LAVA: Long-horizon Visual Action based Food Acquisition

Amisha Bhaskar, Rui Liu, Guangyao Shi, Pratap Tokekar

***Abstract*— Robotic Assisted Feeding (RAF) addresses the fundamental need for individuals with mobility impairments to regain autonomy in feeding themselves. The goal of RAF is to use a robot arm to acquire and transfer food to individuals from the table. Existing RAF methods primarily focus on solid foods, leaving a gap in manipulation strategies for semi-solid and deformable foods. This study introduces Long-horizon Visual Action (LAVA) based food acquisition of liquid, semisolid, and deformable foods. Long-horizon refers to the goal of "clearing the bowl" by sequentially acquiring the food from the bowl. LAVA employs a hierarchical policy for long-horizon food acquisition tasks. The framework uses high-level policy to determine primitives based on food types. At the mid-level, LAVA finds parameters for primitives using vision. To carry out sequential plans in the real world, LAVA delegates action execution which is driven by Low-level policy that uses parameters received from mid-level policy and behavior cloning ensuring precise trajectory execution. We validate our approach on complex real-world acquisition trials involving granular, liquid, semisolid, and deformable food types along with fruit chunks and soup acquisition. Across 46 bowls, LAVA acquires much more efficiently than baselines with a success rate of $89 \pm 4\%$, and generalizes across realistic plate variations such as different positions, varieties, and amount of food in the bowl. Code, datasets, videos, and supplementary materials can be found on our website.**

I. INTRODUCTION

For individuals limited mobility or disabilities, self-feeding can be a daunting task, underscoring the need for Robotic Assisted Feeding (RAF) [1] systems, to enhance independence and quality of life as well as reducing caregiver burden. Dealing with various foods—from granular cereals to semi-solid food such as yogurt and deformable food items such as tofu, without breakage or deformation presents significant challenges for RAF [2], [3]. Traditional RAF methods have relied on pre-set strategies for specific tasks like skewering [4]–[7], bite transfer [4], [8], [9], and scooping [2], [10], which falls short in complex feeding scenarios akin to human feeding actions. This gap highlights the need for replicating nuanced, human-like feeding strategies. This gap in technology prompts the exploration of hierarchical frameworks that break down intricate feeding actions into simpler steps [7], [11]–[13], addressing the challenge of complex food handling. Yet, deploying these frameworks to manage the diverse and changeable nature of food in real-world settings remains a formidable challenge. We aim to leverage hierarchical planning, vision-based control, and flexible adaptation to various food types, addressing the limitations of current RAF approaches.

All authors are from the University of Maryland, College Park, MD 20742 USA. This work is supported by the National Science Foundation under Grant No. 1943368 and an Amazon Research Award. {amishab, ruiliu, gyshi, tokekar}@umd.edu

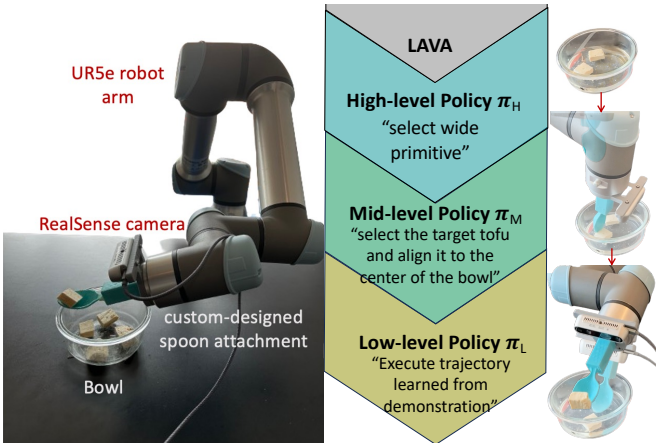


Fig. 1: System setup for LAVA with an illustrative description of the proposed framework with snapshots of task execution.

II. PROBLEM STATEMENT

This study tackles the challenge of sequential bite acquisition to maximize the success rate and efficiency of long-horizon food acquisition for efficient bowl clearance. The focus is on a variety of food types, from granular items such as cereals to semi-solid foods such as yogurt, and deformable substances such as tofu, all within a static bowl and assumed to be scoopable with a spoon. We assume access to bowl image observations $o \in {\mathbf{R}_+}^{W \times H \times C} = \mathcal{O}$ of unknown bowl states S. Here, W, H, and C denote the image dimensions. The image is sourced from a camera attached to the wrist of the robotic arm. We have access to expert demonstration data for robot proprioceptive information (joint positions). Our goal is to learn a policy $\pi(\phi_t|o_t)$ that takes RGB images as input (o_t) and returns output as joint angles θ_t of the arm for efficient long-horizon food acquisition.

III. PROPOSED APPROACH

We formalize the long-horizon food acquisition setting as a hierarchical policy π. To do so we decouple π into separate high, mid, and low-level sub-policies. We assume access to K discrete manipulation primitives P_H^k, $k \in 1, ..., K$, and learn a high-level policy π_H which selects amongst these primitives based on visual input o_t. The mid-level policy π_M further refines this selection, parameterizing the low-level policy π_L based on both the chosen primitive and additional visual inputs. This low-level policy then executes a sequence of actions θ_t^k, aimed at achieving precise food acquisition. The formulation of this hierarchical arrangement is as follows:

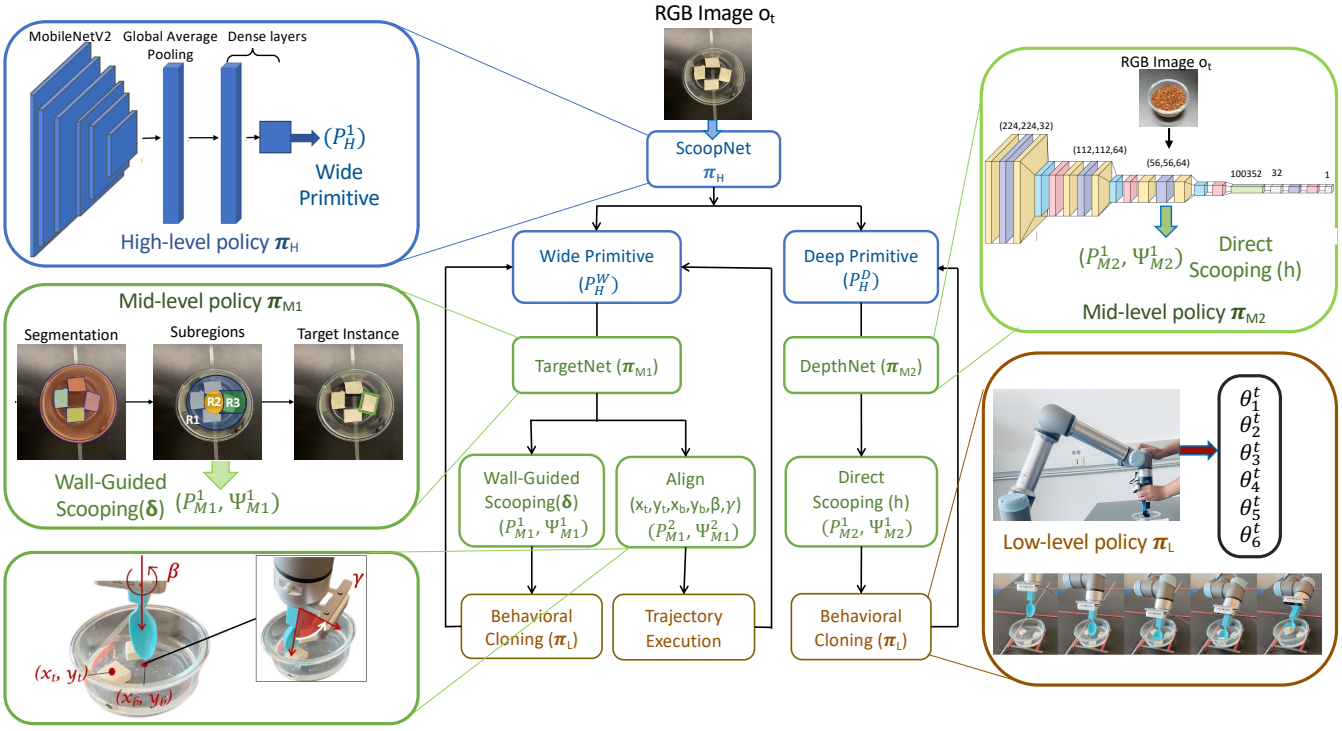


Fig. 2: System Architecture of LAVA, employs a high level policy(blue) π_H to select amongst discrete high level primitives P_H^k, which further gets refined by mid-level policy (green) π_M to select amongst mid-level primitivesP_M^k, low-level vision parametrized policy π_L (brown) executes trajectory learned from Behavioral cloning for long-horizon food acquisition.

- **High-level policy:** $\pi_H(P_H^k|o_t)$ focuses on selecting the manipulation primitive for the current observation.
- **Mid-level policy:** $\pi_M(P_M^k, \psi_M^k|o_t, P_H^k)$ refines this choice by parameterizing actions to the specific food item's characteristics.
- **Low-level policy:** $\pi_L(\theta_t^k|P_M^k, \psi_M^K)$ executes the action sequence, using parameters from the mid-level policy.

We consider low-level actions θ_t, parameterized by the position of the tip of a spoon (x, y) and spoon roll and pitch (γ, β) in the wrist frame of reference. As shown in Figure 2 detailing the LAVA setup, for more detailed description of each module, refer to our paper [14].

A. High-level Policy

At the highest level of our hierarchical model, the high-level policy $\pi_H(P_H^k|o_t)$ uses visual cues to select the most suitable scooping primitive—Wide Primitive (P_H^W) for non-cohesive, deformable foods such as tofu, and Deep Primitive (P_H^D) for cohesive foods such as cereals. The Wide Primitive leverages the bowl's wall for support, creating a mass easy to scoop without causing food breakage, while the Deep Primitive enables direct scooping with precise control over spoon trajectory for minimal disturbance.

ScoopNet (π_H): ScoopNet, built on the MobileNetV2 architecture [15] as the base, distinguishes between these primitives. Trained on a dataset of 5316 images for accurate primitive selection, employing a Global Average Pooling layer and dense layers for refined classification. We use Adam optimizer and binary cross-entropy loss for the Optimizations, producing softmax probabilities for selecting scooping strategies for specific task adaptation.

B. Mid-level Policy

The Mid-level Policy $\pi_M(P_M^k, \psi_M^K|o_t, P_H^k)$ refines and parameterizes the chosen primitive, crucial for translating high-level strategy decisions into low-level action execution.

1) TargetNet (π_{M1}) for Wide Primitive: TargetNet employs Mask R-CNN to identify and segment target items such as tofu for scooping. This model segments food items, enabling the selection of appropriate mid-level primitives: wall-guided scooping and center align using annotations for precise segmentation and transfer learning for accuracy. TargetNet divides the bowl into sub-regions (R1 for rightmost and closest to the wall, R2 for center and R3 otherwise) to guide scooping decisions, whether leveraging wall support or aligning for easier access.

Wall-guided Scooping and Align: This method varies scooping based on food's position—Wall-guided Scooping for foods in subregions R1 and R2 and Align for food positioned in R3 subregion. The alignment step calculates the spoon's orientation and the distance to move food towards the bowl's center, optimizing scooping paths.

2) DepthNet (π_{M2}) for Deep Primitive: DepthNet, with its Sequential model, determines food depth in the bowl, aiding in selecting the depth for deep scooping of cohesive foods. It's trained on diverse cereal images to precisely estimate food volume, adjusting the scooping depth accordingly for effective clearance.

Direct scooping (P_{M2}^1, ψ_{M2}^1) Incorporates real-time feedback to adjust scooping strategies based on DepthNet's depth

information, using behavior cloning to refine the spoon's path, ensuring efficient scooping across different food depths.

C. Low-level policy

At the foundation of our model, we use behavioral cloning (π_L), coupled with kinesthetic teaching [16], to fine-tune the robot's scooping actions across varied food textures, directly informed by expert demonstrations. This method involves learning distinct scooping trajectories for different foods. The goal is to minimize deviations from these optimal paths using a cost function $J(\tau)$, with the Weiszfeld algorithm [17], [18] applied for optimization. This algorithm iteratively adjusts the estimated trajectory $\hat{x}$, improving scooping precision by reducing the sum of distances from demonstrated trajectories until minimal changes are achieved. For a deeper dive into the specifics of our behavioral cloning approach and its application within **LAVA**, refer to our discussion in paper.

IV. Quantitative Results

Our experiments detailed in [14] involve a comprehensive setup (see Figure 1) including a UR5e robot arm with custom spoon attachment, and a RealSense camera, testing on a variety of food types from cereals to tofu in soup. Utilizing two baselines, LAVA-low and Fixed Trajectory Scooping, for comparative analysis, we explore a range of food configurations to assess our hierarchical framework's effectiveness in adaptive food acquisition. Our key findings:

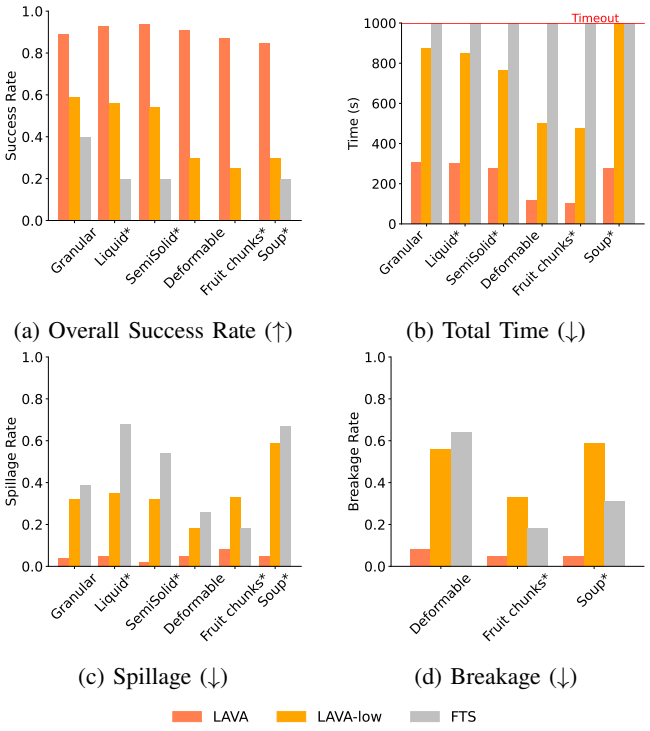
1) Network Performance: ScoopNet achieved 100% accuracy in choosing correct high-level primitives, TargetNet accurately predicted bite targets at 87.9% , and DepthNet successfully determined correct spoon depths for bite sizes at 85.7%, demonstrating the **LAVA** networks' effectiveness in robotic-assisted feeding.

2) Baselines Comparison: **LAVA** outperformed both baseline models, LAVA-low and FTS, in efficiency, scoop size, and minimizing spillage and breakage as visible in Figure 4 and Figure 3.It adeptly managed liquids, significantly minimized breakage with deformable foods such as tofu through strategic scooping, and ensured minimal spillage with solid foods using align-then-scoop strategy.

3) Zero-shot Generalization: **LAVA** effectively handled diverse foods, including soup with tofu and apple chunks, showcasing adaptability in Figures 3 and 5. Its ability to adjust in real-time for both solid and liquid scooping underlines **LAVA**'s robustness across food types. Refer to our paper for a comprehensive overview of **LAVA**'s methodologies.

V. Conclusion, Limitation and Future Work

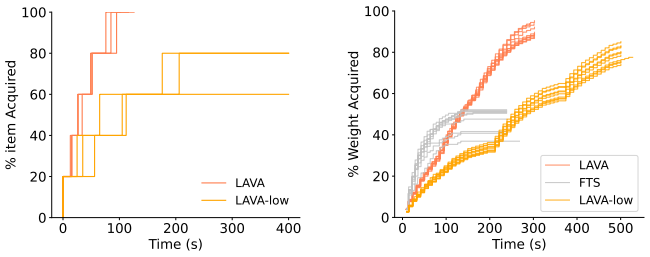
In this study, we introduced a hierarchical policy framework, **LAVA**, that improves robotic food acquisition from liquids to deformable solids. Utilizing **LAVA**'s networks, it addresses the variability in food types, achieving higher efficiency and accuracy with less spillage and breakage than baselines. Despite its success, challenges remain with thin or irregularly shaped foods. Future work aims to expand the action space and explore new data acquisition methods, potentially using online videos for complex food interactions.



(a) Overall Success Rate (↑) (b) Total Time (↓)

(c) Spillage (↓) (d) Breakage (↓)

Fig. 3: Breakdown of experimental performance comparison between **LAVA**, LAVA-low, and Fixed Trajectory Scooping(FTS). $*$ represents zero-shot experiments.



(a) 5 tofu chunks. (b) Cereals.

Fig. 4: Individual trials comparison between **LAVA** and baselines: (a) different tofu configurations, (b) cereals

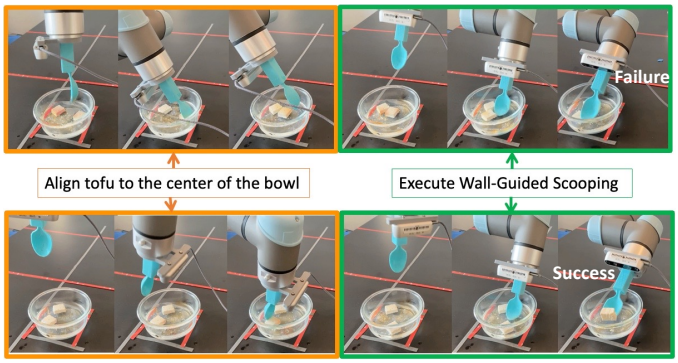


Fig. 5: Zero-shot acquisition with tofu in soup: Top images depict spoon alignment of tofu to the bowl's center, which drifts due to soup's fluidity. Bottom images show realignment and successful scooping.

REFERENCES

[1] S. W. Brose, D. J. Weber, B. A. Salatin, G. G. Grindle, H. Wang, J. J. Vazquez, and R. A. Cooper, "The role of assistive robotics in the lives of persons with disability," *American Journal of Physical Medicine & Rehabilitation*, vol. 89, no. 6, pp. 509–521, 2010.

[2] J. Grannen, Y. Wu, S. Belkhale, and D. Sadigh, "Learning bimanual scooping policies for food acquisition," *arXiv preprint arXiv:2211.14652*, 2022.

[3] P. Sundaresan, J. Wu, and D. Sadigh, "Learning sequential acquisition policies for robot-assisted feeding," *arXiv preprint arXiv:2309.05197*, 2023.

[4] D. Gallenberger, T. Bhattacharjee, Y. Kim, and S. S. Srinivasa, "Transfer depends on acquisition: Analyzing manipulation strategies for robotic feeding," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 267–276.

[5] R. Feng, Y. Kim, G. Lee, E. K. Gordon, M. Schmittle, S. Kumar, T. Bhattacharjee, and S. S. Srinivasa, "Robot-assisted feeding: Generalizing skewering strategies across food items on a plate," in *The International Symposium of Robotics Research*. Springer, 2019, pp. 427–442.

[6] T. Bhattacharjee, G. Lee, H. Song, and S. S. Srinivasa, "Towards robotic feeding: Role of haptics in fork-based food manipulation," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1485–1492, 2019.

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

[8] S. Belkhale, E. K. Gordon, Y. Chen, S. Srinivasa, T. Bhattacharjee, and D. Sadigh, "Balancing efficiency and comfort in robot-assisted bite transfer," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4757–4763.

[9] R. K. Jenamani, D. Stabile, Z. Liu, A. Anwar, K. Dimitropoulou, and T. Bhattacharjee, "Feel the bite: Robot-assisted inside-mouth bite transfer using robust mouth perception and physical interaction-aware control," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 313–322.

[10] R. Liu, A. Bhaskar, and P. Tokekar, "Adaptive visual imitation learning for robotic assisted feeding across varied bowl configurations and food types," *arXiv preprint arXiv:2403.12891*, 2024.

[11] X. Lin, C. Qi, Y. Zhang, Z. Huang, K. Fragkiadaki, Y. Li, C. Gan, and D. Held, "Planning with spatial-temporal abstraction from point clouds for deformable object manipulation," *arXiv preprint arXiv:2210.15751*, 2022.

[12] M. Dalal, D. Pathak, and R. R. Salakhutdinov, "Accelerating robotic reinforcement learning via parameterized action primitives," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 847–21 859, 2021.

[13] S. Nasiriany, H. Liu, and Y. Zhu, "Augmenting reinforcement learning with behavior primitives for diverse manipulation tasks," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 7477–7484.

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

[15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[16] B. Akgun, M. Cakmak, J. W. Yoo, and A. L. Thomaz, "Trajectories and keyframes for kinesthetic teaching: A human-robot interaction perspective," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 391–398.

[17] A. Beck and S. Sabach, "Weiszfeld's method: Old and new results," *Journal of Optimization Theory and Applications*, vol. 164, pp. 1–40, 2015.

[18] E. Weiszfeld and F. Plastria, "On the point for which the sum of the distances to n given points is minimum," *Annals of Operations Research*, vol. 167, pp. 7–41, 2009.