# Research Statement

Yingchen Xu

---

My research aims to build AI agents capable of solving a wide variety of tasks in the physical world. Despite significant successes in specialized applications, such as reinforcement learning agents mastering complex games [6, 8] or precise robotic maneuvers [1], these specialist agents struggle significantly on tasks beyond their training domain. Meanwhile, large-scale multimodal models have demonstrated impressive generalization abilities in language and vision tasks, prompting efforts to translate these successes into embodied environments [7, 5]. However, existing approaches still face substantial challenges, particularly in handling complex, long-horizon tasks requiring *visually and physically grounded reasoning and planning* capabilities. My research addresses this gap through **world modeling**—developing robust internal representations of the physical environment—to enable efficient, adaptive planning and effective control in challenging real-world scenarios [3].

## Completed Research

**Learning Generalizable World Models.** In [9], we introduce a novel framework for scalable and diverse data collection to train world models, which utilizes a population of explorers guided by an information-theoretic objective to maximize trajectory diversity without task-specific rewards. This work tackles the critical challenge of how to efficiently gather the broad, task-agnostic data necessary to build world models that can generalize across a multitude of scenarios.

**Hierarchical World Models.** To address the complexities of long-horizon tasks and sparse rewards, we explore learning and leveraging hierarchical world models from offline trajectories [2]. We introduce an offline model-based RL algorithm that acts as a manager in a hierarchical structure, learning a temporally abstract model of the environment to predict "intent embeddings" or subgoals. These intents then guide a low-level worker policy, significantly improving performance on challenging benchmark tasks by enabling more effective long-term planning and reasoning.

**Generative Models for Continuous Control.** To enable sophisticated control in high-dimensional continuous systems like humanoids, we investigate the use of generative models as powerful world models for complex motor control [4]. Our method, H-GAP, is a trajectory generative model trained on extensive humanoid motion data, capable of zero-shot adaptation to novel downstream control tasks via model predictive control. By learning to represent and generate a wide array of motor behaviors, H-GAP showcases how large-scale generative world models can capture the nuances of complex physical interactions and provide a flexible foundation for general-purpose humanoid control, outperforming even baselines with access to ground-truth dynamics.

## Future Research Directions

**Multimodal Reasoning and Planning.** While current large multimodal models have demonstrated remarkable capabilities, their reasoning often remains heavily anchored in the textual domain. Meanwhile, many real-world embodied tasks necessitates planning and reasoning that are deeply grounded in the visual and physical aspects of the environment. A crucial future research direction, therefore, involves developing new modeling paradigms that enable these powerful multimodal models to "think" directly within the representations of the visual and physical world.

**Computationally Efficient World Models.** An effective world model should be computationally efficient at inference time; for instance, a high-resolution video generation model would be impractical due to its computational cost. A key direction I plan to explore is how principles from large-scale generative models can be adapted to create world models that are both powerful and lean. For instance, can we develop architectures that learn to dynamically allocate computational resources, perhaps by integrating attention mechanisms that selectively predict only crucial future states, key frames, or task-relevant regions within a scene, while strategically eliding redundant or irrelevant details? Beyond selective prediction, I will also investigate other approaches such as learning compact, factorized, or abstract latent representations of the world, and exploring event-driven or sparse prediction mechanisms that update the world state only when significant changes occur.

**A Unified Benchmark for Embodied Generalist Agents.** Many real-world embodied tasks require sophisticated integration of visual understanding, physical reasoning, planning abilities, and precise low-level control. Current benchmarks often test these abilities in isolation. I plan to develop a new collection of tasks to evaluate an agent's synergistic integration of (1) visual planning, by interpreting pixel-level sensory inputs from dynamic environments; (2) fine-grained motor control, enabling agents to execute complex action sequences given subgoals; and (3) understanding and following task instructions specified in natural language without additional abstract or structured guidance.

## References


[1] Forest Agostinelli et al. "Solving the Rubik's cube with deep reinforcement learning and search". In: *Nature Machine Intelligence* 1.8 (Aug. 1, 2019), pp. 356–363. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0070-z. URL: https://doi.org/10.1038/s42256-019-0070-z.

[2] Rohan Chitnis et al. *IQL-TD-MPC: Implicit Q-Learning for Hierarchical Model Predictive Control*. 2023. arXiv: 2306.00867 [cs.LG]. URL: https://arxiv.org/abs/2306.00867.

[3] David Ha and Jürgen Schmidhuber. "World Models". In: (2018). DOI: 10.5281/ZENODO.1207631. URL: https://zenodo.org/record/1207631.

[4] Zhengyao Jiang et al. *H-GAP: Humanoid Control with a Generalist Planner*. 2023. arXiv: 2312.02682 [cs.LG]. URL: https://arxiv.org/abs/2312.02682.

[5] Moo Jin Kim et al. *OpenVLA: An Open-Source Vision-Language-Action Model*. 2024. arXiv: 2406.09246 [cs.RO]. URL: https://arxiv.org/abs/2406.09246.

[6] David Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587 (Jan. 1, 2016), pp. 484–489. ISSN: 1476-4687. DOI: 10.1038/nature16961. URL: https://doi.org/10.1038/nature16961.

[7] Gemini Robotics Team et al. *Gemini Robotics: Bringing AI into the Physical World*. 2025. arXiv: 2503.20020 [cs.RO]. URL: https://arxiv.org/abs/2503.20020.

[8] Oriol Vinyals et al. "Grandmaster level in StarCraft II using multi-agent reinforcement learning". In: *Nature* 575.7782 (Nov. 1, 2019), pp. 350–354. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1724-z. URL: https://doi.org/10.1038/s41586-019-1724-z.

[9] Yingchen Xu et al. *Learning General World Models in a Handful of Reward-Free Deployments*. 2022. arXiv: 2210.12719 [cs.LG]. URL: https://arxiv.org/abs/2210.12719.