
subsequent
acquisition. Always follow the format: Reasoning: <your reason>. Answer: <your
answer>.
```