From Sensor Data to Long-Horizon Plans with Spatial-Temporal Reasoning

Yixuan Huang

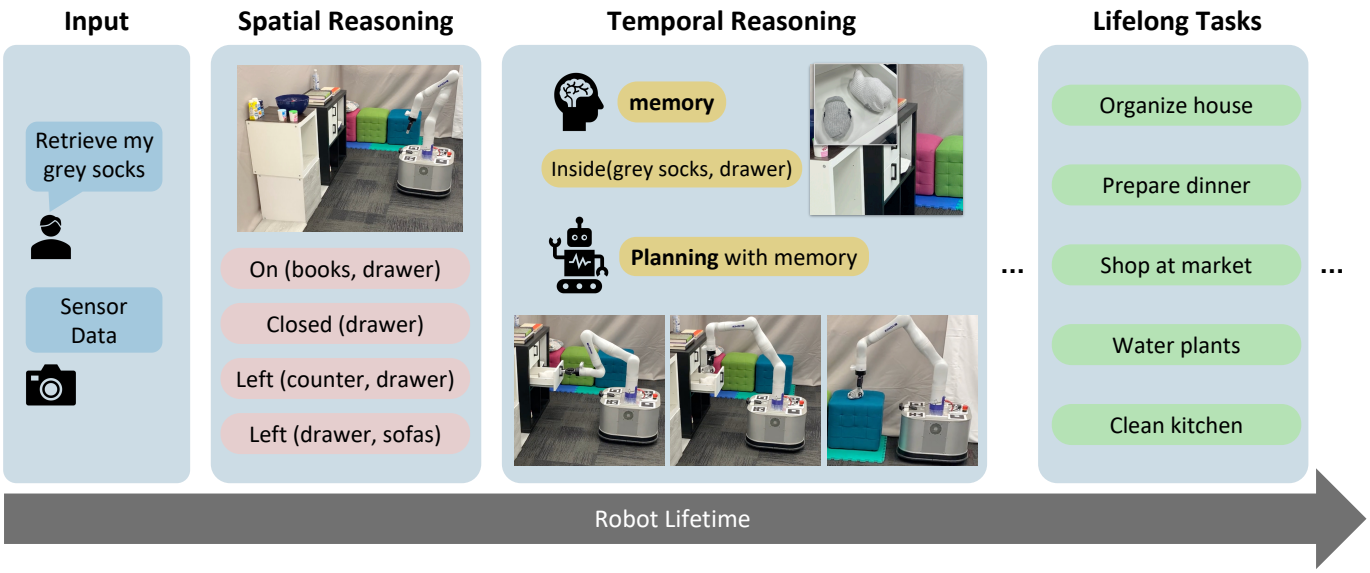

Fig. 1: Spatial-temporal reasoning in a lifelong learning setting. Using language instructions and high-dimensional sensor data (e.g., point clouds), the robot completes a task through spatial-temporal reasoning. Over its lifetime, the robot progressively learns from experiences, such as failure cases, to adapt to new environments and eventually tackle lifelong tasks.

I. Introduction

For robots to function as home assistants or caregivers, they must solve sequential manipulation tasks that require spatial-temporal reasoning and process high-dimensional sensor data as input. For example, as shown in Fig. 1, when assisting with preparing clothing for the next day, a robot must understand spatial relationships among clothes, a drawer, and a counter. Temporal reasoning requires retaining the memory of past objects and employing a future prediction module. If asked to retrieve grey socks inside a drawer, the robot must remember the grey socks' location, recognize that opening the drawer is necessary before grasping the socks, and predict how the socks will move based on diverse robot skills.

A key challenge in this domain is connecting sensory input to a unified representation that supports spatial-temporal reasoning for sequential manipulation tasks. Recent works [1, 2, 3, 4, 5] have addressed this by learning latent space dynamics models for model-based control, but predict state changes on small timescales. Conversely, task and motion planning (TAMP) addresses long-horizon tasks through high-level symbolic planning and low-level geometric reasoning. However, TAMP methods typically rely on explicit 3D object models [6, 7, 8, 9, 10] and symbolic operators with predefined effects [11, 7, 8, 9, 10, 12], limiting their applicability to real-world scenarios with high-dimensional, partial-view point clouds and complex, hard-to-predefine object interactions.

My research aims to bridge this gap by **developing a latent representation that integrates spatial-temporal reasoning with sensory data**. This representation captures both **geometric and symbolic effects of actions** within a shared latent space, enabling robots to perform **complex, long-horizon manipulation tasks in real-world environments**. Fig. 2 provides a high-level overview of the proposed framework.

II. Past and Ongoing Work

Long-horizon planning with learned latent dynamics for sequential manipulation. For robots to fully integrate into human environments and assist in daily life, they must solve sequential manipulation tasks requiring reasoning autonomously about the long-term effects of their actions. To enable spatial-temporal reasoning, I proposed a method to learn a latent space that encodes object-centric information [13, 14, 15]. High-dimensional partial-view point clouds are encoded into latent states, which a decoder translates these latent states into geometric states (e.g., object poses) and symbolic states (e.g., pair-wise relations). A dynamics model then predicts future latent states based on current latent states and parameterized robot skills. Trained in simulation with random robot skills, this system can be deployed on real-world robots (e.g., a 7DOF Kuka Arm with a reflex hand [13, 14] or a custom mobile base equipped with a Kinova arm and gripper [15]) to achieve logical [13, 14] or language goals [15]. Experiments showed that employing graph neural networks and transformers for modeling latent space dynamics yielded the best performance in reasoning about multiple interacting objects and generaliz-

ing to unseen scenarios. This success is attributed to the strong inductive bias inherent in these architectures.

Reasoning about occluded objects with a video tracker. Reasoning about occluded objects is essential for enabling robots to assist effectively in real home environments. To address this challenge, I proposed using an unsupervised video object segmentation (UVOS) algorithm [16] capable of tracking objects consistently while discovering new objects as needed. Leveraging this tracking capability, I devised a method to explicitly manage object-oriented memory in my work [17]. Specifically, I proposed two approaches to incorporate UVOS-based memory into the latent space. The first augments latent states by incorporating predictions for unobserved objects, while the second enhances input point clouds by transforming those of occluded objects using predicted poses. These approaches enable robots to perform challenging real-world tasks, including reasoning with occluded objects, novel objects appearance, and object reappearance. Experimental results demonstrated that the proposed methods outperform a baseline model with implicit memory, validating the effective memory capabilities of my framework.

Lifelong learning with failure cases. Robots often fail in out-of-distribution scenarios, especially in lifelong learning settings [18], where robots must continuously explore and adapt to new environments. To address this challenge, I aim to develop an approach that can detect failures, recover from failures, and learn to reduce future failures.

In ongoing work, I propose a method for detecting failures by evaluating predicted relations and recovering through re-planning. However, replanning alone cannot resolve all failures, especially those arising from errors in the dynamics prediction model. To mitigate this, I enhance the latent space dynamics model by incorporating real-to-sim transfer and generating additional simulation data. When a failure occurs in the real world, the approach creates an approximate simulation environment and generates targeted training data designed to maximize information gain. This iterative process improves the system's robustness and adaptability, ultimately reducing future failures. By continuously learning from failure cases, the proposed system aims to autonomously tackle increasingly complex real-world tasks throughout the robot's lifetime, as illustrated in Fig. 1.

III. Future Directions and Long-term Vision

First, I aim to integrate my latent space dynamics framework with policy learning methods (e.g., diffusion policy [19]), which have shown impressive capabilities in imitation learning tasks. These include predicting sequential actions for receding-horizon control, handling multi-modal action distributions, and maintaining robustness to environmental changes and perturbations. However, policy learning faces fundamental limitations, such as difficulty reasoning in partially observable environments and addressing geometric dependencies in constrained environments. To overcome these challenges, I propose a hierarchical system that interleaves latent space dynamics with a low-level diffusion policy. The high-level latent space dynamics model will reason about partial information and long-term geometric dependencies to generate parameterized robot skills, while the low-level diffusion policy will produce motor actions conditioned on these skills and current visual input. This system will enable the robot to perform long-horizon tasks in partially observable, geometrically constrained environments while remaining robust to environmental changes and disturbances. If a policy fails—for example, attempting to grasp an apple but missing—the high-level planner will detect the failure, replan the skill parameters, and guide the policy to recover.

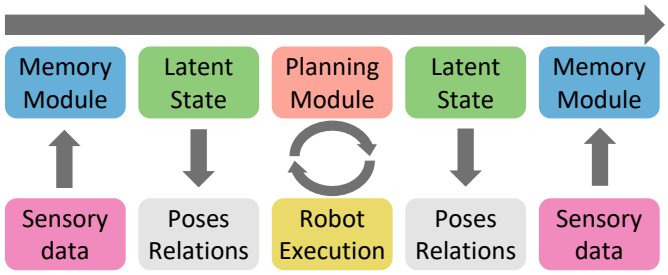

Fig. 2: Planning using a memory module with sensory data as input.

Second, I plan to explore the use of multi-modal inputs for reasoning about occluded objects in complex scenarios. Humans seamlessly integrate sensory inputs from vision, sound, touch, and memory to reason about occluded objects. Inspired by this, I aim to develop a framework that combines historical data as memory, vision for current observations, tactile sensors for touch, and microphones for sound. This framework will learn an integrated latent space that encapsulates information from these modalities, enabling robots not only to memorize the locations of occluded objects but also to update their memory dynamically using multi-modal inputs, as shown in Fig. 2. I believe that reasoning about occluded objects through multi-modal integration is a critical step toward developing robots as reliable daily assistants.

My research goal is to develop robots capable of processing multisensory inputs, encoding them into a latent space representation, performing spatial-temporal reasoning, and continuously adapting to new environments throughout their lifetimes.

To achieve this goal, my proposed approach conceptualizes the robot as an embodied agent [20] within a lifelong learning framework. Equipped with multisensory perception and an initial set of skills, the robot can plan to achieve language instructions through latent space planning. During deployment in uncertain, real-world environments, the robot will detect and correct failures, adapt its model to minimize future failures, and incrementally expand its capabilities. This approach is designed to be transferable across different embodiments, enabling robots to adapt to diverse applications and environments. I believe this framework represents a significant step toward understanding robots as embodied agents capable of reasoning in uncertain, real-world environments. By shifting the robotics community's focus away from optimizing isolated skills for specific tasks and embodiments, this work aims to inspire a new generation of research on versatile, adaptive, and embodied robotic agents.

References

[1] F. Ebert, C. Finn, S. Dasari, A. Xie, A. Lee, and S. Levine, "Visual foresight: Model-based deep reinforcement learning for vision-based robotic control," *arXiv preprint arXiv:1812.00568*, 2018.

[2] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2555–2565.

[3] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," *arXiv preprint arXiv:1912.01603*, 2019.

[4] P. Sundaresan, J. Wu, and D. Sadigh, "Learning sequential acquisition policies for robot-assisted feeding," in *7th Annual Conference on Robot Learning*, 2023. [Online]. Available: https://openreview.net/forum?id=o2wNSCTkq0

[5] Y. Li, S. Li, V. Sitzmann, P. Agrawal, and A. Torralba, "3d neural scene representations for visuomotor control," in *Conference on Robot Learning*. PMLR, 2022, pp. 112–123.

[6] J. Liang, M. Sharma, A. LaGrassa, S. Vats, S. Saxena, and O. Kroemer, "Search-Based Task Planning with Learned Skill Effect Models for Lifelong Robotic Manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2022. [Online]. Available: https://arxiv.org/abs/2109.08771

[7] C. R. Garrett, T. Lozano-Pérez, and L. P. Kaelbling, "Pddlstream: Integrating symbolic planners and blackbox samplers via optimistic adaptive planning," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 30, 2020, pp. 440–448. [Online]. Available: https://arxiv.org/abs/1802.08705

[8] ——, "Sample-based methods for factored task and motion planning." in *Robotics: Science and Systems*, 2017. [Online]. Available: https://dspace.mit.edu/bitstream/handle/1721.1/137701/garrett-rss17.pdf?sequence=2&isAllowed=y

[9] C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, and T. Lozano-Pérez, "Integrated task and motion planning," *Annual review of control, robotics, and autonomous systems*, vol. 4, pp. 265–293, 2021. [Online]. Available: https://arxiv.org/abs/2010.01083

[10] B. Kim, Z. Wang, L. P. Kaelbling, and T. Lozano-Pérez, "Learning to guide task and motion planning using score-space representation," *The International Journal of Robotics Research*, vol. 38, no. 7, pp. 793–812, 2019. [Online]. Available: https://arxiv.org/abs/1807.09962

[11] A. Curtis, X. Fang, L. P. Kaelbling, T. Lozano-Pérez, and C. R. Garrett, "Long-horizon manipulation of unknown objects via task and motion planning with estimated affordances," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 1940–1946.

[12] D. Driess, J.-S. Ha, and M. Toussaint, "Deep visual reasoning: Learning to predict action sequences for task and motion planning from an initial scene image," in *Proceedings of Robotics: Science and Systems*, 2020. [Online]. Available: https://arxiv.org/abs/2006.05398

[13] Y. Huang, A. Conkey, and T. Hermans, "Planning for Multi-Object Manipulation with Graph Neural Network Relational Classifiers," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. [Online]. Available: https://arxiv.org/abs/2209.11943

[14] Y. Huang, N. C. Taylor, A. Conkey, W. Liu, and T. Hermans, "Latent Space Planning for Multi-Object Manipulation with Environment-Aware Relational Classifiers," *IEEE Transactions on Robotics (T-RO)*, 2024. [Online]. Available: https://arxiv.org/pdf/2305.10857.pdf

[15] Y. Huang, C. Agia, J. Wu, T. Hermans, and J. Bohg, "Points2plans: From point clouds to long-horizon plans with composable relational dynamics," *arXiv preprint arXiv:2408.14769*, 2024.

[16] J. Yuan, J. Patravali, H. Nguyen, C. Kim, and L. Fuxin, "Maximal cliques on multi-frame proposal graph for unsupervised video object segmentation," *arXiv preprint arXiv:2301.12352*, 2023.

[17] Y. Huang, J. Yuan, C. Kim, P. Pradhan, B. Chen, L. Fuxin, and T. Hermans, "Out of Sight, Still in Mind: Reasoning and Planning about Unobserved Objects with Video Tracking Enabled Memory Models," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[18] S. Thrun and T. M. Mitchell, "Lifelong robot learning," *Robotics and autonomous systems*, vol. 15, no. 1-2, pp. 25–46, 1995.

[19] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[20] P. Calvo and A. Gomila, *Handbook of cognitive science: An embodied approach*. Elsevier, 2008.