Embodied AI via Internal Planning in Implicit POMDPs

Li NING
Stellaris AI
liam@stellaris.ai

Abstract

Despite the success of Multimodal Large Language Models (MLLMs) in high-level reasoning, a persistent "execution gap" exists between symbolic planning and continuous physical interaction in unstructured environments. Conventional "Plan-then-Execute" paradigms are fundamentally limited by environmental non-stationarity and the lack of physical grounding. This paper addresses these challenges by formalizing the embodied interaction as an **Implicit Partially Observable Markov Decision Process (Implicit POMDP)**. In this framework, state transitions and multimodal observations are governed by high-dimensional latent dynamics rather than explicit analytic models.

We introduce **Internal Planning**, a novel decision-making paradigm where the agent utilizes a **Latent Belief Transformer** to perform continuous mental rollouts, thereby anchoring semantic intent to physically executable motor primitives. To ensure generalizability, we propose a **Hierarchical Task Basis Space** ($\mathcal{V}_T$), which treats embodied tasks as composable vectors within a functional manifold, enabling zero-shot task synthesis via semantic interpolation. Furthermore, we establish the **Hierarchical Ability Evaluation (HAE)** framework—a 6-level metric that quantifies an agent's resilience against counterfactual dynamics and open-world complexity.

1 Introduction

1.1 The Evolution of Embodied Paradigms

The trajectory of Embodied Artificial Intelligence (EAI) has undergone a fundamental transformation, moving from rigid, rule-based systems to fluid, neural-generative architectures. Historically, the field was dominated by **Classical Control Theory**, which relied on explicit analytic models and potential field methods to ensure stability and obstacle avoidance. As computational power scaled, the community transitioned into the era of **Modular Deep Learning**, characterized by disjoint pipelines for visual perception, SLAM (Simultaneous Localization and Mapping), and motion planning. While these systems achieved success in controlled settings, they suffered from "cascading failures", where minor errors in perception led to catastrophic failures in execution.

We are currently witnessing the third wave: **Foundation Model-Driven Embodied AI**. The integration of Multimodal Large Language Models (MLLMs) has provided agents with an unprecedented ability to ground natural language instructions into semantic understanding of the physical world. However, despite the impressive reasoning capabilities of these models, a critical "execution gap" remains between high-level cognition and low-level physical interaction.

1.2 The Brittle Nature of "Plan-then-Execute"

The prevailing methodology in current MLLM-based robotics is the **Plan-then-Execute (PTE)** paradigm. In this scheme, an agent functions as a high-level sequencer, generating a complete

chain of symbolic or linguistic sub-goals before handing them off to a low-level controller. This proposal identifies two fatal flaws in the PTE approach:

1. **Dynamic Fragility:** PTE is inherently open-loop. It assumes that the environment remains stationary during the planning-to-execution window. In unstructured real-world scenarios—where humans move, lighting changes, or objects slip—a pre-generated plan becomes a liability rather than an asset.

2. **The Semantic-Physical Mismatch:** Symbolic plans generated by LLMs often violate the underlying "affordances" of the robot's hardware. A plan may be linguistically logical (e.g., "pick up the table") but physically impossible due to torque limits, reachability constraints, or occluded geometry.

1.3 Defining the Implicit POMDP

Standard Partially Observable Markov Decision Processes (POMDPs) provide a mathematical framework for uncertainty, yet they typically assume an explicit state space where the agent knows *what* variables it is missing (e.g., its coordinates on a map). We argue that the physical world is fundamentally **Implicit**. Many variables essential for successful manipulation—such as the friction coefficient of a surface, the center of mass of a grasped tool, or the fluid dynamics of a pouring task—are never directly observed.

In an **Implicit POMDP**, the state transition $\mathcal{F}_T$ and observation operator $\mathcal{F}_O$ are unknown and non-stationary. This requires the agent to maintain a latent belief state that is constantly refined through active interaction, rather than relying on a static world model.

1.4 Proposed Innovation: Internal Planning and Task Basis Space

To overcome these limitations, this research proposes the **Internal Planning** paradigm. Rather than treating planning as a preliminary step, we view it as a continuous, internal reasoning process. By utilizing MLLMs as learned world models, the agent performs "Mental Simulations" in a latent sandbox, predicting the physical consequences of its actions before committing to them.

Furthermore, we introduce the **Hierarchical Task Basis Space** and **Hierarchical Ability Evaluation (HAE)**. By treating tasks as composable basis vectors in a functional manifold, we move away from binary "success/failure" metrics toward a vectorized understanding of agent intelligence. This allows for a rigorous, level-based quantification of an agent's ability to handle increasing degrees of environmental implicitness and task complexity.

2 Formal Problem Definition: The Implicit POMDP

2.1 Mathematical Framework of $\mathcal{M}_i$

The classical Partially Observable Markov Decision Process (POMDP) assumes an explicit state space and known observation probabilities. However, real-world embodied interaction occurs in a "black-box" physical environment. We formalize this as an **Implicit POMDP** ($\mathcal{M}_i$), defined by the 6-tuple:

$$\mathcal{M}_i = (\mathbb{S}, \mathbb{A}, \mathbb{O}, \Psi, \Omega, \mathcal{R})$$

Unlike the analytic models common in traditional robotics, our framework acknowledges the following constraints:

- **Transition Operator (Ψ):** $s_{t+1} = \Psi(s_t, a_t)$. In our formulation, Ψ represents high-dimensional, non-linear physical dynamics (e.g., contact forces, fluid flow) that are learned as a latent world model rather than explicitly programmed.

- **Observation Operator (Ω):** $o_t = \Omega(s_t)$. This function maps the underlying physical state to multimodal observations (pixel data, depth maps, and tactile feedback), which are inherently noisy and high-dimensional.
- **Implicit Reward ($\mathcal{R}$):** Instead of a scalar reward, we consider $\mathcal{R}$ to be a semantic success condition defined in natural language, requiring a cross-modal alignment between the latent state and the linguistic goal $\mathcal{G}$.

2.2 Latent Belief State and Sufficiency

The agent lacks direct access to the physical state s_t. It must therefore construct a **Latent Belief State** z_t, which serves as a compressed representation of the history $h_t = \{o_0, a_0, \ldots, a_{t-1}, o_t\}$. Following the principles of Variational Information Bottlenecks, we define:

$$z_t = \phi_\theta(z_{t-1}, a_{t-1}, o_t) \tag{1}$$

where ϕ_θ is a learned encoder (often a Transformer-based architecture). To be effective for Internal Planning, z_t must capture **Actionable Affordances**—latent features like object mass or surface friction—that are not explicitly present in the pixels but are essential for predicting future states.

2.3 The Decision Objective

The objective of the embodied agent is to optimize a policy π that maximizes the expected return in this implicit space:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \mathcal{R}(z_t, a_t) \mid z_t \approx \mathbb{S}\right] \tag{2}$$

This objective function necessitates that the agent not only selects the best action but also maintains a high-fidelity "mental simulation" to minimize *surprise* (the divergence between the predicted latent state and the actual observation).

3 Methodology: Internal Planning and Action Anchoring

3.1 The Latent Belief Transformer (LBT)

Internal Planning moves beyond the "Plan-then-Execute" bottleneck by treating the Multimodal Large Language Model (MLLM) not as a text generator, but as a **Latent Belief Transformer**. Existing models often decouple linguistic reasoning from physical state estimation. In contrast, the LBT architecture integrates multimodal observations o_t directly into the hidden state space of the transformer, where the self-attention mechanism performs temporal filtering over the interaction history h_t.

The LBT maintains a latent state z_t that encodes both semantic intent and physical affordances. By operating in the latent space, the model avoids the computational overhead of high-resolution image reconstruction while retaining the capacity to predict future states $\hat{z}_{t+1\ldots t+H}$ under varying action sequences.

3.2 Action Anchoring via Physical Grounding

A fundamental challenge in MLLM-driven embodied AI is the "hallucination of capability"—generating plans that are kinematically or dynamically impossible. We introduce **Action Anchoring** as a closed-loop grounding mechanism. Unlike traditional policy heads that map features to actions in a single pass, Action Anchoring treats the MLLM's internal output α_t as a *semantic*

intention which is then projected onto the robot's physical constraint manifold $\mathcal{C}$. This process involves an iterative refinement loop where the internal plan $\mathcal{P}_{latent}$ is checked against a local stability filter. Action Anchoring ensures that high-level semantic intentions are grounded in the robot's physical constraint manifold, facilitating generalizable manipulation across diverse hardware morphologies.

This process involves an iterative refinement loop where the internal plan $\mathcal{P}_{latent}$ is checked against a local stability filter. If the "mental rollout" predicts a collision or a violation of joint limits, the anchor forces a re-simulation within the LBT, ensuring that the final output a_t is both semantically optimal and physically executable.

3.3 The Planning Algorithm

The following algorithm details the synchronization between the high-level latent reasoning and the low-level anchoring process.

Algorithm 1 Internal Planning & Action Anchoring

Initialize: Latent Belief $\mathcal{B}_0$ (LBT State), Task Goal $\mathcal{G}$ (Natural Language)
while Task Goal $\mathcal{G}$ not satisfied **do**
  $o_t \leftarrow$ Synchronize Multimodal Stream (Vision v_t, Tactile f_t, Proprioception q_t)
  $\mathcal{B}_t \leftarrow$ LBT-Update$(\mathcal{B}_{t-1}, o_t, a_{t-1})$ {Posterior belief update}
  // Mental Sandbox Simulation
  $\mathcal{P}_{latent} \leftarrow \{\hat{z}_{t+k}, \hat{r}_{t+k}\}_{k=1}^{H} \sim$ LBT-Rollout$(\mathcal{B}_t, \mathcal{G})$ {Sample H latent trajectories}
  // Action Derivation and Anchoring
  $\alpha_t \leftarrow \text{argmax}_a \sum_{k=1}^{H} \gamma^k \hat{r}_{t+k}$ {Extract best latent intention}
  $a_t \leftarrow$ Anchor$(\alpha_t, \text{Constraints } \mathcal{C}, \text{Jacobian } \mathbf{J})$ {Map intention to motor torques/velocities}
  Execute a_t and capture transition feedback
  if a_t execution failed **then** Update $\mathcal{C}$ and Trigger Internal Reset
end while

4 Task Basis Space: A Vectorized Approach

4.1 Background: The Compositional Nature of Skill

The challenge of General Embodied Intelligence lies in the infinite variety of potential tasks. Traditional approaches often treat each task as a discrete entity or a unique entry in a PDDL (Planning Domain Definition Language) library. However, human motor control suggests that complex behaviors are synthesized from a finite set of **Dynamic Movement Primitives (DMPs)**.

We extend this concept by proposing the **Task Basis Space** $\mathcal{V}_T$—a functional manifold where the dimensions are not defined by coordinates, but by **Actionable Semantics**. In this space, an embodied agent does not learn a specific "fridge opening" task; rather, it learns the basis vectors of *Reach*, *Grasp*, and *Torque Application*, which can then be vectorized to solve novel manipulation problems.

4.2 Formal Composition and Temporal Dynamics

We define the Task Basis Space $\mathcal{V}_T$ such that any complex embodied task $\mathcal{T}$ can be represented as a weighted combination of basis vectors $\vec{b}_i$. These vectors represent fundamental skills anchored in physical constants (e.g., maintaining the Zero-Moment Point for balance). The decomposition

is defined as:

$$\mathcal{T} = \sum_{i=1}^{n} \omega_i \vec{b}_i + \int_0^T \delta(t) dt \tag{3}$$

where:

- $\vec{b}_i \in \mathcal{V}_T$ are the **orthonormal basis vectors** of the skill space, learned via semantic embedding of motor trajectories.
- ω_i are **composition weights** derived from the MLLM's internal reasoning, reflecting the importance of each primitive to the global goal.
- $\delta(t)$ is the **temporal residual function**, representing the non-linear adjustments required to handle the "implicitness" of the environment (e.g., compensating for unexpected surface friction or wind resistance).

4.3 Generalization via Vector Interpolation

The vectorized approach allows agents to achieve zero-shot generalization through **semantic interpolation**. By representing tasks in a continuous manifold, the agent can solve a "Level 5" open-world task by identifying the closest existing basis components and synthesizing a new vector. This provides a mathematical explanation for how MLLMs can transfer "knowledge of physics" from text to the "execution of physics" in a robot: the MLLM acts as the coefficient generator ω_i for the underlying physical basis $\vec{b}_i$.

5 Hierarchical Ability Evaluation (HAE)

5.1 Beyond the Success Rate: Measuring Intelligence

Current evaluation metrics in Embodied AI primarily focus on the Task Success Rate (SR) within static or narrow-distribution benchmarks. However, SR is an insufficient proxy for general intelligence, as it fails to distinguish between *memorized motion sequences* and *adaptive reasoning*. Drawing inspiration from the "Measure of Intelligence" framework, we propose the **Hierarchical Ability Evaluation (HAE)**.

HAE quantifies an agent's capability based on its efficiency in navigating the "Implicitness Spectrum" of our POMDP model. It measures how effectively an agent utilizes its Task Basis Space $\mathcal{V}_T$ to minimize the surprise between internal rollouts and environmental feedback.

5.2 The 6-Level Hierarchy of Embodied Intelligence

We define a standardized roadmap for evaluating agents, spanning from atomic execution to open-world synthesis:

- **Level 0–2 (Foundational Manipulation):** Assessment of kinematic reachability, multimodal grounding (e.g., "pick the red cube"), and short-horizon sequential logic. These levels represent the current capability of most SOTA Vision-Language-Action (VLA) models.
- **Level 3 (Temporal Reasoning):** Testing the agent's ability to handle long-horizon tasks with sparse rewards, requiring the maintenance of a consistent latent belief state over time.
- **Level 4 (Counterfactual Dynamics):** This level tests **Adaptive Resilience**. The agent must adapt when implicit physical properties are intentionally decoupled from visual

appearances. *Example:* A visually "heavy" steel box is emptied to be feather-light. A Level 4 agent must detect the haptic discrepancy during the "Action Anchoring" phase and re-simulate its internal plan to prevent an over-torqued lift.

- **Level 5 (Open-World Synthesis):** The pinnacle of the HAE, testing the agent's ability to perform **Zero-Shot Composition**. The agent is given a "Level 0" command (e.g., "Tidy up") in a "Level 4" environment (unknown dynamics and unstructured noise). Success requires the agent to decompose the goal into novel combinations of basis vectors $\vec{b}_i$ without prior training on that specific scene.

5.3 HAE Scoring Metric

The HAE score $\mathcal{S}$ is calculated as a weighted product of the **Decomposition Accuracy** (Λ) and the **Anchoring Robustness** (Γ):

$$\mathcal{S}_{HAE} = \sum_{L=0}^{5} w_L \cdot (\Lambda_L \times \Gamma_L) \tag{4}$$

where w_L increases exponentially with the level. This ensures that an agent capable of Level 5 reasoning is ranked significantly higher than a Level 2 specialist, even if both share the same success rate on simple tasks.

6 Scientific and Broader Impact

6.1 Standardizing the Bridge Between LLMs and Physical Actuation

The primary scientific contribution of this framework is the formalization of the transition from "Chatbots" to "World-Reasoning Engines". By introducing the **Internal Planning** paradigm within an **Implicit POMDP** framework, we provide a principled methodology for grounding abstract linguistic reasoning into concrete physical dynamics. This work addresses the critical bottleneck of "hallucinated affordances" in current multimodal models, paving the way for more reliable and autonomous human-robot interaction.

6.2 Broader Societal and Industrial Impact

The implications of a robust, hierarchical task basis space extend beyond academia. In **industrial automation**, the ability to perform zero-shot composition of skills could reduce the downtime required to re-program factory robots for new product lines. In **service robotics**, Level 4 resilience (Counterfactual Dynamics) is essential for safety in human-centric environments, such as eldercare or domestic assistance, where the agent must adapt to unpredictable physical shifts.

Ultimately, this research serves as a blueprint for the next generation of general-purpose embodied agents. By standardizing the evaluation of "Action Anchoring" and "Internal Planning", we aims to lead the industry toward a future where robots possess a genuine, verifiable understanding of the physical world they inhabit.