Embodied Phaneron: A Single-Substrate Architecture for Perception, Affordances, and Safe Skill Composition

Anonymous Preprint

November 15, 2025

Abstract

We present an embodied intelligence architecture that uses a single unlabeled structural substrate—the *Phaneron*—to unify perception, memory, planning, and control. Raw sensor streams arrive as distinctions; local adjacencies become connections; reusable subgraphs (objects, affordances, skills, maps) are induced as patterns when they jointly improve compression and prediction (MDL + predictive gain). Actions are relation motifs that transform the graph via double-pushout rewrites; discourse-like versioning provides replay, counterfactuals, and safety audit. A risk/curiosity objective biases exploration, and a scheduler trades reflex stability against global consolidation. We demonstrate how new sensors cause a transient compute spike that collapses as the dictionary consolidates, and we outline evaluations on manipulation and navigation that test invariance across senses, O.O.D. robustness, and explainable safety.

1 Introduction

Modern robot stacks are mosaics of separate modules. We advocate a single substrate for semantics and memory: a modality-agnostic graph where every nuance—roles, direction, time, goals, safety—lives as *patterns*, not labels. This yields long-range coherence, explicit explanations, and principled cross-modal transfer.

2 From D-C-A to Embodiment

We instantiate Distinction → Connection → Abstraction with an explicit *Consolidation* phase as the runtime: raw samples → local links → induced patterns → global merges/refactors. The substrate stays undirected; asymmetries (cause/effect, before/after) live inside relation motifs (A–R–B). A high-centrality *self* pattern anchors safety and long-horizon consistency. Affect appears as aversive and epistemic fields that bias where cycles are spent.

3 System Architecture

3.1 Representation and actions

Distinctions: time-stamped samples or small spatiotemporal patches. Connections: adjacency and sync. Patterns: parts, objects, scenes, affordances, skills, maps, and operator motifs. An action is a pattern placement R_{skill} connecting {agent, site, object, goal}; effects are versioned events. Direction lives in R roles; base graph remains undirected.

Figure 1: **Embodied pipeline.** Sensors → ingest shims → Phaneron → affect fields → CDA planner → reflex control → actuators.

Figure 2: **Runtime loops.** Reflex (kHz), Semantic (Hz), Planning (Hz), and opportunistic Consolidation.

3.2 Objective and admissible rewrites

We minimize a hybrid objective

$$\Delta\mathcal{L} = \Delta\mathcal{L}_{\text{MDL}} + \lambda_{\text{pred}}\Delta\mathcal{L}_{\text{pred}} + \lambda_{\text{goal}}\Delta\mathcal{L}_{\text{goal}} + \lambda_{\text{fear}}\Phi_{\text{aversion}} - \lambda_{\text{cur}}\Phi_{\text{curiosity}}. \tag{1}$$

A rewrite is admissible if the total is strictly negative under hard safety constraints (force/torque envelopes, keep-out zones).

4 Scheduler: the Cognitive Tradeoff policy

Budget cycles among reflex (stability/safety), semantic (pattern placement), planning, and consolidation. Under risk, prioritize reflex/short horizon; under calm, invest in consolidation and dictionary growth.

5 ROS2 Integration (where shims end)

Minimal shims parse drivers into distinctions and sync links. No semantic adapters: semantics emerge as patterns placed by CDA.

6 Replay and Offline Consolidation

Idle windows run replay: interleave episodes, propose merges/rewrites, validate by prediction and compression, keep/refine/discard, update dictionary.

7 Health and Debt Telemetry

We expose two KPIs: `semantic_fps` (D-C-A cycles/sec at fixed load; higher is better) and `debt_index` (pending merges + conflicts + stale placements; lower is better). The scheduler maintains both within target bands; consolidation windows reduce debt to recover semantic throughput.

8 New Sensor Onboarding: spike -> collapse

A new modality increases update cost; as patterns consolidate and cross-modal merges form, cost collapses and prediction improves. We report time-to-parity and area-under-spike.

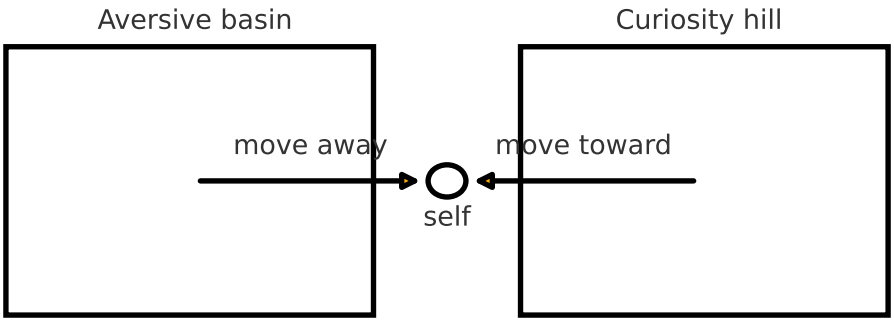

Figure 3: **Affect fields.** Aversive basin repels dangerous futures; curiosity hill attracts probes with high expected information gain.

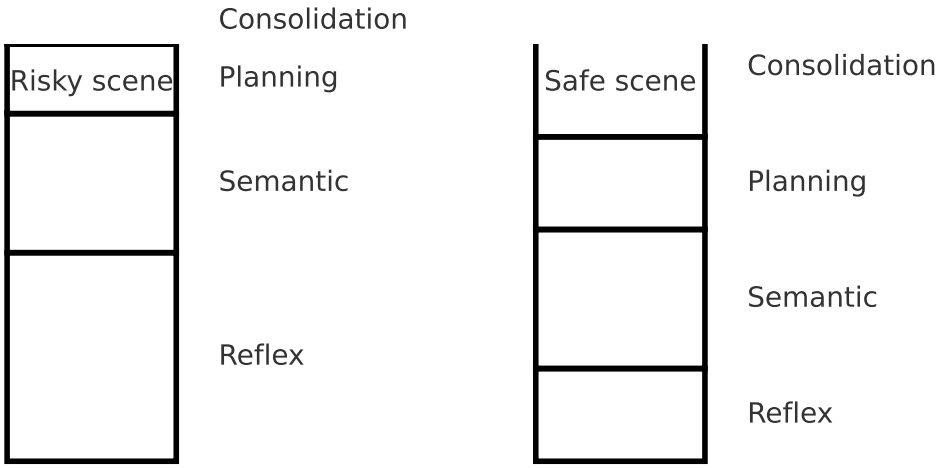

Figure 4: **CTH scheduler.** Risky scenes bias budget toward reflex; safe scenes release budget to consolidation.

9 Safety and Explainability

Hard constraints are subgraph detectors that veto unsafe rewrites. Every decision attaches a Minimal Explanation Subgraph (MES) justifying acceptance or veto.

10 Experiment Plan and Ablations

10.1 Tasks and metrics

Tasks: pick-and-place with distractors; door opening (varied handles); tactile-only identification; visual servoing; mid-run sensor rollout (add tactile).
Metrics: success rate, time-to-success, contact violations, OOD generalization, compression ratio over time, prediction error curves, replan latency, safety incidents, MES fidelity.

Figure 5: **ROS2 integration.** Device drivers → ingest shims → Phaneron → controller adapters → actuators.

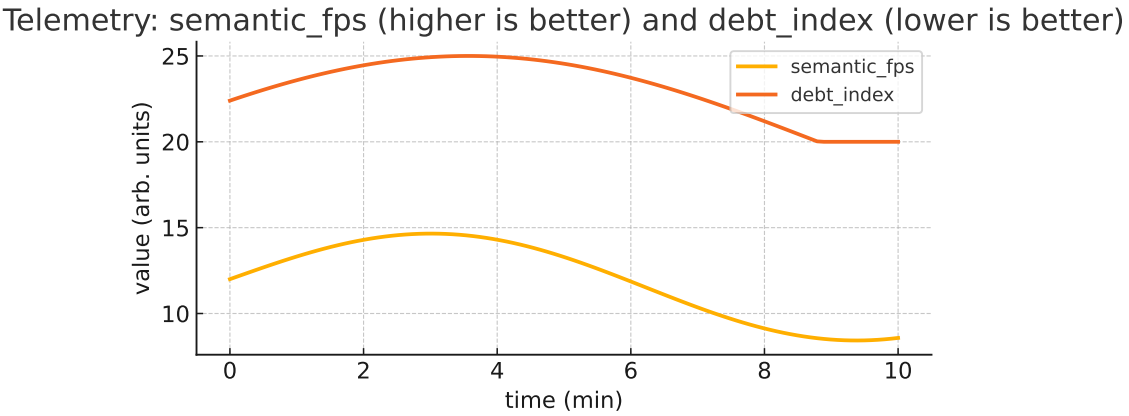

Figure 6: **Telemetry.** Tracking `semantic_fps` and `debt_index` over time.

10.2 Quantitative evaluation table

Task	Metric(s)	Structural check
Pick-and-place	Success, time, near-misses	Affordance motifs reused; MES for grasp/refusal
Door opening	Success, torque violations	Handle motif placement; risk veto explanations
Tactile ID	Accuracy, trials to crit.	Cross-modal merge with vision patterns
Visual servoing	Tracking error, latency	Ego anchor stability under motion
New sense rollout	Time-to-parity, spike area	Dictionary growth; merge rate; prediction gain

10.3 Ablations

Ablation	Expected effect
No aversive field	More near-misses; unstable exploration; higher violations
No curiosity field	Slower pattern discovery; lower compression gain
No replay	Longer onboarding spike; poorer transfer
Fixed pattern cap	Underfitting; brittle behavior in OOD
Directed edges for actions	Conflates base graph; worse cross-modal reuse

11 Related Work (brief)

Graph world models; neurosymbolic control; predictive processing; skill libraries. Our contribution: unlabeled base, patterns for all nuance, MDL+prediction objective, versioning, and

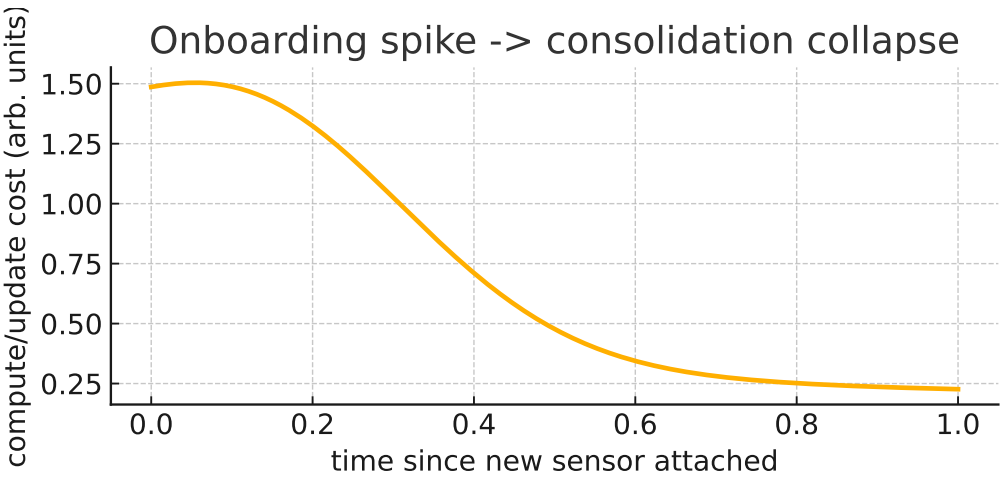

Figure 7: **Onboarding spike.** Compute/update cost spikes then collapses as the dictionary consolidates.

explainable rewrites.

12 Limitations

Pattern search and placement can be heavy; we mitigate with locality (WL-style neighborhoods), dynamic caps, and multi-rate loops. Some phenomena (social cues) require additional motifs and sensors.

13 Stagewise Reflection and Singularities

Reflection parity is not a fixed point but a *stagewise* target. Let W_t be the micro-world, $\pi_{B,G}(t)$ the task-indexed quotient under resource bound B and goals G, and P_t the current Phaneron.

- **Stage k:** an interval $[t_k, t_{k+1})$ where there exists a homomorphism $h_k : P_t \to \pi_{B,G}(t)$ with task error $\leq \varepsilon(B)$ and P_t is MDL-minimal under the equilibrium objective.
- **Singularity at t_{k+1}:** the smallest time where no sequence of local refinements of $P_{t_{k+1}^-}$ can keep task error $\leq \varepsilon(B)$ without (i) raising capacity B, (ii) narrowing G, or (iii) introducing new invariants (a partition re-factor). Equivalently, the optimal partition changes topology/cardinality:

$$\mathcal{P}(B^-, G) \ncong \mathcal{P}(B^+, G) \quad \text{or} \quad |\mathcal{P}(B^-, G)| \neq |\mathcal{P}(B^+, G)|.$$

Predictability horizon. The horizon H at state (P_t, B, G) is the largest τ such that all task queries within $[t, t+\tau]$ admit bounded regret under the current partition; beyond H, any reliable forecast requires a partition transition (capacity increase or new invariants).
Precursors and a practical score. As a singularity approaches, we typically observe: (i) rising conflict curvature despite consolidation, (ii) increasing residual variance and autocorrelation in forecast errors, (iii) accelerated split/merge churn and codebook drift, (iv) longer/variable MES and message-size spikes in multi-agent settings, and (v) a stall in reflection-distance improvement. A simple trigger uses a weighted score $S(t)$ over these signals and initiates a controlled re-factor when $S(t) > \tau$.
Consequences. Intelligence growth is piecewise: long plateaus of reflection parity punctuated by singularities when tasks/evidence demand new invariants. This explains "unknown unknowns"

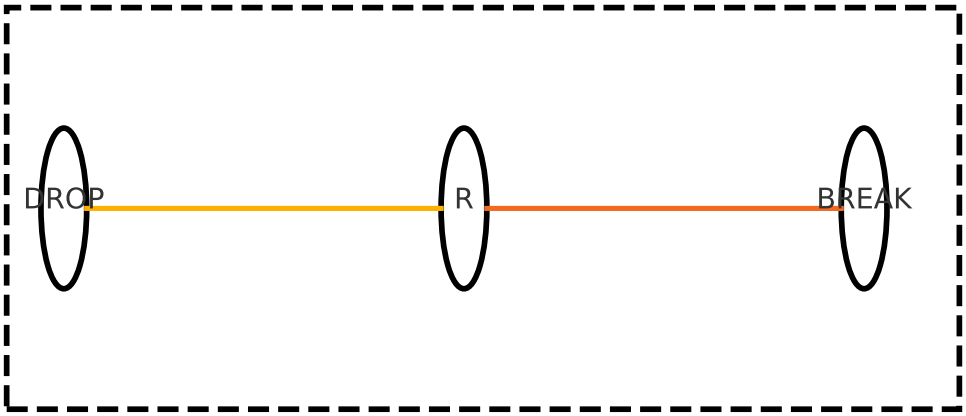

Figure 8: **MES example.** A compact subgraph (DROP -> cause -> BREAK) that justifies a refusal or a prediction.

pre-transition, collective communication cliffs when teams align a finer partition, and subjective time shifts when cognitive debt is reduced across a transition.

14 Conclusion

An embodied Phaneron unifies perception, memory, planning, and control. By reusing one structure for semantics and action, it yields safety-through-structure, explainability, and transfer across sensors and tasks—and it gets faster as it learns.

A Appendix A: Minimal MDL code

Two-part code for dictionary $\mathcal{D}$ and residuals R. For patterns P_i with usages U_i:

$$L(\mathcal{D}) = \sum_i \big(L_{\mathbb{N}}(|V(P_i)|) + L_{\mathbb{N}}(|E(P_i)|) + L_{\text{iso}}(P_i)\big), \tag{2}$$

$$L(U|\mathcal{D}, B) = \sum_i \big(L_{\mathbb{N}}(|U_i|) + \sum_{u \in U_i} L_{\text{place}}(u|P_i, B)\big), \tag{3}$$

$$L(R|\mathcal{D}, U, B) = L_{\text{edges}}(\text{residual edges}|B). \tag{4}$$

Admissible if total $\Delta L + \lambda_{\text{pred}}\Delta L_{\text{pred}} + \lambda_{\text{goal}}\Delta L_{\text{goal}} < 0$ and all safety constraints hold.

B Appendix B: CDA in the control loop (pseudocode)

```
Input: state S_t=(G_t,D_t), proposals P={ G_k }, objective L(  ), budgets
1: // Reflex (always-on): low-level control, safety monitors (hard constraints)
2: // Semantic loop (budget  _sem ): propose candidate placements from sensors
3: score each  G_k  with local  L ; drop non-improving or unsafe
4: greedily build non-overlapping winners W under constraints
5: apply DPO rewrites in W; append versioned events; update telemetry
6: if debt_index > threshold and safe: schedule consolidation (budget  _cons )
7: // Planner loop (budget  _plan ): evaluate options, send high-level targets
```

C Appendix C: ROS2 shim notes

Shims map device messages to distinctions and sync links only; no labels or semantic features are injected. The planner and controllers communicate via skill patterns with role slots (object, site, goal), realized as target poses or force profiles.