REPeat: A Real2Sim2Real Approach for Pre-acquisition of Soft Food Items in Robot-assisted Feeding

Nayoung Ha*[1], Ruolin Ye*[1], Ziang Liu[1], Shubhangi Sinha[1], Tapomayukh Bhattacharjee[1]

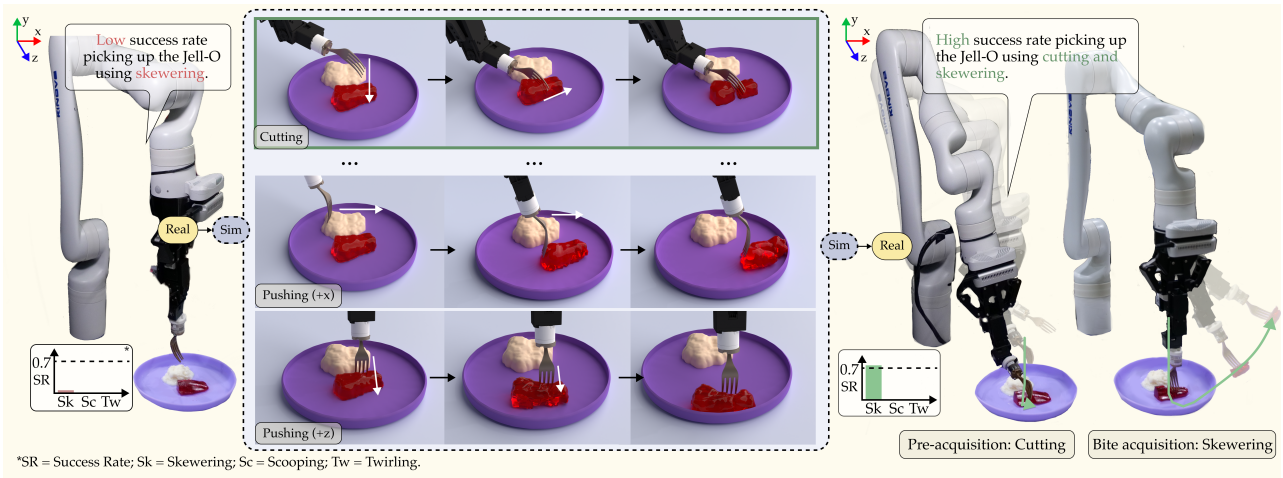


Fig. 1: We propose **REPeat**, a Real2Sim2Real system for pre-acquisition of soft food items. The system evaluates the likelihood of successful bite acquisition; if low, it replicates the setup in simulation to explore various pre-acquisition actions. If a certain pre-acquisition action improves the bite acquisition success rate, the robot executes the pre-acquisition and bite acquisition actions in the real world.

***Abstract*— The paper presents REPeat, a Real2Sim2Real framework designed to enhance bite acquisition in robot-assisted feeding for soft foods. It uses 'pre-acquisition actions' such as pushing, cutting, and flipping to improve the success rate of bite acquisition actions such as skewering, scooping, and twirling. If the data-driven model predicts low success for direct bite acquisition, the system initiates a Real2Sim phase, reconstructing the food's geometry in a simulation. The robot explores various pre-acquisition actions in the simulation, then a Sim2Real step renders a photorealistic image to reassess success rates. If the success improves, the robot applies the action in reality. We evaluate the system on 15 diverse plates with 10 types of food items for a soft food diet, showing improvement in bite acquisition success rates by 27% on average across all plates. See our project website at emprise.cs.cornell.edu/repeat.**

I. Introduction

As of 2021, the World Health Organization [1] reports that approximately 1.3 billion individuals, or 16% of the global population, live with a significant disability. Approximately 142 million of these individuals experience *severe* disabilities [1] that limit their independence in performing basic activities of daily living (ADLs). Among these ADLs, *eating* is especially critical. Robot-assisted feeding systems [2–11] have the potential to improve the quality of life of care recipients and lessen caregiver workload. Previous robot-assisted feeding systems mainly address two subproblems [12]: (i) *bite acquisition* [7, 13, 14], which involves picking up a food item from the plate, and (ii) *bite transfer* [11, 15], which involves moving the food item near or inside the mouth of a care recipient. In this paper, we focus on bite acquisition.

We focus on highly deformable soft diet food items for bite acquisition. A soft diet is essential for care recipients with dysphagia [16, 17], a condition characterized by difficulty in swallowing. This diet includes foods that are easy to chew and swallow [18]. There is a high prevalence of dysphagia with severe mobility limitations, including advanced stages of Amyotrophic Lateral Sclerosis (ALS) (prevalence rate of 80%) [19], and Parkinson's disease (80%) [20].

Soft diet food covers a broad spectrum of rheological properties, including Newtonian fluids with various viscosities (e.g., water), non-Newtonian fluids such as Bingham plastics (e.g., mashed potatoes), and pseudoplastics (e.g., oatmeal) [21, 22], granular solids (e.g., rice) [23], plastic and elastic solids (e.g., banana, avocado, and Jell-O) [24], and composites of the above (e.g., macaroni and cheese) [25]. Bite acquisition is challenging, especially due to the varying rheological properties of soft foods.

Towards developing a robot-assisted feeding system that can handle soft diet food, we take inspiration from human bite acquisition. Humans often use *pre-acquisition* actions to make acquisition easier [7]. For example, pushing consolidates granular items for enhanced support, cutting adjusts food size for easier skewering, and flipping secures a stable surface to prevent the food from rolling during skewering (e.g., flipping a slice of banana on its side to make it rest on the flat surface, preventing it from rolling while skewering).

Similarly, pre-acquisition actions have the potential to

*The two authors contributed equally to the paper.

[1]Computer Science Department, Cornell University {nh285, ry273, zl873, ss3392, tapomayukh}@cornell.edu

This work was partly funded by National Science Foundation IIS #2132846, CAREER #2238792, and DARPA under Contract HR001120C0107. We thank Tom Silver and Rishabh Madan for their feedback.

efficiently enhance robot-assisted bite acquisition success when used intelligently. To explore the large space of pre-acquisition and acquisition actions, one option is to use simulations. Simulation proves particularly valuable for planning and reasoning about irreversible actions, such as cutting food into smaller pieces before performing these actions with an actual robot. Current simulation techniques provide satisfactory models for the dynamics involved in pre-acquisition actions such as moving, tearing, or deforming object [26], thus making simulation a useful tool for developing pre-acquisition control policies. However, simulating bite acquisition is still challenging because it requires accurately modeling friction (e.g. friction between a moist, slippery banana and a fork to determine if banana will slip) which is not trivial and hence, the Sim2Real gap for bite acquisition is high [27].

Our first key insight is that by leveraging a combination of food simulation for exploring pre-acquisition actions (pushing, cutting, flipping) and a real-world data-driven model for estimating bite acquisition success (e.g., whether the food is on the fork or not), we can effectively mitigate this Sim2Real gap. While we realize that SOTA food simulations are far from being accurate [26, 28] in simulating the exact state of food items after manipulation (e.g., determining the exact shape, size, and texture of a lump of mashed potato after pushing), we realize that these simulations are still reliable enough to estimate final approximate food configuration (e.g., determining if a lump of mashed potato is near the plate wall after pushing). This leads to *our second key insight which suggests that food configuration information after pre-acquisition has enough rich information to inform the success rates of future bite acquisition*.

Using this insight, we propose **REPeat**, a Real2Sim2Real approach for the pre-acquisition of soft food items. In the REPeat system, a learned module first estimates the success rate of bite acquisition. We build upon SPANet [3] and develop SPANet-soft, an improved food-action prediction model, particularly for food items from a soft diet. We use this model to decide if it is good enough for direct bite acquisition. If not, the REPeat system performs Real2Sim to create a set of 3D objects in the simulation. In this step, REPeat uses monocular depth estimation to generate 3D models from real-world RGB images. The 3D models are then transferred to a simulation environment (FluidLab [28]) which then uses the Material Point Method (MPM) [29] to simulate the deformation and tearing effect for the exploration of various pre-acquisition actions. We implemented an adaptive sampling module and a render-on-demand module to enhance FluidLab, enabling it to simulate complex food interactions. In the Sim2Real step, we use ControlNet [30], which utilizes simulated depth data to generate realistic RGB images of the final predicted plate configuration after running the pre-acquisition actions. We then evaluate these generated images to determine the bite acquisition success rate and execute the pre-acquisition action that leads to the state with the highest predicted bite acquisition success rate in the real world. Given the lack of high-fidelity food physics simulators, we use an existing MPM-based [31] simulator and design custom simulation environments for soft diet food items. We evaluate the proposed approach on 15 diverse food plates with 10 types of soft food items. Our results demonstrate improvement in bite acquisition success rates, underscoring the effectiveness of our approach in advancing feeding assistance for soft diets.

The main contributions of this paper include:

- REPeat, a Real2Sim2Real framework for physics-informed pre-acquisition of soft diets. It leverages monocular depth estimation for 3D modeling (Real2Sim), employs MPM for realistic food physics simulation, and uses ControlNet (Sim2Real) to generate photorealistic RGB images from simulated outcomes for evaluation of pre-acquisition actions.
- High-fidelity simulation environments for food with various rheological properties supporting pre-acquisition actions with large deformations, fractures, or multi-physics coupling effects.
- An adaptive particle sampling module and a render-on-demand module, enhancing simulation efficiency and enabling the simulation of complex food interactions with the implementation code released.
- SPANet-soft, a network for prediction of food actions tailored to soft diet items.
- Evaluation of our framework on 15 diverse plates with 10 types of soft food items, showing improvement in bite acquisition success rates.

II. Related Work

A. Food Manipulation for Robot-assisted Feeding

Previous work in robot-assisted feeding has predominantly focused on bite acquisition [7, 13, 14] and bite transfer [15, 32], with studies exploring techniques for more effective skewering [3, 33], twirling [9], and scooping [14, 34, 35]. However, research on acquiring foods with different properties such as granular, liquid, and deformable foods [35, 36] has largely been limited to smaller, isolated settings. These studies often fail to capture the complexity of a typical plate setting where food is not isolated, and they lack applicability in caregiving scenarios involving individuals with severe mobility limitations, especially when considering the variety and complexity of soft diet food items. In contrast, our work evaluates a system across realistic combinations of plates consisting of 15 diverse plates with 10 types of food items, based on soft diet recipes, each displaying distinct rheological properties.

Aside from bite acquisition, existing work has also explored non-acquisition actions such as pushing [37], cutting [38], and peeling [39] for food manipulation and meal preparation. Notably, Lee et al. [37] employed pushing as a pre-acquisition action for foods such as lettuce leaves and mashed potatoes. FLAIR [11] uses the few-shot reasoning capabilities of vision-language foundation models to plan and execute bite sequences for various pre-acquisition and bite acquisition skills, taking both user preferences and efficiency into consideration. While FLAIR provides the state-of-the-art

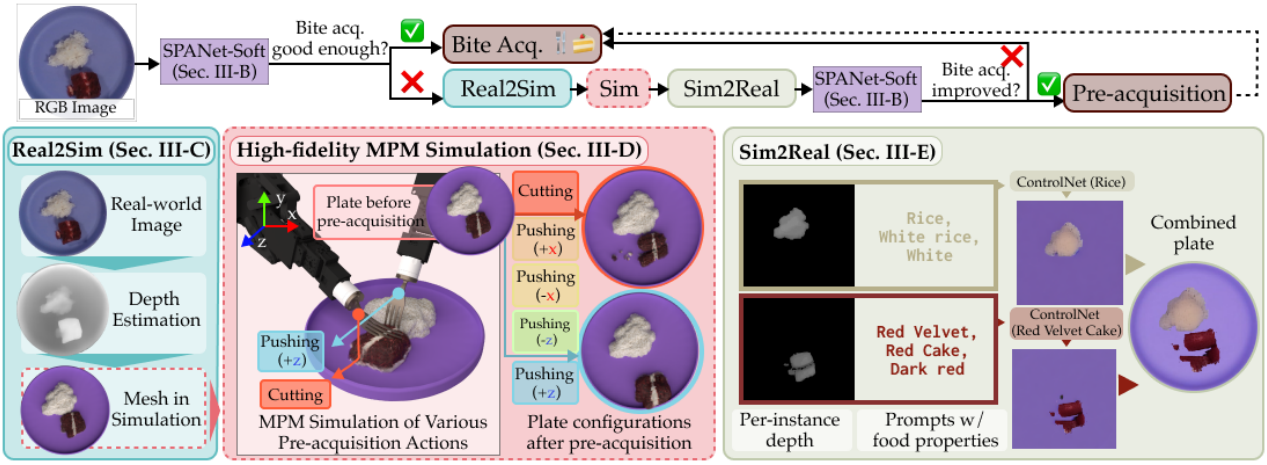


Fig. 2: **Overview of REPeat:** The process begins with SPANet-soft (Sec. III-B) giving an initial estimation of the success rate of bite acquisition. The robot performs direct bite acquisition if the initial estimation of the success rate is higher than a threshold. Otherwise, it enters the Real2Sim2Real loop that consists of: (1) *Real2Sim (Sec. III-C)*: Reconstructing the 3D mesh in real-time with estimated depth as inputs, (2) *Simulation (Sec. III-D)*: Rolling out various pre-acquisition actions using high-fidelity MPM simulation, (3) *Sim2Real (Sec. III-E)*: Rendering a visually realistic picture based on the simulation result. SPANet-soft evaluates the result to compare with the success rate of directly picking up food items without pre-acquisition. If the pre-acquisition action improves the bite acquisition success rate, the robot performs the pre-acquisition action first, followed by the bite acquisition action.

design of a powerful and modular system, due to the lack of accurate physics information in VLMs, the system can sometimes generate unrealistic outputs. Our system builds on these efforts by incorporating a physics-informed evaluation of multiple pre-acquisition actions, including flipping, pushing, and cutting, to effectively perform bite acquisition on a diverse range of soft diet food items.

B. Simulation for Food Manipulation

Simulation techniques for food items fall into two main categories: mesh-based and mesh-free methods. Mesh-based methods predominantly utilize the Finite Element Method (FEM) for food modeling. For example, DiSECt [26] proposes a differentiable FEM-based simulation method to simulate cutting. While these methods offer high physics accuracy, they are computationally intensive due to the need for continuous re-meshing operations. They are also less suited for simulating granular or fluid-like foods. In contrast, mesh-free methods allow more flexibility in deformation modeling, making them the preferred option for simulating fluid-like foods. For instance, FluidLab [28] uses the Material Point Method (MPM) to simulate ice cream, milk, and coffee. MPM has also been used to simulate softer solid foods such as dough [40], given its ability to capture multiphysics coupling, fracture modeling, and large-deformation simulation. This makes it ideal for simulating a variety of soft diet foods with distinct rheological properties. While RCareWorld 1.0 [41], our own developed simulation platform, supports softbody simulation, it does not support MPM simulation yet. Thus, we build our simulation platform upon FluidLab, which allows us to use MPM [31] as the physics backend.

C. Dynamics Modeling for Object Manipulation

Determining the most effective pre-acquisition action requires understanding how each action affects the state of the food. Previous research has explored learning forward predictive dynamics models from interaction data, predicting future system states from current state and action, using different representations such as pixels [42], particles [43], keypoints [44], and latent variables [45]. Although data-driven dynamics models seem sufficient over a short horizon, they typically suffer from error accumulation in long-horizon prediction, and require a substantial volume of physical interaction data for training. Collecting such data in the real world is time-consuming and wastes a lot of food. In addition, it is difficult to revert the food items to their original states once they have been damaged. Physics-based simulation [46] is a viable alternative that provides high-fidelity dynamics modeling. Considering the limitations with real-world food items and the potential need for multiple pre-acquisition steps, we model soft-diet food in a physics-based simulator (MPM), to capture their dynamics.

III. REPeat: Real2Sim2Real Approach for Pre-acquisition of Soft Food Items

REPeat leverages a Real2Sim2Real approach for selecting pre-acquisition actions (Fig. 2). It takes in an RGB image of the food items on the plate, and predicts a pre-acquisition or bite acquisition action as the output for the robot to take. The system begins by using SPANet-soft (Sec. III-B), a data-driven bite acquisition success rate estimation module, to determine whether direct bite acquisition is good enough for a particular piece of food. If direct bite acquisition is predicted to be challenging, the system transitions to a Real2Sim step III-C, creating a simulated environment to replicate the food items on the plate. In simulation III-D, the system explores various pre-acquisition actions by executing each action once. Following these actions, it performs a Sim2Real step III-E to render a photorealistic image for SPANet-soft to estimate the success rate of the bite acquisition actions. After this, the robot executes the pre-acquisition action that leads to the most significant increase in the bite acquisition success rate. We next describe each of these modules in detail (see also the supplementary materials on our website [47]).

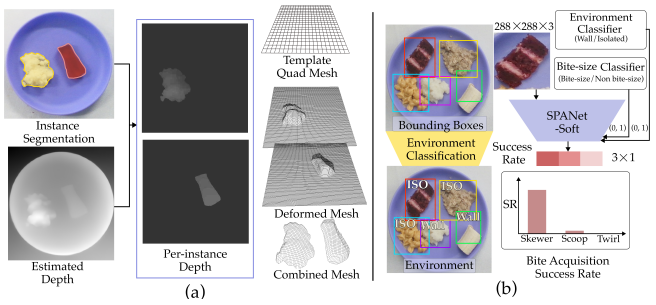


Fig. 3: **(a)** Deformation of the template quad mesh for food mesh reconstruction: Using the RGB image from the camera, we perform instance segmentation, and apply the segmentation mask to the depth map to obtain per-instance depth images. We then use the values of these depth maps as displacement map to deform a template quad mesh. **(b)** The structure of SPANet-soft.

A. Action Space: Pre-acquisition and Bite Acquisition

In the REPeat system, we consider two types of actions: pre-acquisition actions and bite acquisition actions. We derive the actions marked with '*' from the FLAIR system [11] and parameterized them in an identical way. We give a high-level overview of the actions here, and provide further details of the parameterizations for each of these skills on our website.

The pre-acquisition actions include:

- Pushing: We define pushing as a linear motion along the positive or negative direction of the x or z axis, resulting in 4 pushing actions. The pushing action terminates when the food touches other food items or the wall of the plate.
- Cutting*: We bring the fork horizontal, and rotate the tine to be on its side. We execute a swift downward trajectory to cut the food item, and then a quick flick to separate the two pieces of food items.
- Flipping: We bring the tines to the side of the food item, parallel to the major axis. Then, the fork moves quickly perpendicular to the major axis and slightly upward.

The bite acquisition actions include:

- Skewering*: We bring the fork above the center of the food item and then rotate the tines so that the tines are perpendicular to the major axis. We then bring the fork down to let it pierce the food item. After that, the fork performs a scoop-like motion to secure the food item, and brings it up.
- Scooping*: We bring tines to a configuration that is horizontal to the plate, and scoop from the sparsest region to the closest wall support.
- Twirling*: We perform twirling on the densest pile of noodles and actuate the roll joint of the fork to perform 2 full twirls.

B. SPANet-soft: Food-Action Prediction for Soft Diet Food

Due to the Sim2Real gap in bite acquisition, we develop the data-driven SPANet-soft module to provide an empirical estimate of the success rate of each bite acquisition action.

Following the structure of SPANet [3], SPANet-soft takes in an RGB image of the plate as the input, and predicts the success rate of the bite acquisition actions for each piece of food item as the output. The pipeline has the following components:

- Action Space: SPANet-soft predicts the success rate for 3 bite acquisition actions, namely skewering, scooping, and twirling derived from FLAIR [11].
- Food Detection: We replace RetinaNet [48], previously used to generate the bounding boxes for the individual food items on the plate, with Grounded-SAM [49].
- Environment Classifier: We encode the environment surrounding the target food item as a one-hot vector, representing two conditions: (1) Isolated: The target food item is in an empty surrounding. (2) Wall: The target food item is either near the edge of the plate or surrounded by other food pieces.
- Bite-size Classifier: The classifier estimates the volume of the target food item based on the average height and area extracted from the segmentation mask, providing a one-hot output to indicate whether the item is bite-sized.

We show the structure of the pipeline in Fig 3(b). The SPANet-soft module takes in a $288 \times 288 \times 3$ RGB image cropped and resized using the bounding box of a piece of food item, along with a 2×1 vector for environment classification and a 2×1 vector for bite-size classification. We concatenate these two vectors with the image feature vector from the base network. It predicts a 3×1 vector containing the success rate for skewering, scooping, and twirling. To train this network, we collect empirical success rates based on real-robot bite acquisition for the 10 types of food items and provide the corresponding images to the network in a way identical to SPANet. We release the dataset on our website [47]. We use the smooth L1 loss [50] between the ground truth vector and the predicted one to ensure the model learns the success rate distribution as closely as possible. Other training details are identical to SPANet.

C. Real2Sim: Mesh Reconstruction for Food Items

When SPANet-soft predicts that direct bite acquisition is likely to fail, the Real2Sim2Real pre-manipulation pipeline begins. The Real2Sim step reconstructs a high-quality mesh of the food. This task is challenging since the food items are small and can be subject to noise due to moisture and reflective surfaces. This noise makes it almost impossible to reconstruct the meshes accurately with depth from the sensors. To address this issue, we apply DepthAnything [51] to perform monocular depth estimation. We reuse the segmentation masks generated by Grounded-SAM [49] for SPANet-soft and obtain the corresponding depth for each food item. The conventional approaches of reconstructing the mesh with Poisson surface reconstruction [52], alpha shapes [53], or ball pivoting [54] are usually slow to create a high-fidelity mesh. Based on the idea of deforming a template mesh in DepthLab [55], we opt for a more computationally efficient method to create the mesh in real time.

Consider $\mathcal{M}$ as a template quadrilateral mesh, consisting of vertices $V = \{p_1, p_2, \ldots, p_n\}$, where each vertex $p_i \in \mathbb{R}^3$ is represented by $p_i = (x_i, y_i, z_i)$. Given a depth image $D : \mathbb{R}^2 \to \mathbb{R}$, where $D(u, v)$ specifies the displacement at

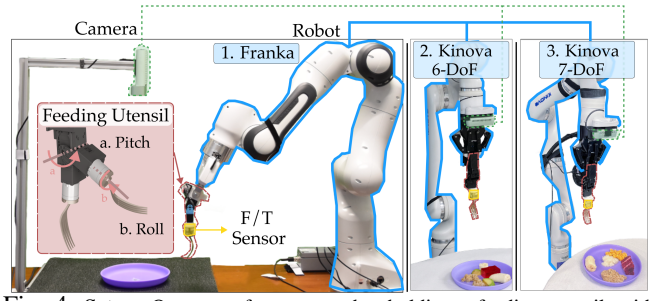


Fig. 4: **Setup:** Our setup features a robot holding a feeding utensil, with a camera for perception and an F/T sensor to detect the end of the pushing action. It is adaptable to various robot embodiments and camera placements (frame or wrist-mounted). The figure shows 3 setups: 1. Franka robot with a camera mounted on a frame 2. Kinova 6-DoF robot with a camera mounted on the wrist 3. Kinova 7-DoF robot with a camera mounted on the wrist. The utensil has 2 DoFs: (a) Pitch, performing a scoop-like motion (b) Roll, performing a twirl-like motion.

the corresponding mesh surface point, and assuming mesh $\mathcal{M}$ and depth image D share the same resolution, each pair (u_i, v_i) maps directly to a point (x_i, y_i). We update the position of each vertex by displacing it in the direction of its normal by the depth value, resulting in a new position v_i' defined as: $p_i' = p_i + D\,(u_i, v_i) \cdot n_i$ where n_i is the normal at vertex p_i and $D\,(u_i, v_i)$ is the displacement value from the depth map.

This deformation process is applied to every vertex in the mesh, effectively deforming the entire mesh according to the depth information. This leads to an updated set of vertices $V' = \{p_1', p_2', \ldots, p_n'\}$, which outlines the newly deformed mesh $\mathcal{M}'$.

D. Sim: High-fidelity Simulation for Food Items using MPM

We use the Moving Least Squares Material Point Method (MLS-MPM) [31] to simulate the pre-acquisition actions. MLS-MPM effectively handles complex phenomena such as deformation, fracture, and multi-physics coupling. We use the mesh of the food items obtained in the Real2Sim step (detailed in Sec. III-C) as the input. In addition to the food item meshes, the environment includes a fork that interacts with the food items, and a silicone plate. We model the food using 3 types of constitutive models, namely the plastic model and elastic model following MLS-MPM [31], and the elastoplastic model following PlasticineLab [56]. While the actual rheological taxonomy of the food items is more complicated, we approximate them using these 3 models by selecting the closest type of constitutive model for each type of food, with a fixed set of Young's modulus and Lamé constants determined following the parameters in FluidLab [28] detailed on our website [47]. We model the fork and the plate as rigid objects using time-varying Signed Distance Fields (SDFs) created using their mesh files. We simulate the frictional interaction between soft food and rigid objects by calculating the surface normals of the SDFs and applying Coulomb friction [57].

FluidLab supports simple interactions with limited objects but cannot simulate the more complex food items on a plate. We implemented 2 modules to enable simulation with multiple objects for a longer horizon. First, we implemented an *adaptive particle sampling* module. It assigns specific densities to each food type instead of using a uniform density for all objects. This allows the simulation of complex food deformations and fractures on a memory-limited GPU. Additionally, we implemented a *render-on-demand* module. It optimizes rendering by generating depth information only after the pre-acquisition action, reducing computational load and enabling long-horizon tasks such as cutting and flipping. These two modules are compatible with FluidLab, enabling it to handle more complex tasks and enhancing its usability. We release this implementation on our project website [47].

E. Sim2Real: Rendering Realistic Images from Sim

After obtaining the predicted final plate configuration for each pre-acquisition action in simulation, we evaluate the predicted plate for success of bite acquisition to select the best pre-acquisition action. To perform this evaluation, we pass the RGB image of the final plate states to SPANet-soft. Although our simulation offers high accuracy in physics, it falls short in visual realism, introducing a Sim2Real gap. We address this gap by generating visually realistic images of the final plate state using the simulated depth data. In particular, we use ControlNet [30], a generative model that takes in a control condition and generates an image. Our inputs include depth for the 3D geometry and food category names to generate texture appearance in the pixel domain. We train a category-level ControlNet for each type of food by collecting a dataset of RGB images of the food items on the plate for each category of food items. We also wrote corresponding prompts to provide food properties to the ControlNet. We then create the mesh of the food items in the simulation, and render the depth image of the food item. Using the depth image and prompt as input, we train it to make the generated RGB images as realistic as possible. The network structure, loss function, and other details are identical to the original ControlNet implementation [30]. We detail other data collection and training process of the ControlNet on our website [47]. Fig. 2 shows the input and example output from ControlNet.

IV. Evaluation

We evaluate the system across 15 unique plates each featuring combinations of 10 different types of food to verify the hypothesis that pre-acquisition actions can help improve the success rate of bite acquisition actions for soft diets. We evaluate REPeat by comparing it to a baseline that is without pre-acquisition actions.

A. Experiment Setup

a) Hardware: We show the real-world hardware setup in Fig. 4. We adopt the utensil proposed by [9]. The utensil gives the robot two extra degrees of freedom: the roll and the pitch, and allows precise control of the utensil for actions such as cutting and pushing, while the robot manages the movement between different points in the workspace using Cartesian position control. This feature helps adapt the actions to different robot embodiments. We adapt the utensil to fit the ATI Nano25 F/T sensor for force/torque

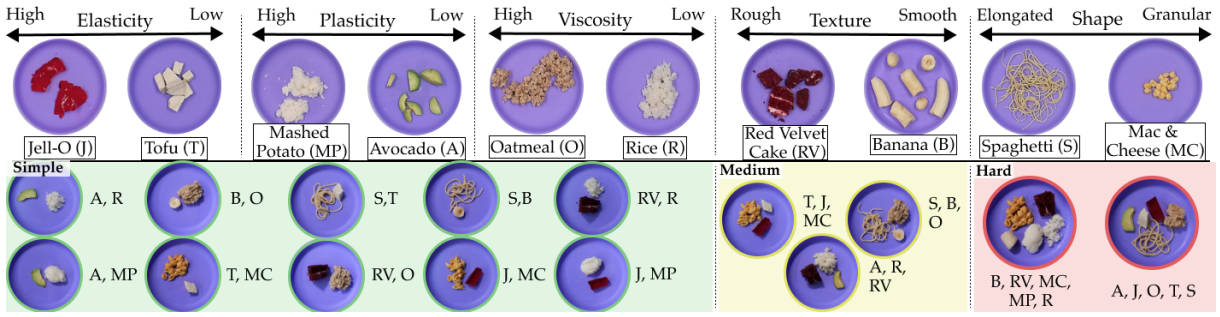


Fig. 5: **Upper** 5 axes corresponding to the characteristics of different food items and 10 food types selected to represent the extremes. **Lower** We evaluate the REPeat system on the following 15 plates containing 10 types of food items. J: Jell-O, MP: Mashed Potato, R: Rice, O: Oatmeal, B: Banana, S: Spaghetti, RV: Red velvet cake, A: Avocado, MC: Mac and cheese, T: Tofu.

sensing. Our framework is robot-and-camera setup-agnostic. We evaluated it using three embodiments: a Kinova Gen3 6-DoF and 7-DoF robot, and a Franka Emika Panda 7-DoF robot. For the two Kinova robots, we use RealSense D435 cameras mounted on their wrists. For the Franka robot, we use an Azure Kinect camera mounted on a fixed frame. We also use a non-slip silicone plate commonly used in caregiving setups [58] to better simulate the environment of feeding care recipients and prevent the plate from moving.

b) Simulation: To simulate soft diet food items, we predefine each food's properties depending on relevant rheological factors. The parameters of the simulation include particle density, elasticity, time step, and shear module, which are calibrated based on food properties and a predetermined set of material properties following Fluidlab [28]. We detail the parameters used for the simulation settings on the project website [47].

c) System Details: The core operations of our system, including data processing activities, are managed by an AMD Ryzen 9 5900X CPU with a base clock speed of 3.7 GHz. For simulation tasks and inference using SAM, DepthAnything, and ControlNet that require acceleration in graphics rendering, we use an NVIDIA GeForce RTX 3090 GPU with 24 GB GDDR6X memory. We utilize 32 GB DDR4 RAM to support faster operations.

d) Food Selection: We select 10 types of food representing a diverse range of rheological properties. As illustrated in Fig. 6 (b), we evaluate our system with Jell-O, tofu, mashed potato, avocado, oatmeal, rice, red velvet cakes, bananas, spaghetti, and macaroni and cheese (also called mac & cheese). These items cover the extremes of the five properties, including elasticity, plasticity, viscosity, texture, and shape, which can affect the bite acquisition success rate. We designed three difficulty levels based on the amount of clutter: simple, medium, and hard. Simple plates have 2 pieces covering 40%, medium plates 3 pieces covering 60%, and hard plates 5 pieces covering 80% of the plate surface area.

B. Evaluation Procedure

We compare our system (w/ pre-acquisition) with a baseline (w/o pre-acquisition) to evaluate its effectiveness. The baseline directly uses SPANet-soft to perform the bite acquisition action that leads to the highest success rate. Based on a bite acquisition success rate set by Gordon et al. [13], if the success rate from SPANet-soft is higher than 70%, we perform a direct bite acquisition. When the success rate is lower, the system performs the pre-acquisition actions that leads to the highest predicted bite acquisition success rate. If the pre-acquisition action fails, we repeat it again once. We perform the bite acquisition action after that.

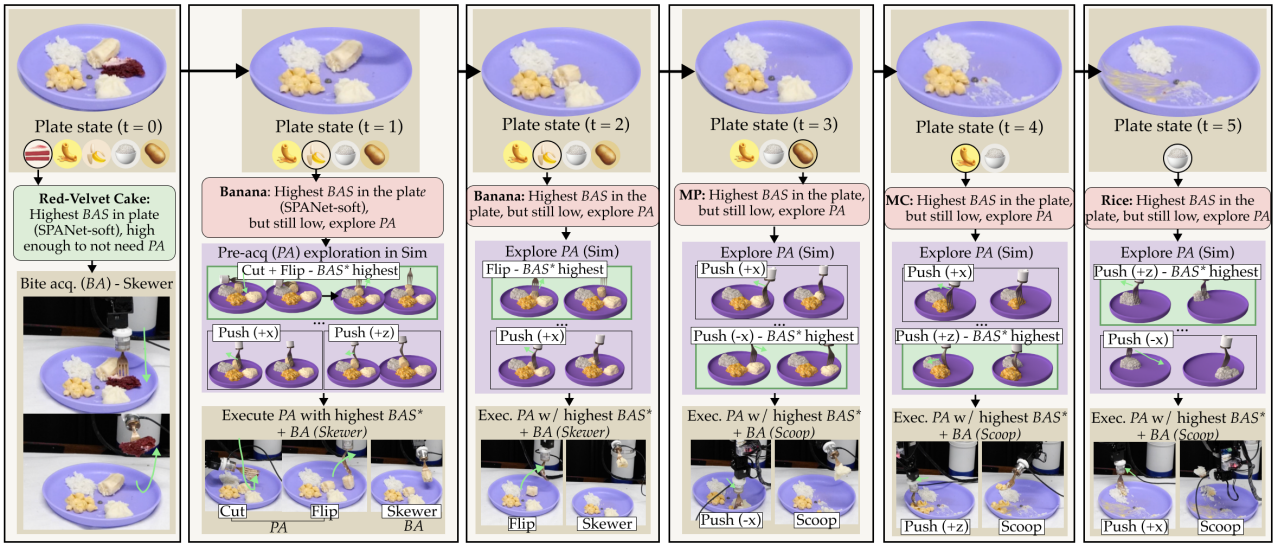
We use the metric defined in [13] for bite acquisition. After acquisition, the food must remain on the fork for 3 seconds, the time needed to move it to a care recipient's mouth. Also, we evaluate whether food item is bite-sized using the quantitative metrics from [13] as the minimum threshold and those from [11] as the maximum threshold.

C. Result

We present the category-level success rate comparison between our method and the baseline in Fig. 7 (a). Results indicate that pre-acquisition actions, on average, enhance the bite acquisition success rate by 27%. We perform chi-square significant tests on the success rates for each food item. The result suggests REPeat performs significantly better for Jell-O, mashed potato, rice, oatmeal, non-bite-sized banana trunk, red velvet cake, mac and cheese, and tofu with p-value < 0.05. We show an execution example for one of the hard plates with 5 pieces of food items in Fig. 5 and show the other examples on our website [47]. We show some of the typical failure cases in Fig. 7 (c).

The effectiveness of pre-acquisition actions in improving bite acquisition can be attributed to the following factors (illustrated in Fig. 7 (b)):

- Pushing consolidates granular food such as mashed potatoes, rice, and mac&cheese, enhancing scooping success by preventing slippage from the fork. Also, it helps to move the food items near a wall (the wall of the plate or other food items), preventing the food from slipping away.
- Flipping exposes flat surfaces, crucial for successful skewering actions, as seen with banana slices, where flipping prevents the slice on the side from rolling away during skewering.
- Cutting fragile food items such as Jell-O into bite-sized pieces helps feeding, and also helps the food items maintain their shape, reducing breakage and the chance of falling off during acquisition.



BA: Bite acquisition action; *PA:* Pre-acquisition action; *BAS*: Bite acquisition success rate; *BAS**: Bite acquisition success rate after pre-acquisition action

Fig. 6: An example robot execution sequence for one of the hard plates with five food items. The robot uses pre-acquisition skills such as cutting, flipping, and pushing.

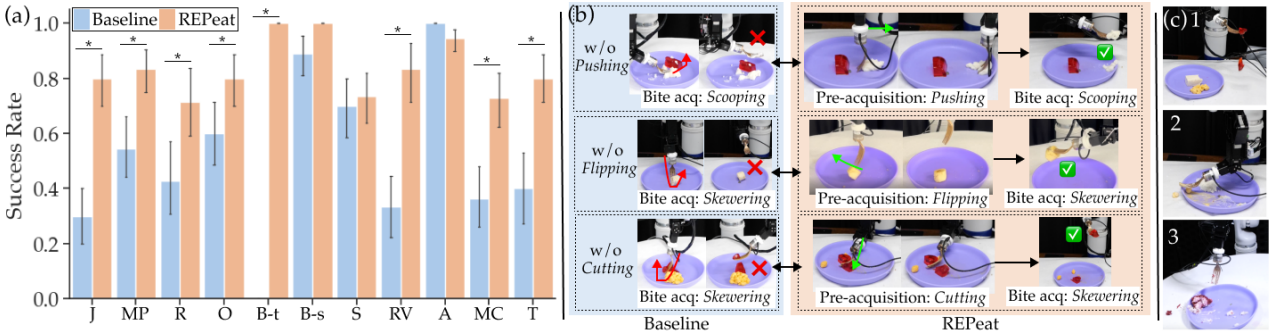


Fig. 7: **(a)** Success rate of bite acquisition for the 10 types of food items using the baseline method compared with REPeat. J: Jell-O, MP: Mashed Potato, R: Rice, O: Oatmeal, B-t: Non-bite-sized banana trunk, B-s: Bite-sized banana slice, S: Spaghetti, RV: Red velvet cake, A: Avocado, MC: Mac and cheese, T: Tofu. **(b)** Examples of how the pre-acquisition actions help the bite acquisition. **(c)** Typical failure cases: 1. Fragile food breaks into pieces. 2. Granular food spills out of the wall of the plate. 3. Multiple food items mixed together and identified as one piece of food item, confusing the perception module. In this case, the white rice is mixed with the white cream on the red velvet cake. Both the cream and the rice are white, making them very similar. Therefore, rice is confused as the cream, making it identified as an entire piece of food.

V. Discussion

REPeat takes the first step towards performing physics-informed pre-acquisition actions for a wide variety of soft diet food items. We evaluate the method across 3 different embodiments, 15 combinations of 10 types of food items with various rheological properties. Our evaluation demonstrated that performing physics-informed pre-acquisition actions can increase the success rate of the bite acquisition of soft diet food items.

Through our evaluation, we identify the following limitations that can potentially be resolved and help improve the system in the future:

- Time-varying food properties: For the food items for soft diets, the rheological properties such as moisture, etc., that affect bite acquisition may vary significantly if placed at room temperature during the course of a meal. In our experiments, we capture food items' properties for each plate beforehand. However, in real-world feeding, strategies to address time-varying food properties would be beneficial.
- Food perception: The VLM (GroundedSAM) can detect and segment various food items in our setup. However, due to the visual variety of the food items, we had to carefully construct the prompts. For example, we specified the red velvet cake as "red velvet; red brick; red cake; burgundy cake; dark red cake; brown cake; maroon cake; dark purple cake". Advancements in open-set detection and segmentation VLMs might improve the perception pipeline.
- Food simulation: The method we use (MPM) is computation-heavy and hard to balance between fidelity and speed. We use adaptive sampling and on-demand rendering, making it possible to simulate various food items, but simulating in-the-wild dishes still remain challenging. Future simulation advancements can improve the REPeat system.

Despite these limitations, our system demonstrated that the use of a Real2Sim2Real framework can help improve the bite acquisition of soft diet food. With future advancements in online food parameter identification, VLMs for perception, and food simulation, we will be able to benefit people with severe mobility limitations who require a soft diet by integrating REPeat with bite-transfer for real-world feeding.

REFERENCES

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

[2] 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)*, pp. 9659–9666, IEEE, 2020.

[3] 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*, pp. 427–442, Springer, 2019.

[4] 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*, pp. 24–33, 2023.

[5] T. Bhattacharjee, E. K. Gordon, R. Scalise, M. E. Cabrera, A. Caspi, M. Cakmak, and S. S. Srinivasa, "Is more autonomy always better? exploring preferences of users with mobility impairments in robot-assisted feeding," in *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 181–190, IEEE, 2020.

[6] D. Park, Y. Hoshi, H. P. Mahajan, H. K. Kim, Z. Erickson, W. A. Rogers, and C. C. Kemp, "Active robot-assisted feeding with a general-purpose mobile manipulator: Design, evaluation, and lessons learned," *Robotics and Autonomous Systems*, vol. 124, p. 103344, 2020.

[7] 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*, vol. 4, no. 2, pp. 1485–1492, 2019.

[8] 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)*, pp. 10528–10535, IEEE, 2021.

[9] P. Sundaresan, J. Wu, and D. Sadigh, "Learning sequential acquisition policies for robot-assisted feeding," in *Conference on Robot Learning*, pp. 1282–1299, PMLR, 2023.

[10] J. Ondras, A. Anwar, T. Wu, F. Bu, M. Jung, J. J. Ortiz, and T. Bhattacharjee, "Human-robot commensality: Bite timing prediction for robot-assisted feeding in groups," in *2022 SoCal Robotics Symposium*, 2022.

[11] R. K. Jenamani, P. Sundaresan, M. Sakr, T. Bhattacharjee, and D. Sadigh, "FLAIR: Feeding via Long-Horizon Acquisition of Realistic dishes," in *Robotics: Science and Systems (RSS)*, 2024.

[12] R. Madan, R. K. Jenamani, V. T. Nguyen, A. Moustafa, X. Hu, K. Dimitropoulou, and T. Bhattacharjee, "Sparcs: Structuring physically assistive robotics for caregiving with stakeholders-in-the-loop," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 641–648, IEEE, 2022.

[13] 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*, pp. 2414–2428, PMLR, 2023.

[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 *Conference on Robot Learning*, pp. 849–865, PMLR, 2023.

[15] H. K. Kim, H. Jeong, J. Park, J. Park, W.-S. Kim, N. Kim, S. Park, and N.-J. Paik, "Development of a comprehensive design guideline to evaluate the user experiences of meal-assistance robots considering human-machine social interactions," *International Journal of Human–Computer Interaction*, vol. 0, no. 0, pp. 1–14, 2022.

[16] C. Jacobsson, K. Axelsson, P. O. Österlind, and A. Norberg, "How people with stroke and healthy older people experience the eating process," *Journal of Clinical Nursing*, vol. 9, no. 2, pp. 255–264, 2000.

[17] Healthdirect Australia, "Dysphagia (difficulty swallowing) – symptoms, causes and treatment." `https://www.healthdirect.gov.au/dysphagia`, 2022. Accessed: 2024-03-03.

[18] T. I. D. D. S. Initiative, 2019.

[19] E. Onesti, I. Schettino, M. C. Gori, V. Frasca, M. Ceccanti, C. Cambieri, G. Ruoppolo, and M. Inghilleri, "Dysphagia in amyotrophic lateral sclerosis: Impact on patient behavior, diet adaptation, and riluzole management," *Frontiers in Neurology*, vol. 8, Mar 2017.

[20] I. Suttrup and T. Warnecke, "Dysphagia in parkinson's disease," *Dysphagia*, vol. 31, no. 1, pp. 24–32, 2016.

[21] R. M. Gresham, "Learning more about viscosity," *Tribology & Lubrication Technology*, vol. 61, pp. 24–27, 2005.

[22] C. Zhou, M. Wu, D. Sun, W. Wei, H. Yu, and T. Zhang, "Twin-screw extrusion of oat: Evolutions of rheological behavior, thermal properties and structures of extruded oat in different extrusion zones," *Foods*, vol. 11, p. 2206, July 2022.

[23] G. Guatemala, F. Santoyo, L. Virgen, R. Corona, and E. Arriola, "Hydrodynamic model for the flow of granular solids in the s-valve," *Powder Technology*, vol. 230, pp. 77–85, 2012.

[24] G. KALETUNC, M. NORMAND, E. JOHNSON, and M. PELEG, ""degree of elasticity" determination in solid foods," *Journal of Food Science*, vol. 56, no. 4, pp. 950–953, 1991.

[25] M. H. Tunick, "Cheese Rheology and Texture," in *Handbook of Cheese Chemistry*, The Royal Society of Chemistry, 07 2023.

[26] E. Heiden, M. Macklin, Y. Narang, D. Fox, A. Garg, and F. Ramos, "Disect: A differentiable simulation engine for autonomous robotic cutting," 2021.

[27] M. Li, Z. Ferguson, T. Schneider, T. Langlois, D. Zorin, D. Panozzo, C. Jiang, and D. M. Kaufman, "Incremental potential contact: intersection-and inversion-free, large-deformation dynamics," *ACM Trans. Graph.*, vol. 39, aug 2020.

[28] Z. Xian, B. Zhu, Z. Xu, H.-Y. Tung, A. Torralba, K. Fragkiadaki, and C. Gan, "Fluidlab: A differentiable environment for benchmarking complex fluid manipulation," in *International Conference on Learning Representations*, 2023.

[29] D. Sulsky, Z. Chen, and H. L. Schreyer, "A particle method for history-dependent materials," *Computer Methods in Applied Mechanics and Engineering*, vol. 118, pp. 179–196, 1993.

[30] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[31] Y. Hu, Y. Fang, Z. Ge, Z. Qu, Y. Zhu, A. Pradhana, and C. Jiang, "A moving least squares material point method with displacement discontinuity and two-way rigid body coupling," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.

[32] 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 *2024 19th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2024.

[33] P. Sundaresan, S. Belkhale, and D. Sadigh, "Learning visuo-haptic skewering strategies for robot-assisted feeding," in *6th Annual Conference on Robot Learning*, 2022.

[34] J. Grannen, Y. Wu, S. Belkhale, and D. Sadigh, "Learning bimanual scooping policies for food acquisition," in *6th Annual Conference on Robot Learning*, 2022.

[35] A. Bhaskar, R. Liu, V. D. Sharma, G. Shi, and P. Tokekar, "Lava: Long-horizon visual action based food acquisition," *arXiv preprint arXiv:2403.12876*, 2024.

[36] P. T. Rui Liu, Amisha Bhaskar, "Adaptive visual imitation learning for robotic assisted feeding across varied bowl configurations and food types," 2024.

[37] G. Lee, T. Bhattacharjee, and S. S. Srinivasa, "Bite acquisition of soft food items via reconfiguration,"

[38] 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)*, pp. 409–416, IEEE, 2019.

[39] "Morpheus: a multimodal one-armed robot-assisted peeling system with human users in-the-loop." `https://emprise.cs.cornell.edu/morpheus/`, 2023. Accessed: 2024-03-06.

[40] X. Lin, Z. Huang, Y. Li, J. B. Tenenbaum, D. Held, and C. Gan, "Diffskill: Skill abstraction from differentiable physics for deformable object manipulations with tools," *Proc. Int. Conf. on Learning Representations*, 2022.

[41] R. Ye, W. Xu, H. Fu, R. K. Jenamani, V. Nguyen, C. Lu, K. Dimitropoulou, and T. Bhattacharjee, "Rcareworld: A human-centric simulation world for caregiving robots," *IROS*, 2022.

[42] C. Allen-Blanchette, S. Veer, A. Majumdar, and N. E. Leonard, "Lagnetvip: A lagrangian neural network for video prediction," 2020.

[43] Y. Li, T. Lin, K. Yi, D. M. Bear, D. L. K. Yamins, J. Wu, J. B. Tenenbaum, and A. Torralba, "Visual grounding of learned physical models," 2020.

[44] L. Manuelli, Y. Li, P. R. Florence, and R. Tedrake, "Keypoints into the future: Self-supervised correspondence in model-based reinforcement learning," *Conf. on Robot Learning (CoRL)*, vol. abs/2009.05085, 2020.

[45] E. Ng, Z. Liu, and M. K. I. au2, "It takes two: Learning to plan for human-robot cooperative carrying," 2023.

[46] T. Howell, N. Gileadi, S. Tunyasuvunakool, K. Zakka, T. Erez, and Y. Tassa, "Predictive sampling: Real-time behaviour synthesis with mujoco," 2022.

[47] E. Lab, "Repeat: A real2sim2real approach for pre-acquisition of soft food items in robot-assisted feeding." `https://emprise.cs.cornell.edu/repeat`, 2024. (Accessed: 17th March, 2024).

[48] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[49] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded sam: Assembling open-world models for diverse visual tasks," 2024.

[50] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, IEEE, 2015.

[51] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *CVPR*, 2024.

[52] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, SGP '06, (Goslar, DEU), p. 61–70, Eurographics Association, 2006.

[53] H. Edelsbrunner and E. P. Mücke, "Three-dimensional alpha shapes," *ACM Transactions on Graphics*, vol. 13, no. 1, pp. 43–72, 1994.

[54] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin, "The ball-pivoting algorithm for surface reconstruction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 5, no. 4, pp. 349–359, 1999.

[55] R. Du, E. Turner, M. Dzitsiuk, L. Prasso, I. Duarte, J. Dourgarian, J. Afonso, J. Pascoal, J. Gladstone, N. Cruces, S. Izadi, A. Kowdle, K. Tsotsos, and D. Kim, "Depthlab: Real-time 3d interaction with depth maps for mobile augmented reality," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, (New York, NY, USA), p. 829–843, Association for Computing Machinery, 2020.

[56] Z. Huang, Y. Hu, T. Du, S. Zhou, H. Su, J. B. Tenenbaum, and C. Gan, "Plasticinelab: A soft-body manipulation benchmark with differentiable physics," in *International Conference on Learning Representations*, 2021.

[57] A. Stomakhin, C. Schroeder, L. Chai, J. Teran, and A. Selle, "A material point method for snow simulation," *ACM Trans. Graph.*, vol. 32, jul 2013.
[58] 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 realistic plate," 2019.