Robotics: Science and Systems 2024
Delft, Netherlands, July 15-July 19, 2024

FLAIR: Feeding via Long-horizon AcquIsition of Realistic dishes

Rajat Kumar Jenamani*[1], Priya Sundaresan*[2], Maram Sakr[3], Tapomayukh Bhattacharjee†[1], Dorsa Sadigh†[2]
*Equal Contribution, †Equal Advising
[1]Cornell University, [2]Stanford University, [3]University of British Columbia

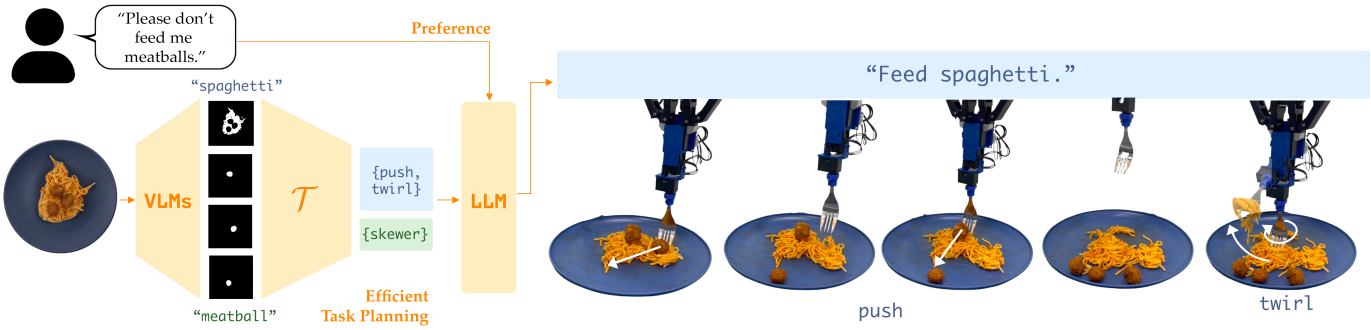

Fig. 1: We propose FLAIR, a system for long-horizon robot-assisted feeding that combines the commonsense and few-shot reasoning capabilities of foundation models with a library of parameterized skills. Above, FLAIR takes visual observations and a given user preference ("Please don't feed me any meatballs") to plan a sequence of actions that pushes aside meatballs and twirls spaghetti.

***Abstract*—Robot-assisted feeding has the potential to improve 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 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 and few-shot 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 varied library of skills for efficient food pickup, while adhering to the diverse preferences of 42 participants without mobility limitations as evaluated in a user study. We demonstrate the seamless integration of FLAIR with existing bite transfer methods [19, 28], and deploy it across 2 institutions and 3 robots, illustrating its adaptability. Finally, we illustrate the real-world efficacy of our system by successfully feeding a care recipient with severe mobility limitations. Supplementary materials and videos can be found at: emprise.cs.cornell.edu/flair.**

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 [54]. These individuals often rely on caregivers for meal assistance, which impacts their sense of independence, daily routines, and the social experience of dining [27, 41, 50]. Moreover, feeding is one of the most time-consuming Activities of Daily Living (ADL) for caregivers [14]. A system for autonomous mealtime assistance holds promise for improving the quality of life for those requiring assistance [9], and reducing the physical workload on caregivers [6, 28].

Robot-assisted feeding entails first performing *bite acquisition* [18, 21–24, 32, 51, 52], where the robot must manipulate a utensil to pick up a bite of food, followed by *bite transfer* [5, 19, 28, 43, 48], or bringing a bite of food to the mouth for consumption. In this paper, we primarily focus on bite acquisition. Several prior works in bite acquisition develop individual skills targeting specific food groups. This includes policies for skewering firm foods [18, 21–23, 51], scooping soft foods [24, 53], or rearranging and twirling noodles [52]. These works, however, mostly operate over a single bite horizon or consider plates with a homogeneous type of food, such as only noodles or only bite-sized fruits and vegetables. However, the challenge of achieving reliable bite acquisition for dishes encountered *in-the-wild*, which contain multiple different food types within the same meal and require strategic skill sequencing over many timesteps, persists.

Consider a robot tasked with feeding a meal with a fruit appetizer—bananas, celery, and watermelon with chocolate sauce and ranch dressing—and spaghetti and meatballs for the main course. The robot must not only execute specialized strategies, such as cutting bananas, skewering fruits, dipping in sauces, and grouping and twirling noodles, but also infer how to sequence them over a *long horizon*, considering:
Efficiency: For the main course, if meatballs incidentally rest on top of the spaghetti, the robot should prioritize efficiency by serving the meatball first. This sequencing exposes the spaghetti for subsequent bites, avoiding the inefficiency of pushing the meatball aside to access the spaghetti initially.

User Preferences: However, if the user prefers to not eat meatballs, the robot must adjust the bite sequence accordingly. Commonsense Reasoning: In the absence of explicit user preferences, the robot must employ commonsense reasoning to correctly order / combine bites for human-like feeding. For the appetizer, it should pair celery with ranch, bananas with chocolate, and feed watermelon standalone, reflecting typical food pairings. For the main course, it should vary the serving order between spaghetti and meatballs to avoid repetitions. We desire a system that considers all these criteria to achieve long-horizon bite acquisition via a library of skills, and finally integrates with frameworks for bite transfer [5, 19, 28, 43, 48] 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 meal to a care recipient. Given a plate image, and an optional 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 executes a sequence of actions that efficiently feeds the items on the plate while adhering to the preference. 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 hierarchical 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 optional user's preference, and per-item efficiences – to a Large Language Model (LLM)-based planner which outputs the next bite to feed. The few-shot 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. We carry out these action sequences via a library of parameterized food manipulation skills implemented on custom hardware. Finally, FLAIR's modular approach to long-horizon bite acquisition enables seamless integration with existing outside-mouth bite transfer [19] and inside-mouth bite transfer [28] frameworks.

We deploy FLAIR across two institutions and three robots: a Kinova 6-DoF at Cornell University and a Franka Emika Panda and a Kinova 7-DoF at Stanford University, demonstrating its adaptability to various robotic platforms. We validate FLAIR for long-horizon food pickup across six diverse plates, ranging from DoorDash orders and 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. Moreover, we compare FLAIR's hierarchical task planner against three state-of-the-art baselines [3, 39, 52] on two different datasets, demonstrating that it significantly outperforms these baselines. Finally, 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 system for long-horizon feeding which leverages foundation models to sequence a library of diverse skills towards in-the-wild long-horizon bite acquisition.
- Deployment of FLAIR across two institutions and three different robots, demonstrating its versatility.
- A user study with 42 individuals without mobility limitations across 6 diverse plates validating the effectiveness of considering both preferences and efficiency for feeding.
- Demonstration of the real-world efficacy of our system by 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 teleoperation. 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 individual food manipulation skills for specific food types. Various works [18, 21, 22, 51] tackle acquisition of solid bite-sized foods, and demonstrate effective skewering strategies based on the food item's pose and material properties. Sundaresan et al. [52] propose visually parameterized primitives for twirling and grouping noodle-like dishes, and show generalization to unseen noodles. Beyond fork-based manipulation, Grannen et al. [24] plan bimanual scooping actions with two custom utensils, while Tai et al. [53] and Zhang et al. [58] develop specialized strategies for scooping with a spoon and cutting with a knife, respectively. However, no prior work in robot-assisted feeding considers complete, in-the-wild meals containing various food types (noodles, semisolids, sauces, cuttable food items, etc.) within the same plate, as typically encountered in everyday scenarios.

In this work, we leverage insights from the aforementioned state-of-the-art food manipulation works to develop a large library of bite acquisition skills, and use foundation models to sequence these skills for efficiently feeding realistic dishes while obeying user preferences. To the best of our knowledge, FLAIR is the first of any autonomous feeding system to tackle in-the-wild meals containing various food types, and incorporate bite sequencing preferences for long-horizon feeding.

Various works have shown joint bite acquisition with transfer [6, 19, 28]. However, they typically consider bite acquisition actions over a single timestep and not over the complete meal. In contrast, we illustrate that our long-horizon bite acquisition framework can seamlessly integrate with existing methods for bite transfer [19, 28], and demonstrate feeding of a full meal to a care recipient.

Foundation Models for Robotic Manipulation. Two of the most challenging aspects associated with feeding are planning over available skills, and developing a library of food manipulation skills themselves. To this end, several recent works in robotic manipulation use foundation models such as vision-language models (VLMs) [3, 31, 35, 46] or large language

models (LLMs) [10, 15, 16, 44, 45] towards both high-level task planning and skill instantiation. A standard approach is to prompt foundation models with context including available skills, object states, etc., and to use them to plan action sequences either for long-horizon manipulation [4, 13, 25, 55] or grounded exploration [29]. However, these works focus on exploiting the commonsense reasoning capabilities of these models [20, 34, 47], such as inferring a sequence of skills that is feasible or user-preferred based on the provided visual and semantic context. In the setting of feeding, we additionally care about planning *efficient* skill sequences. Reasoning about efficiency and skill affordances in a few-shot manner remains brittle and challenging for these models, due to hallucinations [30] and a lack of priors about embodied agents. Instead, we propose inferring the efficiencies of skills separately, and providing this as additional context to aid in planning.

Besides skill sequencing, several recent works show the benefits of using foundation models towards inferring the parameters of low-level skills themselves, rather than data-driven approaches to learning skill policies from scratch [7, 8, 42]. Recent approaches include instantiating skills via code skeletons generated by LLMs [26, 36], or implementing skills parameterized by open-vocabulary object detectors [55, 57] or keypoint affordances from VLMs [37]. These approaches have mainly been applied to simple quasi-static actions such as pick and place. We instead apply this paradigm towards estimating the visual state of food items, and using this to parameterize a diverse library of skills such as twirling, scooping, and cutting.

User Preferences in Assistive Robotics. The inclusion of user preferences in the design and operation of assistive robots is essential for significantly enhancing user satisfaction [12]. These preferences can be identified either implicitly through data-driven methods [56] or explicitly stated by users [11]. Canal et al. [11] 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. Madan et al. [40] propose training a hidden Markov Model with user-demonstrated sequencing data for the same meal collected over multiple days to learn preferred bite sequences, enhancing user satisfaction. However, this proof of concept did not involve a robot-assisted feeding system and is impractical to extend to the diverse meals an individual might consume. Recently, TidyBot [55] showcased that LLMs can summarize information from limited examples and extrapolate general user preferences for determining the proper place to put each object while tidying a room. 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. FLAIR instead factors in both user preference and acquisition efficiency for long-horizon feeding.

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

In this section, we present FLAIR, 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 our custom system hardware, then outline our approach to long-horizon bite acquisition, and finally discuss integration of our method with existing bite transfer frameworks [19, 28] for feeding of in-the-wild dishes.

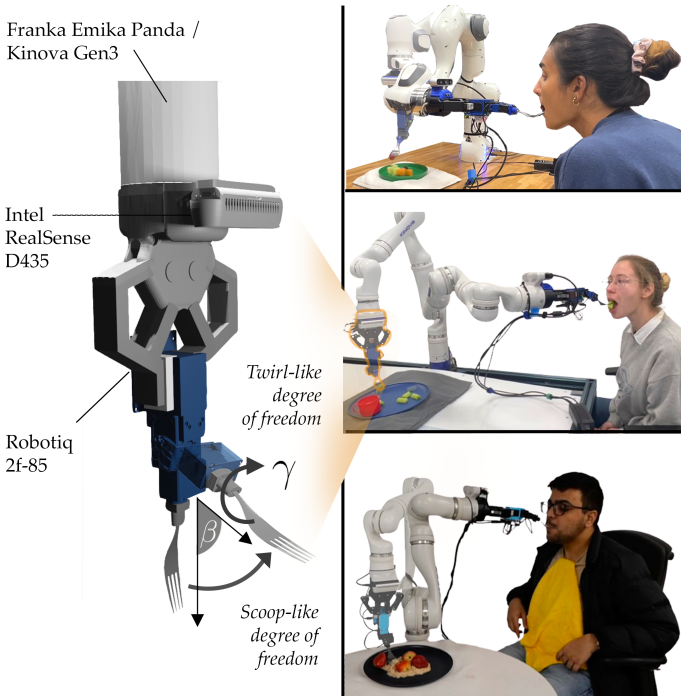

Fig. 2: We implement our skill library using a custom feeding utensil (adapted from [48]) having two degrees of freedom for easy twirling and scooping at the end effector. We deploy the full feeding stack on three robots and two institutions: the 7-DoF Franka Emika Panda (top) and 7-DoF Kinova Gen 3 (middle) at Stanford University, and the 6-DoF Kinova Gen 3 (bottom) at Cornell University.

A. Hardware System

We 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 specialized, dynamic manipulation strategies that typical 6 or 7-DoF robots struggle with due to their limited workspace. We thus implement FLAIR on Kinova and Franka robot arms equipped with a motorized feeding utensil mounted at the end-effector, adapting the design from [48]. 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 position control. We also use a wrist-mounted RGB-D Realsense camera with a known end-effector to camera 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 and one Franka Emika Panda, each with their separate fork attachment and sensors across two different institutions (detailed in Appendix), demonstrating the reproducibilty of our method and hardware (Fig. 2).

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 assume access to an RGB-D plate image observation $o_t \in \mathcal{O} = \mathbb{R}_+^{W \times H \times 4}$ of width W and height H, and an optional natural language instruction ℓ_{pref} from the user, representing their preferred feeding strategy at a high-level (i.e., ℓ_{pref} = "Feed me alternating bites of X and Y" or "Only feed me X"). X and Y can denote an arbitrary food item semantic label (i.e. "spaghetti", "strawberry", "caramel") or category (i.e. "noodles", "fruit", "sauce").

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 $\phi^i(p)$ represents a parameterized manipulation primitive that takes in parameters p and outputs low-level motor commands. We represent a low-level action at time t by $a_t = (x, y, z, \beta, \gamma, \psi)$, where (x, y, z) denotes the position of the feeding utensil tip, β and γ denote pitch and roll of the utensil respectively, and ψ denotes the robot's end effector roll angle. Thus, the output of any skill is a sequence of T actions $\{a_t, a_{t+1}, \ldots, a_{t+T}\}$ 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 and execute a sequence of parameterized skills $\{\phi_1(p_1), \phi_2(p_2), \ldots, \phi_H(p_H)\}$ which results in efficient and user-preferred bite acquisition, where H is the total number of skills to execute to complete feeding a plate and p_h refers to the parameters of skill $\phi_h \in \mathcal{L}$.

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 at time t, we first query GPT-4V [3] in a few-shot manner to recognize which food items are present. We prompt the model with a few 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', 'cuttable']`). These categories are relevant for associating the appropriate skill to each food item for bite acquisition. We then pass the recognized semantic labels to GroundingDINO [38], an open-vocabulary VLM, for bounding box detection. For each bounding box, we use SegmentAnything (SAM) [33] 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 we obtain from VLMs provide a useful way to parameterize food manipulation skills, which we split into *acquisition* and *pre-acquisition* skills. Fig. 3 visualizes all skill parameterizations.

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 to coat it in sauce. We parameterize them as follows, assuming access to a segmentation mask m_t^i for the item of interest:

- `skewer`(x_c, y_c, z_c, γ): We detect the centroid of m_t^i and deproject this 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^i analytically. Following [18, 51], 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. If the tines align parallel to the item's major axis, they may run along its longer length and miss the shorter breadth due to slight calibration challenges, leading to unsuccessful skewering.
- `twirl`(x_d, y_d, z_d, γ): We adopt the parameterization from VAPORS [52], 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^i, 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 [52]). We actuate the roll joint of the fork to complete two full twirls, 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 and scoops from the *sparsest* region (x_s, y_s, z_s) to the densest region (x_d, y_d, z_d) on the plate, up to a pre-defined maximum distance empirically selected to pick up a bite-sized amount. We define the sparsest region as the point on the boundary of the food item mask m_t^i that is furthest from the densest region, with the condition that the line connecting these points is not intersected by other food items, such as toppings.
- `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. We initially orient the fork with tines horizontal to the plate to avoid the food item slipping off the utensil during dipping.

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 helps prevent items from slipping off the fork after being picked up.

2) Pre-acquisition skills: When the above acquisition skills are not immediately feasible due to occlusion from other items or the anticipated amount of food to be picked up being insufficient, 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

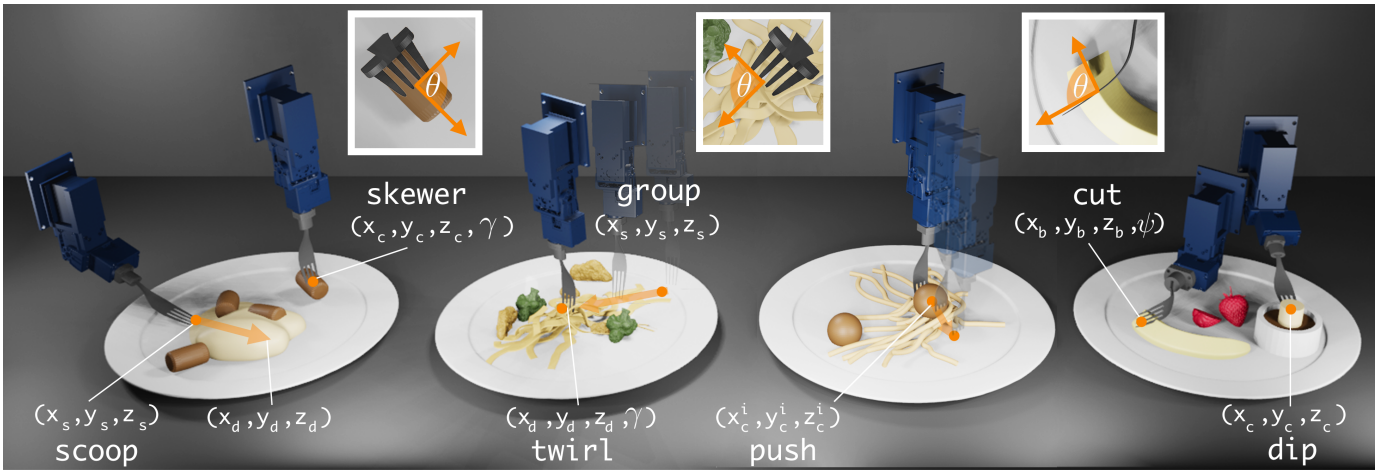

Fig. 3: Our skill library consists of 7 parameterized manipulation skills: 4 acquisition (skewer, twirl, scoop, dip) and 3 pre-acquisition (group, push, cut).

facilitate subsequent acquisition. Grouping noodles into a pile before twirling, pushing a meatball off of a bed of spaghetti before twirling, or cutting banana into a bite-sized piece before pickup are all examples of pre-acquisition. We parameterize them 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^i and execute a linear push with the fork from the sparsest to densest point.
- `push`(x_c^i, y_c^i, z_c^i): For a food item with mask m_t^i obstructing a pile of food (such as noodles or a semisolid) with mask m_t^j, we can execute a linear push motion starting at the centroid of the obstructing item, to the nearest boundary point of the underlying food bed m_t^j.
- `cut`(x_b, y_b, z_b, ψ): 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^i and traverse a fixed unit length from the one end of the axis to estimate this. We then bring the fork horizontal ($\beta = 90°$) and with sideways tines ($\gamma = 90°$). Finally, we set the end-effector roll ψ such that the lateral side of the fork is orthogonal to the major axis angle θ as in skewering and twirling. Then, we 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 Appendix.

Task Planning for Acquisition. We 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 hierarchical task planner $\mathcal{T}$. Our task planner relies on vision modules which post-process the segmented plate observations to quantify the density and spread of food items, along with checking for appropriate bite sizes and collision with other food items. While it uses a few key parameters, the overall pipeline and these parameters are shared across different categories like `'noodles'`, `'semisolid'`, and `'cuttable'`, rendering the approach versatile for diverse plates.

The task planner takes as input a particular food item category c_t^i along with the detected segmentation mask m_t^i 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'`, `'sauce'`, `'noodles'`, `'semisolid'`, `'cuttable'`}.

For most 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 to manipulate the i-th food item, sensed from m_t^i and o_t:

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

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 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:

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

However, when twirling is obstructed by another item mask m_t^j, such as a meatball too close to a planned spaghetti twirling action, the obstructing food must be pushed aside first. Similarly, if toppings such as sausages block all viable scooping actions for mashed potatoes, the sausage nearest to the boundary of the mashed potato mask should be pushed.

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

For food items in the `'noodles'` category, we also consider grouping actions. If the entropy exceeds a pre-defined threshold `ENTROPY_THRESH`, indicating that the food item is spread out on the plate, grouping can be helpful. We can directly execute unobstructed grouping actions. However, if all viable grouping actions are blocked by toppings, we instead push the topping closest to the boundary of the underlying food aside, group the food, and then acquire.

- $\mathcal{T}(\texttt{`noodles'}, m_t^i) = \{\texttt{group}(p^i_{t,\texttt{group}}), \texttt{twirl}(p^i_{t,\texttt{twirl}})\}$ if grouping is unobstructed.
- $\mathcal{T}(\texttt{`noodles'}, m_t^i) = \{\texttt{push}(p^i_{t,\texttt{push}}), \texttt{group}(p^i_{t,\texttt{group}}), \texttt{twirl}(p^i_{t,\texttt{twirl}})\}$ if the robot must push aside an obstructing topping before grouping.

If neither acquisition nor grouping is feasible according to the set thresholds, we push the topping within the food item mask nearest to its boundary, hoping to expose more of the food item for future acquisition. If no such topping is available, we default to acquisition.

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

- $\mathcal{T}(\texttt{`cuttable'}, m_t^i) = \{\texttt{cut}(p^i_{t,\texttt{cut}}), \texttt{skewer}(p^i_{t,\texttt{skewer}})\}$, if the length of major axis of $m_t^i$ exceeds `BITE_LENGTH`,
- $\mathcal{T}(\texttt{`cuttable'}, m_t^i) = \{\texttt{skewer}(p_t^i)\}$, otherwise.

We provide further details on task planning for acquisition in the Appendix.

C. Bite Sequencing via Foundation Models

We introduce a unified framework for planning and executing bite sequences that are efficient and adhere to user preferences. With access to a library of skills $\mathcal{L}$, task planner $\mathcal{T}$, and user preference ℓ_{pref}, 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_{pref}. We augment this context with: (i) a *history* of bites taken so far, (ii) an estimate of the *portions* of each food type remaining, and (iii) 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'`]. For `'cuttable'` food items, we use the major axis length of m_t^i divided by `BITE_LENGTH`. Otherwise for food items where the ground truth quantity is not countable, such as for `'noodles'` or `'semisolid'`, we use $\lceil |m_t^i| / \texttt{PORTION\_SIZE} \rceil$, where `PORTION_SIZE` is empirically determined for approximating the number of bite portions in a mask.

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]
Efficiencies: [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 intensity 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 $\mathcal{T}(\texttt{`noodles'}, m_t^{\texttt{`fettuccine'}})$ and execute it via $\mathcal{L}$.

D. Integration of Acquisition and Transfer

The self-contained nature of our *bite acquisition* framework allows for straightforward integration with *bite transfer* frameworks, and is agnostic to the exact approach used. We demonstrate this easy combination with two existing methods: (i) an outside-mouth bite transfer method [19] that features visual servoing capabilities, and (ii) a recent method for inside-mouth transfer [28] that leverages robust mouth tracking and physical interaction-aware control.

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 [43] 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 (Fig. 2) enables us to uniquely circumvent this challenge. We leverage its roll (γ) and pitch (β) 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 monitor 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.

IV. Experiments

We evaluate the effectiveness of FLAIR for feeding diverse plates each containing various types of food items. We first conduct a user study to assess FLAIR's ability to perform long-horizon bite acquisition of in-the-wild plates, while adhering to user preferences and efficiently feeding bites. For all acquisition experiments, we interchangeably use 2 Kinova Gen3 arms (one 6-DoF, and another 7-DoF), and a 7-DoF Franka Emika Panda. We then ablate our hierarchical task planner $\mathcal{T}$ against various state-of-the-art baselines [3, 39, 52]. Finally, we evaluate the real-world efficacy of our system for feeding a complete plate to a care recipient with mobility limitations.

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 refer to the Preference-Only baseline as *Commonsense-Only*.

Evaluation Plates: We consider an evaluation suite of 6 diverse plates of food spanning a wide range of food categories, visualized in Fig. 4. 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 Applebee's via Doordash. We also consider 2 homemade semisolid dishes: 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 mobility limitations (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. In the survey, we specify the capabilities of our skill library to the participants of our user study, and ask them to note preferences over their preferred order of bites, or pairings of food items with sauces. Details on the reported user preferences are provided in the Appendix. Since evaluating each submitted 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). We focus on cases where users have either no preference or strong preferences, as slight preferences are not informative for comparing 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 bite sequencing will lead to more number of bites across all settings.
- **H2**: Compared to the Efficiency-Only baseline, FLAIR's consideration of user preferences in bite sequencing in presence of strong preferences will lead to more perceived adherence to preferences, and more human-like feeding, rated based on the statement "This method is similar to the strategy I would use to feed myself."
- **H3**: Compared to the Efficiency-Only baseline, FLAIR's consideration of commonsense reasoning in bite sequencing in the absence of preferences will lead to more perceived adherence to bite variety and common food item pairings, and more human-like feeding.

This user study was approved by the Institutional Review Boards of both Cornell University and Stanford University.

Food Pickup Results: Fig. 7 displays the results of food pickup over time across methods in the no preference scenario. We provide additional results for food pickup efficiency for all methods averaged across all plates (both strong preferences and no preferences) in Appendix, noting a similar trend. Due to its consideration of efficiency in bite sequencing, FLAIR executes a greater number of pickup skills compared to Preference-Only, validating **H1**. This is because when faced with multiple valid candidate bites, FLAIR, informed with efficiency scores for each bite, is able to choose the bite that optimizes for efficiency. In contrast, Preference-Only randomly selects one bite from this set, often leading to inefficient acquisition trajectories (Fig. 5). The efficiency

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

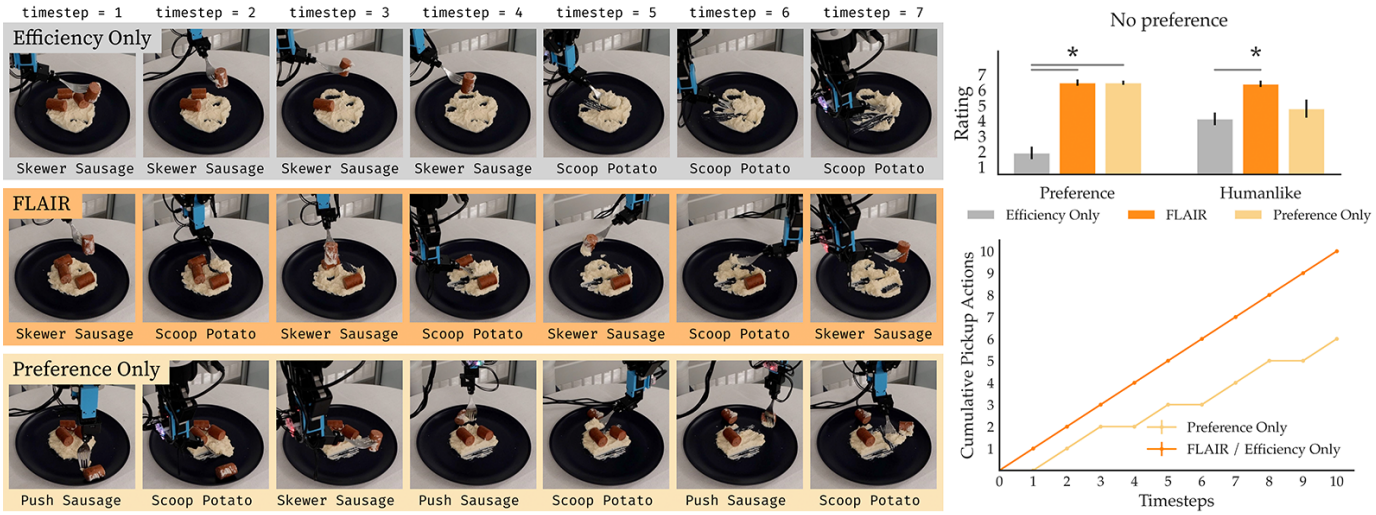

Fig. 5: Example run on a plate with mashed potatoes and sausages where the user specified no preference. FLAIR, which balances user preferences (bite variety) and efficiency, is judged by users to better adhere to preferences than Efficiency-Only and outperforms Preference-Only (Commonsense-Only) in plate clearance. Consequently, FLAIR is considered to provide a more human-like feeding experience compared to the baseline methods. Note that * indicates statistical significance (p-value < 0.05), determined via a Mann-Whitney U test.

disparity between Efficiency-Only and FLAIR can be linked to settings where bite variety or strong preferences require the robot to pickup bites 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 a bed of spaghetti hidden beneath multiple meatballs, methods that consider preferences must undertake multiple pre-acquisition skills to push away the meatballs.

User Evaluation: Fig. 6 presents average participant ratings comparing FLAIR with baseline approaches for settings with strong user preferences. We conduct a Mann-Whitney U test for statistical significance, and indicate pairs of methods for which the average participant ratings were significant (p-value < 0.05). This is a non-parametric test compatible with ordinal Likert data, and without specific assumptions on the normality or variance of the data distributions. 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, as hypothesized (**H2**). The exceptions, where the performance difference between Efficiency-Only and FLAIR is not significant, occur in settings where the bite sequence, generated based solely on efficiency, inadvertently matches the user's preferences. Fig. 7 shows participant ratings that compare FLAIR with baseline approaches in settings where no user preferences were specified. By leveraging commonsense reasoning, FLAIR significantly outperforms the Efficiency-Only baseline across most plates by ensuring bite variety and appropriately pairing food items with dips, resulting in a more human-like feeding experience (**H3**).

B. Comparisons with Task Planning Baselines

For evaluating necessity of pre-acquisition actions, FLAIR first estimates the distribution of food items by sensing density and entropy metrics from segmented observations, and then uses a hierarchical decision-tree style approach. In this section, we compare this approach against other task planning baselines. The closest relevant work, VAPORS [52], concentrates on noodle dishes and employs physics-based simulations for decision-making between twirling and grouping noodles. Our work, however, encompasses a broader spectrum of food textures and types (such as solids like meatballs, semi-solids like mashed potatoes, and noodle-like items like spaghetti), making direct adaptation of VAPORS challenging due to the complex physics simulations required for accurately representing their varied interactions. Taking this gap into account, we compare FLAIR's task planning accuracy against 3 established baselines: (i) VAPORS, (ii) VLM-TaskPlanner, which queries a VLM (GPT-4V [3]) using 10 in-context examples from the training set to decide between candidate actions, and (iii)

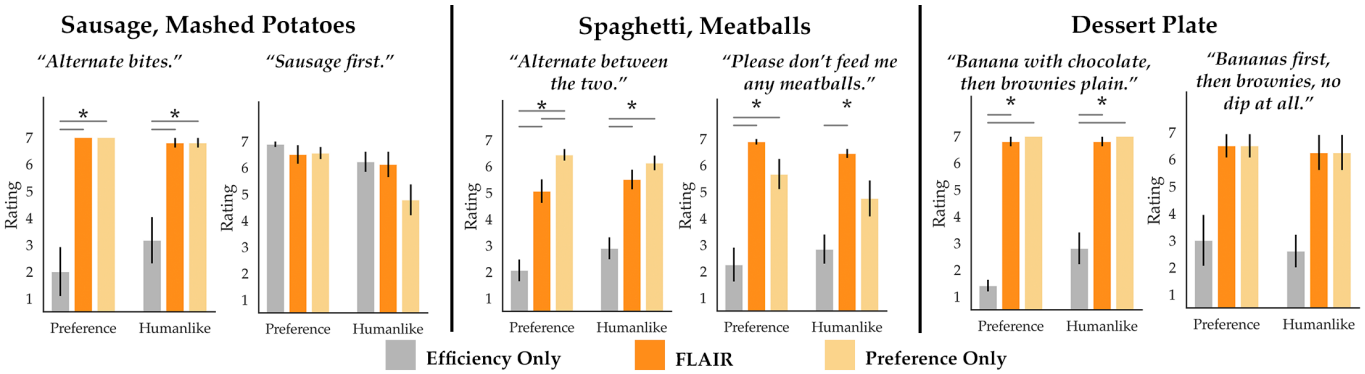

Fig. 6: Average participant ratings for settings with strong user preferences show FLAIR significantly outperforms Efficiency-Only baseline in aligning with user preferences and achieving human-like feeding in all scenarios, except cases where the efficiency-based bite sequence coincidentally aligns with user preferences. Note that * indicates statistical significance (p-value < 0.05), determined via a Mann-Whitney U test.

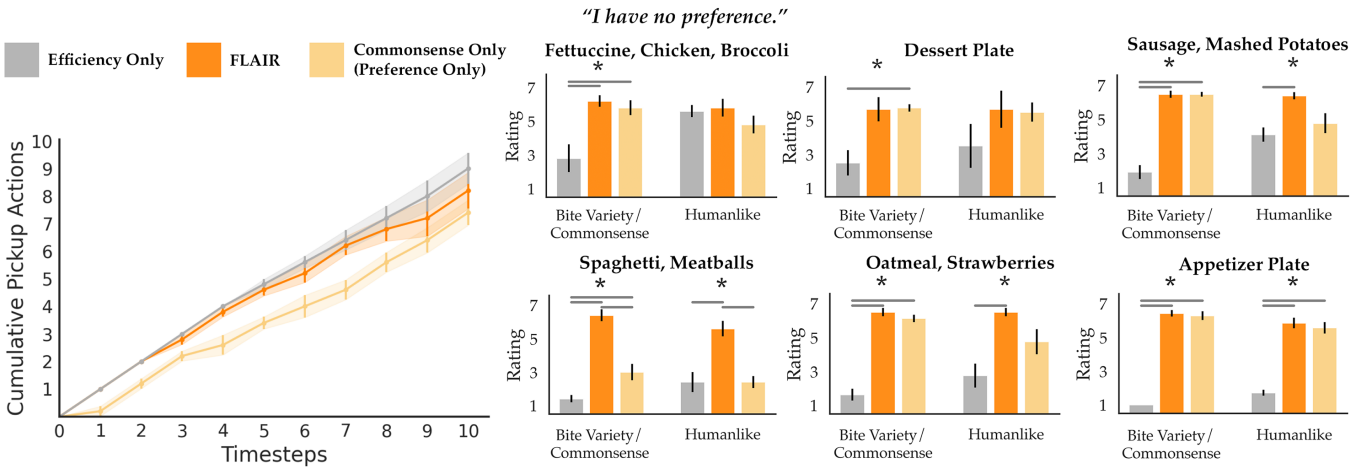

Fig. 7: *Left:* FLAIR picks up more bites than Preference-Only (Commonsense-Only) accumulated across all plates for no preference scenarios. *Right:* For most no preference scenarios, average participant ratings show FLAIR significantly outperforms the Efficiency-Only baseline in aligning with user preferences (bite variety) and achieving human-like feeding. Note that * indicates statistical significance (p-value < 0.05), determined via a Mann-Whitney U test.

image classification using a pre-trained Swin-Transformer [39] fine-tuned on the training set. We leverage 2 datasets of plate images with next action ground truth labels as a test-bed:

Evaluation on noodle-only plates from VAPORS ([52]). We use the $\sim$ 100 held-out noodle-only image dataset from VAPORS and have two third-party human annotators label each image with "Twirl" or "Group," corresponding to the action they deem appropriate. We extract the images with intercoder agreement and evenly split them into training and test sets. We compare our task planning approach (FLAIR), against the three baselines on this set.

Evaluation on logged user study plates. We further compare FLAIR to baselines on the $\sim$ 100 images per plate logged from our user study. Each image is labeled by two third-party human annotators with appropriate ground truth labels for food items where pre-acquisition is relevant: twirl/group/push for noodles, push/scoop for semisolids (oatmeal/mashed potatoes), and cut/acquire for a whole banana. We extract the images with intercoder agreement, split them evenly into training and test sets, and report the mean accuracy across plates.

TABLE I: Comparison of FLAIR's task planner with baselines

	User Study Plates	Noodle-Only Plates
FLAIR	**0.817**	**0.854**
VAPORS [52]	-	0.415
VLM-TaskPlanner [3]	0.518	0.683
Swin-Transformer [39]	0.720	0.785

FLAIR uses identical parameters for all plates, whereas other baselines use plate-specific parameters inferred from their respective training data. Table I shows the results. FLAIR significantly outperforms all other baselines on both datasets: the noodle-only and user study plates. We posit that VAPORS may suffer due to the sim-to-real gap present in real vs. simulated observations, and the black-box VLM-TaskPlanner struggles without reasoning in a hierarchical manner. The Swin-Transformer classifier is the most competitive baseline, but likely suffers due to a lack of large-scale training data.

C. Demonstration of Real-World Feeding

We demonstrate FLAIR's effectiveness in helping a care recipient with severe mobility restrictions 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 with Multiple Sclerosis for 19 years, has a severely limited range of motion in their head and neck. Consequently, they require inside-mouth transfer [28] of acquired bites for successful feeding.

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

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

through natural language, and when caregivers adhere to this preference, it "definitely enhances my eating experiences." For the considered 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." Adhering to this preference, FLAIR begins by skewering celery, dipping them in ranch dressing, and transferring them inside the mouth of the user. Once it detects there is no more celery on the plate, it switches to skewering watermelons and feeding them without dips as instructed. Finally, it skewers the strawberries, dips them in chocolate sauce, and feeds them to complete the meal (Fig. 8). Following successful feeding, we posed two questions on a seven-point Likert scale on the necessity of a robot-assisted feeding system to (i) have a 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 refrain from sharing these preferences with their human caregivers. They expressed concern that such requests might impose additional burdens on caregivers who are already assisting them with feeding. However, they were hopeful that a robot designed to assist with feeding could accommodate their preferences seamlessly, thus enhancing their mealtime experience by respecting their autonomy and enabling them to better enjoy their meals.

V. Discussion

FLAIR is a first step towards robot-assisted feeding in real-world scenarios, adeptly handling various in-the-wild meals composed of diverse food items. We deploy FLAIR across 2 institutions and 3 different embodiments with a library of 7 dexterous skills. Our evaluations include both bite acquisition and bite transfer, along with a demonstration feeding a complete plate to a care recipient. FLAIR showcases the ability to abide by preferences across 42 individuals and a range of diverse plates, without compromising on efficient food pickup. Through our extensive evaluations, we identify the following limitations to guide future work in robot-assisted feeding.

Limitations of Food Perception using VLMs. While current VLMs are capable of identifying food items on a plate, using these generated identifiers with open-set object detectors can sometimes lead to inaccuracies. FLAIR addresses this challenge by enriching the identifiers with a set of hand-coded descriptors tailored to the typical type of the food item, for example, specifying `'banana'` as `['banana piece', 'sliced banana']` and `'fettuccine'` as `['fettuccine pasta', 'fettuccine noodles']`. In the future, advancements in open-set object detection may eventually make such specific enhancements unnecessary.

Limitations of Food Manipulation Skills. FLAIR leverages a library of skills inspired by state-of-the-art food manipulation methods, but open challenges that occasionally occur include: slippage during skewering, failing to twirl noodles or scoop mashed potatoes into reasonable bite sizes, failing to cut tough items, and errors due to perception (erroneous depth sensing or imprecise food detection) which can cause manipulation imprecision. Although some of these failures can be addressed by re-trying (as long as the item is re-detected), these challenges can be mitigated in the future by making the skills themselves reactive, enabling adaptive utensil trajectories that adjust to food slippage or deformation on the fly.

Limitations of Bite Sequencing. We harness LLMs to plan efficient bite sequences that adhere to user preferences. However, today's language models can sometimes generate unrealistic or irrelevant outputs ("hallucinations"). In FLAIR, we reduce hallucinated artifacts in bite sequencing by using prompt-engineering strategies which we detail in the Appendix. However, even with templated prompts, FLAIR is still limited by the tendency of language models to occasionally neglect context, such as the manipulation efficiency of food items and their remaining portions. In the future, we are excited by structured prompting strategies [17] and incorporation of real-time corrections from the user [49] to address these challenges.

Although these are current limitations, **FLAIR's modular system design allows for easy interchange of the perception/planning stacks or even skills themselves.** Thus, it will be able to take full advantage of future advances in VLMs or better low-level skill policies that are learned or engineered.

VI. Acknowledgement

This work was partly funded by NSF IIS #2132846, CAREER #2238792, and DARPA under Contract HR001120C0107. It was additionally supported by NSF awards #2132847, #2218760, the Office of Naval Research award #N00014-21-1-2298, and AFOSR YIP. Priya Sundaresan is supported by an NSF GRFP and Maram Sakr is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). We would like to acknowledge Anthony Song, Pranav Thakkar, Eric Hu, and Karan Jha for their assistance with user studies and annotations for our task planning experiments.

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] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[9] 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.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

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

[13] Yongchao Chen, Jacob Arkin, Yang Zhang, Nicholas Roy, and Chuchu Fan. Autotamp: Autoregressive task and motion planning with llms as translators and checkers. *arXiv preprint arXiv:2306.06531*, 2023.

[14] 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.

[15] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24 (240):1–113, 2023.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[17] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.

[18] 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.

[19] 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.

[20] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. *arXiv preprint arXiv:2309.02561*, 2023.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[26] 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.

[27] 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.

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

[29] Hanxiao Jiang, Binghao Huang, Ruihai Wu, Zhuoran Li, Shubham Garg, Hooshang Nayyeri, Shenlong Wang, and Yunzhu Li. Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation. *arXiv preprint arXiv:2402.15487*, 2024.

[30] Subbarao Kambhampati. Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 2024.

[31] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023.

[32] Maya N Keely, Heramb Nemlekar, and Dylan P Losey. Kiri-spoon: A soft shape-changing utensil for robot-assisted feeding. *arXiv preprint arXiv:2403.05784*, 2024.

[33] 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.

[34] Minae Kwon, Hengyuan Hu, Vivek Myers, Siddharth Karamcheti, Anca Dragan, and Dorsa Sadigh. Toward grounded social reasoning. *arXiv preprint arXiv:2306.08651*, 2023.

[35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[36] 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.

[37] Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. *arXiv preprint arXiv:2403.03174*, 2024.

[38] 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.

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[40] 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.

[41] 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.

[42] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[43] 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.

[44] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[45] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9, 2019.

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya

Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[47] Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, 2020.

[48] 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.

[49] Lucy Xiaoyang Shi, Zheyuan Hu, Tony Z Zhao, Archit Sharma, Karl Pertsch, Jianlan Luo, Sergey Levine, and Chelsea Finn. Yell at your robot: Improving on-the-fly from language corrections. *arXiv preprint arXiv:2403.12910*, 2024.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.

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

[57] 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.

[58] 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)*, 2019.