Research Statement: Day-One Unboxing

Vishnu Sashank Dorbala

Background:

With the industry shipping out more and more robots to human-centric environments, (households, offices, hospitals, etc.) a central problem is that of generalization under partial priors. Unlike factory settings, human environments tend to be diverse, dynamic, and unstructured. A *deployable* robot agent must be able to adapt to these conditions to gain the trust of the end-user and show competence, right from *day-one*. I pose this as the **day-one unboxing** challenge, which asks the fundamental question:

> *"What minimum set of perception, reasoning, and interaction capabilities must an agent showcase right after unboxing to be deemed generally intelligent in a new environment?"*

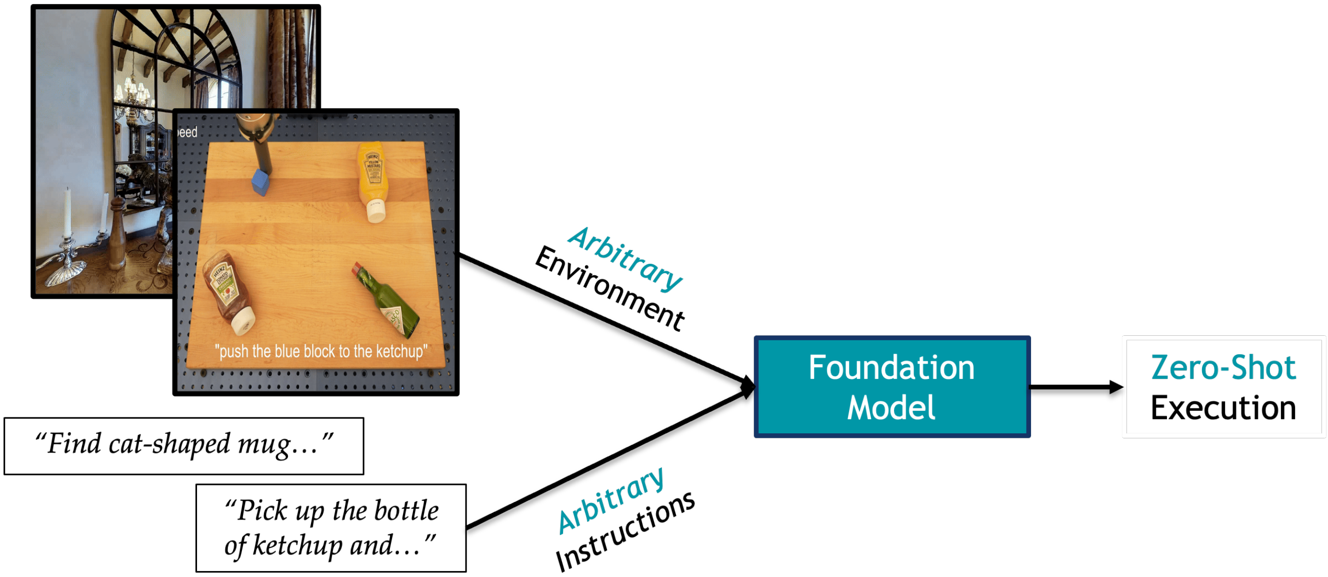
My PhD thesis attempts to address this question through the lens of foundation models (see figure below), and asks:

> *"How effective are large foundation models (VLMs, LLMs, Diffusion World Models) as priors for solving the Day-One Unboxing Challenge, and how can they be grounded using situated, multimodal context?"*

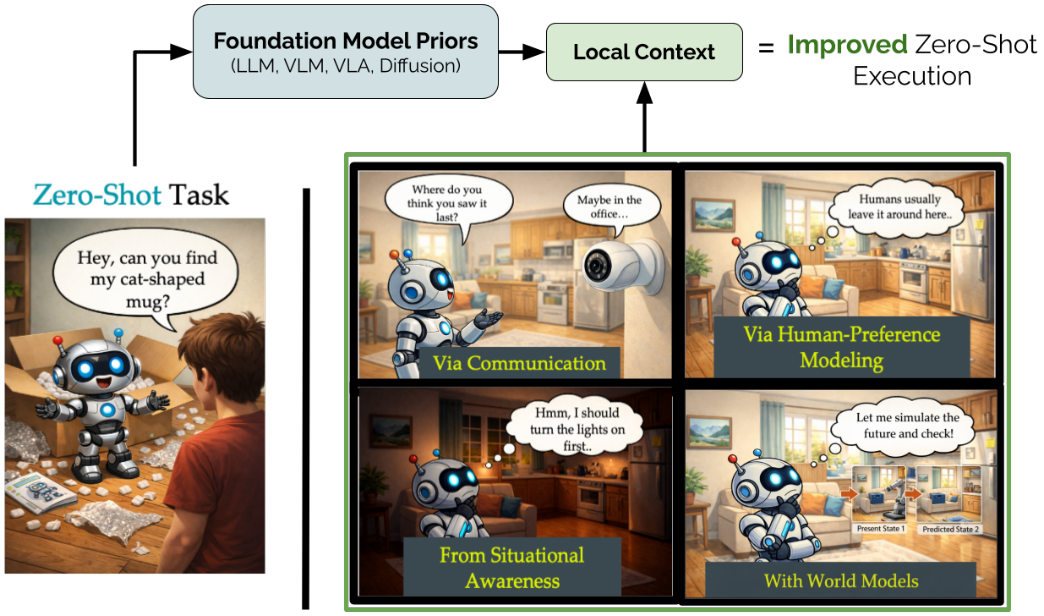
My research uncovers that large pre-trained foundation models show a poor physical understanding of the world, and are hence ineffective at zero-shot robot tasks. This necessitates research into novel methods to better extract information from the local environment to improve sequential decision making.

Current Interests:

I am deeply interested in developing techniques that improve the "zero-shot" performance of embodied robot agents. I view this as a problem of **generalizing under partial environmental knowledge**.



Summary of Prior Research: My PhD studied the zero-shot capabilities of Foundation Models (FM) for on embodied agents. This involved developing FM-driven methods, data synthesis tools and realistic tasks. One of the biggest challenges I tackled was to quantify the zero-shot capability of robot agents under partial knowledge, or what I like to call **day-one unboxing** problem.


Foundation Model Priors
(LLM, VLM, VLA, Diffusion)
Local Context
= Improved Zero-Shot Execution
Zero-Shot Task
Hey, can you find my cat-shaped mug?
Where do you think you saw it last?
Maybe in the office...
Via Communication
Humans usually leave it around here..
Via Human-Preference Modeling
Hmm, I should turn the lights on first..
From Situational Awareness
Let me simulate the future and check!
Present State 1
Predicted State 2
With World Models

Current Interests: I am interested in improving the zero-shot performance of foundation models on robot tasks. This involves incorporating with **local context** in different ways, including interactive communication, human preference modeling, future state prediction with world models and improved situational awareness.

Given the unpredictable nature of human-populated environments, I believe that training large foundation models (like VLA's) is only the first step. We require robust techniques that utilize these models only as "priors", requiring the robot agent to gather "local context" of the environment to improve zero-shot performance (See Figure above). I have identified three potential problems to work on as follows:-

1. **Memory Efficiency on Edge**: Foundation models, while powerful, rely heavily on the context given to them for decision-making. Further, their real-time deployment on edge poses many challenges with current hardware. Towards solving this, I am interested in developing efficient neural architectures that can work in conjunction with foundation models to improve both context efficiency and memory footprint. Beyond RAG, I would explore filtration techniques that capture and store only relevant observational data during exploration as context for sequential decision making. My previous work explores this direction via detachable memory *heads* to greatly improve the efficiency of low-parameter MLLMs.

2. **The Day-One Unboxing Problem**: I define this as the problem of determining the perceived intelligence of a robot agent on day-one of unboxing. A zero-shot problem by nature, I am interested in developing methods to evaluate and improve foundation model performance on these tasks. This would involve identifying what information from the environment would serve as local context to "prime" the foundation model, developing interactive methods to gather these multi-modal cues, as well as better modeling of human priors to determine how an agent can best adapt to a scene. My prior work on modeling human object-placement habits for personalized navigation is a step in this direction.

3. **World models for Zero-Shot Decision Making**: Embodied agents using foundation models make decisions as follows:

$$\text{Reasoning/Action Decision} = \mathcal{F}(\text{Pre-trained Foundation Model (Past Knowledge)}, \\ \text{Local Context (Current Knowledge)})$$

   World models *complete* this decision-making function by also providing a way to model future states. This would complete the temporal dependency of data (past, present and future) involved in sequential decision making. In this direction, I am eager to explore novel physics-based grounding losses to model realism. My current work on developing embodied world models is tangential to this goal, and going forward, I am very excited to explore ideas in geometric modeling and physically grounded simulators to improve the accuracy of future state predictions.