

## 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)	$2.59 \pm 0.32$	$3.69 \pm 0.82$	$0.21 \pm 0.01$	$3.58 \pm 0.74$	$8.54 \pm 1.21$

#### 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 success rate 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 Type				Methods			
Plate	Items	Visual	Haptic	PF	PF-wo	MF	MF-wo
1	strawberries*, watermelon, carrots	Similar	Diverse	10/15	9/19	11/15	11/15
2	tomatoes*, broccoli, carrots	Similar	Diverse	7/13	4/17	6/15	5/17
3	avocado*, banana, sauce	Diverse	Diverse	7/11	5/15	7/10	7/9
4	muffin*, cake*, jello	Diverse	Diverse	6/13	6/13	7/11	7/11
5	cookie*, bread*, cheese	Diverse	Diverse	6/13	7/12	7/12	7/12
6	roasted turkey, chicken nuggets*, mashed potatoes, green beans*	Diverse	Diverse	10/18	8/21	10/17	10/18
7	salmon, mushrooms	Diverse	Similar	5/17	6/17	7/17	7/15
8	chicken breast, tofu, mushrooms	Diverse	Diverse	5/9	3/12	6/9	5/11
9	chicken, broccoli*, noodles*	Diverse	Diverse	7/10	7/10	6/12	6/12
10	steak*, broccoli*, mashed potatoes	Diverse	Diverse	6/16	5/19	8/13	7/13
Average Success Rate (%)				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 SR for crackers and a 95% average SR for tofu. These results suggest that our method can adapt to diverse food items when informed by interaction feedback.

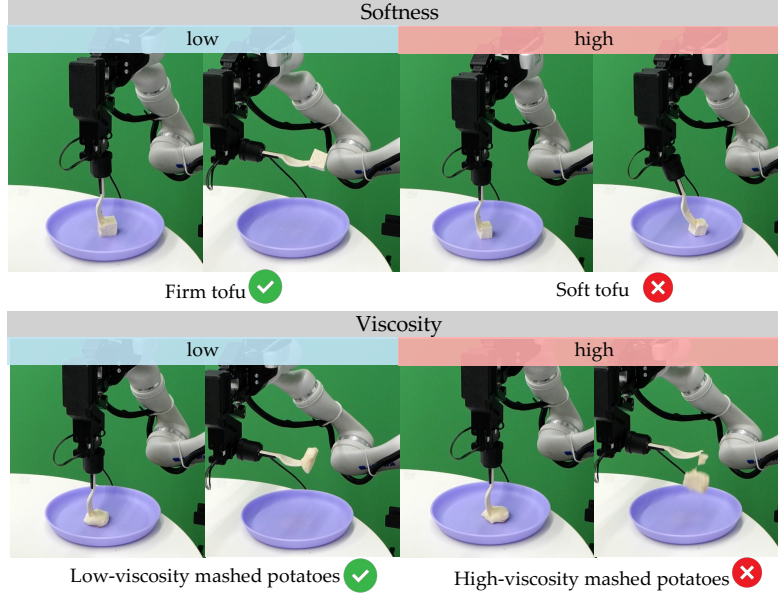


Figure 8: **Effect of food physical properties on utensil interactions.** The robot skewers food items of varying *softness* (top) and *viscosity* (bottom). Soft tofu and low-viscosity mashed potatoes are successfully acquired, while firm tofu and high-viscosity mashed potatoes lead to failure, illustrating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

#### 456 A.4 Open-loop Nature of Skill Execution

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

#### 463 A.5 Multi-food Interactions on Cluttered Plates

464 Our study addresses multi-food interactions on cluttered plates. 8 out of 10 dishes contain overlap-  
 465 ping food items, where interactions with one food item affect another food item. In such challenging  
 466 cases, SAVOR-Net often gives low confidence for push actions, but when the food is pushed toward  
 467 a cluster of rigid items, it provides high confidence scores and meaningful property estimates.

## 468 B Implementation Details

### 469 B.1 Baselines

470 We provide implementation details for the baselines as follows:

471 **(i) SayCan [31]:** This method selects a skill by combining two scores: the skill’s relevance to the  
 472 instruction and its predicted likelihood of success. As the original work does not release the value  
 473 function, we train a value function using our SAVOR dataset. The model takes a single RGB image  
 474 as input and outputs success probabilities for each skill in our predefined skill library. For rele-  
 475 vance estimation, we follow the original SayCan setup and use a vision-language model to compute  
 476 instruction-skill alignment scores.

477 **(ii) End2End:** We train an end-to-end model for action selection in bite acquisition as a baseline.  
 478 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  $3 \times 3$  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  $\psi_t$ .

### B.3.2 Training

SAVOR-Net is trained using cross-entropy loss with the hyperparameters listed in Table 4. Training is conducted on an NVIDIA RTX 3070 GPU and completes in approximately 50 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 [37] to generate masks for each food item. The prompt we use for this application is:

```

515 For the given image, please list the food items on the plate in a Python list
516 format.
517 Here are three examples:
518 Example Image 1; Answer: ['chicken', 'broccoli', 'sausage']
519 Example Image 2; Answer: ['steak', 'mushroom']
520 Example Image 3; Answer: ['carrot', 'watermelon', 'strawberries']
521 <Given Image>, please list down all the food items in the plate. Follow this
522 format: Answer: ['first_food', 'second_food', ..., 'last_food']
523

```

## 525 B.4.2 Calibration

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

```

531 The robot interacts with various food items using a plastic fork. We summarize
532 the history as follows:
533 Food Item: Nuts
534 Shape: Oval, Size: Bite-sized, Softness: 1, Moisture: 1, Viscosity: 2
535 Skill with Success Rate: Skewer 0/5, Scoop 3/5, Cut 0/5, Push 5/5, Dip 5/5
536
537 Food Item: Cheese
538 Shape: Block, Size: Bite-sized, Softness: 3, Moisture: 2, Viscosity: 4
539 Skill with Success Rate: Skewer 5/5, Scoop 3/5, Cut 5/5, Push 5/5, Dip 5/5
540
541 Food Item: Raw Carrot
542 Shape: Cylindrical, Size: Bite-sized, Softness: 2, Moisture: 2, Viscosity: 1
543 Skill with Success Rate: Skewer 0/5, Scoop 3/5, Cut 0/5, Push 5/5, Dip 5/5
544
545 Food Item: Cooked Carrot
546 Shape: Cylindrical, Size: Bite-sized, Softness: 2, Moisture: 3, Viscosity: 1
547 Skill with Success Rate: Skewer 5/5, Scoop 3/5, Cut 4/5, Push 5/5, Dip 5/5
548
549 Food Item: Soft Tofu
550 Shape: Cubic, Size: Large, Softness: 4, Moisture: 3, Viscosity: 2
551 Skill with Success Rate: Skewer 1/5, Scoop 4/5, Cut 5/5, Push 5/5, Dip 5/5
552
553

```

## 554 B.4.3 State Estimation

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

```

558 <Image on the target food item>
559 This is a plate of <food item>.
560 Please estimate the physical properties of the food item, including Shape, Size
561 , Softness, Moisture, and Viscosity, based on commonsense reasoning. For
562 Softness, Moisture, and Viscosity, provide a score ranging from 0 to 5, similar
563 to a 5-point Likert scale (e.g., a softness score of 1 indicates very hard,
564 while 5 indicates very soft).
565
566 Always follow this format:
567 Answer: Shape: <shape> ; Size: <size>; Softness: <softness score>; Moisture: <
568 moisture score>; Viscosity: <viscosity score>
569
570

```

#### 571 B.4.4 Skill Selection

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

```
576 < Calibration Summary >
577 The robot is using a plastic fork to pick up the food. Please select an
578 appropriate skill by considering the food's category, shape, size, softness,
579 moisture, and viscosity.
580
581 We briefly describe the skills as follows:
582 < Skill description >
583
584 The attempt history is summarized as follows:
585 Steak:
586 shape: round
587 size: bite-sized
588 softness: 2
589 moisture:2
590 viscosity: 1
591 scoop: success
592
593 Example Prompt 1: <Image on sausage>
594 This is a food item: Sausage Slice.
595 The robot uses a plastic fork to try picking up the food.
596 The estimated food physical properties are as follows. The scores range from 0
597 to 5, similar to a 5-pt Likert scale. For example, a softness score of 1
598 indicates very hard, while a score of 5 indicates very soft.
599 Shape: cylinder
600 Size: bite-sized
601 Softness: 3
602 Moisture: 2
603 Viscosity: 1
604
605 Please select an action from ['skewer', 'scoop', 'twirl', 'dip'] to pick up the
606 food item. Notice that if all acquisition skills are not immediately feasible,
607 please select an action from ['cut', 'push'] to rearrange or manipulate items
608 to facilitate subsequent acquisition. Always follow the format: Reasoning: <
609 your reason>. Answer: <your answer>.
610
611 Example Answer 1:
612 Reasoning: Sausage slices are moderately firm to maintain their structure.
613 Skewering is suitable as the fork can easily pierce them without breaking them
614 apart.
615 Answer: skewer
616
617 Example Prompt 2: <Image on mashed potatoes>
618 This is a food item: Mashed Potatoes.
619 The robot uses a plastic fork to try picking up the food.
620 The estimated food physical properties are as follows. The scores range from 0
621 to 5, similar to a 5-pt Likert scale. For example, a softness score of 1
622 indicates very hard, while a score of 5 indicates very soft.
623 Shape: Amorphous
624 Size: bite-sized
625 Softness: 4
626 Moisture: 3
627 Viscosity: 2
628
```

629  
630 Please select an action from ['skewer', 'scoop', 'twirl', 'dip'] to pick up the  
631 food item. Notice that if all acquisition skills are not immediately feasible,  
632 please select an action from ['cut', 'push'] to rearrange or manipulate items  
633 to facilitate subsequent acquisition. Always follow the format: Reasoning: <  
634 your reason>. Answer: <your answer>.

635  
636 Example Answer 2:  
637 Reasoning: The food is soft and moist, making it suitable to scoop rather than  
638 skewer or cut. The viscosity indicates it will adhere moderately to the fork.  
639 Answer: scoop

640  
641 This is a food item: Tofu. <image>  
642 The robot uses a plastic fork to try picking up the food.  
643 Food Item: Tofu  
644 Shape: Cubic  
645 Size: bite-sized  
646 Softness: 4  
647 Moisture: 3  
648 Viscosity: 2

649  
650  
651 Please select an action from ['skewer', 'scoop', 'twirl', 'dip'] to pick up the  
652 food item. Notice that if all acquisition skills are not immediately feasible,  
653 please select an action from ['cut', 'push'] to rearrange or manipulate items  
654 to facilitate subsequent acquisition. Always follow the format: Reasoning: <  
655 your reason>. Answer: <your answer>.