



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

## Long-horizon manipulation through hierarchical motion planning with subgoal prediction

**Citation for published version:**

Saito, N, Pousa De Moura, J & Vijayakumar, S 2024, Long-horizon manipulation through hierarchical motion planning with subgoal prediction. in *Proceedings of 40th Anniversary of the IEEE Conference on Robotics and Automation*. Institute of Electrical and Electronics Engineers, 40th Anniversary of the IEEE Conference on Robotics and Automation, Rotterdam, Netherlands, 23/09/24.

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Peer reviewed version

**Published In:**

Proceedings of 40th Anniversary of the IEEE Conference on Robotics and Automation

**General rights**

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



# Long-horizon Manipulation through Hierarchical Motion Planning with Subgoal Prediction

Namiko Saito<sup>1</sup>, João Moura<sup>1</sup>, and Sethu Vijayakumar<sup>1</sup>

*Abstract*—The research on long-horizon manipulation in environments with numerous objects and subtasks falls under the framework of task and motion planning (TAMP). One effective solution for TAMP is to separate higher-level discrete short-horizon subgoals, and lower-level continuous motion generation to enhance robustness, scalability, and generalizability. We propose a concept of hierarchical framework combining deep neural networks (DNN) for higher-level subgoal decisions and optimization for lower-level motion control. This will be evaluated on a latent state box transport and stacking task – where the robot needs to change the order of actions and speed to control during motion execution. Additionally, we can apply this framework to daily tasks such as cooking, where the robot needs to recognise the states of ingredients, select appropriate tools and subtasks, and adjust its motions accordingly.

## I. INTRODUCTION

Many daily tasks require long-horizon reasoning consisting of multiple short-term primitive tasks (subtasks). One solution for long-horizon manipulation containing a number of objects and potential actions is to separate higher-level discrete short-horizon subgoals, and lower-level continuous motion generation to enhance robustness, scalability, and generalizability. To achieve each short-horizon subgoal, the robot must recognise the target object characteristics such as weight, stiffness and fragility, and flexibly adjust its motions such as speed, acceleration, and trajectory to the specific characteristics. We propose a new concept of a hierarchical TAMP framework that combines deep neural networks (DNN) for representation learning of target objects and higher-level subgoal decisions, and optimized Model Predictive Control (MPC) for lower-level motion plans and executions.

Prior research designed decision trees to break tasks into short-term subgoals [1], [2], or learned sequences of subtasks from demonstrations to infer latent vectors that represent the subtasks [3], [4]. However, their higher-level modules lack online adaptation of subgoal generation and ability to incorporate physical constraints in the lower-level motion generation. Consequently, they are unable to (1) update subgoals, (2) adjust timing of subgoal completion, and (3) update motions online. We will realize:

- Perceive target object’s characteristics (kinematic, dynamic) during motion execution,
- Update higher-level subgoals adapting to the online feedback,
- Plan optimal, robust motions following updated subgoals.

<sup>1</sup>Authors are with the School of Informatics, The University of Edinburgh, Edinburgh, U.K., and with The Alan Turing Institute, London, U.K.

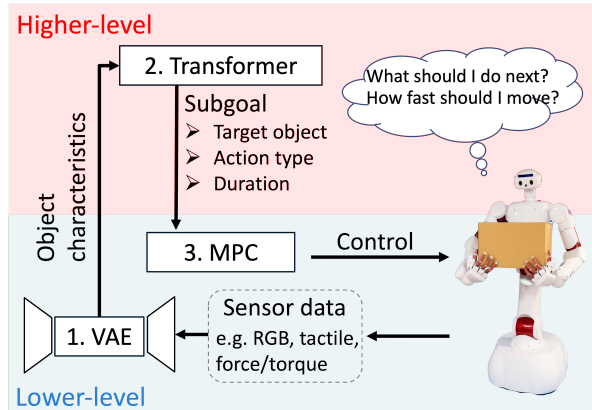


Fig. 1: The hierarchical TAMP framework which combines DNN for object representation and higher-level subgoal decisions, and MPC for lower-level motion plans and executions.

## II. PROPOSED CONCEPT AND METHOD

The proposed framework shown in Fig. 1 integrates deep learning and optimal control methods, where deep learning handles object characteristics recognition and subgoal prediction at the higher hierarchy, while optimal control manages trajectory planning and execution at the lower hierarchy. The hierarchical approach is expected to reduce learning costs, improve generalization performance, and improve transferability for other tasks and objects. Additionally, by using optimal control for trajectory planning, we can explicitly define safety constraints such as workspace and velocity limits, as well as behavioural goals such as robustness, smoothness and time efficiency when generating the robot’s motions. We can also improve the explainability of the generated output by the framework. The proposed hierarchical framework comprises three modules.

### A. Object characteristics perception

We equip some sensors to the robot such as RGB camera, force-tactile and tactile sensors. A Variational Auto Encoder (VAE) performs self-supervised learning to represent features in its latent space, such as weight, centre of mass, instability (how easily it moves), directionality of movement (in which direction it moves), etc. Building upon prior work [5], we identify object characteristics dynamically as the robot moves.

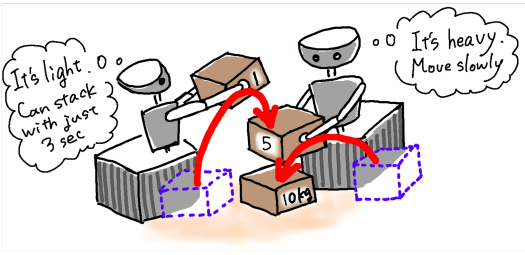


Fig. 2: Application showcase: Box transfer and stacking

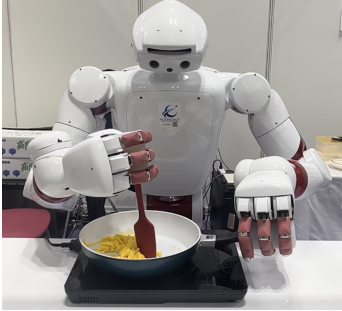


Fig. 3: Application showcase: cooking scrambled egg. The live demonstration was shown in ICRA2024 at Yokohama.

### B. Higher-level: Updating subgoals adjusting to object characteristics

We employ the Transformer model [6] to predict a series of subgoals towards the final goal based on the recognized characteristics. The Transformer model is adept at processing sequential data and are proficient in discerning semantic relationships and capturing the long-term dependencies, and allows flexibility in adjusting the number of output time steps. Each subgoal is represented by the next target object to be operated on, the action type to be conducted, and the duration of the operation. The speed, tilt angles and trajectories are adjusted according to the object characteristics.

### C. Lower-level: Online planning and control

Leveraging our previous research on online, multi-contact MPC [7], we realise fast and stable trajectory and manipulation motions while avoiding dynamic obstacles based on the provided subgoals. Formulated as an optimal control problem, MPC generates a sequence of trajectories from the subgoals and environmental information, considering parameters such as the target end-effector’s position, operation time, maximum speed, and obstacles to be avoided.

## III. APPLICATION SHOWCASES

### A. Box transfer and stacking

As an example to demonstrate the effectiveness of the proposed method, we use a whole-body robot with a mobile base and dual arms for the task of transferring and stacking boxes with hidden contents including several types of objects. The robot is required to recognize the contents of the boxes and dynamically adjust the speed and trajectory of the transport accordingly. Furthermore, it must consider the weight of the boxes and adhere to stacking principles, placing heavy or stable boxes at the bottom and lighter or less stable ones

at the top. We aim to transport and stack all boxes at the designated location in a stable and efficient manner.

### B. Cooking scrambled egg

The concept can be applied to daily tasks such as cooking, cleaning and doing laundry. Prior to implementing our concept, we demonstrated cooking scrambled eggs using only DNN in a previous study [8]. We also showcased the task at Moonshot goal3 exhibition booth during IEEE International Conference on Robotics and Automation (ICRA) 2024 as depicted in Fig. 3. In this previous work, the robot adjusted its stirring method and direction based on the egg’s status: initially stirring throughout the pot and later transitioning to flipping and splitting motions as the egg heated. With the introduction of our novel concept featuring a hierarchical framework, longer and more flexible sequential manipulations involving multiple subtasks become feasible. These manipulations may include tasks such as changing tools, adjusting the cooking duration based on personal preferences for egg stiffness, integrating additional subtasks such as cracking eggs, grating cheese, stirring egg mixture and the cheese and poring it to the pan.

## IV. IMPACT

Our TAMP framework covers object characteristics recognition, subgoal prediction, and motion planning, by combining higher and lower-level modules. The novelty lies in updating subgoals through learning and executing motion via trajectory optimization. Our proposal has the potential to be applied to a wide range of tasks involving diverse objects, enabling semi-autonomous operation, which will have an impact on the practical use of robots in daily spaces.

## REFERENCES

- [1] C. R. Garrett, T. Lozano-Perez, and L. P. Kaelbling, “Pddlstream: Integrating symbolic planners and blackbox samplers via optimistic adaptive planning,” in *International Conference on Automated Planning and Scheduling*, 2018.
- [2] R. Holladay, T. Lozano-Pérez, and A. Rodriguez, “Robust planning for multi-stage forceful manipulation,” in *The International Journal of Robotics Research*, vol. 43, 2024, pp. 330–353. DOI: [10.1177/02783649231198560](https://doi.org/10.1177/02783649231198560).
- [3] X. Lin, Z. Huang, Y. Li, J. B. Tenenbaum, D. Held, and C. Gan, “Diffskill: Skill abstraction from differentiable physics for deformable object manipulations with tools,” in *Proceeding of International Conference on Learning Representations*, 2022.
- [4] K. Pertsch, O. Rybkin, F. Ebert, C. Finn, D. Jayaraman, and S. Levine, “Long-horizon visual planning with goal-conditioned hierarchical predictors,” 2020, pp. 17321–17333.
- [5] N. Saito, J. Moura, H. Uchida, and S. Vijayakumar, “Latent object characteristics recognition with visual to haptic-audio cross-modal transfer learning,” in *arXiv*, 2024. DOI: [10.48550/arXiv.2403.10689](https://doi.org/10.48550/arXiv.2403.10689).
- [6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, E. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [7] J. Moura, T. Stouraitis, and S. Vijayakumar, “Non-prehensile planar manipulation via trajectory optimization with complementarity constraints,” in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 970–976. DOI: [10.1109/ICRA46639.2022.9811942](https://doi.org/10.1109/ICRA46639.2022.9811942).
- [8] N. Saito, M. Hiramoto, A. Kubo, K. Suzuki, H. Ito, S. Sugano, and T. Ogata, “Realtime motion generation with active perception using attention mechanism for cooking robot,” in *arXiv*, 2023. DOI: [10.48550/arXiv.2309.14837](https://doi.org/10.48550/arXiv.2309.14837).