FLAIR: Feeding via Long-Horizon AcquIsition of Realistic dishes

Author Names Omitted for Anonymous Review. Paper-ID 472

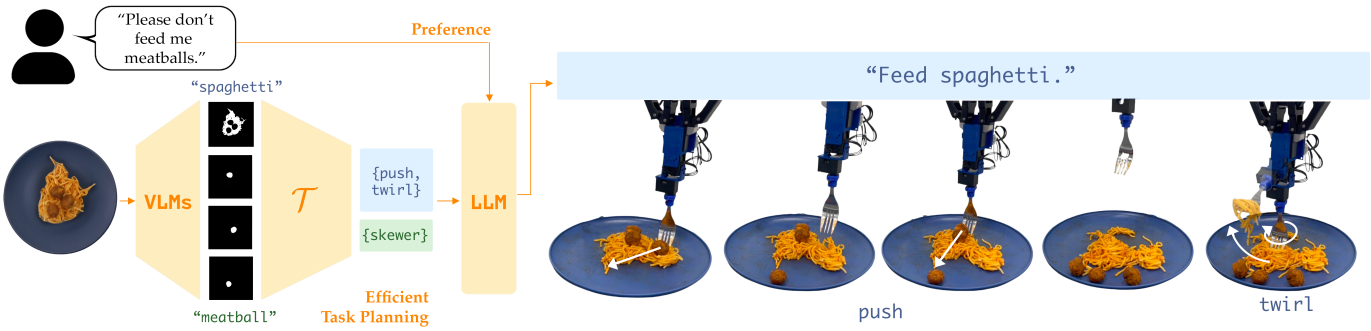

Fig. 1: We introduce FLAIR, a system for long-horizon feeding that combines foundation models' commonsense reasoning with a library of parameterized skills. In a user study with 42 participants considering 6 realistic plates, FLAIR demonstrates its ability to efficiently clear plates while meeting user preferences. We also showcase FLAIR's real-world application by successfully feeding a care recipient with severe mobility limitations.

***Abstract*—Robot-assisted feeding holds immense promise for improving the quality of life for individuals with mobility limitations who are unable to feed themselves independently. However, there exists a large gap between the homogeneous, curated plates existing assistive feeding systems can handle, and truly in-the-wild meals. Feeding realistic plates is immensely challenging due to the sheer range of food items that a robot may encounter, each requiring specialized manipulation strategies which must be sequenced over a long-horizon to feed an entire meal. An assistive feeding system should not only be able to sequence different strategies *efficiently* in order to feed an entire meal, but also be mindful of user *preferences* given the personalized nature of the task. We address this with FLAIR , a system for long-horizon feeding which leverages the commonsense reasoning capabilities of foundation models, along with a library of parameterized skills, to plan and execute user-preferred and efficient bite sequences. In real-world evaluations across 6 realistic plates, we find that FLAIR can effectively tap into a library of dexterous skills for efficient plate clearance, while adhering to the diverse preferences of 42 participants without mobility limitations as evaluated in a user study. We demonstrate the real-world efficacy of our approach by deploying our system with an in-mouth bite transfer framework for successfully feeding a care recipient with mobility limitations.**

I. Introduction

Eating is a vital part of everyday life, yet millions worldwide struggle to feed themselves independently due to mobility limitations caused by conditions such as neurological disorders, injuries, the effects of aging, or other health complications [32]. These individuals often rely on caregivers for meal assistance, which impacts their sense of independence, daily routines, and the social experience of dining [18, 24, 27]. Moreover, feeding is one of the most time-consuming Activities of Daily Living (ADL) for caregivers [10]. A system for autonomous mealtime assistance holds promise for improving the quality of life for those requiring assistance [7], and reducing the physical workload on caregivers [6, 19].

Robot-assisted feeding entails first performing *bite acquisition* [11, 13–16, 28, 30], where the robot must manipulate a utensil to pick up a bite of food, followed by *bite transfer* [5, 12, 19, 25, 26], or bringing a bite of food to the mouth for consumption. Transfer is thus critically not possible without a pre-acquired bite, highlighting the need to tackle reliable bite acquisition in the first place. Several recent assistive feeding systems demonstrate joint bite acquisition and transfer [6, 12, 19], but largely focus on self-curated plates of fruits and vegetables. The challenge of achieving truly reliable bite acquisition for dishes encountered *in-the-wild*, however, persists. For example, consider a robot tasked with feeding a user a restaurant-ordered meal of spaghetti and meatballs, mashed potatoes, and tiramisu. Not only must the robot perform highly specialized strategies like twirling noodles, scooping mashed potatoes, and cutting cake, it must also infer how to chain these strategies over a *long-horizon*. Feeding a complete meal with vastly different food items thus proves to be a challenging interplay of dexterous low-level manipulation, along with nuanced long-horizon sequential decision-making.

Several prior works in bite acquisition have focused on developing individual skills targeting specific food groups. This includes policies for skewering firm foods [11, 13–15, 28], scooping soft foods [16, 31], or rearranging and twirling noodles [30]. These works, however, mostly operate over a single bite horizon or consider plates with a homogenous type of food (e.g. only noodles or only fruits). They do not consider planning long-horizon actions sequences to handle multiple different food types within the same meal. Imagine a robot feeding the aforementioned meal to a user who prefers to not eat the meatballs. The robot may first need to push aside any meatballs on top of the noodles, then group the

remaining noodles into a pile, and finally twirl for a bite. On the other hand, an ambivalent user might be perfectly satisfied with the more efficient choice of skewering a meatball and eating it immediately. These scenarios force the robot to jointly reason across multiple foods and timesteps while tapping into a diverse set of skills. The key idea for sequencing bites over an extended horizon is to choose *efficient* actions, without compromising a user's *preferred* ordering of bites. We desire a system that considers both criteria, achieves long-horizon bite acquisition via a library of skills, and finally integrates with frameworks for bite transfer [5, 12, 19, 23, 26] to effectively feed complete meals.

In this work, we introduce FLAIR (Feeding via Long-horizon AcquIsition of Realistic dishes), a robot-assisted feeding system capable of feeding a complete plate to a care recipient. Given a plate image, and a user-provided natural language preference specifying their desired feeding strategy (i.e. 'I prefer to alternate bites of X and Y' or 'Don't feed me X'), FLAIR aims to plan and execute a sequence of actions that efficiently clears the plate while feeding bites in a preferred order. The framework starts by detecting food items and their semantic labels (i.e. `'spaghetti'`) via Vision-Language Models (VLMs). We then pass the visual state estimate and semantic label for all items to a task planner, which outputs per-item efficiencies by proxy of inferring a sequence of skills to achieve acquisition for each item. Finally, we pass all of this context – the food item labels, the user's preference, and per-item efficiences — to an LLM-based planner which outputs a sequence of bites to feed. The few-shot common-sense reasoning capabilities of LLMs allows for reasoning about the available context in a chain-of-thought manner, and planning sequences of bites that cater to both preference and efficiency. Finally, we carry out these action sequences via a library of parameterized food manipulation skills implemented on custom hardware.

We validate FLAIR for long-horizon food pickup across 6 diverse plates. These plates range from DoorDash orders to prepared grocery store meals to homemade meals. In a user study across 42 individuals without mobility limitations, we use FLAIR to demonstrate the necessity of balancing between both preferences and efficiency for feeding complete, realistic meals, as compared to an efficiency-only or preference-only approach. Finally, FLAIR combines our modular approach to long-horizon bite acquisition with an existing bite transfer framework [19]. We demonstrate the real-world effectiveness of our system in feeding a care recipient with Multiple Sclerosis a meal consisting of various fruits and dips.

Overall, our contributions include:

- FLAIR: A robot-assisted feeding system which leverages a library of skills towards in-the-wild long-horizon bite acquisition.
- Comprehensive evaluations across 6 diverse plates evaluated by 42 individuals, validating the effectiveness of considering both preferences and efficiency for feeding.
- Demonstration of the real-world efficacy for feeding a care recipient with mobility limitations.

II. Related Work

Robot-Assisted Feeding. While various commercial robot-assisted feeding systems [1, 2] have been introduced, they typically rely on pre-programmed trajectories or user teleportation. This limited autonomy has hindered their widespread adoption and retention, and inspired autonomous methods for bite acquisition and transfer. Prior work in bite acquisition has focused on developing specialized individual food manipulation skills for specific food types. Gallenberger et al. [12], Gordon et al. [13], and Sundaresan et al. [28] demonstrate skewering strategies based on food item pose, and followed by angled pickup, are effective for various solid bite-sized foods. Sundaresan et al. [30] propose visually parameterized primitives for twirling and grouping noodle-like dishes, and show generalization to unseen noodles. Beyond fork-based manipulation, Grannen et al. [16] plan bimanual scooping actions with two custom utensils, while Tai et al. [31] and Zhang et al. [36] develop specialized strategies for scooping with a spoon and cutting with a knife, respectively. We leverage insights from these works to develop a comprehensive bite acquisition skill library, implemented with a custom feeding utensil and validated across two different robotic systems with in-the-wild plates containing diverse food items.

Various prior works demonstrate joint bite acquisition with transfer [6, 12, 19]. However, these works consider bite acquisition actions over a single timestep and not over the complete meal. In contrast, we demonstrate that our long-horizon bite acquisition framework can seamlessly integrate with existing methods for bite transfer [19], and demonstrate feeding of a full meal to a care recipient.

Foundation Models for Robotic Manipulation. There has been a wave of progress in leveraging visual-language foundation models for robot manipulation in recent years. Several recent works use LLMs to directly output low-level motor commands by exposing a robot API through prompting [17, 21, 35]. While these approaches have been shown to work well for small action spaces, food manipulation requires a degree of dexterity and precision that is difficult to achieve with these approaches. Other recent works use LLMs and VLMs models at higher layers of abstraction such as for zero-shot task planning [4], summarization of user preferences [33], and conditioning manipulation primitives or policies via the outputs of open-vocabulary object detectors [29]. We draw inspiration from these towards building a library of vision parameterized food manipulation skills in our setting, and also using foundation models for planning bite sequences.

Personalization in Assistive Robotics. The inclusion of user preferences in the design and operation of assistive robots is essential for significantly enhancing user satisfaction [9]. These preferences can be identified either implicitly through data-driven methods [34] or explicitly stated by users [8]. Canal et al. [8] explore task planning adhering to user preferences for an assistive shoe dressing experiment. However, they explore user specification only in form of post-hoc scoring of executed actions which is restrictive for various safety

critical applications. Recently, TidyBot [33] showcased LLMs can summarize information from limited examples and infer general user preferences for broader application in future interactions. However, their approach to task planning lacks consideration of additional metrics, such as efficiency, which is crucial in our context of feeding a complete meal.

III. FLAIR: Feeding via Long-horizon Acquisition of Realistic dishes

In this section, we present FLAIR: Feeding via Long-horizon Acquisition of Realistic dishes. FLAIR is a system for feeding complete meals which combines existing foundation models in a novel way towards personalized and efficient bite sequencing. We first give an overview of the hardware which enables our system, then outline our approach to long-horizon bite acquisition, and finally discuss integration of our method with an existing in-mouth bite transfer framework for feeding complete plates.

A. Hardware System

We aim to tackle a wide range of food categories in this work such as fruits, vegetables, noodles, meat, soft foods, dipping sauces, and non-bite-sized items that require cutting. Many of these foods require dexterous manipulation strategies that typical 6 or 7-DoF robots struggle with, especially due to joint limit constraints for dynamic actions. We thus implement FLAIR on a Kinova robot equipped with a motorized feeding utensil mounted at the end-effector, adapting the design from [26]. The utensil contains a fork attachment and has two degrees of freedom corresponding to the orientation of the fork tines and the tilt angle. This allows for directly controlling the utensil to perform dynamic movements like twirling and scooping, while the robot handles moving between waypoints in the workspace via Cartesian impedance control. We also use a wrist-mounted RGB-D Realsense camera with a known end-effector tocamera transformation. This enables perceiving plates of food and localizing food items in the 3D workspace. We note that the same hardware was replicated on two different Kinova arms with their separate fork attachment and sensors at two different institutions (detailed in Appendix), demonstrating the reproducibilty of our method and hardware.

B. Long-Horizon Bite Acquisition Framework

With access to a hardware platform that supports dexterous food manipulation strategies, our goal is to plan and execute long-horizon bite sequences that cater to a user's preference while efficiently feeding a meal.

Problem Formulation. We start by assuming access to an RGB-D plate image observation $o_t \in \mathcal{O} = \mathbb{R}_+^{W \times H \times 4}$. Additionally, we assume that a user optionally provides a natural language instruction ℓ representing their preferred feeding strategy at a high-level (i.e., ℓ = "Feed me alternating bites of X and Y" or "Only feed me X").

We further assume access to a library $\mathcal{L} = \{\phi^1, \ldots, \phi^N\}$ of N skills that the robot can use to manipulate food items. Each skill $\phi^n(p)$ represents a parameterized manipulation primitive that takes in parameters p and outputs low-level motor commands. We represent low-level actions by $a_t = (x, y, z, \beta, \gamma)$, where (x, y, z) denotes the position of the feeding utensil tip and β and γ denote pitch and roll, respectively. Thus, the output of any skill is a sequence of actions $\{a_\tau, a_{\tau+1}, \ldots, a_{\tau+T}\}$ that the robot takes to execute the particular strategy. For instance, a skewering skill may take the position and orientation of a desired food item as input, and output a trajectory that skewers the item of choice. Our goal is to plan a sequence of parameterized skills $\{(\phi_1, p_1), (\phi_2, p_2), \ldots, (\phi_H, p_H)\}$, $\phi_h \in \mathcal{L}$ which results in efficient and user-preferred bite acquisition.

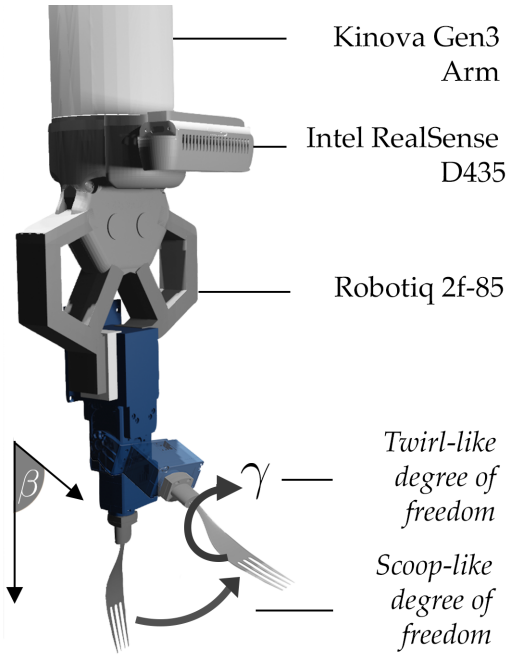

Fig. 2: Our skill library is implemented around a custom feeding utensil (adapted from [26]) having two degrees of freedom for easy twirling and scooping at the end effector.

State Representations for Food. Our approach addresses the main challenges in long-horizon bite acquisition — parameterizing low-level skills and sequencing them — by integrating state-of-the-art visual-language models. We use visual state estimates and semantic features of food items to guide skill parameterization and sequencing.

For a given plate observation o_t, we first query GPT-4V [3] in a few-shot manner to recognize which food items are present. We prompt the model with 5 in-context examples of plate images and their corresponding ground truth food item semantic labels, and ask the model to complete the prompt for the test image o_t. GPT-4V outputs a list of semantic labels l_t that are present, (i.e., l_t = `['fettuccine', 'chicken', 'broccoli']`) along with their corresponding categories c_t (i.e., c_t = `['noodles', 'meat/seafood', 'vegetable']`). These categories are relevant for associating the appropriate skill to each food item during actual bite sequencing. We then pass the

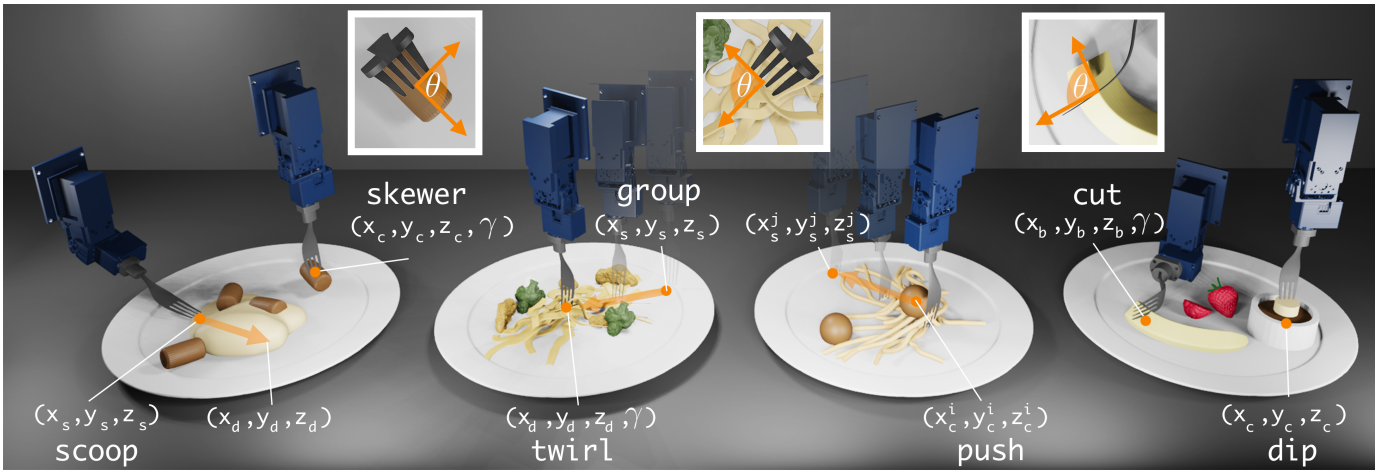

Fig. 3: Our skill library consists of 4 acquisition (skewer, twirl, scoop, dip) and 3 pre-acquisition (group, push, cut) skills, each a parameterized manipulation primitive.

recognized semantic labels to GroundingDINO [22], an open-vocabulary VLM, for bounding box detection. For each bounding box, we use SegmentAnything (SAM) [20] to refine these bounding boxes into segmentation masks $\{m_t^1, m_t^2, \ldots, m_t^D\}$ for all D items detected.

Skill Library The segmented representations of food which we obtained from VLMs provide a useful way to parameterize food manipulation skills, which we split into *acquisition* and *pre-acquisition* skills.

1) Acquisition skills: Acquisition skills refer to those that pick up food, such as skewering a food item, twirling a pile of noodles, scooping a soft pile of food, or dipping an item into a sauce. We parameterize each of these as follows, assuming access to a segmentation mask m_t for the item of interest:

- `skewer`(x_c, y_c, z_c, γ): We detect the centroid of m_t and deproject ths 2D pixel coordinate to a 3D coordinate (x_c, y_c, z_c) representing the center of a food item in the robot's frame of reference. We also estimate the *major-axis* orientation θ of an item from m_t analytically. Following [11, 30], we bring the utensil above the the food item center with $\gamma = 90° + \theta$ and execute a swift downward trajectory skewering perpendicular to the main axis of the item. This encourages the tines of the fork to pierce the item. The fork can easily slip if the tines run parallel to the main axis.
- `twirl`(x_d, y_d, z_d, γ): We adopt the parameterization from VAPORS [30], a long-horizon system for noodle acquisition. Specifically, we twirl noodles by bringing the fork to the sensed *densest* pile (x_d, y_d, z_d) on the plate, estimated via 2D Gaussian filtering on m_t, and with γ identical to the parameterization for skewering (orthogonal to the major axis of the noodle pile sensed via a pose estimation network from [30]). We actuate the roll joint of the fork to complete two full twirls in 10 seconds, wrapping noodles on the fork.
- `scoop`$(x_s, y_s, z_s, x_d, y_d, z_d)$: The fork starts with tines horizontal to the plate at $\beta = 80°$, instead of upright, and scoops from the *sparsest* region (x_s, y_s, z_s) to the *densest* region on the plate. This helps consolidate soft foods while picking up the largest bite.
- `dip`(x_c, y_c, z_c): Finally, dipping entails bringing a fork containing a food item into the center x_c, y_c, z_c of a small dish containing sauce.

Immediately following each of these actions, the robot moves the fork tines in a scooping motion by actuating the utensil's pitch joint. The resulting horizontal fork $(\beta = 90°)$ prevents items from slipping off the fork after being picked up.

2) Pre-acquisition skills: When the above acquisition skills would be suboptimal, or not immediately feasible due to occlusion from other items, we employ a number of auxiliary strategies which we refer to as *pre-acquisition* skills. These actions do not directly pick up food but rearrange or manipulate items to facilitate subsequent acquisition. Grouping noodles into a pile before twirling, pushing a meatball off of a bed of spaghetti before twirling, or cutting cake into a bite-sized piece before eating are all examples of pre-acquisition. We parameterize these as follows:

- `group`$(x_s, y_s, z_s, x_d, y_d, z_d)$: For a pile of food distributed on the plate, we sense the densest (x_d, y_d, z_d) and sparsest (x_s, y_s, z_s) regions via m_t and execute a linear push with the fork from the sparsest to densest point as in [30].
- `push`$(x_c^i, y_c^i, z_c^i, x_s^j, y_s^j, z_s^j)$: For a food item with mask m_t^i obstructing a pile of food (noodle or semisolid) with mask m_t^j, we can execute a linear push motion starting at the centroid of the first item, to the sparsest point of the bed of food underneath.
- `cut`(x_b, y_b, z_b, γ): To cut a food item, we estimate a

point on the object (x_b, y_b, z_b) that would result in a bite-sized portion once cut. In practice, we detect the major axis of m_t and traverse a fixed unit length from the one end of the axis to estimate this. We bring the fork horizontal with $\beta = 90°$, with γ orthogonal to the major axis as in skewering and twirling, and execute a swift downward trajectory to slice the soft item.

$\mathcal{L} = \{\texttt{skewer}, \texttt{twirl}, \texttt{scoop}, \texttt{dip}, \texttt{group}, \texttt{push}, \texttt{cut}\}$ forms the library of vision-parameterized skills at the core of FLAIR. We provide further details on the vision-based parameterizations for each of these skills in the supplementary material.

Task Planning for Acquisition. We aim to plan a sequence of bites that both satisfies the preference of the user, and is efficient for the robot to acquire. The latter consideration requires reasoning over the sequence of pre-acquisition and acquisition skills needed to pick up an item, for which we introduce a task planner $\mathcal{T}$.

The task planner takes as input a particular food item category c_t^i, the detected segmentation mask m_t^i, the RGB-D plate image o_t, and outputs a sequence of skills to acquire the item. The skill library in this work addresses the following categories of food items: $\{$`'meat/seafood'`, `'fruit'`, `'vegetable'`, `'dip'`, `'noodles'`, `'semisolid'`, `'cuttable'`$\}$.

For most of these categories, acquisition tends to be immediately possible. Food items such as a bite of $\{$`'meat/seafood'`, `'fruit'`, `'vegetable'`$\}$ tend to be isolated on a plate and immediately acquirable. Thus, we plan the following acquisition skills, where p_t^i denotes the parameters of the skill, sensed from m_t^i and o_t:

- $\mathcal{T}(c_t^i, m_t^i) = \{\texttt{skewer}(p_t^i)\}$ for $c_t^i \in \{$`'meat/seafood'`, `'fruit'`, `'vegetable'`$\}$
- $\mathcal{T}($`'dip'`$, m_t^i) = \{\texttt{dip}(p_t^i)\}$

Food items that are instead in the `'noodles'`, `'semisolid'`, or `'cuttable'` category require more nuanced reasoning about pre-acquisition depending on the distribution of the food on the plate, and whether other food items are intermixed, on top, or to the side. We critically observe that the segmentation mask m_t^i obtained from the VLM provides a useful prior over the spread of food on the plate, which can guide action selection. We apply a Gaussian smoothing kernel over m_t^i which has the effect of producing a normalized density heatmap of the food, and use simple pre-conditions to determine a sequence of skills to pick up a bite of noodles or a semisolid. Specifically, we measure the maximum *density* and the 2D *entropy* of the the heatmap and plan actions as follows.

If the density exceeds a pre-defined threshold `DENSITY_THRESH`, this indicates the presence of a large pile of food that can be immediately acquired:

- $\mathcal{T}($`'noodles'`$, m_t^i) = \{\texttt{twirl}(p_t^i)\}$
- $\mathcal{T}($`'semisolid'`$, m_t^i) = \{\texttt{scoop}(p_t^i)\}$.

Otherwise, if the entropy exceeds a pre-defined threshold `ENTROPY_THRESH`, this indicates that the items are spread out over the plate and grouping is necessary before acquiring. If the grouping action is unobstructed, we simply execute it; if instead another item mask m_t^j intersects with a planned a grouping action, such as a meatball blocking a path to grouping (i.e. Figure 3), we instead push a topping aside, group the food, and then acquire.

- $\mathcal{T}($`'noodles'`$, m_t^i) = \{\texttt{group}(p_{t,\texttt{group}}^i), \texttt{twirl}(p_{t,\texttt{twirl}}^i)\}$ if grouping is unobstructed.
- $\mathcal{T}($`'noodles'`$, m_t^i) = \{\texttt{push}(p_{t,\texttt{push}}^i), \texttt{group}(p_{t,\texttt{group}}^i), \texttt{twirl}(p_{t,\texttt{twirl}}^i)\}$ if the robot must push aside a topping before twirling.
- $\mathcal{T}($`'semisolid'`$, m_t^i) = \{\texttt{push}(p_{t,\texttt{push}}^i), \texttt{scoop}(p_{t,\texttt{scoop}}^i)\}$ if the robot must push aside a topping before scooping.

Finally, for `cuttable` items like cake, we use a simple area threshold to determine if m_t^i is bite-sized or not, and either cut and then skewer, or skewer a bite immediately:

- $\mathcal{T}($`'cuttable/non-bite-sized'`$, m_t^i) = \{\texttt{cut}(p_{t,\texttt{cut}}^i), \texttt{skewer}(p_{t,\texttt{skewer}}^i)\}$, if the area of m_t^i exceeds `BITE_SIZE`
- $\mathcal{T}($`'cuttable/non-bite-sized'`$, m_t^i) = \{\texttt{skewer}(p_t^i)\}$, otherwise

The only parameters the task planner relies on are the density, entropy, and bite size thresholds. Even these are shared across categories of foods like noodles, semisolids, and cuttable items, rendering the approach general enough to handle many different foods.

C. Bite Sequencing via Foundation Models

We introduce a unified framework for planning and executing preference-informed and efficient bite sequences. With access to a library of skills $\mathcal{L}$, task planner $\mathcal{T}$, and user preference ℓ, we show how the commonsense-reasoning capabilities of LLMs enable them to act as few-shot planners for bite sequencing, inherently balancing preference and efficiency.

We prompt an LLM, in our case GPT-4V with relevant context about the meal. This includes the semantic food item labels l_t and the user's preference ℓ. We augment this context with a *history* of bites taken so far, an estimate of the *portions* of each food type remaining, and lastly the *per-item efficiencies* which correspond to the number of actions required to pick up a food item (i.e. $|\mathcal{T}(c_t^i, m_t^i)|$). A higher number indicates a less efficient bite option since more pre-acquisition skills are required to pick up the item. To estimate quantities, we simply count the number of instances detected of the food item present if the item category is [`'fruit'`, `'meat/seafood'`, `'vegetable'`]. Otherwise, we use $\lceil m_t / \texttt{BITE_SIZE} \rceil$ as a proxy for quantifying the number of bites present, which approximates the number of bite portions in a mask based on a pre-defined portion size. This

is necessary for food items where the ground truth quantity is not countable, such as for ‘noodles’ or ‘semisolid’ items.

Below, we show an example input prompt in gray with the immediate next bite planned by asking the LLM to perform completion. We abridge the prompt here for brevity, but include our full prompting strategy in Appendix.

```
Items remaining: ["fettuccine", "chicken", "broccoli"]
Preference: "Alternating bites of each"
History: ["chicken"]
Portions remaining: [5, 1, 2]
Effiencies: [3, 1, 1]
---
Decide what bite to feed me next.
Format your response as follows:
Strategy: Sentence describing your high-level strategy
Next bite: Phrase describing the next bite you will feed
Next bite as list: ['item'] # Or ['item', 'dip'] or []
---
Output your response here.
```

```
Strategy: Given that you want to alternate amongst the
three types of items present, and you just ate chicken,
I will feed you either fettucine or broccoli.
Even though broccoli is a more efficient option, with only
1 action required, I will feed you fettuccine since
there are more portions of it.
Next bite: Feed fettuccine.
Next bite as list: ['fettuccine']
```

By reasoning about the provided context in a chain-of-thought style, the LLM generates a subsequent bite. Importantly, there are no explicit tradeoffs between efficiency, portion size, or preference which we embed into the prompt, allowing the LLM to reason about the most sensible strategy based on the strength of the user’s preference and available plate context. Given a next bite, in this case ‘fettuccine’, we then plan the appropriate skill sequence via $\mathcal{T}(\text{‘noodles’}, m_t^{\text{‘fettuccine’}})$ and execute it via $\mathcal{L}$. FLAIR performs long-horizon bite acquisition in an MPC-style approach where after each acquired bite, we update last bite in the history h_t and re-prompt the LLM to obtain a next bite.

D. Integration of Acquisition and Transfer

The self-contained nature of our *bite acquisition* framework presented in Section III-B allows for straightforward integration with *bite transfer* frameworks, and is agnostic to the exact approach used. We specifically combine our acquisition framework with a recent method for transfer [19], which leverages robust mouth tracking and physical interaction-aware control for in-mouth transfer of an acquired food item.

A significant challenge in this integration is ensuring that food, particularly semi-solid items such as mashed potatoes or noodle-like items such as spaghetti, does not spill while it moves from above the plate to the pre-transfer pose infront of the mouth. Prior works with non-actuated utensils [25] use an MPC-based approach to generate robot trajectories that constrain the orientation of the utensil to remain upright. However, these methods often require complex tuning and can be prone to getting trapped in local minima. Our feeding utensil enables us to uniquely circumvent this challenge. We leverage its roll and pitch degrees of freedom, distinct from the robot’s own degrees of freedom, to consistently keep the fork’s tines horizontal regardless of the robot’s motion. We continuously monitoring the robot’s end-effector pose at 10 Hz and adjust the feeding utensil’s joints accordingly, ensuring a smooth and spill-free transfer of food to the user’s mouth.

FLAIR’s modular composition of acquisition with transfer allows for feeding a wide range of complete meals with a personalized and efficient bite sequence.

IV. Experiments

We aim to evaluate the effectiveness of FLAIR as a holistic system for feeding diverse plates. We first independently assess FLAIR’s ability to perform long-horizon bite acquisition of in-the-wild plates, while adhering to user preferences and efficiently clearing plates. We then evaluate the real-world efficacy of our system for feeding a care recipient a complete plate.

A. Bite Acquisition Experiments

Baselines: FLAIR presents a unique approach of taking into account both preference and efficiency considerations for bite sequencing. This naturally begs the question of how an *Efficiency-Only* or *Preference-Only* approach would compare. We implement an Efficiency-Only baseline which greedily selects the next bite as the item which requires the least number of pre-acquisition and acquisition skills for pickup in the current instant, as dictated by the task planner $|\mathcal{T}(\cdot,\cdot)|$. The Preference-Only baseline is identical to FLAIR in implementation, but notably omits efficiency scores when prompting the LLM to generate a next bite. This encourages the LLM to only respect a user’s preference without consideration for how efficient a particular bite may be. In the case that a user has no preference for feeding, we refere to the Preference-Only baseline as *Commonsense-Only.*

Experimental Setup: We consider an evaluation suite of 6 diverse plates of food spanning a wide range of food categories, visualized in Fig. 7. We include 2 in-the-wild noodle dishes: a spaghetti and meatballs plate which is a prepared frozen meal from a grocery store, and a fettuccine alfredo dish with chicken and broccoli ordered from Applebees on Doordash. We also consider 2 semisolid dishes including mashed potatoes with sausage and oatmeal with strawberries. Lastly, we evaluate an appetizer plate of strawberry, watermelon, celery, ranch, and chocolate dipping sauce, as well as a dessert plate of a whole banana, brownie bites, and chocolate dipping sauce.

User Study Design: We evaluate FLAIR’s ability to cater to user preferences via a two-phase user study across 42 individuals without any disabilities (Ages: 19-64, Genders: 22F, 20M). In the first phase, we present participants with a survey showing images of all 6 evaluation plates, and solicit their natural language preference over how they would prefer to be fed each plate. Since evaluating each submitted

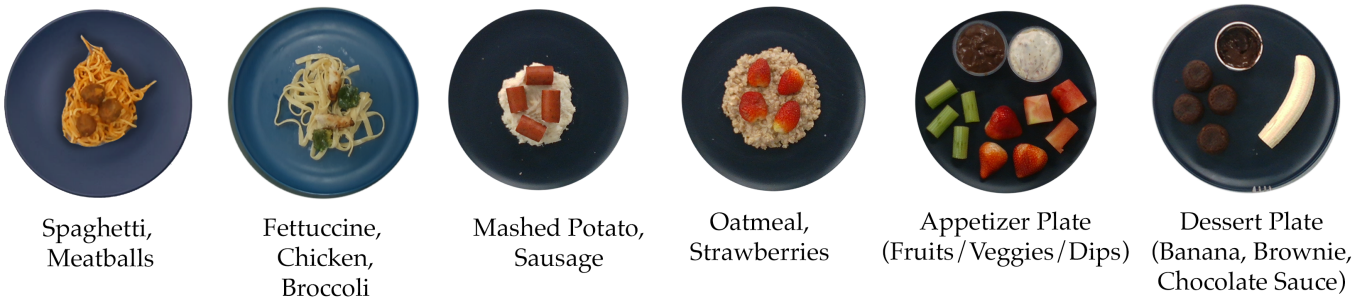

Fig. 4: **Plates**: We evaluate our system on the following six plates containing a variety of food items, each necessitating highly different manipulation skills.

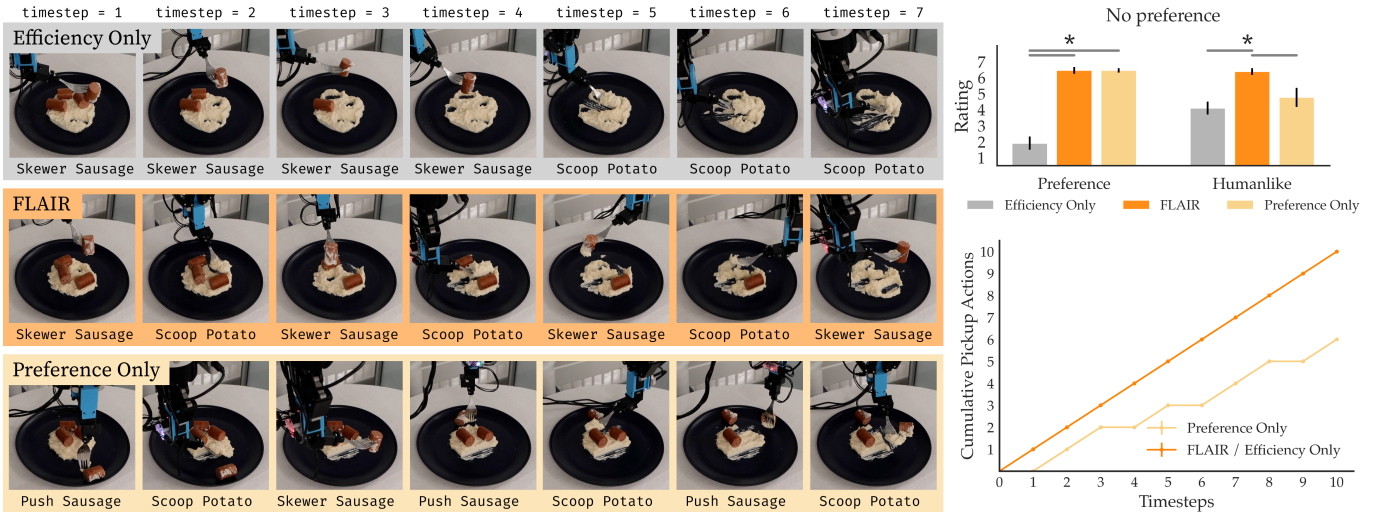

Fig. 5: Example run on a plate with mashed potatoes and sausages where the user specified no preference. FLAIR, which balances user preferences and efficiency, is judged by users to better adhere to preferences than Efficiency Only and outperforms Preference Only in plate clearance. Consequently, FLAIR is considered to provide a more human-like feeding experience compared to the baseline methods.

preference across all of the plates and baselines is not scalable, we cluster the submitted preferences into common shared responses via LLM summarization (GPT-4V), grouped by user. We are particularly interested in cases when a user has either no preference, or a strong preference, since slight preferences do not provide an informative way to compare method behaviors. Thus, we specifically prompt GPT to filter for strong preferences (i.e. 'Always feed me alternating bites of X and Y' or 'Please do not feed me X') and group them accordingly. For each of the six plates, we then evaluate our system on the 2 most popular strong preferences summarized per plate, as well as a 'I have no preference' setting for completeness.

We hypothesized the following:

- **H1**: Compared to the Preference Only baseline, FLAIR's consideration of efficiency in task planning will lead to more number of pickup actions across all settings.
- **H2**: Compared to the Efficiency Only baseline, FLAIR's consideration of common sense reasoning in task planning in the absence of preferences will lead to more perceived adherence to bite variety and common food item pairings, and more human-like feeding.
- **H3**: Compared to the Efficiency Only baseline, FLAIR's consideration of user preferences in task planning in the presence of strong preferences will lead to more perceived adherence to user preferences, and more human-like feeding.

Plate Clearance Results: Figure **??** displays the results of plate clearance. Due to its consideration of efficiency in task planning, FLAIR executes a greater number of pickup actions compared to Preference Only. This is because when faced with multiple candidate actions satisfying the user's preference, FLAIR , informed with efficiency scores for each action, is able to choose the action that optimizes for efficiency. In contrast, Preference Only randomly selects one action from this set, often leading to inefficient acquisition trajectories (Figure 5). The performance disparity between Efficiency Only and FLAIR can be linked to settings with strong preferences that require the robot to perform actions that are less efficient than those of a method which does not take such preferences into account. For instance, in a scenario where a robot is instructed to feed spaghetti hidden beneath a bed of noodles, methods that consider preferences must undertake multiple pre-acquisition actions to clear away the noodles. This performance difference between FLAIR and Efficiency Only baseline is mitigated in the no preference scenario, as we detail in the Appendix.

User Evaluation: Figure 4 presents participant ratings com-

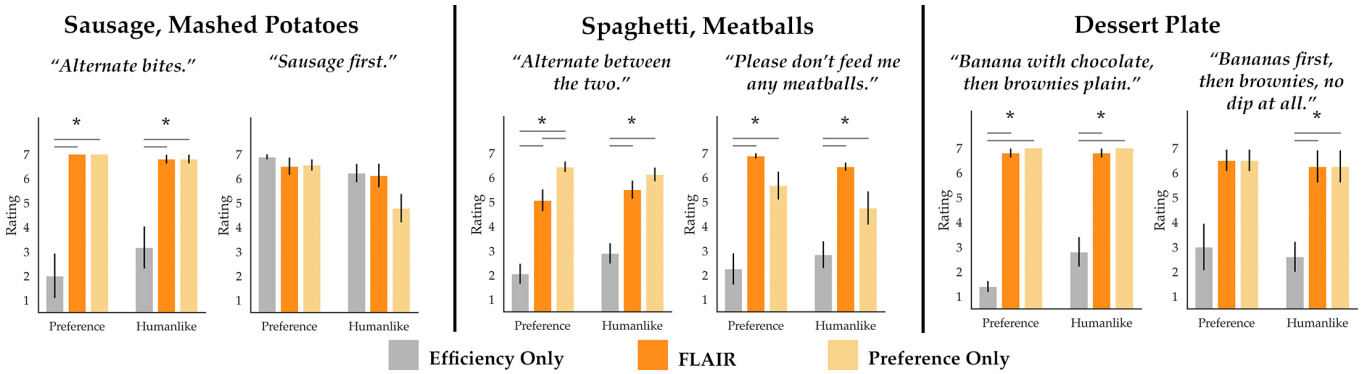

Fig. 6: Participant ratings for settings with strong user preferences show FLAIR significantly outperforms the Efficiency Only baseline in aligning with user preferences and achieving human-like feeding in all scenarios, except cases where the efficiency-based bite sequence coincidentally aligns with user preferences.

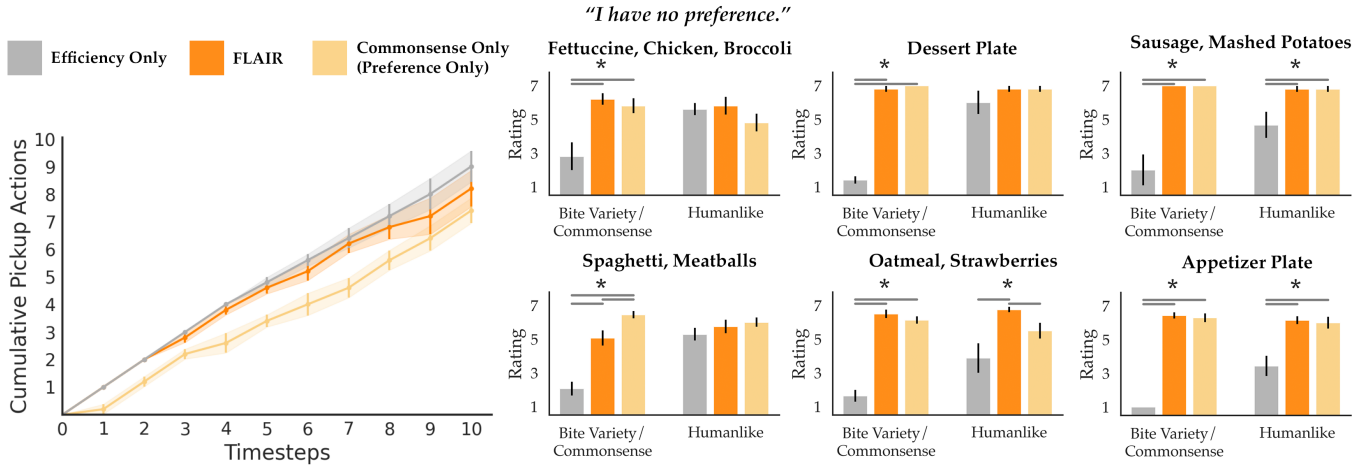

Fig. 7: Participant ratings for settings with strong user preferences show FLAIR significantly outperforms the Efficiency Only baseline in aligning with user preferences and achieving human-like feeding in all scenarios, except cases where the efficiency-based bite sequence coincidentally aligns with user preferences.

paring FLAIR with baseline approaches for settings with strong user preferences. By integrating user preferences into task planning, FLAIR substantially surpasses the Efficiency Only baseline in terms of adherence to user preferences and human-like feeding across various settings. The exceptions, where the performance difference between Efficiency Only and FLAIR is not statistically significant, occur in settings where the bite sequence, generated based solely on efficiency, inadvertently matches the user's preferences. Figure **??** presents participant ratings comparing FLAIR with baseline approaches for settings where users did not specify any preferences. By leveraging common-sense reasoning, FLAIR substantially surpasses the Efficiency Only baseline across all settings in terms of ensuring a variety in the bites fed, and combining food items with appropriate dips. However, in the absence of user preferences, while participants cumulatively rate FLAIR to provide better human-like feeding, this result is not statistically significant.

B. Demonstration of Real-World Feeding

We demonstrate the effectiveness of FLAIR in assisting a care-recipient with significant mobility restrictions to eat an entree dish comprising boiled baby carrots, watermelon, strawberries, ranch dressing, and chocolate sauce. The care-recipient, a 44-year-old Caucasian/White female, has been living with Multiple Sclerosis for 19 years, resulting in a severely limited range of motion in their head and neck. Consequently, the bite sizes of the food items picked up must be small such that they can be safely placed inside their mouth.

In the pre-study questionnaire, the care-recipient mentioned that they typically have a preferred order in which they like to eat their meal. They convey this preference to their caregivers through natural language, and when caregivers adhere to this preference, it "definitely enhances my eating experiences." For the consider plate, the care-recipient specified "I want to first finish all the celery with ranch dressing, then eat watermelon without any dips, and finally end with strawberries dipped in chocolate sauce." Following successful feeding with FLAIR, we posed two questions on a seven-point Likert scale on the necessity of a robot-assisted feeding system to (i) have a



Fig. 8: We demonstrate the real-world effectiveness of our method by feeding an entree dish to a care-recipient with severe mobility limitations.

diverse bite acquisition skill library, and (ii) adherence to meal preferences, for acceptance for day to day usage. The care-recipient strongly agreed (rating = 7) with both, emphasizing the core contributions of our paper as critical aspects for an in-the-wild feeding system.

In the post-study questionnaire, the care-recipient noted that while they often have specific preferences for the order in which they eat their food, they often refrains from sharing these preferences with their human caregivers. They worry that making such requests might be burdensome, considering that their caregivers are already providing assistance with feeding. However, they expressed an expectation that a robot designed to help with feeding would follow their meal order preferences without issue, enabling them to enjoy meals better.

V. Discussion

In this work, we present FLAIR, a system that achives long-horizon bite acquisition and feeding of complete plates with both user preference and efficiency in mind. We first develop an extensive library of vision-parameterized food manipulation skills capable of handling a range of food items present in in-the-wild dishes. FLAIR then employs the commonsense reasoning capabilities of vision-language foundation models to plan and execute bite sequences which balance efficient plate clearance with user-provided preferences. We evaluate our system with a comprehensive user study that spans a range of plate dishes and preferences, finding that users perceive our system to satisfy their preferences with high alignment without compromising on efficiency, which we validate quantitatively. We finally integrate our bite acquisition approach with an in-mouth bite transfer framework and feed a care recipient with mobility limitations an entire plate of fruits, vegetables and dipping sauces.

References

[1] Meet Obi. https://meetobi.com/. [Online; accessed 6-June-2022].

[2] Neater eater robot, 2024. URL https://www.neater.co.uk/neater-eater-robotic. (Accessed: 1st January, 2024).

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

[4] Christopher Agia, Toki Migimatsu, Jiajun Wu, and Jeannette Bohg. Taps: Task-agnostic policy sequencing. *arXiv preprint arXiv:2210.12250*, 2022.

[5] Suneel Belkhale, Ethan K Gordon, Yuxiao Chen, Siddhartha Srinivasa, Tapomayukh Bhattacharjee, and Dorsa Sadigh. Balancing efficiency and comfort in robot-assisted bite transfer. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 4757–4763. IEEE, 2022.

[6] Tapomayukh Bhattacharjee, Ethan K Gordon, Rosario Scalise, Maria E Cabrera, Anat Caspi, Maya Cakmak, and Siddhartha 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)*, pages 181–190. IEEE, 2020.

[7] Steven W Brose, Douglas J Weber, Ben A Salatin, Garret G Grindle, Hongwu Wang, Juan J Vazquez, and Rory A Cooper. The role of assistive robotics in the lives of persons with disability. *American Journal of Physical Medicine & Rehabilitation*, 89(6):509–521, 2010.

[8] Gerard Canal, Guillem Alenyà, and Carme Torras. Adapting robot task planning to user preferences: an assistive shoe dressing example. *Autonomous Robots*, 2019.

[9] Gerard Canal, Carme Torras, and Guillem Alenyà. Are preferences useful for better assistance? a physically assistive robotics user study. *THRI*, 2021.

[10] Adriano Chiò, A Gauthier, A Vignola, Andrea Calvo, Paolo Ghiglione, Enrico Cavallo, AA Terreni, and Roberto Mutani. Caregiver time use in als. *Neurology*, 67(5):902–904, 2006.

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

[12] Daniel Gallenberger, Tapomayukh Bhattacharjee, Youngsun Kim, and Siddhartha S Srinivasa. Transfer depends on acquisition: Analyzing manipulation strategies for robotic feeding. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 267–276. IEEE, 2019.

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

[14] Ethan K Gordon, Sumegh Roychowdhury, Tapomayukh

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

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

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

[17] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[18] Catrine Jacobsson, Karin Axelsson, Per Olov Österlind, and Astrid Norberg. How people with stroke and healthy older people experience the eating process. *Journal of Clinical Nursing*, 9(2):255–264, 2000. doi: https://doi.org/10.1046/j.1365-2702.2000.00355.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2702.2000.00355.x.

[19] 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.

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[21] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

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

[23] Rishabh Madan, Rajat Kumar Jenamani, Vy Thuy Nguyen, Ahmed Moustafa, Xuefeng Hu, Katherine Dimitropoulou, and Tapomayukh 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)*, pages 641–648. IEEE, 2022.

[24] Amal Nanavati, Patricia Alves-Oliveira, Tyler Schrenk, Ethan K Gordon, Maya Cakmak, and Siddhartha S Srinivasa. Design principles for robot-assisted feeding in social contexts. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 24–33, 2023.

[25] Daehyung Park, Yuuna Hoshi, Harshal P Mahajan, Ho Keun Kim, Zackory Erickson, Wendy A Rogers, and Charles C Kemp. Active robot-assisted feeding with a general-purpose mobile manipulator: Design, evaluation, and lessons learned. *Robotics and Autonomous Systems*, 124:103344, 2020.

[26] Lorenzo Shaikewitz, Yilin Wu, Suneel Belkhale, Jennifer Grannen, Priya Sundaresan, and Dorsa Sadigh. In-mouth robotic bite transfer with visual and haptic sensing. *arXiv preprint arXiv:2211.12705*, 2022.

[27] Samantha E. Shune. An altered eating experience: Attitudes toward feeding assistance among younger and older adults. *Rehabilitation nursing : the official journal of the Association of Rehabilitation Nurses*, 2020.

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

[29] Priya Sundaresan, Suneel Belkhale, Dorsa Sadigh, and Jeannette Bohg. Kite: Keypoint-conditioned policies for semantic manipulation. *arXiv preprint arXiv:2306.16605*, 2023.

[30] Priya Sundaresan, Jiajun Wu, and Dorsa Sadigh. Learning sequential acquisition policies for robot-assisted feeding. In *Conference on Robot Learning*, pages 1282–1299. PMLR, 2023.

[31] Yen-Ling Tai, Yu Chien Chiu, Yu-Wei Chao, and Yi-Ting Chen. Scone: A food scooping robot learning framework with active perception. In *Conference on Robot Learning*, pages 849–865. PMLR, 2023.

[32] Danielle M Taylor. Americans with disabilities: 2014. *US Census Bureau*, pages 1–32, 2018.

[33] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *arXiv preprint arXiv:2305.05658*, 2023.

[34] Guang Yang, Shuoyu Wang, Junyou Yang, and Peng Shi. Desire-driven reasoning considering personalized care preferences. *Transactions on Systems, Man, and Cybernetics: Systems*, 2021.

[35] Lihan Zha, Yuchen Cui, Li-Heng Lin, Minae Kwon, Montserrat Gonzalez Arenas, Andy Zeng, Fei Xia, and Dorsa Sadigh. Distilling and retrieving generalizable knowledge for robot manipulation via language corrections. *arXiv preprint arXiv:2311.10678*, 2023.

[36] Kevin Zhang, Mohit Sharma, Manuela Veloso, and Oliver Kroemer. Leveraging multimodal haptic sensory data for

robust cutting. In *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*, pages 409–416. IEEE, 2019.