Research Statement

Yingchen Xu

My research is driven by the conviction that for AI agents to solve diverse tasks in the physical world, they must possess a robust World Model—an internal representation of the environment that allows them to simulate futures, reason about causality, and plan actions [2]. While specialized agents have mastered specific domains like games or single-task robotics, and large multimodal models have achieved broad passive understanding, we have yet to achieve true embodied generalists.

My work to date has focused on establishing the three pillars required to build these **generalist world models**: *Exploration* (how to get data), *Abstraction* (how to structure data), and *Control* (how to act on data).

- **Exploration (Data)**: A world model is only as good as its training data. In Learning General World Models, I addressed the challenge of gathering diverse data without task-specific rewards. We developed a population of explorers driven by information-theoretic objectives, demonstrating that population diversity and intrinsic motivation are keys to learning generalizable representations at a scalable fashion [4].
- **Hierarchy (Abstraction)**: Real-world tasks span long time horizons that flat models cannot handle. In my work on Hierarchical World Models, I introduced offline model-based algorithms that learn temporal abstractions. By predicting "intent embeddings" (subgoals) rather than immediate actions, we enabled agents to reason and plan over long horizons [1].
- **Generation (Control)**: Finally, a world model must enable precise interaction. With H-GAP, I demonstrated that generative models can serve as effective world models for high-dimensional continuous control, allowing zero-shot adaptation to novel tasks via model predictive control [3].

The Next Leap: Active World Models

While these pillars provide a strong foundation, I believe the current paradigm faces a fundamental bottleneck: we treat agents as *Passive Observers*. Current state-of-the-art vision systems (VLMs, video generators) consume massive, redundant grids of pixels to learn representations. Embodied agents with pre-trained vision models inherent these limitations. My current research passion is to bridge this gap by moving to **Active World Models**.

I posit that intelligence is not just about processing static data, but about actively deciding what data to process. Inspired by biological vision—which relies on foveation (high detail in tiny regions) and saccades (rapid movement) to construct a stable internal world—I aim to build agents that treat perception as a dimension of action. By restricting an agent's access to full frames and forcing it to actively query the environment, we compel the learning of robust abstractions like Object Permanence and Physics, ultimately enabling agents that are both more data-efficient and capable of deeper visual and physical reasoning.

Perception as Action. I propose a framework where the "World Model" separates the "eye" from the "brain."

- The Belief State: A latent world model summarizes a sequence of partial observations into an internal belief state.

- The Policy: The agent learns a policy π that outputs both control actions (locomotion/manipulation) and perception actions (where to look/saccade next).
- Adaptive Inference (Test-Time Compute): Unlike fixed-pass Vision Transformers, this framework supports test-time compute scaling. For simple scenes, the agent glances once; for complex reasoning tasks, the policy chooses to "look longer" and gather more context before acting.

Learning Progress as Intrinsic Motivation. To train these active perceivers in a self-supervised setting, I propose using Learning Progress (the rate of prediction improvement) rather than simple Uncertainty Reduction. Uncertainty reduction fails in stochastic environments (the "Noisy TV" problem), where agents get stuck staring at unlearnable noise because it maximizes entropy. Instead, We define the intrinsic reward r_t^{int} as the KL-divergence between the model's belief state before and after an update:

$$r_t^{int} = D_{KL}\big(P_{\theta'}(\cdot|x)\,||\,P_{\theta}(\cdot|x)\big)$$

This incentivizes the agent to focus on data that is learnable but not yet learned, theoretically aligning with Bayesian Surprise.

Concrete Research Ideas

- **Active Masking for Efficient Pre-training.** Many modern video models use random masking, which is computationally inefficient. I propose replacing random masking with a learned Active Masking Policy. By training a "saccading" mechanism to select patches that maximize the predictor's loss reduction, we can achieve a higher performance-to-compute ratio. This effectively introduces a computational attention mechanism during pre-training.
- **Embodied Active Exploration.** I am excited to test this framework on mobile agents in simulated environments. The goal is to demonstrate that an agent driven by *Learning Progress* can map a new environment and learn about environment transitions significantly faster than agents driven by alternative exploration strategies.
- **Visual Reasoning via Active Accumulation.** To address the lack of reasoning in current large vision models, I will focus on tasks that require gathering evidence over time (e.g., "Is the object in the drawer red?"). In this setting, the agent is given a "budget" of glimpses. This forces the model to learn hierarchical planning—first locating a landmark, then a container, then the target—proving that the World Model is performing causal reasoning rather than pattern matching

By treating perception as an active decision-making process, we can build agents that are not only more efficient but also capable of the deep physical understanding required for general-purpose robotics. My research will provide the theoretical and practical foundations for this next generation of Active World Models.

References

[1] 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.

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

[3] 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.

[4] 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.