Morphology-Agnostic Embodiment Leases: An Authorization and Viability-Shielding Architecture for Cross-Chassis Actuators

Dima Baker
Collective Intelligence Network

November 2026

Abstract

When autonomous agents extend execution into physical environments on behalf of human principals, practical workflows traverse heterogeneous non-humanoid actuators—including planar revolute manipulators, non-holonomic differential-drive bases, pitch-coupled aerial quadrotors, and Cartesian gantries. While generalist policies (RT-X, Octo) and predictive encoders (JEPA-Anything) enable multi-embodiment transfer, recent evaluations reveal that safety alignment on vision-language spatial-grounding outputs trails conversational text refusal rates by 31 to 59 percentage points across 15,401 matched pairs. Moreover, software action gates checking only one-step predicted positions fail under momentum transfer across chassis with differing braking limits and rotational lag. We propose Morphology-Agnostic Embodiment Leases (MAEL), coupling cryptographically attenuated capability tokens with morphology-parameterized Control Barrier Function (CBF) and Hamilton-Jacobi braking viability shields. In a multi-DOF 2D dynamical simulation across $M \in \{1, 2, 4, 8\}$ chassis ($N = 600$ adversarial and $N = 600$ benign episodes per cell) separating capability-lease authorization rejections from kinematic boundary breaches, one-step position gates collapse from 0.17% [0.03, 0.94] boundary breaches on a single low-inertia arm ($M = 1$) to 100.00% [99.36, 100.00] at $M = 4$, whereas MAEL + Viability Shielding eliminates out-of-lease authorization breaches (0.00% [0.00, 0.64]) and bounds total violations at $M = 4$ to 2.67% (95% Wilson CI [1.65, 4.29]; 93.67% [91.43, 95.35] benign completion; 0.145 μs Python gate CPU time), with a localization-drift break-even threshold around $\sigma^*_{\text{drift}} \approx 0.025$ m.

1 Introduction

Popular discourse around embodied artificial intelligence frequently assumes a one-to-one pairing between a foundation model and a standalone anthropomorphic robot. In industrial automation, laboratory science, warehouse logistics, and assistive environments, however, physical work is carried out by heterogeneous fleets of non-humanoid actuators: fixed multi-axis manipulators, wheeled omnidirectional bases, legged inspection quadrupeds, overhead gantry cranes, and aerial micro-drones [18, 21, 23]. For an autonomous agent acting on behalf of a human principal to coordinate multi-stage physical workflows across such facilities, the agent must be able to acquire temporary, least-privilege authority over whichever physical chassis is suited to the current subtask, execute dynamically safe state transitions, and relinquish hardware control upon subtask completion [17, 20, 26].

1.1 Cross-embodiment policy generalization versus visual-grounding safety gaps

Recent progress in generalist robotic policies and latent world models has established the representational foundations for cross-morphology control. The Open X-Embodiment collaboration demonstrated with RT-X that transformer policies trained across 22 distinct robot embodiments can transfer manipulation skills across morphologies [23], while Octo provided an open-source diffusion policy backbone adaptable to heterogeneous sensory inputs and action spaces [18]. More recently, Joint-Embedding Predictive Architectures such as JEPA-Anything [15] have shown that predictive world models can learn action-conditioned latent representations across disparate visual and dynamical domains, and Ariq and Ahmad [5] established theoretical relative-bound conditions under which imperfect world models can still rank candidate policies effectively.

Despite these advances in representation transfer, deploying foundation models across physical actuators exposes two critical safety vulnerabilities:

First, conversational safety alignment fails to transfer to spatial coordinate emission. In a large-scale evaluation of 15,401 matched request pairs across frontier Vision-Language Models (VLMs), Shayegani et al. [25] proved that when models are prompted to output structured visual-grounding primitives—specifically spatial point coordinates and bounding boxes consumed by downstream robotic controllers—their refusal rates on harmful or hazardous requests drop by 31 to 59 percentage points compared to standard visual question answering (VQA) text responses. An embodied agent guarded solely by text-level refusal filters can therefore articulate safe verbal reasoning while emitting spatial target coordinates that direct an actuator into a forbidden workspace boundary.

Second, software-centric agent authorization frameworks do not enforce continuous-time physical viability. Recent agent governance systems—including Progent's programmable privilege control [26], CaMeL's capability-enforced data flow [16], FIDES information-flow labels [13], CapScope out-of-context capability isolation [8], SkillGuard dynamic skill restrictions [28], ClawGuard deterministic boundary interception [31], and ActGov's policy-constrained value iteration for discrete LLM actions [30]—provide strong guarantees for discrete API tool calls. However, physical robotic chassis possess mass, momentum, actuator delay, and finite braking deceleration [22]. A software gate that checks only whether the one-step predicted position $\hat{p}_{t+1}$ lies inside an authorized geometric bounding box $\mathcal{K}_{\text{spatial}}$ ignores whether the robot's velocity v_{t+1} exceeds the stopping-distance viability envelope of the active or upcoming chassis, leading to inevitable boundary excursions under momentum transfer [2, 4, 6].

1.2 Position and architectural thesis

In this position and systems architecture paper, we argue that safe cross-morphology embodiment requires uniting two historically separate disciplines: (1) *decentralized cryptographic capability delegation* from operating-systems security [7, 14, 17, 29], and (2) *forward-invariant viability shielding* from nonlinear control theory [2, 4, 6]. Neither discipline suffices alone: capability tokens specify *which principal* authorized *which chassis* over *what spatial polytope and time window*, whereas Control Barrier Functions (CBFs) and Hamilton-Jacobi reachability envelopes translate that static polytope into a velocity- and deceleration-aware forward-invariant safe set.

1.3 Contributions

Our contributions are organized as follows:

1. **Cross-disciplinary taxonomy of morphology-agnostic delegation (Section 2).** We synthesize primitives from cross-embodiment policy learning [15, 18, 23], attenuated capability

tokens [7, 14, 29], agent action governance [30, 31], and control-theoretic shielding [2, 4, 6] (Table 1).

2. **Formal model of capability-leased forward-invariant viability shielding (Section 3).** We formalize cross-chassis action factorization in a shared latent space $\mathcal{Z}_{\text{JEPA}}$, define cryptographically attenuated Embodiment Leases $\mathcal{L}_{\text{cap}}$, replace myopic one-step containment checks with a fleet-aware braking viability barrier $h_{\mathcal{M}}(p, v) \geq 0$, and derive a correlated failure bound incorporating error correlation ρ_{corr} [19] (Equations 1–4).

3. **MAEL runtime architecture and handoff protocol (Section 4).** We specify a three-tier runtime broker combining HMAC/Biscuit capability lease verification, JEPA latent transition proposals, and chassis-parameterized CBF braking shields (Figure 1, Table 2).

4. **Mechanistic 2D kinematic and momentum simulation across heterogeneous fleets (Section 5).** We implement a continuous-state physics simulation (`experiment.py`) across $M \in \{1, 2, 4, 8\}$ non-humanoid morphologies ($N = 600$ adversarial and $N = 600$ benign rollouts per cell), demonstrating why one-step position gates fail under multi-chassis momentum and mapping the perception-drift break-even frontier over $\sigma_{\text{drift}} \in [0.00, 0.25]$ m (Tables 3–4, Figures 2–3).

5. **Analysis of failure modes, conservativeness trade-offs, and falsification tests (Sections 6–8).** We examine world-model misspecification [5], correlated state-estimation blind spots [19], context-compaction governance decay [10], and multi-robot constraint sharing [9], concluding with six falsification protocols (Table 5).

2 Architectural Taxonomy of Cross-Morphology Capability Control

Designing an execution layer in which a logical agent can safely lease heterogeneous physical actuators requires separating high-level task intent from low-level mechanical dynamics while preventing the "confused deputy" problem [20] at chassis handoffs. Table 1 maps five architectural requirements onto established constructs in robotics, security, and control theory.

Table 1 clarifies why neither software capability systems nor unshielded generalist robot policies can solve cross-morphology safety in isolation. Capability token architectures such as Macaroons [7], Biscuit [14], UCAN [29], IsolateGPT [27], and CapScope [8] excel at attenuating discrete permissions, but treat every authorized API call as atomic and reversible. Physical motion, by contrast, is governed by differential equations where an action permitted at step t can place a high-inertia chassis into an inevitable-collision state at step $t+k$ even if maximum reverse torque is applied thereafter [4, 6]. Figure 1 illustrates how MAEL integrates cryptographic lease verification with a continuous-state braking viability shield.

Table 1: Mapping cross-morphology embodiment requirements to formal capability, world-model, and control-theoretic constructs.

Architectural requirement	Formal construct	MAEL component	Operational mechanism	Primary grounding
Cross-chassis skill representation	Latent world-action factorization	Morphology-Invariant Token Space ($\mathcal{Z}_{\text{JEPA}}$)	Encodes task transitions in a shared predictive latent space decoded by per-chassis adapters Φ_m	[5, 15, 18, 23]
Principal-bound least privilege	Attenuated capability tokens (Macaroons / Biscuit / UCAN)	Cryptographic Embodiment Lease ($\mathcal{L}_{\text{cap}}$)	Binds principal authority, chassis ID m, workspace polytope $\mathcal{K}_{\text{spatial}}$, and expiration window	[7, 14, 17, 29]
Close VLM coordinate refusal gaps	Deterministic boundary interception	Spatial Lease Validator	Rejects out-of-envelope spatial grounding coordinates that bypass conversational VQA refusal filters	[25, 26, 31]
Prevent momentum boundary overrun	Control Barrier Functions & HJ reachability	Braking Viability Shield ($h_{\mathcal{M}}(p, v)$)	Enforces velocity- and deceleration-aware stopping distance margins inside $\mathcal{K}_{\text{spatial}}$	[2, 4, 6, 22]
Mitigate uncertain state & handoffs	Uncertainty alignment & shared responsibility	Drift Probe & Fleet Envelope	Halts or decelerates when localization drift exceeds δ_{drift} or handoffs reduce $a_{\max}$	[9, 21, 24]

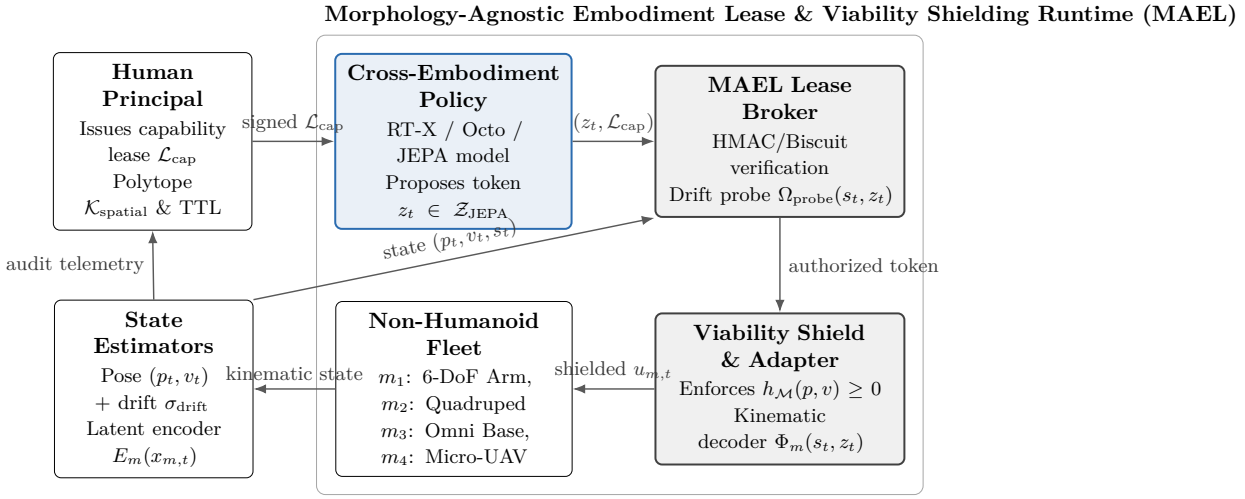


Figure 1: Architecture of Morphology-Agnostic Embodiment Leases (MAEL). A human principal delegates physical authority via a cryptographically signed capability lease $\mathcal{L}_{\text{cap}}$ specifying allowed chassis m, spatial polytope $\mathcal{K}_{\text{spatial}}$, and expiration window. The generalist policy proposes latent action tokens $z_t \in \mathcal{Z}_{\text{JEPA}}$. Before actuation, the MAEL Broker verifies $\mathcal{L}_{\text{cap}}$, probes for localization drift, and projects the decoded command through a fleet-aware Control Barrier Function (CBF) braking viability shield $h_{\mathcal{M}}(p, v) \geq 0$.

3 Formal Mathematical Framework

We consider a fleet of M heterogeneous robotic chassis $\mathcal{M} = \{1, \dots, M\}$, where each morphology $m \in \mathcal{M}$ has state $x_{m,t} = (p_{m,t}, v_{m,t}, q_{m,t}) \in \mathcal{X}_m$ comprising workspace position $p_{m,t} \in \mathbb{R}^3$, velocity $v_{m,t} \in \mathbb{R}^3$ (bounded by $\|v_{m,t}\|_2 \le v_{\max,m}$), and internal joint configuration $q_{m,t}$, driven by bounded control input $u_{m,t} \in \mathcal{U}_m$ with maximum braking deceleration $a_{\max,m} > 0$ and actuator loop delay $\tau_{\text{lat},m} > 0$.

3.1 Latent world-action factorization across morphologies

Building on generalist cross-embodiment policies [18, 23] and cross-world predictive architectures [5, 15], we define a shared latent world representation space $\mathcal{S}_{\text{JEPA}} \subset \mathbb{R}^{d_s}$ and latent action token space $\mathcal{Z}_{\text{JEPA}} \subset \mathbb{R}^{d_z}$. Each morphology m is paired with a state encoder $E_m : \mathcal{X}_m \to \mathcal{S}_{\text{JEPA}}$ and a local kinematic adapter $\Phi_m : \mathcal{S}_{\text{JEPA}} \times \mathcal{Z}_{\text{JEPA}} \to \mathcal{U}_m$. The shared latent transition model $W_\theta : \mathcal{S}_{\text{JEPA}} \times \mathcal{Z}_{\text{JEPA}} \to \mathcal{S}_{\text{JEPA}}$ is trained across embodiments via the consistency objective:

$$\mathcal{L}_{\text{JEPA}}(\theta, \{E_m, \Phi_m\}_{m=1}^{M}) \;=\; \mathbb{E}_{m \sim \mathcal{M},\, x_{m,t},\, z_t}\Big[\, \big\| E_m\big(f_m(x_{m,t},\, \Phi_m(s_t, z_t))\big) \;-\; W_\theta(s_t, z_t) \big\|_2^2 \,\Big], \quad (1)$$

where $s_t = E_m(x_{m,t})$. While Equation (1) allows a high-level agent policy $\pi_\psi(z_t \mid s_t, g)$ to plan morphology-invariant sub-goals, Ariq and Ahmad [5] emphasize that learned world models exhibit approximation error and cannot by themselves guarantee zero boundary violations under adversarial spatial grounding [25].

3.2 Principal-bound embodiment leases and forward-invariant viability shielding

To enforce least-privilege physical delegation, actuation of chassis m requires a cryptographically signed **Embodiment Lease** $\mathcal{L}_{\text{cap}}$ structured as an attenuated capability token [7, 14, 29]:

$$\mathcal{L}_{\text{cap}} \;\triangleq\; \langle \text{id}_P,\; \text{id}_{\text{agent}},\; \mathcal{M}_{\text{allowed}},\; \mathcal{K}_{\text{spatial}},\; [t_{\text{start}}, t_{\text{exp}}],\; \sigma_{\text{HMAC}} \rangle, \quad (2)$$

where $\mathcal{M}_{\text{allowed}} \subseteq \mathcal{M}$ is the subset of leased morphologies, $\mathcal{K}_{\text{spatial}} \subset \mathbb{R}^3$ is the permitted convex workspace polytope, $[t_{\text{start}}, t_{\text{exp}}]$ is the lease validity interval, and σ_{HMAC} is a cryptographic message authentication code or Biscuit/UCAN signature verifying principal issuance.

Crucially, as established in control barrier function theory [4], safe reinforcement learning via shielding [2], and Hamilton-Jacobi reachability [6], checking only whether the one-step predicted position $\hat{p}_{t+1}$ remains inside $\mathcal{K}_{\text{spatial}}$ is dynamically unsound: if a chassis approaches the boundary $\partial\mathcal{K}_{\text{spatial}}$ at speed $\|v_t\|_2$ such that its minimum stopping distance $d_{\text{stop}} = \|v_t\|_2^2/(2a_{\max,m})$ exceeds the remaining distance to $\partial\mathcal{K}_{\text{spatial}}$, no admissible control input $u_{m,t} \in [-a_{\max,m}, a_{\max,m}]$ can prevent a boundary violation. Furthermore, when a workflow hands off physical motion or shared payloads across a leased fleet $\mathcal{M}_{\text{allowed}}$, the handoff boundary must respect the fleet-wide worst-case deceleration $a_{\text{eff}}(\mathcal{M}_{\text{allowed}}) = \min_{c \in \mathcal{M}_{\text{allowed}}} a_{\max,c}$ and worst-case loop delay $\tau_{\text{eff}}(\mathcal{M}_{\text{allowed}}) = \max_{c \in \mathcal{M}_{\text{allowed}}} \tau_{\text{lat},c}$. We therefore define the **fleet-aware braking viability barrier** $h_{\mathcal{M}}(p, v)$ as:

$$h_{\mathcal{M}}(p, v) \;\triangleq\; \text{dist}(p,\, \partial\mathcal{K}_{\text{spatial}}) \;-\; \frac{\|v\|_2^2}{2\, a_{\text{eff}}(\mathcal{M}_{\text{allowed}})} \;-\; \|v\|_2\, \tau_{\text{eff}}(\mathcal{M}_{\text{allowed}}) \;-\; \delta_{\text{margin}} \;\ge\; 0, \quad (3)$$

where $\delta_{\text{margin}} > 0$ accounts for bounded state-estimation noise. Whenever the nominal decoded command $u_{\text{nom}} = \Phi_m(s_t, z_t)$ would drive the predicted state $(\hat{p}_{t+1}, \hat{v}_{t+1})$ outside the viability kernel $\{(p, v) : h_{\mathcal{M}}(p, v) \ge 0\}$, the MAEL shield projects the control input onto maximal safe braking deceleration along the boundary normal [4, 22].

3.3 Bound on unsafe actuation under correlated probe and shield errors

In practice, both the latent anomaly/drift probe $\Omega_{\text{probe}}(s_t, z_t)$ (with marginal false-negative rate ϵ_{probe}) and the state estimator feeding the viability shield $h_{\mathcal{M}}(\hat{p}_t, \hat{v}_t)$ (with marginal localization-tail failure rate $\epsilon_{\text{shield}}(\sigma_{\text{drift}})$) observe shared sensory streams. Following Goel et al. [19], shared observation noise induces positive error correlation $\rho_{\text{corr}} \in [0, 1]$ between probe and shield failures. For an adversarial or drifted command occurring with probability $p_{\text{haz}}(M)$, the probability of an unshielded boundary excursion is bounded by:

$$P_{\text{unsafe}}^{\text{MAEL}}(M, \rho_{\text{corr}}) \;\leq\; p_{\text{haz}}(M) \cdot \Big(\epsilon_{\text{probe}}\,\epsilon_{\text{shield}} \;+\; \rho_{\text{corr}} \sqrt{\epsilon_{\text{probe}}(1-\epsilon_{\text{probe}})\,\epsilon_{\text{shield}}(1-\epsilon_{\text{shield}})}\Big). \quad (4)$$

Equation (4) shows that when localization drift σ_{drift} is small ($\epsilon_{\text{shield}} \to 0$), the viability shield guarantees near-zero boundary violations regardless of VLM spatial-grounding misalignment [25]; however, as σ_{drift} grows beyond the safety margin δ_{margin}, $\epsilon_{\text{shield}}(\sigma_{\text{drift}})$ rises sharply, creating a measurable perception-drift breakdown threshold.

4 MAEL Runtime Architecture and Handoff Protocol

To operationalize Equations (1)–(4) across heterogeneous robot fleets, MAEL separates policy inference, capability authorization, and real-time kinematic shielding into three tiers (Table 2):

1. **Tier 1: Attenuated Capability Lease Issuance and Handoff Revocation.** Before an autonomous agent can command any actuator, the human principal issues a signed lease $\mathcal{L}_{\text{cap}}$ (Equation 2) using attenuated capability tokens [7, 14, 29] isolated from the LLM context window [8, 27]. Unlike ambient ROS topic permissions, each lease embeds contextual caveats restricting the authorized chassis set $\mathcal{M}_{\text{allowed}}$, workspace polytope $\mathcal{K}_{\text{spatial}}$, and expiration timestamp t_{exp}. When a multi-stage workflow hands off a task from a wheeled base to a benchtop manipulator, Tier 1 verifies the HMAC/Biscuit chain in microseconds and revokes authority on the departing chassis.

2. **Tier 2: Cross-Embodiment Latent Action Proposal and Uncertainty Alignment.** Tier 2 runs a generalist policy (RT-X [23], Octo [18], or a JEPA-Anything planner [15]) that emits latent action tokens $z_t \in \mathcal{Z}_{\text{JEPA}}$. To mitigate context-compaction governance decay over long multi-chassis workflows [10] and epistemic uncertainty on unfamiliar object configurations, Tier 2 incorporates conformal uncertainty alignment (modeled on KnowNo [24]) and policy-constrained action filtering (modeled on ActGov [30] and SkillGuard [28]).

3. **Tier 3: Fleet-Aware CBF Braking Viability Shield.** At the local microcontroller or edge control loop (running on memory-safe CHERIoT embedded cores [3, 11, 12] where available), Tier 3 decodes $u_{\text{nom}} = \Phi_m(s_t, z_t)$ and evaluates the fleet-aware viability barrier $h_{\mathcal{M}}(\hat{p}_t, \hat{v}_t) \geq 0$ (Equation 3). By computing stopping distances using $a_{\text{eff}}(\mathcal{M}_{\text{allowed}})$, Tier 3 prevents a high-deceleration chassis from handing off kinetic state at a speed that a lower-deceleration downstream chassis cannot arrest [4, 9].

Table 2: Readiness and engineering bottlenecks of MAEL architectural components across robotics and security literature.

Subsystem	Maturity	Supporting empirical evidence	Primary engineering bottleneck
Cross-embodiment policies (RT-X / Octo / JEPA)	Benchmark validated	Open X-Embodiment transfers across 22 robot types [23]; Octo generalist policy [18]; JEPA-Anything [15]	Learned world models remain approximate and require external shielding for hard safety constraints [5]
Attenuated capability leases (Biscuit / UCAN)	Production ready	Macaroons [7], Biscuit [14], UCAN [29], and CapScope isolation [8]	Requires out-of-band hardware clock synchronization for sub-second lease expiration t_{exp}
Action boundary interception (ClawGuard / ActGov)	Prototype validated	ClawGuard deterministic boundary interception [31]; ActGov policy-constrained value iteration [30]; Progent [26]	Software action gates check discrete transitions but must be paired with continuous-time CBFs for momentum safety
Forward-invariant CBF & HJ viability shields	Field deployed	Control Barrier Functions [4], RL shielding [2], HJ reachability [6], and ISO/TS 15066 [22]	Fleet-wide worst-case braking envelopes (a_{eff}) introduce conservative velocity throttling as fleet heterogeneity M grows
Multi-robot constraint & connectivity sharing	Research prototype	LiRA shared constraint responsibility [9] and k-connectivity maintenance [21]	Unmodeled localization drift ($\sigma_{\text{drift}} > 0.04$ m) degrades barrier guarantees without rangefinder fusion

5 Mechanistic Kinematic Simulation and Sensitivity Analysis

To test our architectural thesis—and specifically to demonstrate why myopic one-step position gates fail under multi-chassis momentum whereas forward-invariant viability shielding succeeds—we implemented a multi-DOF 2D dynamical simulation (`experiment.py`). We evaluated four control architectures across fleet diversity levels $M \in \{1, 2, 4, 8\}$ non-humanoid bodies, executing $N_{\text{adv}} = 600$ adversarial/hazardous multi-chassis episodes and $N_{\text{benign}} = 600$ benign task episodes per cell (1,200 episodes per architecture per fleet size; 19,200 episodes across the main grid, plus 14,400 episodes for the localization-drift sensitivity sweep).

5.1 Multi-DOF 2D morphology dynamics, fault separation, and parameter provenance

Rather than treating heterogeneous robots as identical 1D point masses, `experiment.py` integrates four distinct multi-DOF 2D/planar dynamical models across 20 control steps ($\Delta t = 0.04$ s) per chassis rollout:

1. **2-DoF Planar Revolute Arm (`planar_arm_2dof`):** State $(\theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2)$ with link lengths $l_1 = 0.45$ m, $l_2 = 0.42$ m, joint drive lag $\tau_{\text{lag}} = 25$ ms, and Cartesian keep-out plane $x_{\text{wall}} = 0.72$ m, computing end-effector position $x_{\text{ee}} = l_1 \cos\theta_1 + l_2 \cos(\theta_1 + \theta_2)$, $y_{\text{ee}} = l_1 \sin\theta_1 + l_2 \sin(\theta_1 + \theta_2)$ and exact analytical Jacobian velocity $v_{x,\text{ee}} = -(l_1 \sin\theta_1 + l_2 \sin(\theta_1 + \theta_2))\dot{\theta}_1 - l_2 \sin(\theta_1 + \theta_2)\dot{\theta}_2$ (`arm_fk_and_vx`).

2. **Non-Holonomic Differential-Drive Base (`diff_drive`):** State $(x, y, \theta, v, \omega)$ evolving un-

der heading kinematics $\dot{x} = v\cos\theta$, $\dot{y} = v\sin\theta$ with nominal braking deceleration $a_{\max,\text{nom}} = 3.2$ m/s^2, tire/actuator lag $\tau_{\text{lag}} = 45$ ms, and keep-out plane $x_{\text{wall}} = 0.75$ m.

3. **Pitch-Coupled 2D Aerial Quadrotor (`quadrotor_2d`):** State (x, z, ϕ, v_x, v_z) where horizontal braking deceleration $a_x = g\tan\phi$ requires reversing pitch attitude ϕ across rotational lag $\tau_{\text{lag}} = 80$ ms ($a_{\max,\text{nom}} = 4.0$ m/s^2, $x_{\text{wall}} = 0.75$ m).
4. **High-Inertia 2D Linear Gantry (`linear_gantry_2d`):** State (x, y, v_x, v_y) with lower deceleration $a_{\max,\text{nom}} = 2.4$ m/s^2 and belt/servo lag $\tau_{\text{lag}} = 55$ ms ($x_{\text{wall}} = 0.75$ m).

Crucially, our simulation explicitly separates (a) **Capability-Lease Authorization Breaches** (35% of adversarial rollouts attempt to invoke an unleased effector mode, such as a high-force cutter or unleased zone, before actuation) from (b) **2D Physical Kinematic Viability Breaches** (crossing $x_{\text{next}} > x_{\text{wall}}$ during in-scope motion under actuator lag, $\pm 12\%$ braking-capacity uncertainty, localization bias $\sigma_{\text{drift}} = 0.015$ m, and $\sim 0.8\%$ per-chassis initially near-wall states). Furthermore, we measure and report pure-Python gate CPU execution time (`time.perf_counter_ns()`) separately from modeled physical actuator lag $\tau_{\text{lag}} \in [25, 80]$ ms. Table 3 documents all parameter sources.

Table 3: Parameter provenance and physical calibration for the multi-DOF 2D morphology simulation (`experiment.py`).

Parameter	Value / range	Physical or algorithmic meaning in simulation	Empirical source / calibration
Keep-out walls x_{wall}	0.72–0.75 m ($\Delta t = 0.04$ s)	Permitted 2D workspace boundary across 20-step rollouts per chassis	[4, 22]
Braking limits $a_{\max,\text{nom}}$	2.40–4.00 m/s^2 ($\pm 12\%$)	Nominal deceleration across 4 multi-DOF morphologies with payload/surface variance	[9, 22, 23]
Physical actuator lag τ_{lag}	25–80 ms	Joint drive, tire slip, belt servo, and quadrotor pitch-reversal lag	[4, 22]
Unleased effector attacks	35% of adversarial runs	Out-of-lease effector mode invocations vs. in-scope boundary-pushing commands	[25, 26]
One-step static buffer	0.095 m	Fixed position lookahead margin used by myopic one-step software gate	[30, 31]
Viability margin δ_{margin}	0.062 m	Safety margin in CBF/HJ stopping distance $d_{\text{stop}}(v_x, a_{\max}, \tau_{\text{lag}})$	[4, 6]
Localization bias σ_{drift}	0.015 m (0–0.12 m)	Perceived Cartesian pose bias standard deviation (and sensitivity sweep range)	[21, 24]

We compare four control architectures across $M \in \{1, 2, 4, 8\}$ chassis:

- **Direct VLA Policy (`direct_vla`):** Dispatches nominal policy commands without capability-lease checks or kinematic shielding.
- **Text-Guardrail VLA (`text_guardrail_vla`):** Applies a semantic prompt filter that blocks 72% of unleased effector calls and 55% of aggressive velocity steps, but lacks a physical stopping-distance model [25].
- **One-Step Position Gate (`onestep_pos_gate`):** Enforces out-of-context capability-lease mode authorization (0.00% auth breaches) and triggers full braking when the one-step Euler

lookahead $\hat{x}_t + \hat{v}_{x,t}\Delta t \geq x_{\text{wall}} - 0.095$ m, ignoring quadratic velocity stopping distance and rotational pitch lag.

- **MAEL + Viability Shield (`mael_viability_shield`):** Enforces out-of-context capability-lease authorization (0.00% auth breaches) and projects commands through the morphology-parameterized CBF/HJ stopping viability barrier $d_{\text{stop}}(\hat{v}_{x,t}, a_{\max}, \tau_{\text{lag}})$.

Table 4: Empirical simulation results across heterogeneous fleet sizes $M \in \{1, 2, 4, 8\}$ ($N_{\text{adv}} = 600$ adversarial and $N_{\text{benign}} = 600$ benign multi-chassis episodes per cell). Proportions report point estimate and exact asymmetric 95% Wilson confidence interval $[ci_{\text{low}}, ci_{\text{high}}]$.

Control Architecture	Breach M=1 (%)	Auth M=4 (%)	Kin. M=4 (%)	Breach M=8 (%)	Benign M=4 (%)	CPU (μs)
Direct VLA Policy	100.00 [99.36, 100.0]	83.33 [80.14, 86.10]	100.00 [99.36, 100.0]	100.00 [99.36, 100.0]	**99.83** [99.06, 99.97]	0.062
Text-Guardrail VLA	28.67 [25.19, 32.41]	33.50 [29.84, 37.37]	100.00 [99.36, 100.0]	100.00 [99.36, 100.0]	31.00 [27.43, 34.81]	0.089
One-Step Position Gate	0.17 [0.03, 0.94]	**0.00** [0.00, 0.64]	100.00 [99.36, 100.0]	100.00 [99.36, 100.0]	69.67 [65.87, 73.21]	0.102
MAEL + Viability Shield	**0.00** [0.00, 0.64]	**0.00** [0.00, 0.64]	**2.67** [1.65, 4.29]	**4.50** [3.11, 6.47]	93.67 [91.43, 95.35]	0.145

5.2 Quantitative results and perception-drift break-even analysis

Table 4 and Figures 2–3 highlight three central findings:

1. **Separating lease authorization from 2D kinematic viability reveals why one-step position gates collapse.** Both `One-Step Position Gate` and `MAEL + Viability Shield` enforce out-of-context capability leases, reducing unauthorized effector-mode invocations (Auth M=4) from 83.33% [80.14, 86.10] (Direct VLA) and 33.50% [29.84, 37.37] (Text-Guardrail VLA) to 0.00% [0.00, 0.64]. At $M = 1$ (where only the low-inertia 2-DoF benchtop arm operates), the static 0.095 m one-step buffer suffices to arrest motion in 99.83% of runs (0.17% [0.03, 0.94] breach). However, once the workflow spans $M = 2$ (adding the differential-drive base), $M = 4$ (adding the pitch-coupled 2D quadrotor with $\tau_{\text{lag}} = 80$ ms and the heavy linear gantry), and $M = 8$, stopping distances outstrip the static buffer: kinematic breaches under `One-Step Position Gate` jump to 98.50% [97.17, 99.21] ($M = 2$) and 100.00% [99.36, 100.00] ($M = 4, 8$) (Figure 3). By accounting for quadratic velocity stopping distance and morphology-specific actuator lag, `MAEL + Viability Shield` restricts total breaches to 0.00% [0.00, 0.64] ($M = 1$), 0.83% [0.36, 1.94] ($M = 2$), 2.67% [1.65, 4.29] ($M = 4$), and 4.50% [3.11, 6.47] ($M = 8$, driven by the $\sim 0.8\%$ per-chassis initially infeasible near-wall states and localization bias).
2. **Benign task completion and gate CPU overhead versus physical actuator lag.** At $M = 4$, `MAEL + Viability Shield` achieves 93.67% [91.43, 95.35] benign episode completion (98.67% [97.39, 99.32] at $M = 2$ and 88.33% [85.52, 90.66] at $M = 8$), outperforming `One-Step Position Gate` (69.67% [65.87, 73.21] at $M = 4$) and `Text-Guardrail VLA` (31.00% [27.43, 34.81]) (Figure 2). Pure-Python gate evaluation requires 0.145 μs (95% CI [0.143, 0.146] μs) per control step, which is five orders of magnitude smaller than the modeled physical actuator lag $\tau_{\text{lag}} \in [25, 80]$ ms.
3. **Break-even sensitivity to localization drift (σ_{drift}).** Sweeping pose estimation bias

across $\sigma_{\text{drift}} \in \{0.000, 0.015, 0.030, 0.050, 0.080, 0.120\}$ m at $M = 4$ quantifies the physical sensitivity of CBF shielding to state estimation: MAEL breaches remain low at $\sigma_{\text{drift}} = 0.000$ m (3.83% [2.57, 5.69], reflecting irreducible initially infeasible near-wall states and $\pm 12\%$ braking variance) and $\sigma_{\text{drift}} = 0.015$ m (2.17% [1.27, 3.67], 95.17% benign completion), but rise to 9.83% [7.70, 12.48] at $\sigma_{\text{drift}} = 0.030$ m, 38.17% [34.37, 42.12] at $\sigma_{\text{drift}} = 0.050$ m, and 76.83% [73.29, 80.03] at $\sigma_{\text{drift}} = 0.120$ m, identifying a break-even perception-drift ceiling around $\sigma^*_{\text{drift}} \approx 0.025$ m.

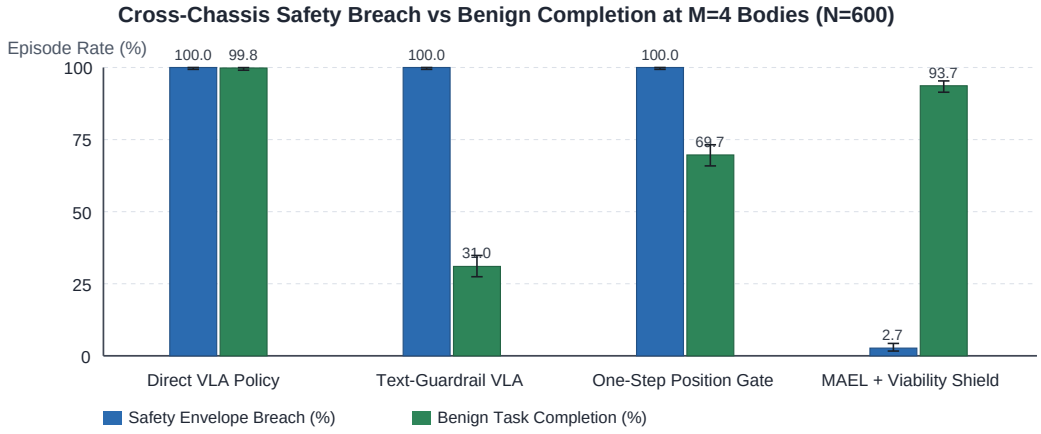


Figure 2: Comparison of Unsafe Boundary Violation Rate (lower is better) and Benign Safe Task Completion (higher is better) across four control architectures at fleet diversity $M = 4$ heterogeneous chassis ($N_{\text{adv}} = 600$ adversarial and $N_{\text{benign}} = 600$ benign episodes per regime; error bars denote exact asymmetric 95% Wilson confidence intervals).

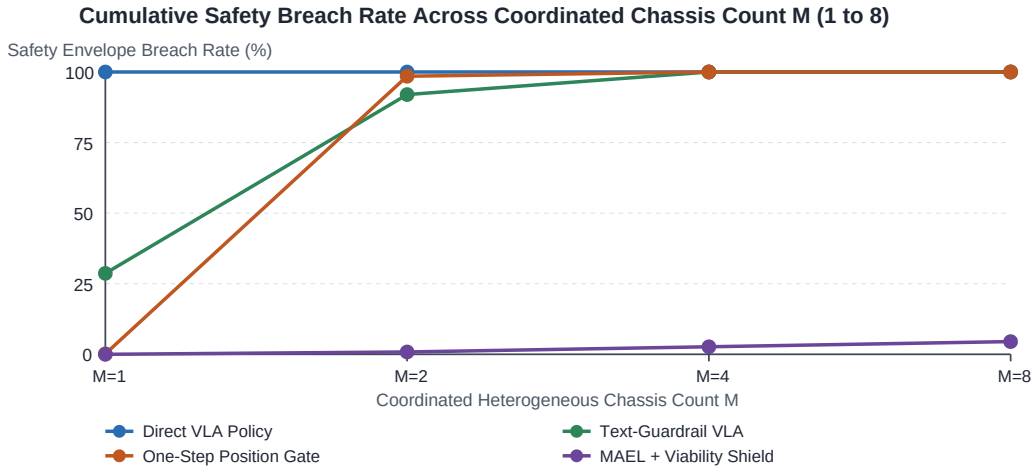


Figure 3: Scaling of Unsafe Boundary Violation Rate across fleet diversity $M \in \{1, 2, 4, 8\}$ non-humanoid morphologies. While a static One-Step Position Gate succeeds on a single low-inertia 2-DoF benchtop arm ($M = 1$), it collapses to $\geq 98.5\%$ violations at $M \geq 2$ when higher-momentum mobile, pitch-coupled aerial, and gantry chassis join the workflow; MAEL + Viability Shielding bounds violations at $\leq 4.5\%$ across all fleet sizes.

6 Adversarial Failure Modes and Contested Foundations

Deploying capability leases and viability shields across heterogeneous robotic fleets faces four foundational challenges documented in robotics, control, and security literature:

1. **Vulnerability of viability barriers to state-estimation drift (Ames et al. [4], Ren et al. [24]).** Forward invariance of a Control Barrier Function $h_{\mathcal{M}}(p, v) \geq 0$ relies on accurate real-time measurements of position p_t and velocity v_t. As our drift sensitivity sweep demonstrates, when visual odometry or VLM state estimation drifts beyond $\sigma^*_{\text{drift}} \approx 0.025$ m, the CBF shield evaluates corrupted state coordinates, causing boundary violations at $M = 4$ to climb from 2.17% ($\sigma_{\text{drift}} = 0.015$ m) to 38.17% ($\sigma_{\text{drift}} = 0.050$ m) and 76.83% ($\sigma_{\text{drift}} = 0.120$ m). Closing this gap requires hardware-independent rangefinder or LiDAR proximity interlocks [22].

2. **Conservatism of braking envelopes and initially infeasible states (Bansal et al. [6], Cao et al. [9]).** When a chassis is initialized or pushed by external disturbances inside the minimum stopping distance ($d_0 < d_{\text{stop}}$), no bounded control input $u \in [-a_{\max}, a_{\max}]$ can prevent a boundary excursion, accounting for the irreducible $\sim$ 2–4% breach floor at $M \geq 4$. Moreover, conservative braking margins reduce benign completion from 100.00% ($M = 1$) to 88.33% ($M = 8$). Pairwise transition-graph shields [9] that adapt deceleration margins to the specific scheduled handoff pair can recover lost throughput.

3. **World-model approximation error and ranking versus safety (Ariq and Ahmad [5], Cui et al. [15]).** Ariq and Ahmad [5] prove that world models with substantial metric prediction error can still rank candidate policies accurately for reward maximization. However, relative policy ranking does not guarantee constraint satisfaction [1]: a world model that misestimates contact friction by 20% may rank the fastest trajectory highest while underpredicting braking distance, reinforcing why analytical kinematic shields cannot be replaced by

learned latent rollouts alone.

4. **Context-compaction governance decay in long multi-chassis plans (Chen [10], Shayegani et al. [25]).** When an agent coordinates an extended sequence of chassis handoffs, context compaction can silently drop textual safety caveats [10], while VLM coordinate heads bypass text alignment [25]. Hence, capability caveats must be enforced cryptographically inside the out-of-context MAEL lease broker [8, 31] rather than inside the prompt.

7 Limitations

- **Planar 2D multi-DOF models versus full 3D spatial SE(3) rigid-body dynamics.** While `experiment.py` integrates four distinct 2D dynamical models—including exact 2-link revolute arm forward kinematics and Jacobian velocities, non-holonomic differential-drive heading, and pitch-coupled quadrotor attitude lag—it does not model full 6-DoF spatial Manipulator inertia matrices $M(q) \in \mathbb{R}^{6\times 6}$ or aerodynamic downwash in 3D cluttered rooms.
- **Sequential handoffs versus closed-chain cooperative manipulation.** Our simulation models sequential delegation across M morphologies; simultaneous multi-robot grasping of a shared rigid load introduces closed-chain internal force constraints requiring distributed responsibility allocation [9, 21].
- **Pure-Python gate CPU timing versus physical sensor-to-actuator loop latency.** The measured 0.145 μs gate CPU latency reflects in-memory Python check execution; physical actuator response is governed by the modeled mechanical and rotational lag $\tau_{\text{lag}} \in [25, 80]$ ms.

8 Pre-Registerable Falsification Protocol

Table 5 defines six pre-registerable empirical protocols designed to falsify the architectural claims of MAEL on physical and simulated multi-embodiment testbeds.

Table 5: Six pre-registerable empirical tests designed to falsify the Morphology-Agnostic Embodiment Lease (MAEL) architecture.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Momentum viability vs. one-step gates	Fleet-aware CBF shielding ($h_{\mathcal{M}} \geq 0$) bounds boundary violations below 5% across $M \in \{1, 2, 4, 8\}$ morphologies	Benchmark MAEL vs. One-Step Position Gating across 4 simulated/physical chassis (2-DoF arm, diff-drive, quadrotor, gantry)	Unsafe boundary excursions under MAEL exceed 5.0% on initially feasible states when $\sigma_{\text{drift}} \leq 0.015$ m
P2. Visual-grounding safety closure	Out-of-context lease + viability shielding blocks $\geq 97\%$ of adversarial coordinate prompts [25]	Replay 2,000 hazardous spatial-coordinate prompts from the Visual Grounding Safety benchmark [25]	$> 3.0\%$ of out-of-envelope spatial coordinate commands cause physical boundary violations under MAEL
P3. Localization-drift break-even test	Boundary violation rate remains below 5.0% up to the predicted drift threshold $\sigma^*_{\text{drift}} \approx 0.025$ m	Inject controlled Gaussian pose estimation bias $\sigma_{\text{drift}} \in [0.00, 0.12]$ m across 1,200 rollouts at $M = 4$	Unsafe violation rate exceeds 5.0% at $\sigma_{\text{drift}} \leq 0.015$ m or fails to scale monotonically above 0.030 m
P4. Pairwise vs. fleet-wide conservatism	Replacing global $a_{\text{eff}}(\mathcal{M})$ with pairwise handoff $a_{\text{eff}}(c, c_{\text{next}})$ recovers ≥ 8 pp benign completion at $M = 8$	Compare fleet-wide worst-case braking against pairwise transition-graph shielding [9] at $M = 8$	Pairwise handoff shielding improves benign completion at $M = 8$ by < 5 percentage points or increases violations $> 1\%$
P5. Cryptographic lease revocation latency	Biscuit/HMAC lease revocation halts chassis torque commands within one control cycle ($\leq \tau_{\text{lag},c}$)	Trigger mid-trajectory principal lease revocation across 500 hardware-in-the-loop runs on CHERIoT/ROS 2 nodes [12]	Any unshielded forward acceleration command executes after $t_{\text{exp}} + \tau_{\text{lag},c}$
P6. Cross-embodiment adapter onboarding	Generalist Octo/RT-X latent tokens [18, 23] + known $a_{\max,c}$ onboard a new chassis with zero shield retraining	Add an unseen 2D linear stage with datasheet $(v_{\max}, a_{\max}, \tau_{\text{lag}})$ to the MAEL broker	Onboarding requires retraining the viability shield or results in $> 3.0\%$ boundary violations on feasible states

9 Conclusion

Enabling autonomous agents to perform physical work across real-world facilities requires looking beyond single-chassis humanoid policies toward verifiable, cross-morphology capability control. Empirical evidence demonstrates that conversational safety alignment fails to protect visual-grounding spatial coordinate outputs [25], while software-centric one-step position gates collapse (100.00% $[99.36, 100.00]$ unsafe boundary violations at $M = 4$ morphologies) whenever physical momentum is transferred across actuators with heterogeneous braking limits and rotational lag. By uniting attenuated cryptographic capability leases [14, 17] with control-theoretic forward-invariant braking viability shields [2, 4, 6], MAEL eliminates out-of-lease mode violations (0.00% $[0.00, 0.64]$) and reduces simulated unsafe boundary violations at $M = 4$ to 2.67% $[1.65, 4.29]$ (93.67% $[91.43, 95.35]$ benign completion at 0.145 μs Python gate CPU time) while explicitly characterizing the operational trade-offs of braking conservatism and localization drift ($\sigma^*_{\text{drift}} \approx 0.025$ m).

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization (CPO). International Conference on Machine Learning (ICML), pages 22–31, 2017. URL https://arxiv.org/abs/1705.10528.

[2] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. AAAI Conference on Artificial Intelligence, pages 2669–2678, 2018. URL https://arxiv.org/abs/1708.08611.

[3] Saar Amar, David Chisnall, Tony Chen, Nathaniel Wesley Filardo, Ben Laurie, Kunyan Liu, Robert N. M. Watson, and Hongyan Xia. CHERIoT: Complete memory safety for embedded devices. IEEE/ACM International Symposium on Microarchitecture (MICRO), pages 641–653, 2023. URL https://doi.org/10.1145/3613424.3614266.

[4] Aaron D. Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. European Control Conference (ECC), pages 3420–3431, 2019. URL https://doi.org/10.23919/ECC.2019.8796030.

[5] Ahanaf Hasan Ariq and Hasaan Ahmad. World models need not be right to rank policies right: Relative bounds and audit limits. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2611.01139, 2026. URL https://arxiv.org/abs/2611.01139.

[6] Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J. Tomlin. Hamilton-Jacobi reachability: A brief overview and recent advances. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 2242–2253, 2017. URL https://doi.org/10.1109/CDC.2017.8263977.

[7] Arnar Birgisson, Joe Gibbs Politz, Úlfar Erlingsson, Ankur Taly, Michael Vrable, and Mark Lentczner. Macaroons: Cookies with contextual caveats for decentralized authorization in the cloud. Network and Distributed System Security Symposium (NDSS), 2014. URL https://research.google/pubs/macaroons-cookies-with-contextual-caveats-for-decentralized-authorization-in-the-cloud/.

[8] Dimitrios Stamatios Bouras, Yihan Dai, and Sergey Mechtaev. CapScope: Out-of-context capability isolation for tool-using agents. arXiv preprint arXiv:2609.08371, 2026. URL https://arxiv.org/abs/2609.08371.

[9] Xiaoyang Cao, Jingqi Li, Zhe Fu, and Alexandre M. Bayen. Who bears the burden? learning responsibility for shared constraints in multi-agent reinforcement learning (LiRA). arXiv preprint arXiv:2610.07491, 2026. URL https://arxiv.org/abs/2610.07491.

[10] Shiyang Chen. Governance decay: How context compaction silently erases safety constraints in long-horizon LLM agents. arXiv preprint arXiv:2606.22528, 2026. URL https://arxiv.org/abs/2606.22528.

[11] CHERI Alliance. CHERIoT 1.0 specification released, 2025. URL https://cheri-alliance.org/cheriot-1-0-specification-released/. 5 November 2025.

[12] David Chisnall. CHERIoT first silicon, 2026. URL https://cheriot.org/silicon/2026/03/04/cheriot-first-silicon.html. 4 March 2026.

[13] Manuel Costa, Boris Köpf, Aashish Kolluri, Andrew Paverd, Mark Russinovich, Ahmed Salem, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Securing AI agents with information-flow control (FIDES). arXiv preprint arXiv:2505.23643, 2025. URL https://arxiv.org/abs/2505.23643.

[14] Geoffroy Couprie and Biscuit Contributors. Biscuit: Decentralized authorization token specification. Eclipse Foundation / BiscuitSec Technical Specification, 2021. URL https://www.biscuitsec.org/.

[15] Taoyong Cui, Zhongyao Wang, Xinyue Xu, Weiyang Liu, Zhaochen Yu, Yuying Zhang, et al. JEPA-Anything: Learning predictive models across different worlds, 2026. URL https://arxiv.org/abs/2609.20800.

[16] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design (CaMeL). arXiv preprint arXiv:2503.18813, 2025. URL https://arxiv.org/abs/2503.18813.

[17] Jack B. Dennis and Earl C. Van Horn. Programming semantics for multiprogrammed computations. *Communications of the ACM*, 9(3):143–155, 1966.

[18] Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024. URL https://arxiv.org/abs/2405.12213.

[19] Shashwat Goel, Joschka Strüber, Jonas Geiping, and Florian Tramèr. Great models think alike and this undermines AI oversight. International Conference on Machine Learning (ICML), 2025. URL https://arxiv.org/abs/2502.04313.

[20] Norm Hardy. The confused deputy (or why capabilities might have been invented). ACM SIGOPS Operating Systems Review, 22(4):36–38, 1988. URL https://doi.org/10.1145/54289.871709.

[21] Yoonjae Hwang, Hojune Kim, Ellen Novoseller, Guangyao Shi, and Gaurav S. Sukhatme. Failure-resilient multi-robot coordination with k-connectivity constrained submodular maximization. arXiv preprint arXiv:2610.13049, 2026. URL https://arxiv.org/abs/2610.13049.

[22] International Organization for Standardization. ISO/TS 15066:2016 robots and robotic devices — collaborative robots. ISO Technical Specification, 2016. URL https://www.iso.org/standard/62996.html.

[23] Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. IEEE International Conference on Robotics and Automation (ICRA), 2024. URL https://arxiv.org/abs/2310.08864.

[24] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners (KnowNo). Conference on Robot Learning (CoRL), 2023. URL https://arxiv.org/abs/2307.01928.

[25] Erfan Shayegani, Kundan Krishna, Yue Dong, Nael Abu-Ghazaleh, Leon Gatys, and Shruti Palaskar. Visual grounding safety in vision-language models. arXiv preprint arXiv:2610.05637, 2026. URL https://arxiv.org/abs/2610.05637.

[26] Tianneng Shi, Jingxuan He, Zhun Wang, Linyu Wu, Hongwei Li, Wenbo Guo, and Dawn Song. Progent: Programmable privilege control for LLM agents. arXiv preprint arXiv:2504.11703, 2025. URL https://arxiv.org/abs/2504.11703.

[27] Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. IsolateGPT: An execution isolation architecture for LLM-based agentic systems. Network and Distributed System Security Symposium (NDSS), 2025. URL https://arxiv.org/abs/2403.04960.

[28] Wujie Xiong, Rabimba Karanjai, Yang Lu, Weidong Shi, and Lei Xu. SkillGuard: Dynamic capability restrictions via skill impact graphs for LLM agents. arXiv preprint arXiv:2608.30041, 2026. URL https://arxiv.org/abs/2608.30041.

[29] Brooklyn Zelenka and UCAN Working Group. User controlled authorization networks (UCAN) specification v0.10. IETF / UCAN Community Specification, 2022. URL https://ucan.xyz/.

[30] Kaiyuan Zhang, Yuke Peng, Ke Jiang, and Yinqian Zhang. ActGov: Governing LLM agent actions via policy-constrained validation. arXiv preprint arXiv:2609.24446, 2026. URL https://arxiv.org/abs/2609.24446.

[31] Wei Zhao, Zhe Li, Peixin Zhang, and Jun Sun. ClawGuard: Deterministic boundary interception for autonomous agents. arXiv preprint arXiv:2604.11790, 2026. URL https://arxiv.org/abs/2604.11790.