PhD Research Statement

Applicant: Puru Ojha

Proposed Title: Hierarchical Embodied Intelligence: Unifying Semantic Reasoning and Differentiable Control

1. Problem Definition: The Case for Learned Hierarchical Abstraction

Complex intelligent behavior in nature does not emerge from a single, monolithic model; it emerges from a "Tower of Abstractions." Just as biology builds upon chemistry, and chemistry upon physics, natural intelligence relies on distinct hierarchical layers—from high-level semantic reasoning down to low-level motor control—that evolve independently but interlock coherently. My research is grounded in the observation that the current stagnation in general-purpose robotics stems from a failure to respect this fundamental structure.

Currently, the field attempts to solve embodied intelligence through two competing philosophies, both of which face asymptotic limits:

1. **The End-to-End Scaling Approach:** Models like RT-2 and VLA attempt to map high-level language directly to low-level motor tokens. Philosophically, this is akin to deriving biology directly from quantum mechanics; it ignores the necessary intermediate abstractions. Consequently, these models struggle with geometric precision, causal reasoning, and long-horizon error compounding.
2. **The Classical Hybrid Approach:** Systems like VLM-TAMP correctly acknowledge the need for hierarchy, effectively gluing a "semantic brain" (VLM) to a "geometric body" (TAMP). However, the interface between them is brittle and ungrounded. The VLM plans in 1D text space, unaware of physical affordances, while the planner executes rigid, hand-engineered primitives that cannot adapt to new environments.

The Scientific Gap: The Missing Grounded Representation The fundamental bottleneck preventing robust autonomy is not the capability of the VLM or the controller, but the **interface** between them. We lack a **Unified Probabilistic Representation** that is semantic enough for high-level reasoning yet geometric enough for physical execution. Current systems force a choice between hallucinating in text or failing in code; they lack a shared, learned belief state that captures object permanence, physical affordances, and uncertainty.

Research Proposal I propose to investigate the **structural hierarchy of embodied intelligence**. My hypothesis is that general-purpose autonomy requires learning a **Probabilistic Symbol Grounding** layer—a unified, object-centric world model that bridges the chasm between semantic intent and differentiable control. My research will move beyond rigid "contracts" or fixed skill libraries to explore how **learned, spatially-grounded abstractions** can enable a VLM to reason about geometry and a controller to execute semantic goals, effectively closing the loop on hierarchical robotic agency.

2. Literature Review: The Convergence of Learning and Structure

My proposal is grounded in the observation that robotics is currently split between two asymptotic limits: the semantic breadth of large-scale learning and the physical robustness of formal control. A new synthesis is emerging—driven by advances in World Modeling and Spatial Supersensing—that suggests the solution lies in **learned, probabilistic abstractions**.

2.1 The Limits of End-to-End Scaling While Vision-Language-Action (VLA) models like RT-2 have demonstrated emergent semantic reasoning, recent benchmarks reveal critical limitations in their physical understanding. As argued in *Cambrian-S*, current MLLMs excel at "Semantic Perception" (naming objects) but fail at "Implicit 3D Spatial Cognition" and "Predictive World Modeling" . Scaling context length alone does not solve this; truly general agents require "Spatial Supersensing"—the ability to maintain an internal, evolving model of the 3D world rather than just reacting to pixel patterns . My research addresses this by enforcing a structural bottleneck that compels the VLM to reason about geometry, not just semantics.

2.2 The Limits of Classical Hybrid Systems The current alternative, exemplified by VLM-TAMP and LLM-GROP, integrates VLMs with classical planners. While these systems provide long-horizon guarantees, they suffer from the "Discrete Skill Bottleneck." They rely on brittle, human-engineered predicates (e.g., `is_graspable`) that cannot adapt to novel environments. They treat the VLM as a text generator rather than a sensor, ignoring the "unconscious inference" required to handle physical uncertainty.

2.3 The "Middle Layer" Revolution: Spatial Reasoning & World Models Recent work has begun to dismantle the black box of end-to-end learning by introducing explicit intermediate representations—a "System 2" layer for robotics.

- **Probabilistic World Models:** Approaches like *Probabilistic Structure Integration (PSI)* and *Cambrian-S* demonstrate that general intelligence requires predicting future states and "intermediate structures" (e.g., optical flow, depth) rather than just reacting to current pixels.
- **Action Reasoning Models (ARMs):** A new class of models, exemplified by **MolmoAct**, argues that agents must "Reason in Space". MolmoAct explicitly generates **Visual Reasoning Traces** (spatial overlays) and **Depth Perception Tokens** before predicting actions, bridging the gap between semantic instruction and physical control.
- **Guided Diffusion:** Parallel work reinforces this hierarchy. **VLM-TDP** and **AnchorDP3** utilize VLMs to generate coarse 3D trajectories or "Affordance Anchors," which then guide a low-level diffusion policy. This confirms my hypothesis that the VLM's output should be a **geometric constraint** (a "Spatial Contract"), not a raw motor command

2.4 The Rise of Reflective Planning The most recent advance, **ReflectVLM**, introduces a "System 2" reflection mechanism where a VLM proposes actions and a diffusion model "imagines" the visual outcome to critique the plan. This successfully reduces errors in long-horizon tasks. However, ReflectVLM relies on **pixel-space hallucination**. This is computationally expensive (increasing inference time by ~20x) and prone to "physics

hallucinations" where the model generates visually plausible but physically impossible interactions (e.g., interpenetration).

2.5 The Remaining Gap: Closing the Loop While *MolmoAct* and *VLM-TDP* successfully generate spatial plans, they essentially act as "open-loop" reasoners. They generate a trace, and a policy follows it. We still lack a **differentiable feedback mechanism** where the downstream controller can critique the VLM's plan based on dynamic feasibility (e.g., "I cannot follow this trace because of torque limits"). My research proposes to close this loop, treating the "Spatial Contract" not just as an output, but as a probabilistic interface for bidirectional negotiation between reasoning and control.

---

3. Proposed Research: Hierarchical Action Reasoning Models (HARMs)

My central hypothesis is that robust robotic autonomy cannot be solved by "System 1" reflexes alone. It requires a "System 2" architecture where a **Probabilistic Spatial Contract** serves as the differentiable interface between high-level semantic reasoning and low-level kinematic control.

Recent works like **MolmoAct** [1] and **ReflectVLM** [2] have demonstrated the power of "Action Reasoning" and "Reflection." However, they leave critical gaps: *MolmoAct* is largely open-loop [3], and *ReflectVLM* relies on slow, pixel-space hallucination [4]. I propose to unify these concepts into a closed-loop system that validates spatial intent against physical constraints.

Thrust 1: Probabilistic Spatial Contracts

- **Research Question:** How can we represent high-level intent as a **spatial distribution** that captures uncertainty, rather than a deterministic command?
- **The Gap:** Current models like *MolmoAct* predict deterministic "Visual Reasoning Traces" (2D polylines) [5] or discrete "Depth Tokens" [6]. This forces the model to commit to a single solution too early, collapsing the probability distribution of valid actions.
- **Proposed Approach:** I will investigate **Probabilistic Spatial Contracts**. Instead of predicting a single trace, the VLM will output a **belief state** (e.g., a flow field or density map) representing the manifold of valid interactions (inspired by the probabilistic prediction in *PSI* [7]). This allows the downstream planner to sample feasible trajectories rather than struggling to follow a rigid, potentially erroneous line.

Thrust 2: Differentiable Geometric Verification (The "Critic")

- **Research Question:** How can we implement a "System 2" reflection loop that verifies kinematic feasibility *without* the computational cost of pixel-level video generation?
- **The Gap: ReflectVLM** introduces a "reflection" mechanism where a diffusion model imagines future video frames to critique a plan [8]. While effective for semantic logic, this

approach is computationally expensive (11s inference time vs 0.45s) [9] and prone to "physics hallucinations" where generated pixels look plausible but violate contact dynamics.

- **Proposed Approach:** I propose to replace *pixel imagination* with **Latent Geometric Verification**. I will explore training a lightweight "Physics Critic" that evaluates the *Spatial Contract* (Thrust 1) against learned kinematic constraints (e.g., joint limits, collision hulls) in a latent state-space. This moves verification from "Does the video look right?" (*ReflectVLM*) to "Is the trajectory physically executable?", enabling fast, gradient-based rejection of invalid plans.

Thrust 3: Self-Supervised Adaptation via Predictive Sensing

- **Research Question:** How can a robot autonomously refine its internal world model using "surprise" as a learning signal?
- **The Gap:** Truly general agents must adapt to new environments without human annotation. *Cambrian-S* posits that "Predictive Sensing"—using prediction error to drive memory—is essential for scaling beyond fixed contexts[101010].
- **Proposed Approach:** I will investigate a self-supervised loop where the robot continuously compares its *predicted* Spatial Contract against the *actual* executed trajectory. Large deviations (prediction errors) will be treated as "surprise signals" [11] to automatically curate training data, refining the VLM's understanding of physical dynamics in a "lifelong learning" setup.

---

4. Relevance and Suitability of Applicant's Background

My research agenda is not derived from abstract theory, but from the specific failure modes I observed while building state-of-the-art hierarchical and generative robotic systems. I have extensively engineered both the "top-down" planner and the "bottom-up" policy, identifying the exact disconnect that **Hierarchical Action Reasoning Models (HARMs)** aim to solve.

From PHELPS to Thrust 1 (The Representation Gap): In developing *PHELPS* (Planning Hierarchical Execution for Long Horizon Possibilities), I designed a system where a VLM generated symbolic contracts for a low-level executor. While the logic was sound, I observed that the VLM frequently "hallucinated" object states and spatial relations across frames because it lacked a persistent world model. The system failed not because of poor planning, but because the **deterministic contract** could not capture the VLM's perceptual uncertainty. This direct experience drives **Thrust 1**: the realization that the interface between reasoner and controller must be a **Probabilistic Spatial Contract** (similar to the belief states in *PSI*) rather than a brittle text command.

From DCROSS to Thrust 2 (The Control Gap): In my work on *DCROSS* (Diffusion-enabled Cross-Embodiment Policy Transfer), I trained and evaluated leading VLA architectures, including **OpenVLA** and **PiZero**. I found that despite their scale, these "end-to-end" models remained remarkably brittle to camera viewpoint shifts and environmental perturbations. They lacked a mechanism for "System 2" verification; if the diffusion model predicted a collision, there was no feedback loop to correct it. This highlighted the necessity of **Thrust 2**: we need a **Differentiable Geometric Critic** (improving upon the visual reflection in *ReflectVLM*) to verify that the "middle layer" plan is physically executable before the robot moves.

My background combines the systems engineering required to build *MolmoAct*-style pipelines with the generative modeling experience needed to implement *PSI*-style probabilistic inference. I am uniquely positioned to bridge these disciplines.

---

6. Conclusion

Current robotic foundation models face a computational dilemma: they must choose between the fast but hallucination-prone execution of end-to-end policies (*OpenVLA*) or the slow, pixel-heavy verification of reflective models (*ReflectVLM*). My research offers a third path.

By formalizing a **Probabilistic Spatial Contract**, we can decouple semantic reasoning from kinematic verification without breaking the differentiability of the system. This allows us to reject infeasible plans in the latent space—milliseconds before execution—rather than waiting for a slow video generation or a physical failure. This architecture does not just improve performance; it makes the "System 2" reasoning required for general autonomy computationally viable for real-world deployment.

---

7. References

Publications & Manuscripts in Preparation

- **Ojha, P.**, et al. "PHELPS: Planning Hierarchical Execution for Long Horizon Possibilities." *Unpublished Manuscript*. [1]
- **Ojha, P.**, et al. "DCROSS: Diffusion enabled Cross Embodiment Policy Transfer." *Unpublished Manuscript*. [2]

Key Literature

- **Abbatematteo, B.**, et al. (2024). "Composable Interaction Primitives: A Structured Policy Class for Efficiently Learning Sustained-Contact Manipulation Skills." *ICRA*. [3]

- **Black, K.**, et al. (2024). "π_0: A Vision-Language-Action Flow Model for General Robot Control." *arXiv preprint*. [4]
- **Brohan, A.**, et al. (2023). "RT-2: Vision-Language-Action Models Transfer Web-Scale Knowledge to Robotic Control." *arXiv preprint*. [5]
- **Chi, C.**, et al. (2023). "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion." *Robotics: Science and Systems (RSS)*. [6]
- **Bear, D.**, et al. (Stanford NeuroAI Lab). (2025). "World Modeling with Probabilistic Structure Integration (PSI)." *arXiv preprint*. [7]
- **Feng, Y.**, et al. (2025). "Reflective Planning: Vision-Language Models for Multi-Stage Long-Horizon Robotic Manipulation." *arXiv preprint*. [8]
- **Huang, K.**, et al. (2025). "VLM-TDP: VLM-guided Trajectory-conditioned Diffusion Policy." *arXiv preprint*. [9]
- **Lee, J.**, Duan, J., et al. (2025). "MolmoAct: Action Reasoning Models that can Reason in Space." *Allen Institute for AI Technical Report*. [10]
- **Tong, P.**, et al. (2025). "Cambrian-S: Towards Spatial Supersensing in Video." *arXiv preprint*. [11]
- **Wang, Y.**, et al. (2024). "GenDP: 3D Semantic Fields for Category-Level Generalizable Diffusion Policy." *arXiv preprint*. [12]
- **Wu, Z.**, et al. (2022). "SlotFormer: Unsupervised Visual Dynamics Simulation with Object-Centric Models." *ICLR*. [13]
- **Yang, Z.**, et al. (2024). "VLM-TAMP: Guiding Long-Horizon Task and Motion Planning with Vision Language Models." *arXiv preprint*. [14]
- **Zhao, Z.**, et al. (2025). "AnchorDP3: 3D Affordance Guided Sparse Diffusion Policy." *arXiv preprint*. [15]
