On Day-One Unboxing

Vishnu Sashank Dorbala

Abstract:

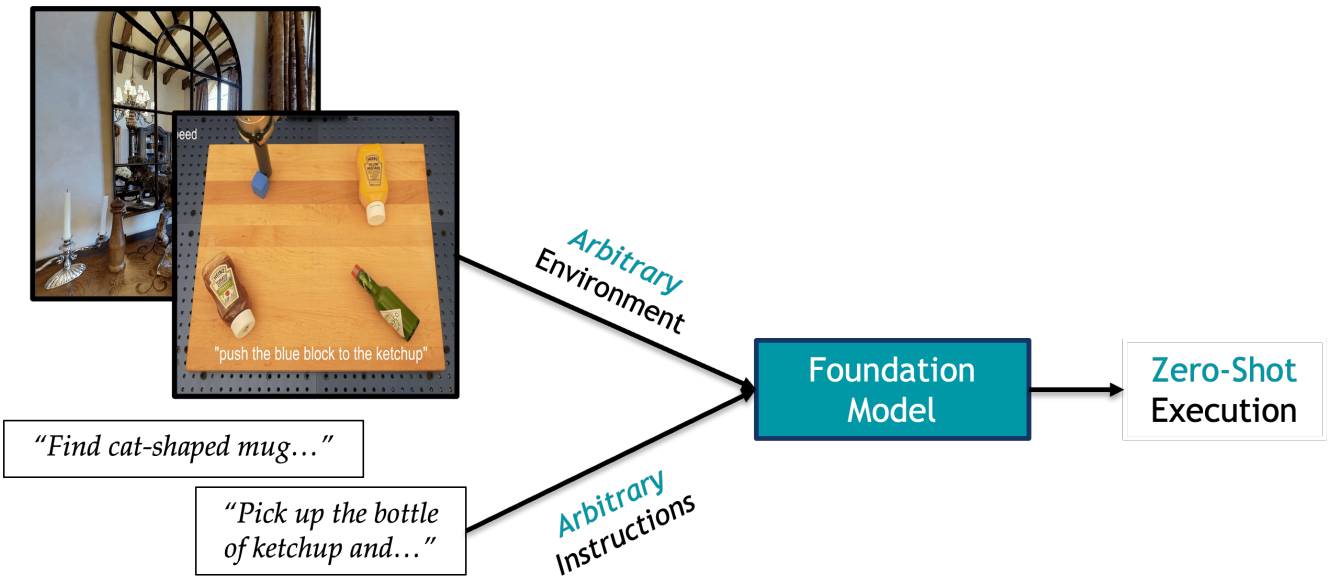
Data scarcity is a key issue in robotics and this has prevented learning-based robot models from generalizing to novel, unseen settings. Foundation Models (FM) like LLM/VLM/VLA's are trained on large, internet-scale data, but FM-based methods for robotics still fail to perform reliably during real-world deployment, and is reflective of the physical/real-world context needed to solve robotics tasks. To improve the ***deployability*** *of FM-based robot agents, I propose tackling this issue in two ways: 1) By developing novel memory architectures to enable better utilization of foundation models on robot devices, and 2) by building better representations of the physical world via world modeling, which can help improve the accuracy of sequential decision making. I provide a brief overview of my efforts towards this in the statement below.*

Motivation & Prior Work:

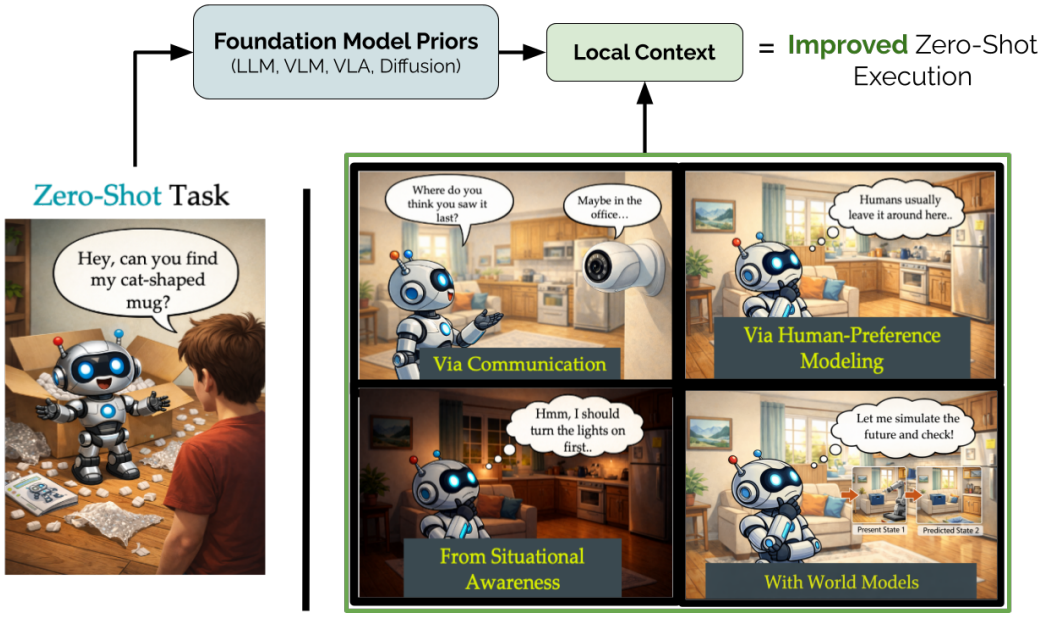
With the industry shipping out more and more robots to human-centric environments, (households, offices, hospitals, etc.) a central problem is that of generalization under partial priors. Unlike factory settings, human environments tend to be diverse, dynamic, and unstructured. A *deployable* robot agent 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?"*

Solving this challenge is a career goal of mine, and my PhD thesis attempts to address this question through the lens of foundation models (see figure below), and asks:



Summary of Prior Research: My PhD studied the zero-shot decision making capability of Foundation Models (FMs) on embodied robot tasks. I view this context of a **day-one unboxing** problem, where I seek to quantify the perceived intelligence of embodied agents on day-one of unboxing. For this, I developed various FM-driven methods ([7],[4]), data synthesis tools ([1],[2]), and realistic tasks ([6],[5],[2]) to benchmark their performance.



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 [6], human preference modeling [5], future state prediction with world models and improved situational awareness [2].

> *"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 showcase a poor understanding of the physical world, making them largely ineffective at solving zero-shot tasks. This necessitates research into novel methods to improve the decision making capacity of these models, including techniques to enrich them with *local context* of the environment.

Current Interests:

I am deeply interested in developing techniques that improve the "zero-shot" performance of embodied robot agents, and view this as a problem of **generalizing under partial priors**.
Given the unpredictable nature of human-populated environments, I believe that training large foundation models (like VLA's) is only the first step. To generalize, we require developing robust techniques that utilize these models only as "priors", allowing 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 recent work explores this direction via detachable memory *heads* to greatly improve the efficiency of low-parameter MLLMs [3].

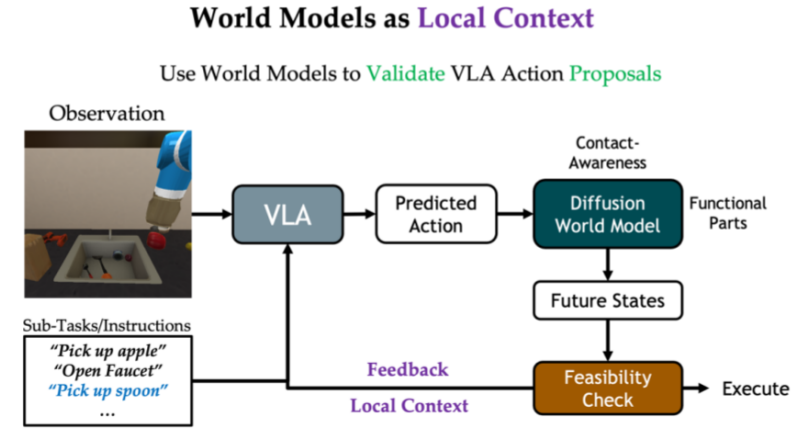
2. **The Day-One Unboxing Problem**: I define this as the problem of determining the perceived intelligence of a robot agent on day-one of unboxing. A zero-shot problem by nature, I am interested in developing methods to evaluate and improve foundation model performance on these tasks. This would involve 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 recent work on modeling human object-placement habits for personalized navigation [5] is a step in this direction.

3. **World models for Zero-Shot Decision Making**: Embodied agents using foundation models can be modeled as a Bayesian system as follows:

$$\text{Reasoning/Action Decision (Posterior)} = \mathcal{F}(\text{Pre-trained Foundation Model (Prior)}, \text{Local Context (Likelihood \& Evidence)})$$

Local context includes *world models* here, as a way to model future states, and not just past or current ones. Unlike prior work that looks only at past and current observations for decision making, the inclusion of world models would complete the temporal dependency of data (past, present and future) for better sequential decision making. The figures below highlights this pipeline in conjunction with VLA's.

Finetuning Diffusion World Models. In my experiments, I LORA-finetuned Stable Diffusion 2.1 with curated MP3D data to generate future images on a robot's trajectory assuming it went straight ahead from a particular viewpoint. While the visuals resemble the household in the sample, they are *temporally* inconsistent in generating the next viewpoint. I am interested in exploring scene graph reconstruction ideas in this direction, and improving world model performance by injecting **language details of the scene**.



Using World Models as Local Context. I propose a training scheme that jointly improves VLA and World Model performance via feedback from local context. A VLA predicts future actions that are validated using a diffusion world model modeled on contact-awareness, realism and functional-parts. The feasibility of these actions is then evaluated and fed back as local context (or *scene affordance*) to both the VLA and World Model as a joint training objective.

References

[1] Vishnu Sashank Dorbala, Sanjoy Chowdhury, and Dinesh Manocha. Can llm's generate human-like wayfinding instructions? towards platform-agnostic embodied instruction synthesis. In *NAACL (Short Papers)*, 2024.

[2] Vishnu Sashank Dorbala, Prasoon Goyal, Robinson Piramuthu, Michael Johnston, Reza Ghanadhan, and Dinesh Manocha. Is the house ready for sleeptime? generating and evaluating situational queries for embodied question answering. *arXiv preprint arXiv:2405.04732*, 2024.

[3] Vishnu Sashank Dorbala and Dinesh Manocha. Memctrl: Using mllms as active memory controllers on embodied agents. *arXiv preprint arXiv:2601.20831*, 2026.

[4] Vishnu Sashank Dorbala, James F. Mullen, and Dinesh Manocha. Can an embodied agent find your "cat-shaped mug"? llm-based zero-shot object navigation. *IEEE Robotics and Automation Letters*, 9(5):4083–4090, 2024.

[5] Vishnu Sashank Dorbala, Bhrij Patel, Amrit Singh Bedi, and Dinesh Manocha. Tas: A transit-aware strategy for embodied navigation with non-stationary targets, 2025.

[6] Vishnu Sashank Dorbala, Vishnu Dutt Sharma, Pratap Tokekar, and Dinesh Manocha. Improving zero-shot objectnav with generative communication. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12818–12825, 2025.

[7] Vishnu Sashank Dorbala, Gunnar A Sigurdsson, Jesse Thomason, Robinson Piramuthu, and Gaurav S Sukhatme. Clip-nav: Using clip for zero-shot vision-and-language navigation. In *Workshop on Language and Robotics at CoRL 2022*.