STATEMENT OF PURPOSE

Applicant: Puru Ojha **Target:** Ph.D. in Robotics

Embodiment-Aware Causal Verification: Grounding Long-Horizon Semantic Plans in Physical Reality

Modern robotics is currently fractured between two paradigms: end-to-end Vision–Language–Action (VLA) models that possess semantic common sense but often hallucinate physics, and classical Task and Motion Planning (TAMP) systems that are geometrically rigorous but rely on brittle, hand-engineered abstractions. My research goal is to bridge this divide to enable Generalizable Autonomy. I believe the solution lies in a synergistic architecture that decouples intent from execution: a **Generative Planning layer** that proposes a probabilistic distribution of semantic strategies, grounded by an **Embodiment-Aware Causal Verification** module. My PhD agenda focuses on formalizing this interaction, enabling robots to sample from manifolds of feasible plans rather than adhering to rigid symbolic tokens, thereby ensuring long-horizon reasoning adapts to the specific physical constraints of the robot and its environment.

Foundations in Spatial Interaction

In work published at CHI PLAY, I engineered a lightweight navigation protocol that replaced expensive ray-casting with graph-based relative orientation indices. This abstraction enabled the generation of infinite, non-Euclidean mazes on constrained hardware. My empirical validation revealed that users navigated these ”physically impossible” architectures without disorientation, proving that navigation relies on local topological coherence, not global metric consistency.

Research Trajectory: Semantic Planning vs. Embodied Reality

PHELPS: Long-Horizon Manipulation and Symbolic Bottlenecks

In PHELPS, I led the development of a long-horizon manipulation system from its inception. I co-defined the problem, designed the system architecture, integrated a VLM with a PDDL-inspired contract representation, modified RLBench to create composite tasks, evaluated multiple policies, and trained the primitive skills invoked in the contracts. I later implemented motion-planning controllers and coordinated the project with two collaborators.

The failures were more instructive than the successes. The VLM frequently **hallucinated affordances**; issuing semantically valid but physically impossible plans, such as commanding a ”reach and grasp” when the robot’s base was out of kinematic range. Our symbolic interface amplified this issue: As the predicates were static and open-loop, the planner lacked a mechanism to verify embodiment constraints. This experience crystallized my core research insight: treating the VLM as an isolated commander is insufficient. We need a **Synergistic Planning framework** where the high-level planner is grounded in a learned feasibility model, enabling it to propose trajectories that are not just semantically logical, but causally executable by the specific robot.

DCROSS: Cross-Embodiment Transfer and Representation Mismatch

In DCROSS, I aimed to achieve zero-shot policy transfer by designing a mechanism for deploying policies trained on a Franka Panda directly onto an XArm7 without fine-tuning. I started with engineered digital twins in ROS and Gazebo to benchmark foundation policies like OpenVLA and PiZero, but found that simulation-to-real gaps rendered them ineffective. To bridge this, I developed a hybrid pipeline in which a Generative Visual Adapter (using a bridge diffusion model to translate live XArm7 observations into the Franka Panda’s visual domain) coupled with kinematic retargeting to map the action space.

The failures were decisive. I found that foundation policies exhibit extreme sensitivity to visual domain shifts; even slight artifacts in the diffusion output caused catastrophic control failures, and kinematic retargeting failed to capture the dynamic nuances of the target hardware. When our XArm7 physically failed, I pivoted to collecting paired data on an XArm6 Lite. This constraint became a catalyst for insight: it demonstrated that trying to ”hallucinate” one robot’s pixels onto another is a dead end. Instead, I steered the project toward 3D Semantic Features and end-effector control spaces. This experience proved that generalizable autonomy requires Object-Centric 3D Priors that disentangle the task from the agent’s specific morphology, rather than brittle 2D pixel-space adaptations.

Proposed Research: Closing the Loop via Embodiment-Aware Verification

My research will focus on formalizing Embodiment-Aware Causal Verification, a framework to ground high-level semantic planning in low-level physical reality. This agenda consists of two synergistic components that form a ”Generative-Critical” architecture:

First, I aim to develop Probabilistic Object-Centric World Models that serve as the generative interface for planning. Moving beyond rigid VLM tokens or deterministic pixel-space predictions, these models will learn multimodal 3D latent state abstractions (e.g., object graphs or flow fields) that capture the stochastic nature of interaction. **By modeling a distribution of possible future states rather than a single outcome the representation allows planners to sample multiple feasible trajectories and reason about physical uncertainty**. Inspired by my findings in DCROSS, this approach disentangles ”task dynamics” from ”robot morphology,” creating a structured, generalizable space for robust decision-making.

Second, I aim to construct a **Latent Causal Verifier**—a lightweight ”Geometric and Physics Critic” for millisecond-scale feasibility checking. Instead of relying on computationally expensive rollout imagination. **This module will operate in the learned latent space to verify whether a sampled trajectory is executable by the specific robot’s embodiment.** By training this verifier on failures from the motion planner (building on my insights from PHELPS), we can create a closed-loop generative planner. This system filters the probabilistic proposals from the world model against the hard constraints of the hardware, enabling the robot to adapt long-horizon plans to unstructured environments without hallucinating physical capabilities.

Conclusion

The fracture between semantic hallucination and geometric reality is the primary bottleneck in modern robotics. My proposed research addresses this by replacing brittle symbolic tokens with probabilistic, causal representations that respect the constraints of embodiment. With experience ranging from writing low-level controllers to fine-tuning modern VLAs, from coming up with a Hybrid architecture to dissecting its bottlenecks and shortcomings, I possess the technical depth and research maturity required to execute this vision. I am eager to pursue a PhD where I can continue developing physically grounded, semantically aware embodied intelligence