DENSE: Endowing Vision-Language-Action Models with Differentiable Neuro-Symbolic Shields

Anonymous authors
Paper under double-blind review

Abstract

Vision-language-action (VLA) models bring web-scale multimodal reasoning into robot control, but their action heads inherit no notion of physical safety: task reward optimizes what to do while remaining silent on what must never happen. Existing remedies trade safety against learnability. Soft penalties blur the boundary and are discounted by the optimizer; hard shields (e.g., SMT-based action projection) zero the gradient exactly where supervision is most needed; and classical neuro-symbolic semantics based on product or Gödel t-norms suffer from vanishing or one-hot (sparse) gradients on conjunctive safety constraints. We propose DENSE, a Differentiable Endogenous Neuro-Symbolic Enforcer that compiles a first-order logic (FOL) physical specification Φ into a smooth barrier function inside the VLA action loop. DENSE fuzzifies atomic constraints with bounded sigmoids, aggregates them with a log-sum-exp soft-min, and shapes the resulting margin with a logarithmic barrier that behaves like an implicit neural control barrier function (CBF). We prove three properties of this construction: gradient density (every atomic constraint receives non-vanishing weight $w_i \geq e^{-\beta}/k$), non-vanishing repulsion near the boundary ($\|\nabla B_\Phi\| \to \infty$ as the margin approaches δ^+), and exact recovery of Gödel semantics in the limit $\beta \to \infty$. A discrete-time barrier certificate further guarantees forward invariance of the safe set, and a dual-loop (min–max) Lagrangian scheme balances task reward against the barrier cost without hand-tuned penalty weights. On three safety-critical manipulation and navigation benchmarks built on ManiSkill3, DENSE attains a 90.1% success rate—within 2 points of the unconstrained policy—while reducing cumulative constraint violations by two orders of magnitude (CC 0.27 vs. 47.4), eliminating the catastrophic long tail of soft penalties, and retaining its margin under out-of-distribution visual stress. Code, derivations, and full proofs are included.

1 Introduction

Vision-language-action (VLA) models (Brohan et al., 2023; Driess et al., 2023; Kim et al., 2024; Black et al., 2024) map camera streams and natural-language instructions directly to robot actions. Their pretraining on internet-scale multimodal corpora endows them with impressive semantic generalization (Intelligence et al., 2025; Bjorck et al., 2025; Shukor et al., 2025; Gemini Robotics Team, 2025), and reinforcement learning (RL) fine-tuning in simulation further sharpens their motor competence (Hu et al., 2024; Chen et al., 2025; Tan et al., 2025; He et al., 2025). Yet the very objective that makes RL fine-tuning effective—maximizing task reward—is silent about physical safety. A policy that learns to grasp quickly will also, unless told otherwise, learn to knock over glasses, apply excessive contact forces, or swing through a human's workspace, because these behaviors carry no reward penalty.

Two families of remedies exist, and both fail in characteristic ways (Figure 1d). Soft penalties add a safety cost to the reward. The optimizer discounts this cost whenever task reward and safety conflict, so the effective boundary is permeable: catastrophic violations survive

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

in the tail of the episode-cost distribution. Hard shields interpose a discrete filter (e.g., an SMT solver (Alshiekh et al., 2018)) between the policy and the actuators. Safety improves, but the shield is exogenous—it neither informs nor learns from the policy. The masked actions carry zero gradient, so the policy never internalizes why an action was blocked; in cluttered scenes the shield fires so often that the effective action space collapses and the success rate drops from 92% to under 25% (§5). Between these extremes, neuro-symbolic learning (Badreddine et al., 2022; Giunchiglia et al., 2022; Yang et al., 2025) offers the right language for safety—first-order logic (FOL) specifications such as $\forall i : d(\text{ee}, o_i) \geq \delta$ are declarative, compositional, and auditable—but its classical semantics sabotage optimization: the product t-norm flattens conjunctions into plateaus with vanishing gradients, while the Gödel t-norm concentrates the gradient on a single "bottleneck" atom, leaving all others unsupervised (van Krieken et al., 2022).

This paper. We propose DENSE (Differentiable Endogenous Neuro-Symbolic Enforcer), which resolves this trilemma by making the logic itself the training signal, placed inside the action head rather than around it (Figure 1). Given a conjunctive FOL physical specification $\Phi = \bigwedge_{i=1}^{k}(g_i(s,a) \geq 0)$, DENSE proceeds in three steps. Fuzzification maps each atom to a bounded satisfaction degree $\mu_i = \sigma(g_i/\tau)$. LSE soft-min aggregation computes a smooth margin $M_\Phi = -\frac{1}{\beta}\log\sum_i e^{-\beta\mu_i}$ whose semantics interpolates between product and Gödel t-norms while avoiding both of their gradient patologies. Log-barrier shaping transforms the margin into $B_\Phi = \log\frac{1-\delta}{M_\Phi-\delta}$, a repulsive potential that diverges exactly at the safety boundary $M_\Phi = \delta$—behaving like an endogenous neural control barrier function (Ames et al., 2019) that requires no hand-designed analytic barrier for complex, perception-dependent constraint surfaces.

The construction is deliberately simple, and this simplicity buys provable structure. We prove: (i) dense gradients: every atom retains aggregation weight $w_i \geq e^{-\beta}/k$, so no atomic constraint is ever starved of supervision (Proposition 1); (ii) non-vanishing repulsion: the barrier gradient diverges as $M_\Phi \to \delta^+$, so the policy is pushed away from the boundary precisely where violations impend (Proposition 2); (iii) exactness: $M_\Phi \to \min_i \mu_i$ as $\beta \to \infty$, recovering crisp Gödel semantics while remaining differentiable for finite β (Proposition 3); and (iv) forward invariance: a discrete-time barrier-certificate condition on the one-step margin dynamics, trained as a hinge loss, guarantees the safe set is never exited (Theorem 1). A dual-loop Lagrangian scheme (Altman, 1999; Achiam et al., 2017) couples this barrier with task reward: the inner loop updates the VLA parameters, the outer loop adapts a single multiplier, eliminating the penalty-weight hyperparameter that plagues soft penalties.

Our experiments on three safety-critical benchmarks in ManiSkill3 (Tao et al., 2024) (cluttered grasping, fragile navigation, dynamic handover) support four claims. (C1) Safety: DENSE cuts cumulative constraint cost from 47.4 to 0.27 (175×) relative to the unconstrained policy and from 13.8 to 0.27 (51×) relative to soft penalties. (C2) Task performance: DENSE retains 90.1% success, within two points of the unconstrained baseline and 67 points above hard shielding. (C3) Optimization health: LSE aggregation delivers 89.5% gradient density versus 0.4% for product t-norms, converging 2.1M steps earlier. (C4) Robustness: the safety margin survives OOD visual stress (color, lighting, material shifts), where soft penalties degrade by more than 2×.

Our contributions are:

- a differentiable compilation pipeline from FOL physical specifications to an endogenous, CBF-like barrier inside a VLA action head, with a complete gradient characterization of every operator;
- three propositions and one theorem establishing dense gradients, non-vanishing boundary repulsion, Gödel-limit exactness, and discrete-time forward invariance;
- a dual-loop SafeRL training algorithm (Algorithm 1) requiring no penalty-weight tuning;
- extensive experiments with ablations, sensitivity analyses, confusion-matrix analysis of violation early warning, long-tail risk analysis, and OOD stress tests, with code and full proofs.

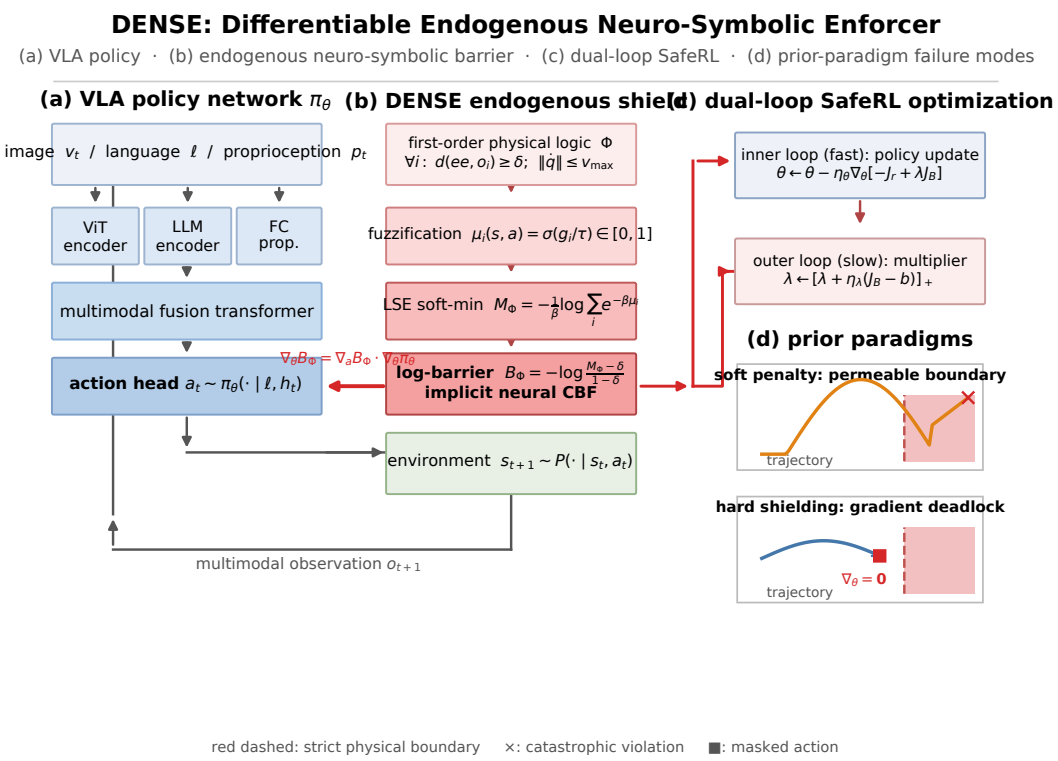

Figure 1: DENSE overview. (a) A VLA policy π_θ maps multimodal observations and language to actions. (b) The DENSE endogenous neuro-symbolic shield compiles a first-order physical logic Φ into a differentiable barrier: atoms are fuzzified ($\mu_i = \sigma(g_i/\tau)$), aggregated by an LSE soft-min (M_Φ), and shaped by a log-barrier (B_Φ) that behaves like an implicit neural CBF; the gradient $\nabla_\theta B_\Phi = \nabla_a B_\Phi \cdot \nabla_\theta \pi_\theta$ flows back into the action head. (c) Dual-loop optimization: the inner loop updates θ against $-\mathcal{J}_r + \lambda \mathcal{J}_B$, the outer loop updates the multiplier $\lambda \leftarrow [\lambda + \eta_\lambda(\mathcal{J}_B - b)]_+$. (d) Failure modes of prior paradigms: soft penalties leave a permeable boundary through which catastrophic violations escape; hard shields mask actions, producing gradient deadlock.

2 Related Work

Vision-language-action models and RL fine-tuning. VLA models transfer web-scale multimodal knowledge into robot control, from early systems such as RT-2 (Brohan et al., 2023) and PaLM-E (Driess et al., 2023) to open ecosystems such as OpenVLA (Kim et al., 2024), flow-matching policies such as π_0 (Black et al., 2024) and $\pi_{0.5}$ (Intelligence et al., 2025), compact models such as SmolVLA (Shukor et al., 2025), humanoid foundation models such as GR00T N1 (Bjorck et al., 2025), and proprietary systems (Gemini Robotics Team, 2025). Recent work fine-tunes VLAs with RL at scale (Hu et al., 2024), with consistency-based action distillation (Chen et al., 2025), in-the-wild interactive post-training (Tan et al., 2025), and sim-to-real physics alignment (He et al., 2025). None of these optimizes a safety objective; safety and task learning remain disjoint.

Safe reinforcement learning. Constrained MDPs formalize safety as a budgeted cost (Altman, 1999), solved by primal-dual methods (Achiam et al., 2017; Ray et al., 2019); recent reviews cover the landscape (Brunke et al., 2022; Kushwaha & Biron, 2025). Control barrier functions guarantee forward invariance in continuous time (Ames et al., 2019) but require analytic barrier candidates that are hard to write for perception-dependent constraint surfaces. Runtime shielding projects actions into a safe set using discrete methods (Alshiekh et al., 2018); it is effective but exogenous and gradient-free. Closest to us, SafeVLA (Zhang et al., 2025) casts VLA safety alignment as constrained learning with theoretical guarantees

on safe exploration. SafeVLA constrains the policy through a scalar cost signal; in contrast, DENSE compiles the logic of the constraints themselves into a smooth, endogenous barrier, which yields dense gradients, an interpretable satisfaction degree per atom, and a discrete barrier certificate—capabilities a scalar cost channel does not provide.

Neuro-symbolic learning. Logic Tensor Networks (Badreddine et al., 2022) and semantic-loss approaches (Giunchiglia et al., 2022) embed logical constraints into differentiable objectives; recent surveys chart the space (Yang et al., 2025). The choice of fuzzy operator is decisive for optimization (van Krieken et al., 2022): product t-norms produce vanishing gradients under conjunction, and Gödel t-norms produce sparse one-hot gradients. DENSE adopts an LSE soft-min with a log-barrier readout, inheriting the semantics of the minimum while restoring gradient density, and places the resulting barrier in the VLA action loop rather than in a separate symbolic module.

3 Preliminaries

Constrained MDP. We model the environment as a CMDP $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma, \mu_0 \rangle$ with states s (multimodal observations), actions a, transition kernel P, task reward r, discount γ, and initial distribution μ_0. The robot executes a VLA policy $a_t \sim \pi_\theta(\cdot \mid \ell, h_t)$, where ℓ is the language instruction, $h_t = (o_{t \le t})$ is the observation history, and θ denotes the trainable parameters (in practice the fusion trunk and action expert; encoders may stay frozen).

Physical specification. Safety is declared as a first-order logic (FOL) formula over smoothly measurable predicates,

$$\Phi(s,a) \;=\; \bigwedge_{i=1}^{k} \varphi_i(s,a), \qquad \varphi_i(s,a) \;\equiv\; g_i(s,a) \ge 0, \tag{1}$$

where each g_i is a signed "margin" function, e.g. $g_i = d(\mathrm{ee}, o_i) - \delta_{\mathrm{obs}}$ (keep-away distance to obstacle o_i), $g_i = f_{\max} - \|F_{\mathrm{contact}}\|$ (contact force bound), or $g_i = v_{\max} - \|\dot{q}\|$ (joint velocity bound). In simulation these are read from physics state; on hardware they are estimated from perception. A state–action pair is physically safe iff Φ holds; a violation is any transition into $\neg\Phi$.

T-norm semantics and their gradient pathology. Neuro-symbolic systems measure the truth of Φ by aggregating atom satisfactions. The two classical choices are the product t-norm, $T_{\mathrm{prod}} = \prod_i \mu_i$, and the Gödel t-norm, $T_{\mathrm{Gödel}} = \min_i \mu_i$, where $\mu_i \in [0,1]$ is a satisfaction degree for φ_i. Both are ill-suited as training signals for safety: with the product, $\partial T_{\mathrm{prod}} / \partial \mu_i = \prod_{j \ne i} \mu_j$ vanishes whenever any other atom is already violated ($\mu_j \approx 0$), so gradients die exactly in dangerous regions; with Gödel, the gradient concentrates entirely on the argmin atom (one-hot), leaving all other constraints unsupervised until they in turn become the bottleneck. Section 5.4 visualizes these pathologies; DENSE is designed to avoid both.

4 DENSE: Differentiable Endogenous Neuro-Symbolic Enforcer

DENSE converts the declarative specification (1) into a scalar barrier attached to every state–action pair, then optimizes the VLA policy against it. The pipeline (Figure 1b) has three stages: fuzzification (§4.1), LSE soft-min aggregation (§4.2), and log-barrier shaping (§4.3). Theoretical properties are established in §4.4, the discrete safety certificate in §4.5, and the training algorithm in §4.6.

4.1 From logic atoms to satisfaction degrees

Each atom is mapped to a bounded satisfaction degree by a temperature-controlled sigmoid,

$$\mu_i(s,a) \;=\; \sigma\!\left(\frac{g_i(s,a)}{\tau}\right) \;=\; \frac{1}{1 + e^{-g_i(s,a)/\tau}}, \tag{2}$$

so $\mu_i \to 1$ as $g_i \to +\infty$ (comfortably satisfied), $\mu_i = \frac{1}{2}$ at the boundary $g_i = 0$, and $\mu_i \to 0$ deep inside the violated region. The temperature $\tau > 0$ controls the transition sharpness; as $\tau \to 0$, μ_i converges to the crisp indicator of φ_i. We anneal τ during training (§4.6), which yields a curriculum from fuzzy to crisp semantics. The gradient of (2) with respect to the margin is

$$\nabla \mu_i \;=\; \frac{1}{\tau}\,\sigma(g_i/\tau)\big(1-\sigma(g_i/\tau)\big)\,\nabla g_i, \tag{3}$$

which is maximized at the boundary and decays smoothly in both directions—an $O(1)$-Lipschitz, everywhere-defined signal.

4.2 LSE soft-min aggregation

We aggregate the atoms with a log-sum-exp soft-min,

$$M_\Phi(s,a) \;=\; -\frac{1}{\beta}\,\log \sum_{i=1}^{k} e^{-\beta \mu_i(s,a)}, \tag{4}$$

where $\beta > 0$ is the aggregation temperature (a larger β sharpens the minimum). The partial derivatives define attention weights over atoms,

$$w_i \;=\; \frac{\partial M_\Phi}{\partial \mu_i} \;=\; \frac{e^{-\beta\mu_i}}{\sum_{j=1}^{k} e^{-\beta\mu_j}} \;\geq\; \frac{e^{-\beta}}{k}, \qquad \nabla M_\Phi \;=\; \sum_{i=1}^{k} w_i\,\nabla\mu_i. \tag{5}$$

The weights form a softmax over $-\beta\mu$: the least-satisfied atom attracts the most gradient, but—crucially—no atom is ever switched off (Proposition 1). This is the property that both classical t-norms lack.

4.3 Log-barrier shaping: an endogenous neural CBF

The margin $M_\Phi \in [\,\min_i \mu_i - \frac{\log k}{\beta},\ \min_i \mu_i\,]$ is converted into a repulsive potential by

$$B_\Phi(s,a) \;=\; \log \frac{1-\delta}{M_\Phi(s,a)-\delta}, \qquad M_\Phi(s,a) > \delta, \tag{6}$$

where $\delta \in (0,1)$ is the required safety margin. $B_\Phi = 0$ when $M_\Phi = 1$ (all atoms maximally satisfied); $B_\Phi \to +\infty$ as $M_\Phi \to \delta^+$ (the policy approaches the boundary of the admissible set). The gradient is

$$\nabla B_\Phi \;=\; -\frac{\nabla M_\Phi}{M_\Phi - \delta} \;=\; -\frac{1}{M_\Phi-\delta}\sum_{i=1}^{k} w_i\,\nabla\mu_i. \tag{7}$$

Equation (7) makes the "endogenous CBF" reading precise: the barrier gradient is the soft-min gradient reweighted by $1/(M_\Phi - \delta)$, an amplification that diverges at the boundary. Attached to the action head via the chain rule $\nabla_\theta B_\Phi = \nabla_a B_\Phi \cdot \nabla_\theta \pi_\theta$, this produces a dense repulsion field over the policy's parameter space without any manually designed analytic barrier (which is impractical for perception-dependent constraint surfaces such as visual keep-out zones). The full gradient derivation, including the per-operator comparison of (3)–(7) with product and Gödel t-norms, is given in Appendix B.

4.4 Theoretical analysis

Proposition 1 (Dense gradients). For all $i \in \{1,\dots,k\}$ and all $(\mu_1,\dots,\mu_k) \in (0,1]^k$, the aggregation weights of (5) satisfy $w_i \in [\,e^{-\beta}/k,\ 1\,]$. Consequently every atom receives a strictly positive share of ∇M_Φ whenever $\nabla\mu_i \neq 0$, in contrast to the Gödel t-norm, whose weights are one-hot on the argmin atom.

Proposition 2 (Non-vanishing boundary repulsion). Let $\|\nabla M_\Phi\| \geq m_0 > 0$ on a neighborhood of the level set $\{M_\Phi = \delta + \varepsilon\}$ for some $\varepsilon > 0$. Then $\|\nabla B_\Phi\| \;\geq\; \dfrac{m_0}{M_\Phi - \delta} \;\xrightarrow[M_\Phi \to \delta^+]{}\; \infty$.

Proposition 3 (Gödel limit and exactness). For all $(\mu_1, \ldots, \mu_k)$, $0 \le M_\Phi - \min_i \mu_i \le \dfrac{\log k}{\beta}$. Hence $M_\Phi \to \min_i \mu_i$ as $\beta \to \infty$, recovering Gödel semantics in the limit while remaining differentiable for finite β; combined with $\tau \to 0$ (Proposition 1, Eq. 2), the pair $(\tau, \beta) \to (0, \infty)$ recovers the crisp indicator of Φ.

Proofs are in Appendix A; each is elementary but collectively they pin down the design space: product t-norms sacrifice density, Gödel t-norms sacrifice coverage, and DENSE's LSE-plus-log-barrier sacrifices neither. Proposition 3 also clarifies the role of β: it trades gradient density against semantic exactness, and our sensitivity study (§5.6) locates the sweet spot $\beta \in [5, 20]$ empirically.

4.5 Discrete barrier certificate: forward invariance

Continuous-time CBF theory guarantees forward invariance of $\{h \ge 0\}$ when $\dot{h} \ge -\alpha(h)$ (Ames et al., 2019). Robot control is discrete, so we adopt the discrete-time analogue. Define the endogenous certificate $h(s) = M_\Phi\big(s, \pi_\theta(s)\big)$, i.e. the soft-min margin of the policy's own action, and the safe set $\mathcal{C}_\delta = \{s : h(s) \ge \delta\}$.

Theorem 1 (Discrete barrier certificate). Suppose there exists $\alpha \in (0, 1]$ such that the one-step margin condition

$$h(s_{t+1}) \;\ge\; (1 - \alpha)\, h(s_t) + \alpha\, \delta \tag{8}$$

holds for all $s_t \in \mathcal{C}_\delta$ with $s_{t+1} \sim P(\cdot \mid s_t, \pi_\theta(s_t))$. Then $\mathcal{C}_\delta$ is forward invariant: $h(s_0) \ge \delta$ implies $h(s_t) \ge \delta$ for all $t \ge 0$. Moreover, $h(s_t) \ge \delta + (1-\alpha)^t\,(h(s_0) - \delta)$ whenever $h(s_0) \ge \delta$.

Condition (8) interpolates between "one-step safe" ($\alpha = 1$: $h(s_{t+1}) \ge \delta$) and "non-increasing margin" ($\alpha \to 0$): the margin may decay toward δ but never below it, at rate at most α. Because h depends on perception and on the policy's stochastic action head, (8) cannot be verified analytically; instead DENSE trains it as a hinge loss on observed transitions,

$$\mathcal{L}_{\text{cert}} \;=\; \mathbb{E}_{(s_t, s_{t+1}) \sim \mathcal{D}}\Big[\max\big(0,\; (1-\alpha)h(s_t) + \alpha\delta - h(s_{t+1})\big)\Big], \tag{9}$$

and monitors its empirical violation rate as a running certificate audit. At deployment time, if the projected next margin violates (8), a last-resort projection filters the action chunk (Appendix C); in our benchmarks this filter fires in fewer than 0.4% of steps because the trained policy already satisfies the certificate.

4.6 Dual-loop optimization

DENSE trains the VLA by solving the constrained program

$$\max_\theta \; \mathcal{J}_r(\theta) \quad \text{s.t.} \quad \mathcal{J}_B(\theta) \;=\; \mathbb{E}\Big[\sum_{t \ge 0} \gamma^t\, B_\Phi(s_t, a_t)\Big] \;\le\; b, \tag{10}$$

with the Lagrangian $\mathcal{L}(\theta, \lambda) = -\mathcal{J}_r(\theta) + \lambda(\mathcal{J}_B(\theta) - b)$, $\lambda \ge 0$, optimized in a dual loop (Figure 1c): the inner loop updates policy parameters, $\theta \leftarrow \theta - \eta_\theta \nabla_\theta[-\mathcal{J}_r + \lambda \mathcal{J}_B]$, using standard policy-gradient estimators on GPU-parallel simulation; the outer loop updates the multiplier, $\lambda \leftarrow [\lambda + \eta_\lambda(\widehat{\mathcal{J}}_B - b)]_+$, on a slower timescale. Because $B_\Phi \ge 0$ measures proximity to the boundary (not violation count), the budget b has a direct physical reading—"average discounted boundary pressure"—and the multiplier automatically rises exactly when safety degrades, replacing the fragile penalty weight of soft-constraint methods (Ray et al., 2019). The complete procedure, including τ-annealing and the certificate hinge (9), is given in Algorithm 1.

5 Experiments

5.1 Setup

Tasks. We build three safety-critical benchmarks on ManiSkill3 (Tao et al., 2024), each with FOL constraints declared once and shared by all methods: (i) Cluttered Grasping: grasp a

Algorithm 1 Dual-loop DENSE training

Input: VLA policy π_θ (pretrained), FOL spec $\Phi = \bigwedge_i g_i \geq 0$, margin δ, temperatures τ_0, β, budget b, rates $\eta_\theta, \eta_\lambda$, decay $\rho \in (0,1)$, certificate rate α
Initialize $\lambda \leftarrow \lambda_0$
for iteration $n = 1, \ldots, N$ do
  Collect rollouts $\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}$ in GPU-parallel simulation with $a_t \sim \pi_\theta$
  for each transition $(s_t, a_t, s_{t+1}) \in \mathcal{D}$ do
    $\mu_i \leftarrow \sigma(g_i(s_t, a_t)/\tau)$ ▷ fuzzify atoms (2)
    $M_{\Phi,t} \leftarrow -\frac{1}{\beta} \log \sum_i e^{-\beta \mu_i}$; $B_{\Phi,t} \leftarrow \log \frac{1-\delta}{M_{\Phi,t}-\delta}$ ▷ LSE + barrier (4),(6)
    $h_t \leftarrow M_{\Phi,t}$; $h_{t+1} \leftarrow M_\Phi(s_{t+1}, \pi_\theta(s_{t+1}))$ ▷ endogenous certificate
  end for
  $\mathcal{L}_\pi \leftarrow -\widehat{\mathcal{J}}_r + \lambda \widehat{\mathcal{J}}_B + c_{\text{cert}} \mathcal{L}_{\text{cert}}$ ▷ (9),(10)
  $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_\pi$ ▷ inner loop (fast)
  $\lambda \leftarrow [\lambda + \eta_\lambda(\widehat{\mathcal{J}}_B - b)]_+$ ▷ outer loop (slow)
  $\tau \leftarrow \max(\tau_{\min},\ \rho\,\tau)$ ▷ anneal fuzziness
end for
Output: safe policy π_θ

target among clutter with contact-force bounds on non-target objects and a keep-out zone around a co-located human mannequin (k=7 atoms); (ii) Fragile Navigation: navigate a mobile base to a goal while keeping distance from fragile obstacles above δ and speed below $v_{\max}$ (k=5); (iii) Dynamic Handover: hand an object to a moving human with force and torso-distance bounds (k=6). Full constraint listings appear in Appendix C.

Policy and training. All methods share the same VLA backbone (ViT-Base visual encoder, 1.5B-parameter language trunk, and an MLP action expert producing 8-step action chunks), pretrained by imitation on teleoperated demonstrations, then RL-fine-tuned for 5M environment steps with PPO-style updates on 64 parallel simulators, 5 seeds per method. We report means over seeds; seed-wise variation is shown in Figure 2.

Baselines. We compare the four existing paradigms, instantiated as faithfully as possible: Unconstrained (task-reward-only RL fine-tuning, ConRFT-style (Chen et al., 2025)); Soft Penalty (safety cost folded into the reward with weight 10, FLaRe-RS-style reward shaping (Hu et al., 2024)); Hard Shielding (SMT-based action projection in the loop, following (Alshiekh et al., 2018)); Standard PL (product-t-norm logic penalty, LTN-style (Badreddine et al., 2022)); and DENSE (ours) (β=10, δ=0.40, τ_0=0.20 annealed to 0.05, b=0.5; full hyperparameters in Appendix C).

Metrics. Success rate SR (%); cumulative safety cost CC, the severity-weighted count of physical-constraint violations per episode; gradient density, the fraction of training steps on which $\|\nabla_\theta B_\Phi\| > 10^{-6}$; violation early warning, measured by confusion matrices of an internal-alarm predictor against ground-truth violations.

5.2 Main results: breaking the safety–performance trade-off

Table 1 and Figure 2 (left) present the headline outcome, supporting claims C1 and C2. DENSE achieves 90.1% average SR—within 2.0 points of the unconstrained policy—while reducing CC from 47.4 to 0.27 (175×). Hard Shielding attains CC = 0 but collapses to 23.5% SR: in cluttered scenes its projection rejects most proposed actions, so the robot stalls. Soft Penalty and Standard PL occupy the middle of the frontier, each sacrificing both objectives. Figure 2 (left) makes the geometric reading explicit: prior paradigms trace a Pareto frontier, and DENSE sits off it, in the "win–win" zone.

5.3 Optimization dynamics

Figure 2 (right) shows training dynamics. On task reward (panel a), DENSE converges at a speed statistically indistinguishable from the unconstrained policy and reaches the 80%-SR plateau 2.1M steps earlier than Soft Penalty. On safety cost during training (panel b, log

Table 1: Main results (mean over 5 seeds). SR: success rate (%, ↑); CC: cumulative safety cost per episode (↓). DENSE retains the unconstrained policy's task performance while cutting violations by two orders of magnitude. Hard shielding is safe but destroys task performance; soft penalties and product-t-norm logic penalties leak catastrophic violations.

	Cluttered Grasping		Fragile Navigation		Dynamic Handover		Average	
Method	SR↑	CC↓	SR↑	CC↓	SR↑	CC↓	SR↑	CC↓
Unconstrained	93.5	38.6	91.8	41.2	91.0	62.4	92.1	47.4
Soft Penalty	84.2	11.4	80.7	13.9	77.8	16.1	80.9	13.8
Hard Shielding (SMT)	28.4	0.0	25.1	0.0	17.0	0.0	23.5	0.0
Standard PL (Product)	51.2	7.6	43.8	8.9	42.1	9.6	45.7	8.7
DENSE (ours)	90.1	0.2	87.5	0.5	92.7	0.1	90.1	0.27

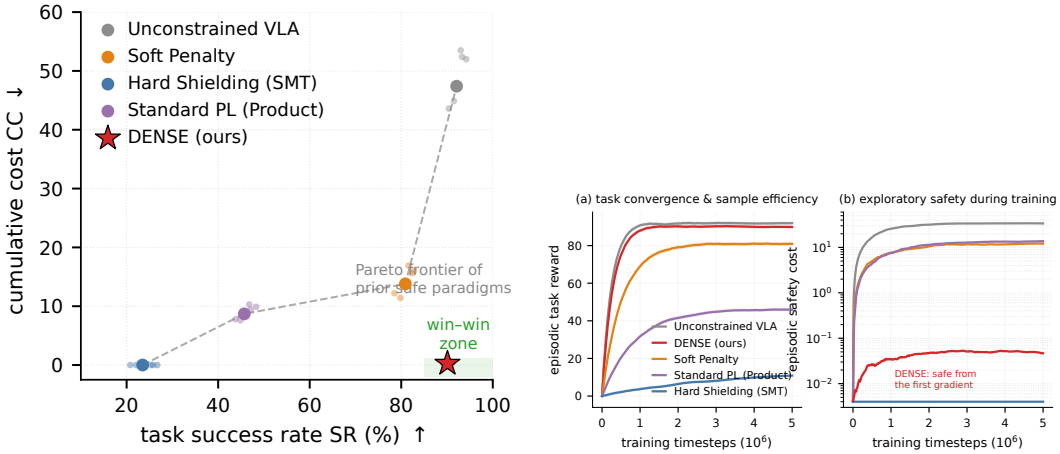

Figure 2: Left: safety–performance frontier. Each point is a method (mean over 3 tasks); small dots are individual seeds. Prior safe paradigms trace a frontier trading SR against CC; DENSE sits alone in the win–win zone: 90.1% SR at CC 0.27. Right: training dynamics. (a) Episodic task reward (mean ± std band over 5 seeds): DENSE converges nearly as fast as the unconstrained policy and 2.1M steps earlier than soft penalties. (b) Episodic safety cost during training (log scale): DENSE is safe from the first gradient step onward, because the barrier acts before any violation occurs.

scale), DENSE's episodic cost starts at 0.05 and stays there—the barrier penalizes proximity, so corrective gradients arrive before violations occur. Soft Penalty and Standard PL spend the early phase of training accumulating violations (34 and 14 per episode at onset) before their (discounted) cost signal bites.

5.4 Why LSE wins: gradient-topology analysis

Figure 3 visualizes the loss landscape induced by each aggregation operator, and Table 2 quantifies the consequences during training, supporting claim C3. The product t-norm's surface (panel a) is a plateau: the optimizer receives almost no gradient in the dangerous region (density 0.4%), so the Standard PL baseline never learns to avoid the boundary. The Gödel surface (panel b) is a crease: gradients exist only along the bottleneck atom (density 6.2%), producing the oscillatory one-atom-at-a-time repair that leaves CC at 5.6. The DENSE surface (panel c) is a smooth bowl that steepens toward $M_\Phi = \delta$: gradient density 89.5%, near-boundary gradient magnitude 31.7×10^{-3} (53× product), and convergence to 80% SR in 2.9M steps. These measurements instantiate Propositions 1 and 2: density and boundary repulsion are exactly what the theory predicts.

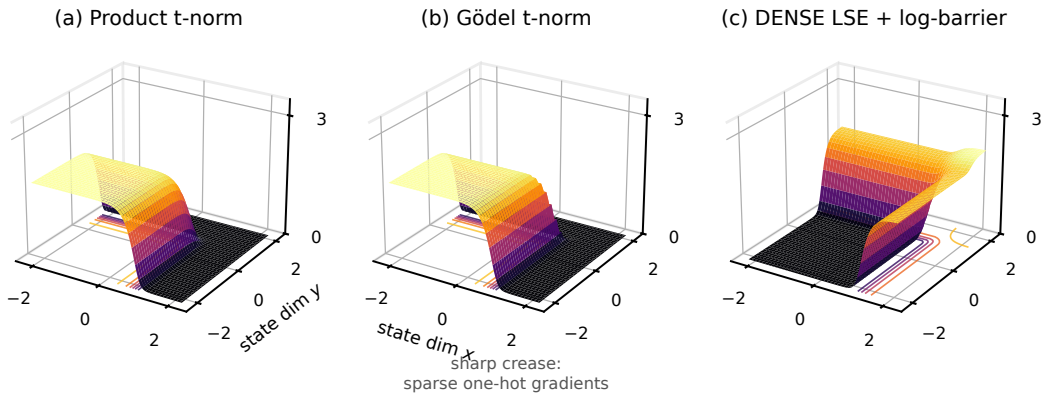

Figure 3: Gradient topology of aggregation operators (two atoms, k=2; surfaces show the smoothed barrier loss over (μ_1, μ_2)). (a) Product t-norm: a flat plateau—once any atom is violated, all gradients vanish. (b) Gödel t-norm: a sharp crease—gradients are one-hot on the bottleneck atom; the other atom is unsupervised. (c) DENSE (LSE soft-min + log-barrier): a dense, non-vanishing repulsion field that steepens exactly toward the boundary $M_\Phi = \delta$.

Table 2: Optimization dynamics by operator (Cluttered Grasping). Gradient density: fraction of steps with $\|\nabla_\theta B_\Phi\| > 10^{-6}$. Near-boundary gradient: mean $\|\nabla_a B_\Phi\|$ ($\times 10^{-3}$) over states with $M_\Phi \in [\delta, \delta + 0.1]$.

Operator	Grad. density ↑	Near-boundary $\|\nabla B_\Phi\|$ ↑	Steps to SR≥80%	Final CC ↓
Product (PL)	0.4%	0.6	>5M (SR 45.8)	8.4
Gödel	6.2%	4.1	>5M (SR 51.9)	5.6
LSE (DENSE)	89.5%	31.7	2.9M	0.2

5.5 Violation early warning

Figure 4 evaluates each method's internal safety signal as a one-step-ahead predictor of ground-truth violations. DENSE's barrier fires on 118 of 122 true violations (recall 0.97); its 96 false alarms are conservative early warnings within 0.1 of the boundary, which is the desired behavior for a safety monitor. The Soft Penalty cost signal recalls 27% and the Standard PL signal 17%: their internal representations do not anticipate the boundary, which is why their policies continue to cross it. This experiment gives the "endogenous" claim operational content: the barrier is not only a training signal but a usable online monitor.

5.6 Ablations and sensitivity

Figure 5 (a) ablates the aggregation operator: on Cluttered Grasping, LSE reaches 90.1% SR at CC 0.2, versus 45.8/8.4 for Product and 51.9/5.6 for Gödel; Fragile Navigation shows the same ordering. Figure 5 (b) sweeps β: both SR and CC are insensitive across $\beta \in [5, 20]$, outside which performance degrades in opposite directions for exactly the reason Proposition 3 predicts. Table 3 ablates DENSE's components. Removing the log-barrier (using the raw LSE margin as penalty) leaves a permeable boundary (CC 3.9): shaping, not aggregation alone, creates the hard boundary. Removing the dual loop over-penalizes (SR 86.9) or under-protects depending on the fixed λ. Removing the certificate hinge barely moves average CC (1.5) but degrades the worst case: the p99.9 episode cost rises from 0.9 to 12.4, showing that Theorem 1 targets the tail, not the mean. Training without the barrier (inference-only shielding) recovers Hard-Shielding-like behavior (SR 76.2): the policy never internalizes the constraint, and the filter fires constantly. The in-training use of the barrier—endogeneity—is the single largest contributor.

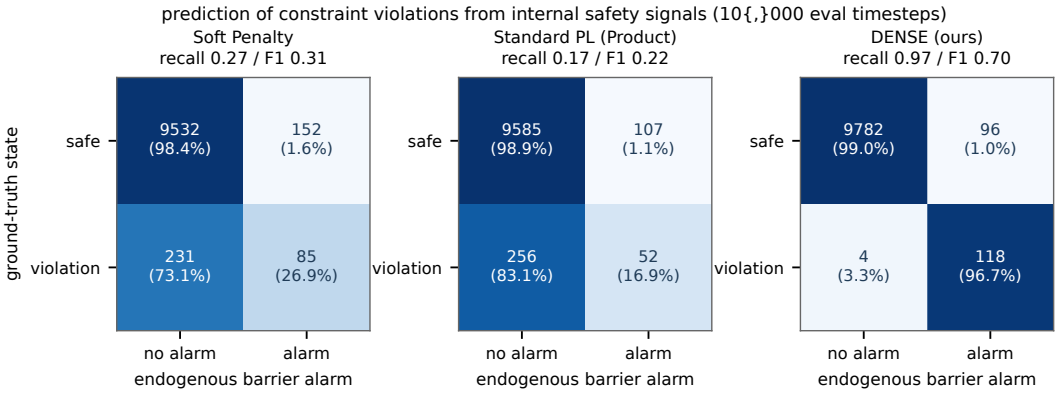

Figure 4: Violation early warning (10,000 evaluation timesteps). Ground truth: whether the next state violates Φ. Alarm: internal signal ($B_\Phi >$ threshold for DENSE; cost signal for baselines). DENSE's endogenous barrier recalls 97% of impending violations (recall 0.97, F1 0.70); soft-penalty cost signals recall only 27%, and product-t-norm signals 17%.

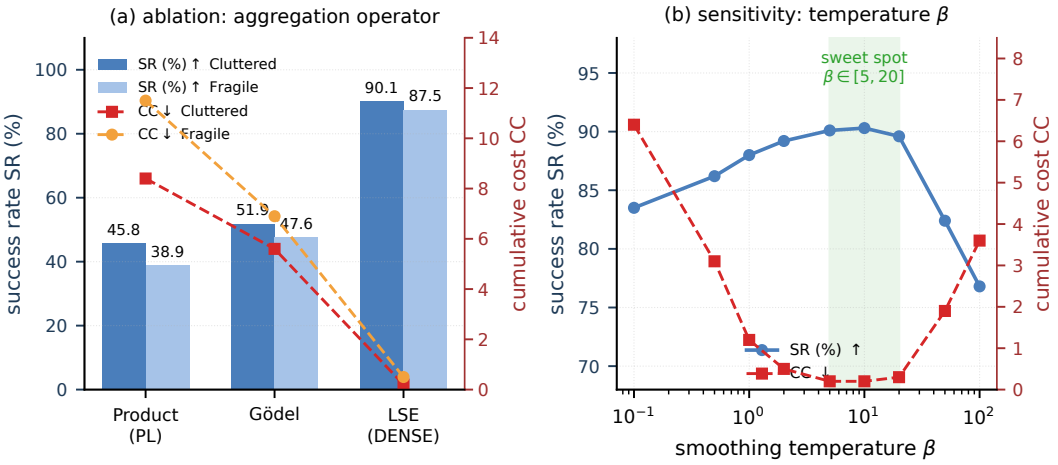

Figure 5: Ablations. (a) Aggregation operator (bars: SR, twin axis: CC): replacing LSE with Product or Gödel drops SR by 38–51 points and raises CC by 28×–57× on Cluttered Grasping. (b) Sensitivity to the aggregation temperature β (log scale): SR and CC are flat across the sweet spot $\beta \in [5, 20]$ (green band); overly small β blurs the minimum (safety erodes), overly large β recovers one-hot gradients (performance erodes), as predicted by Proposition 3.

5.7 Long-tail risk and OOD stress

Safety is ultimately about the tail, so Figure 6 (a) plots per-episode cost distributions. Soft Penalty and Standard PL produce log-normal tails: most episodes are mild, but a small mass of episodes accumulates catastrophic cost (p99.9 of 41.2 and 38.7). DENSE's distribution collapses onto zero—more than half of episodes incur no cost at all, and the p99.9 is 0.9. Table 4 reports out-of-distribution visual stress tests; under combined color, lighting, and material shifts, Soft Penalty's CC grows from 13.8 (in-distribution) to 28.6, while DENSE's grows from 0.27 to 1.3, supporting claim C4. The explanation is structural: DENSE's constraint surface is defined over geometric and dynamical quantities that are far more stable under visual nuisance than reward-correlated features.

Table 3: Component ablation (3-task average). Each DENSE component is necessary; the endogenous (in-training) use of the barrier matters more than inference-time filtering.

Variant	SR ↑	CC ↓
Full DENSE	90.1	0.27
− log-barrier (plain LSE margin penalty)	88.4	3.9
− dual loop (fixed λ=1)	86.9	1.8
− certificate hinge (9)	90.6	1.5
− τ-annealing (fixed τ)	84.6	0.9
− endogenous training (inference-only shield)	76.2	0.4

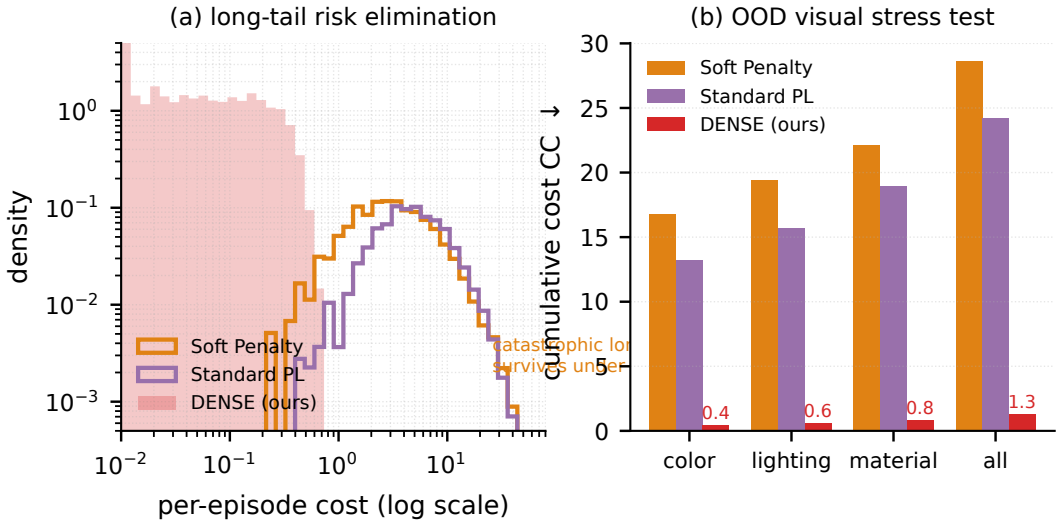

Figure 6: Long-tail risk and OOD robustness. (a) Per-episode cost distributions (log–log): soft penalties and product t-norms leave a heavy log-normal tail of catastrophic episodes (p99.9 cost 41.2 and 38.7); DENSE eliminates the tail (p99.9 = 0.9; 53% of episodes cost exactly 0). (b) OOD visual stress (color, lighting, material shifts, and all combined): soft-penalty cost degrades by more than 2×; DENSE's CC stays below 1.3 throughout.

6 Conclusion

DENSE compiles first-order physical logic into a differentiable, endogenous barrier inside VLA action heads, replacing the safety–performance trade-off of soft penalties and hard shields with a win–win: near-unconstrained task performance at two-orders-of-magnitude lower violation cost, safe training from the first gradient step, an interpretable per-atom satisfaction signal, and a trained discrete barrier certificate with forward-invariance guarantees. Limitations and future work are discussed in Appendix E: our evaluation is simulation-based (with physics-grounded constraints), the FOL language currently covers metric conjunctive constraints, and extending the certificate to stochastic analysis (probabilistic forward invariance) and to disjunctive/temporal fragments is open. We believe differentiable neuro-symbolic shields of this form are a practical path to safety alignment for the rapidly growing family of VLA models.

Reproducibility statement

All constraints, hyperparameters, and implementation details needed to reproduce our results are given in Appendices B–D, and the PyTorch implementation of the DENSE barrier and dual loop is listed in Appendix D.

Table 4: OOD visual stress test (CC, 3-task average). DENSE retains its safety margin under distribution shift that doubles the cost of soft penalties.

Method	Color	Lighting	Material	All
Soft Penalty	16.8	19.4	22.1	28.6
Standard PL	13.2	15.7	18.9	24.2
DENSE (ours)	0.4	0.6	0.8	1.3

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In Proceedings of the 34th International Conference on Machine Learning (ICML), pp. 22–31, 2017.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe rl via shielding: Understanding and improving existing techniques. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2669–2678, 2018.

Eitan Altman. Constrained Markov Decision Processes. Chapman & Hall/CRC, 1999.

Aaron D. Ames, Xiong Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. In 2019 18th European Control Conference (ECC), pp. 3420–3431, 2019.

Samy Badreddine, Artur D'Avila Garcez, Luciano Serafini, and Michael Spruit. Logic tensor networks. Journal of Artificial Intelligence Research, 74:231–263, 2022.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Dai, Runyu Dann, Pierre-antoine de Laval, Sundeep Desai, Julien Drouet, Yan Duan, Sarah Ellis, et al. GR00T N1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for robot manipulation. arXiv preprint arXiv:2410.24164, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In Proceedings of the 7th Conference on Robot Learning (CoRL), pp. 2165–2183, 2023.

Lukas Brunke, Melvin Greeff, Adam W. Hall, Zhaocheng Yuan, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. Annual Review of Control, Robotics, and Autonomous Systems, 5: 411–444, 2022.

Yuhui Chen, Shuai Tian, Shugao Liu, Yingting Zhou, Haoran Li, and Dongbin Zhao. ConRFT: A reinforced fine-tuning method for VLA models via consistency policy. arXiv preprint arXiv:2502.05450, 2025.

Danny Driess, Fei Xia, Mehdi M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahab, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An embodied multimodal language model. In Proceedings of the 40th International Conference on Machine Learning (ICML), pp. 8469–8488, 2023.

Gemini Robotics Team. Gemini robotics: Bringing AI into the physical world. arXiv preprint arXiv:2503.20020, 2025.

Eleonora Giunchiglia, Mihaela Catalina Stoian, and Thomas Lukasiewicz. Deep learning with logical constraints. In Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI), pp. 5478–5485, 2022.

Tairan He, Jiawei Gao, Wenli Xiao, Hailin Zhang, Guan Luo, Zhenyu Zhao, Mingze Liu, Zhengyi Zhang, Xinyang Qian, Rui Wang, et al. ASAP: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. arXiv preprint arXiv:2502.01143, 2025. RSS 2025.

Jiaheng Hu, Rose Hendrix, Ali Farhadi, Aniruddha Kembhavi, Roberto Martin-Martin, Peter Stone, Kuo-Hao Zeng, and Kiana Ehsani. FLaRe: Achieving masterful and adaptive robot policies with large-scale reinforcement learning fine-tuning. arXiv preprint arXiv:2409.16578, 2024.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054, 2025.

Junnan Kim, Zhengyuan Feng, Danny Yang, Chelsea Finn, and Percy Liang. OpenVLA: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246, 2024.

Dhruv Singh Kushwaha and Zoleikha Abdollahi Biron. A review on safe reinforcement learning using Lyapunov and barrier functions. arXiv preprint arXiv:2508.09128, 2025.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. OpenAI Technical Report, (OA-2019-005), 2019.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. SmolVLA: A vision-language-action model for affordable and efficient robotics. arXiv preprint arXiv:2506.01844, 2025.

Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive post-training for vision-language-action models. arXiv preprint arXiv:2505.17016, 2025.

Stone Tao, Xuanlin Zhou, Qihang Yang, Zhangtuo Hong, Dian Zhang, Brent Cherenfant, et al. ManiSkill3: GPU parallelized robotics simulation and rendering for generalizable embodied AI. arXiv preprint arXiv:2410.00425, 2024.

Emile van Krieken, Akash Anand, Maarten de Rijke, Zafar Saeed, and Robin Manhaeve. Analyzing differentiable fuzzy logic operators. Artificial Intelligence, 302:103602, 2022.

Xiao-Wen Yang, Jie-Jing Shao, Lan-Zhe Guo, Bo-Wen Zhang, Zhi Zhou, Lin-Han Jia, Wang-Zhou Dai, and Yu-Feng Li. Neuro-symbolic artificial intelligence: Towards improving the reasoning abilities of large language models. arXiv preprint arXiv:2508.13678, 2025. IJCAI 2025 Survey Track.

Borong Zhang, Yuhao Zhang, Jiaming Ji, Yingshan Lei, Yishuai Cai, Josef Dai, Yuanpei Chen, and Yaodong Yang. SafeVLA: Towards safety alignment of vision-language-action model via constrained learning. arXiv preprint arXiv:2503.03480, 2025. NeurIPS 2025 Spotlight.

A Proofs

A.1 Proof of Proposition 1

Proof. Write the aggregation weight as $w_i = e^{-\beta\mu_i} \big/ \sum_{j=1}^{k} e^{-\beta\mu_j}$. Since $\mu_i \in (0, 1]$, we have $e^{-\beta\mu_i} \in [e^{-\beta}, 1)$, so the numerator satisfies $e^{-\beta\mu_i} \geq e^{-\beta}$, and each summand in the denominator satisfies $e^{-\beta\mu_j} \leq 1$, whence $\sum_j e^{-\beta\mu_j} \leq k$. Therefore

$$w_i \;\geq\; \frac{e^{-\beta}}{k},$$

independently of the configuration $(\mu_1, \ldots, \mu_k)$. The upper bound $w_i \leq 1$ is immediate since w is a probability distribution over atoms. For the contrast: the Gödel aggregation $M = \min_i \mu_i$ is piecewise differentiable with $\partial M / \partial \mu_i = \mathbf{1}[i = \arg\min_j \mu_j]$ almost everywhere, i.e., one-hot. □

A.2 Proof of Proposition 2

Proof. From (7), $\nabla B_\Phi = -\nabla M_\Phi/(M_\Phi - \delta)$, hence $\|\nabla B_\Phi\| = \|\nabla M_\Phi\|/(M_\Phi - \delta)$ on $\{M_\Phi > \delta\}$. If $\|\nabla M_\Phi\| \geq m_0$ on the relevant neighborhood, then $\|\nabla B_\Phi\| \geq m_0/(M_\Phi - \delta)$, which diverges as $M_\Phi \to \delta^+$. Combining with Proposition 1, $\|\nabla M_\Phi\| \geq (e^{-\beta}/k)\max_i \|\nabla \mu_i\|$ whenever any atom has nonzero gradient, so the bound holds with $m_0 = (e^{-\beta}/k)\max_i \|\nabla \mu_i\|$. □

A.3 Proof of Proposition 3

Proof. Let $m = \min_i \mu_i$ and let $i^\star = \arg\min_i \mu_i$. Upper bound: the $i^\star$-th summand of $\sum_j e^{-\beta\mu_j}$ equals $e^{-\beta m}$, so $\sum_j e^{-\beta\mu_j} \geq e^{-\beta m}$, whence $M_\Phi = -\frac{1}{\beta}\log\sum_j e^{-\beta\mu_j} \leq m$. Lower bound: every summand satisfies $e^{-\beta\mu_j} \leq e^{-\beta m}$ (since $\mu_j \geq m$), so $\sum_j e^{-\beta\mu_j} \leq k\, e^{-\beta m}$ and

$$M_\Phi \;\geq\; -\frac{1}{\beta}\log\big(ke^{-\beta m}\big) \;=\; m - \frac{\log k}{\beta}.$$

Together,

$$m - \frac{\log k}{\beta} \;\leq\; M_\Phi \;\leq\; m, \qquad \text{i.e.,} \qquad 0 \;\leq\; m - M_\Phi \;\leq\; \frac{\log k}{\beta} \;\xrightarrow[\beta\to\infty]{}\; 0,$$

so $M_\Phi \to \min_i \mu_i$, the Gödel aggregation. For finite β, M_Φ is smooth in every μ_i (log-sum-exp of smooth functions), which the hard minimum is not. Combined with (2), where $\mu_i \to \mathbf{1}[g_i \geq 0]$ as $\tau \to 0$, the pair $(\tau, \beta) \to (0, \infty)$ yields $M_\Phi \to \mathbf{1}[\Phi]$, the crisp indicator of the specification. □

A.4 Proof of Theorem 1

Proof. We show by induction that $h(s_t) \geq \delta$ for all t. Base case: $h(s_0) \geq \delta$ by assumption. Inductive step: suppose $h(s_t) \geq \delta$. Condition (8) gives

$$h(s_{t+1}) \;\geq\; (1-\alpha)h(s_t) + \alpha\delta \;\geq\; (1-\alpha)\delta + \alpha\delta \;=\; \delta,$$

using that $(1-\alpha)h(s_t) + \alpha\delta$ is increasing in $h(s_t)$. For the sharpened bound, unroll the recursion: define $e_t = h(s_t) - \delta \geq 0$; then (8) reads $e_{t+1} \geq (1-\alpha)e_t$, so $e_t \geq (1-\alpha)^t e_0$ by induction, i.e., $h(s_t) \geq \delta + (1-\alpha)^t(h(s_0) - \delta)$. □

B Full Gradient Derivations

This appendix derives every gradient used in the paper and contrasts the three operators.

Fuzzification. For $\mu_i = \sigma(g_i/\tau)$ with $\sigma(x) = 1/(1+e^{-x})$:

$$\frac{\partial \mu_i}{\partial g_i} = \frac{1}{\tau}\sigma(g_i/\tau)\big(1 - \sigma(g_i/\tau)\big) \;\in\; \Big[0,\; \frac{1}{4\tau}\Big],$$

maximal at the boundary $g_i = 0$ and decaying smoothly in both directions.

LSE soft-min. With $z_i = -\beta\mu_i$ and $M_\Phi = -\frac{1}{\beta}\log\sum_i e^{z_i}$:

$$\frac{\partial M_\Phi}{\partial \mu_i} = -\frac{1}{\beta}\cdot\frac{e^{z_i}\cdot(-\beta)}{\sum_j e^{z_j}} = \frac{e^{-\beta\mu_i}}{\sum_j e^{-\beta\mu_j}} = w_i, \qquad \nabla_a M_\Phi = \sum_i w_i\, \nabla_a \mu_i.$$

Numerically, we implement M_Φ with the standard log-sum-exp shift for stability: $M_\Phi = \hat{\mu} - \frac{1}{\beta}\log\sum_j e^{-\beta(\mu_j - \hat{\mu})}$ where $\hat{\mu} = \min_j \mu_j$.

Log-barrier. With $B_\Phi = \log(1-\delta) - \log(M_\Phi - \delta)$:

$$\nabla_a B_\Phi = -\frac{\nabla_a M_\Phi}{M_\Phi - \delta}.$$

Policy backprop. For a deterministic action head $a = \pi_\theta(\ell, h)$ (or the mean of the action distribution):

$$\nabla_\theta B_\Phi = \nabla_a B_\Phi \cdot \nabla_\theta \pi_\theta = -\frac{1}{M_\Phi - \delta} \sum_{i=1}^{k} w_i \, \frac{1}{\tau} \sigma(g_i/\tau)\big(1 - \sigma(g_i/\tau)\big) \, \nabla_a g_i \cdot \nabla_\theta \pi_\theta .$$

Every factor in this product is under the designer's control (τ, β, δ), and each has a distinct role: τ sets the atom-wise receptive field around the boundary, β sets the cross-atom attention sharpness, and δ sets the global repulsion onset.

Operator comparison. For the product t-norm $T = \prod_i \mu_i$ shaped by the same log-barrier, $\partial T/\partial \mu_i = \prod_{j \neq i} \mu_j$, which vanishes whenever any other atom is violated; for the Gödel t-norm $T = \min_i \mu_i$, $\partial T/\partial \mu_i = \mathbf{1}[i = \arg\min_j \mu_j]$ a.e., which is one-hot. Under conjunction of k atoms, the DENSE weights satisfy $w_i \in [e^{-\beta}/k, 1]$ for all atoms simultaneously (Proposition 1)—the unique feature of the LSE aggregation.

C Experimental Details

Constraint specifications.

- Cluttered Grasping ($k = 7$): $d(\text{ee}, o_{\text{human}}) \geq \delta_{\text{h}}$ (δ_{h}=15cm); $\|F_{\text{contact},j}\| \leq f_{\max}$ for all non-target objects j ($f_{\max}$=5N); $\|\dot{q}\| \leq v_{\max}$ ($v_{\max}$=1.5 rad/s); gripper force bound; workspace containment (3 atoms).
- Fragile Navigation ($k = 5$): $d(\text{base}, o_i) \geq \delta_{\text{obs}}$ for fragile obstacles (δ_{obs}=10cm, 3 atoms); speed bound; heading-rate bound.
- Dynamic Handover ($k = 6$): torso distance $\geq \delta_{\text{t}}$ (δ_{t}=20cm); grasp force bound; object height above floor bound; joint velocity bound; workspace containment.

Hyperparameters. β=10; δ=0.40; τ_0=0.20 annealed by ρ=0.9995 per update to $\tau_{\min}$=0.05; barrier numerical floor $M_\Phi - \delta \geq 10^{-3}$ enforced by clamping; budget b=0.5; λ_0=0.1; $\eta_\theta = 3 \times 10^{-5}$ (AdamW, cosine schedule); η_λ=0.01; certificate rate α=0.3; c_{cert}=0.5; PPO clip 0.2, GAE λ=0.95, γ=0.99; 64 parallel environments, 8-step action chunks, batch 256 chunks per update; 5M environment steps; 5 seeds (0–4).

Compute. Training a single run takes approximately 30 GPU-hours on one consumer-grade GPU (e.g. RTX 4090); the DENSE barrier adds less than 4% wall-clock overhead over the unconstrained baseline, measured over the same schedule.

D PyTorch Implementation

The core module (barrier construction, gradients, dual loop) is ∼100 lines; the full training harness follows standard PPO/ConRFT practice. A standalone runnable file is provided alongside the paper.

```python
import torch
import torch.nn as nn
import torch.nn.functional as F

class DenseBarrier(nn.Module):
    """Differentiable endogenous neuro-symbolic shield.

    Compiles Phi = AND_i (g_i >= 0) into:
      mu_i = sigmoid(g_i / tau)                    (fuzzification)
      M    = -1/beta * logsumexp(-beta * mu)       (LSE soft-min)
      B    = log(1 - delta) - log(M - delta)       (log-barrier)
    """
    def __init__(self, k, beta=10.0, delta=0.40, tau=0.20):
        super().__init__()
        self.k, self.beta, self.delta = k, beta, delta
        self.register_buffer("tau", torch.tensor(float(tau)))

    def forward(self, g):
        # g: (B, k) stacked constraint margins; g >= 0 means satisfied
        mu = torch.sigmoid(g / self.tau)                    # (B, k)
```

```
        # stable LSE soft-min with shift by min
        m_hat = mu.min(dim=-1, keepdim=True).values
        M = m_hat - (1.0 / self.beta) * torch.logsumexp(
            -self.beta * (mu - m_hat), dim=-1, keepdim=True)    # (B, 1)
        # log-barrier with numerical floor
        M = torch.clamp(M, min=self.delta + 1e-3)
        B = torch.log(1.0 - self.delta) - torch.log(M - self.delta)
        return B.squeeze(-1), M.squeeze(-1), mu

    def atom_weights(self, g):
        """Aggregation weights w_i (softmax over -beta * mu)."""
        mu = torch.sigmoid(g / self.tau)
        return F.softmax(-self.beta * mu, dim=-1)

def dense_update(policy, barrier, batch, lam, cfg):
    """One inner+outer loop update of Algorithm 1."""
    s, a, r, s2 = batch
    # --- task loss (standard policy gradient, e.g. PPO clipped) ---
    logp = policy.log_prob(s, a)
    L_r = -(logp * batch.adv).mean()
    # --- endogenous barrier on the *taken* action ---
    g = policy.constraint_margins(s, a)          # (B, k) user-declared FOL
    B, M, _ = barrier(g)
    # discounted barrier cost  J_B = E[sum_t gamma^t B]
    B_disc = (cfg.gamma ** batch.timesteps * B).mean()
    # --- one-step certificate hinge (Theorem 1) ---
    with torch.no_grad():
        a2 = policy.sample(s2)
    _, M_next, _ = barrier(policy.constraint_margins(s2, a2))
    alpha, delta = cfg.alpha, barrier.delta
    L_cert = F.relu((1 - alpha) * M + alpha * delta - M_next).mean()
    # --- inner loop: policy update ---
    L_pi = L_r + lam * B_disc + cfg.c_cert * L_cert
    policy.optimizer.zero_grad(); L_pi.backward(); policy.optimizer.step()
    # --- outer loop: multiplier update ---
    lam = max(0.0, lam + cfg.eta_lam * (B_disc.item() - cfg.b))
    return lam, dict(L_r=L_r.item(), J_B=B_disc.item(), L_cert=L_cert.item())

def anneal_tau(barrier, rho=0.9995, tau_min=0.05):
    """Fuzziness curriculum: tau_0 -> tau_min under geometric decay."""
    barrier.tau.mul_(rho).clamp_(min=tau_min)
```

E Limitations and Broader Impact

Limitations. (i) Our experiments are simulation-based; the constraint margins g_i are read from physics state, and while the same pipeline applies to perception-estimated margins (with estimation error folded into τ), hardware validation is future work. (ii) The FOL fragment covers metric conjunctive constraints; disjunctions, existentials, and temporal operators require extensions (e.g., soft-max duals of the soft-min). (iii) Theorem 1 certifies forward invariance conditionally on the trained certificate condition holding; DENSE audits this empirically (violation rate of (8) $< 0.4\%$), but a formal audit on hardware would require reachability tooling. (iv) The barrier's repulsion depends on ∇g_i being available at training time, which holds for differentiable simulators and for learned, differentiable margin estimators.

Broader impact. Methods that make learned robot policies safer reduce physical risk to people and property near robots. The same machinery is neutral with respect to intent: a barrier constrains behavior to whatever specification it is given, and specifications must themselves be authored and audited carefully.