ICCV 2025 Tutorial Time: 2025-10-20 Location: 306B

Foundation Models Meet Embodied Agents



Manling Li
Northwestern



Yunzhu Li Columbia



Jiayuan Mao Amazon FAR and UPenn



Wenlong Huang Stanford

















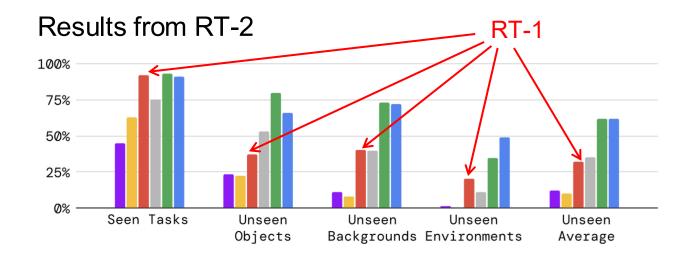


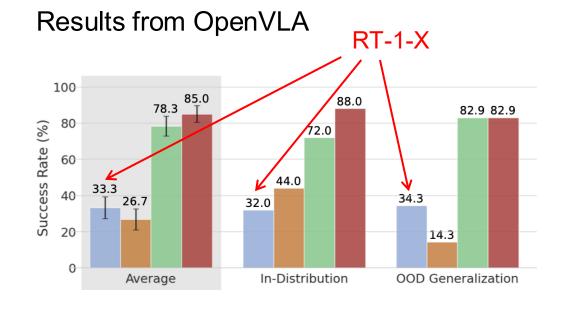


Inconsistent results across papers

Results from RT-1

Model	Seen Tasks	Unseen Tasks	Distractors	Backgrounds
Gato (Reed et al., 2022)	65	52	43	35
BC-Z (Jang et al., 2021)	72	19	47	41
BC-Z XL	56	43	23	35
RT-1 (ours)	97	76	83	59









- Evaluation is primarily conducted in the real world
 - Real-world evaluation is costly and noisy
 - "We have large enough budget such that we can still make progress."
 - Weak correlation between training loss and real-world success rate.
 - Training objectives vs task-specific metrics, training vs testing horizons







Recent efforts in real-world policy evaluation











- What about evaluation in simulation?
 - Sim-to-real gap: rigid / deformable / cloth
 - Efficient asset generation
 - Digitalization of the real world
 - Procedural generation of realistic and diverse scenes
 - Correlation between sim and real

ImageNet in Embodied AI?







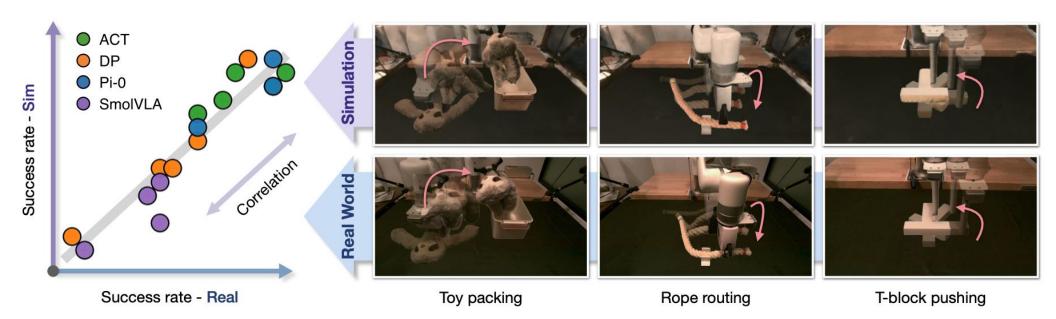
Habitat 3.0





- What about evaluation in simulation?
 - Sim-to-real gap: rigid / deformable / cloth
 - Efficient asset generation
 - Digitalization of the real world
 - Procedural generation of realistic and diverse scenes
 - Correlation between sim and real

ImageNet in Embodied AI?



Kaifeng Zhang*, Shuo Sha*, Hanxiao Jiang, Matthew Loper, Hyunjong Song, Guangyan Cai, Zhuo Xu, Hu Xiaochen, Changxi Zheng, **Yunzhu Li** Real-to-Sim Policy Evaluation with Gaussian Splatting Simulation of Soft-Body Interactions

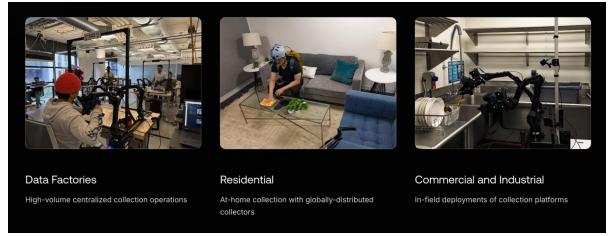


- What to collect? Where to collect? How much to collect?
 - If I have the budget to collect 10K demonstrations?
 - If I want the policy's performance to go from 80% to 99.99%?
 - Still unanswered research questions

Expanding Our Data Engine for Physical Al

by Ben Levin on September 24, 2025

Scale Al



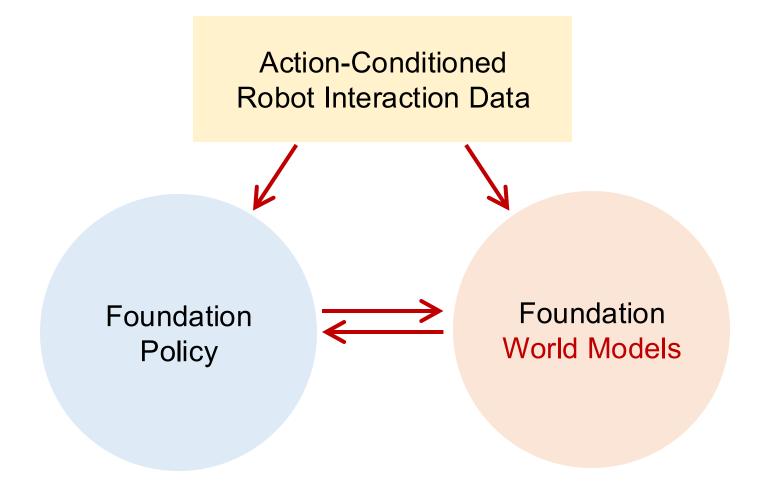
AgiBot-World



Foundation Policy → Foundation World Models



My definition of world models: action-conditioned future prediction



Foundation Policy -> Foundation World Models



My definition of world models: action-conditioned future prediction

SCIENCE ROBOTICS | REVIEW

MANIPULATION

A review of learning-based dynamics models for robotic manipulation

Bo Ai¹*, Stephen Tian², Haochen Shi², Yixuan Wang³, Tobias Pfaff⁴, Cheston Tan⁵, Henrik I. Christensen¹, Hao Su^{1,6}, Jiajun Wu², Yunzhu Li³*

Dynamics models that predict the effects of physical interactions are essential for planning and control in robotic manipulation. Although models based on physical principles often generalize well, they typically require full-state information, which can be difficult or impossible to extract from perception data in complex, real-world scenarios. Learning-based dynamics models provide an alternative by deriving state transition functions purely from perceived interaction data, enabling the capture of complex, hard-to-model factors and predictive uncertainty and accelerating simulations that are often too slow for real-time control. Recent successes in this field have demonstrated notable advancements in robot capabilities, including long-horizon manipulation of deformable objects, granular materials, and complex multiobject interactions such as stowing and packing. A crucial aspect of these investigations is the choice of state representation, which determines the inductive biases in the learning system for reduced-order modeling of scene dynamics. This article provides a timely and comprehensive review of current techniques and trade-offs in designing learned dynamics models, highlighting their role in advancing robot capabilities through integration with state estimation and control and identifying critical research gaps for future exploration.

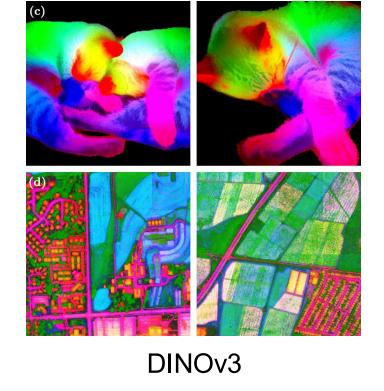
Foundation Models for Embodied Agents



- Current foundation models are not tailored for embodied agents
 - LLM/VLM can fail in embodied-related tasks
 - Limited understanding of geometric / embodied / physical interactions
 - □ Reinforcement learning (RL) from human feedback → RL from Embodied Feedback





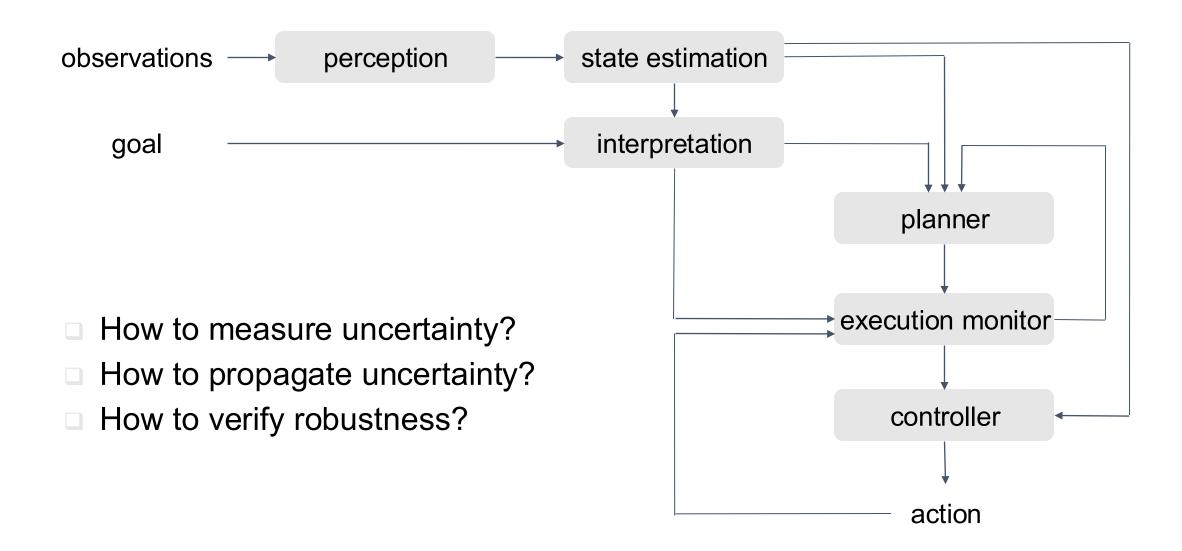


GPT

Segment Anything

Uncertainty Quantification & Propagation





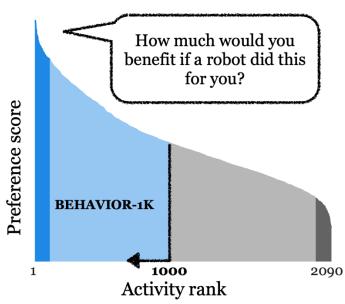
Adaptation / Life-Long Learning



- Adapt to new scenarios
- Adapt to human preferences
- Self improve / life-long learning



Adapt to new scenarios



Adapt to human preferences

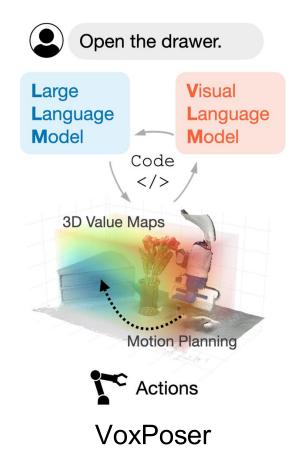


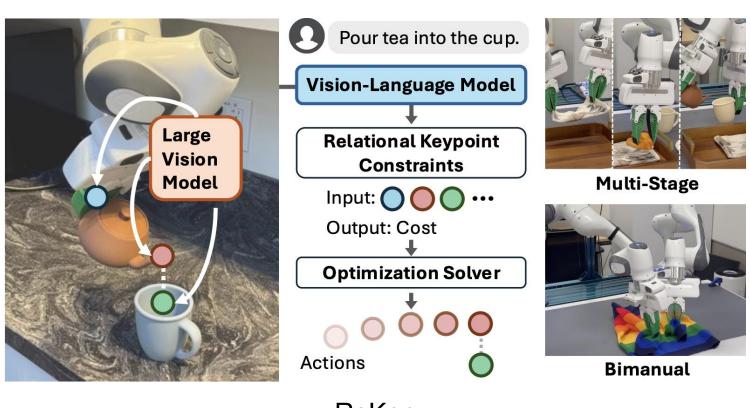
Improve through experience

Practical Considerations of Foundation Models



- Every robotics work is a system work
- System-level considerations:
 - delays / computing / modules talking to each other (high-level vs low-level)





Practical Considerations of Foundation Models



- Every robotics work is a system work
- System-level considerations:
 - delays / computing / modules talking to each other (high-level vs low-level)

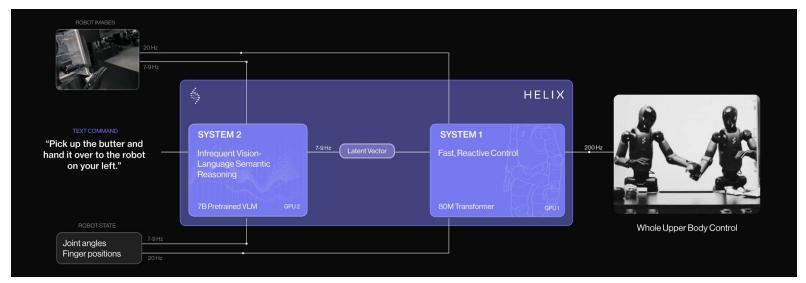
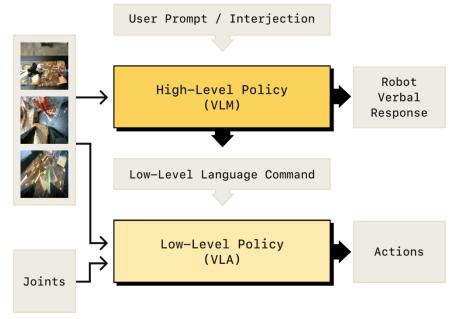


Figure AI: Helix



Physical Intelligence: Hi-Robot

Virtual Agent & Physical Agent



- Bring together researchers working on both virtual & physical agents
- Approaching the problem from a more structured lens: MDP
- Key techniques, emerging opportunities, and notable challenges



ICCV 2025 Tutorial Time: 2025-10-20 Location: 306B

Foundation Models Meet Embodied Agents



Manling Li
Northwestern



Yunzhu Li Columbia



Jiayuan Mao Amazon FAR and UPenn



Wenlong Huang Stanford







