DON'T LEARN WHEN TO LOOK: POST-HOC COUNTERFACTUAL GATING FOR VISUOMOTOR RL

Anonymous authors
Paper under double-blind review

ABSTRACT

A visuomotor policy that also senses its own body pays for the camera at every step (in our architecture, 98.8% of policy computation) whether or not vision is useful right now. The natural fix is a learned gate that decides *when* to look. We show that every in-loop version of this idea fails, and that a simple change works immediately: don't learn when to look; infer it from the frozen policy. The diagnosis comes first. The gradient-corruption premise behind gated fusion is unsupported. Teacher-relative value gaps measure how well the student learned, not whether vision helps, and no threshold repairs them. A corrected counterfactual label produces striking domain separation, which mixed-domain training, a protocol we propose, exposes as gate–critic co-evolution. Three further mechanisms (soft attention, expert routing, a reward-level observation cost) either cannot discriminate or weight vision *inversely* to its utility, a cold-start lock-in that changing visual difficulty does not move. Then the remedy. Train one *any-gate* policy with mask dropout; gate it post hoc with counterfactual signals read from the frozen network. A performance-difference bound attaches to the *action* counterfactual and yields a frontier with guaranteed loss. Empirically, across three seeds and five settings (Cond A/B, Leak, Mixed, and an independent robot), post-hoc gating sits 7–13pp above the always-on/quota convex hull on Cond A at matched budgets, beats a random-episode control by 12–27pp at matched savings where vision is essential, is return-neutral (98%) at 17% savings, and in mixed domains outperforms even a domain-label reference; a distilled proprioception-only predictor gates *without looking* at 39% real savings with no loss where vision is redundant or mixed. No labels, no retraining. When to look is not something a policy should learn; it is something to infer from one after training.

1 INTRODUCTION

Consider a manipulator learning to reach a target. Its proprioceptive stream is cheap, clean, and always informative about its own body; its camera is neither: the visual branch is 98.8% of the policy's forward computation in our measurements, and its content mixes task-critical information with spurious variation (lighting, colors, distractors) that never enters the dynamics or reward. Whether the camera is worth this price varies across tasks and within episodes. A growing line of work in manipulation therefore asks *when* a modality should influence the policy, triggering fusion from contact events, predicted contacts, or force-aware expert routing (Lei et al., 2026; He et al., 2025; Yu et al., 2025). For vision in RL, the analogous question has a seemingly natural answer: learn a per-step gate, supervised by a *value-based* utility signal, for instance the gap between the policy's value and that of a physics-only reference. This paper began as a proposal of exactly such a method.

It does not end as one. Controlled experiments meant to validate the recipe dismantled it at three levels, and the anatomy of that failure points to the remedy. First, the *motivating premise*: gated and asymmetric-learning pipelines are commonly motivated by the claim that naive joint training lets spurious visual gradients corrupt the proprioceptive representation. Stop-gradient controls and encoder probing do not support this. The physics representation survives joint training essentially intact; the collapse the premise is meant to explain has a simpler cause: the task needs vision. Second, the *supervision signal*: teacher-relative value gaps do not measure visual utility. They measure how well the student learned relative to the teacher. So a policy that outperforms its physics-only reference labels vision "useful" even where vision is provably redundant. No threshold repairs a signal that

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.

measures the wrong quantity; five designs (absolute, calibrated, quota, two adaptive anchors) each fail in an instructive way.

Third, and most cautionary: the *most convincing apparent success is an artifact*. Correcting the label to a counterfactual magnitude (how much masking the visual features moves the value estimate) produces striking behavior: with one dimensionless threshold, independently trained runs hold the gate $\sim 80\%$ open where vision is essential and drive it to exactly 0.000 where vision is redundant. This is the result one would publish.

Two independent tests show otherwise. Probing the same label offline, outside the training loop, retains only a fraction of the separation ($4\times$ vs $33\times$); the dramatic gap is manufactured by *gate–critic co-evolution*: the gate's early drift reshapes the critic's training distribution, which amplifies the label that drives the gate. And the discriminating experiment, *mixed-domain training*, where a single policy faces both domains episode-by-episode and genuine discrimination would have to show up within one run, collapses the gate to 0.000 in both domains across all seeds. The per-domain runs had not learned to discriminate; each had simply co-evolved to a different fixed point. We propose mixed-domain training as a discriminating protocol that adaptive-fusion claims should pass. Two continuous mechanisms, soft attention and a two-expert router, *pass* the test the hard gate fails, yet weight vision *inversely* to its utility.

The picture that survives has a remedy. The utility signal is *real at the representation level*: the counterfactual influence of vision on the learned value function separates domains even offline and reproduces under off-policy value learning. What fails is the conversion of representation into decision *inside the training loop*, at three located points: loss-scale dynamics, myopic labels whose per-step correctness provably holds (Thm. 4.2) yet does not imply return-optimal gating, and the coupling above. Applying the same protocol to a reward-level observation cost, the most principled in-loop formulation, finds no cost at which the gate discriminates correctly. So we move the decision out (Fig. 1): one *any-gate* policy trained with mask dropout, gated post hoc by counterfactual signals read from the frozen network. A performance-difference bound (Prop. 4.5) shows the guarantee attaches to the *action* counterfactual, and experiments confirm the choice: post-hoc gating sits above the convex hull of trivial always-on/quota mixtures at moderate budgets and above a random-episode control at every matched budget, whereas every learned gate sat below the hull.

Contributions.

1. **Diagnosis: why in-loop gating fails.** Across four mechanism families (label-supervised hard gates, soft attention, expert routing, and reward-level observation costs), learned modality importance is untrustworthy: gates collapse via gate–critic co-evolution (exposed by *mixed-domain training*, a protocol we propose), and continuous weights invert relative to utility through cold-start credit lock-in, not perceptual difficulty; two common assumptions (gradient corruption, teacher-relative value gaps) are retired on the way.
2. **Method and theory: decouple the decision.** An *any-gate* policy trained with mask dropout, gated post hoc by counterfactual signals from the frozen network. A performance-difference bound (Prop. 4.5) attaches to the *action* counterfactual and gives a frontier with guaranteed loss (Cor. 4.6); label correctness (Thm. 4.2) and a decomposition (Prop. C.1) explain why the value counterfactual alone does not suffice.
3. **Results.** Across three seeds and five settings (including an independent robot), post-hoc gating lies 7–13pp above the always-on/quota convex hull on Cond A at matched budgets, beats a random-episode control by 12–27pp at matched savings where vision is essential, is return-neutral at 17% real savings, and in mixed domains outperforms a domain-label reference; a distilled proprioception-only predictor gates without vision at 39% real savings with no loss where vision is redundant or mixed; a residual *execution-time* loop is exposed and controlled by the choice of signal.

2 Related Work

Adaptive and conditional modality fusion. Recent manipulation work learns *when* a modality should influence the policy: contact-event gating of force–torque streams, future-contact prediction, and force-aware expert routing (Lei et al., 2026; He et al., 2025; Yu et al., 2025), building on

multimodal fusion for contact-rich manipulation (Lee et al., 2019). These trigger fusion from physical events in imitation pipelines; we study the RL analogue for vision, where the trigger must itself be learned. Our central caution for this line: apparent adaptivity can be manufactured by the training loop, and we propose a protocol (Sec. 5.3) such claims should pass.

Privileged and asymmetric learning. Asymmetric actor–critic, distillation, and teacher–student pipelines (Pinto et al., 2018; Chen et al., 2020; Salter et al., 2020; Lee et al., 2020; Kumar et al., 2021; Rusu et al., 2016; Hinton et al., 2015; Andrychowicz et al., 2020) exploit privileged state during training, often alongside domain randomization for transfer (Tobin et al., 2017; Zhao et al., 2020; Truong et al., 2023; Rizzardo et al., 2023; Da et al., 2025). A recurring motivation is that naive joint training lets spurious visual gradients corrupt clean proprioceptive representations; we test this premise directly with stop-gradient controls and encoder probing and do not find support for it (Sec. 5.1).

Robustness to visual distractors. Visual RL is made robust to distractors through invariant representations, augmentation, or bisimulation (Laskin et al., 2020; Kostrikov et al., 2020; Yarats et al., 2022; Schwarzer et al., 2020; Zhang et al., 2020; Higgins et al., 2017; Arjovsky et al., 2019; Ferns et al., 2011; Li et al., 2006), benchmarked on procedurally varied or distracting suites (Cobbe et al., 2019; 2020; Stone et al., 2021), or pretrained visual representations (Xiao et al., 2022; Nair et al., 2022; Majumdar et al., 2023): *what*-solutions; we study the complementary *when*-decision.

Conditional computation and active perception. Gating, adaptive computation, and expert routing (Jacobs et al., 1991; Shazeer et al., 2017; Bengio et al., 2013; Jang et al., 2017; Maddison et al., 2017; Graves, 2016; Wang et al., 2018; Wu et al., 2018) supply our mechanisms, and learned input gating for robust control (Tomar et al., 2023) is the published in-loop form of the idea we diagnose; modality dropout (Neverova et al., 2016) underlies any-gate training; when to sense is the classical concern of active perception, value of information, and partially observable control (Bajcsy, 1988; Howard, 1966; Kaelbling et al., 1998; Hausknecht & Stone, 2015; Ni et al., 2022); we add a case study of a router co-evolving with the statistics it thresholds.

3 Problem Setup

Mixed-observation RL. The state is $s_t = (s_{P,t}, s_{V,t})$: a proprioceptive part $s_P \in \mathbb{R}^{d_P}$ (cheap, clean, always task-relevant for the body) and an RGB image s_V (expensive, since the visual branch is 98.8% of policy forward multiply–accumulates (MACs) and $7\times$ the per-step CPU cost of the proprioceptive path in our architecture, and a mixture of task-critical and spurious content). Spurious factors are *exogenous by construction*: visual factors enter neither the transition nor the reward, and the two auxiliary physical factors lie outside the causal path from action to reward (Appendix E).

Evaluation domains. (i) **Cond A/B** (vision-essential): Franka reach with the target excluded from s_P, at low/high spurious variation; vision is required to localize the target. (ii) **Leak** (vision-redundant by construction): identical task, but cube and target coordinates are appended to s_P, so vision is provably uninformative beyond proprioception (Lemma J.1). (iii) **PiPER**: an independent 6-DoF arm (different robot, actuation, and camera geometry) with strong domain randomization. (iv) **Mixed**: each episode samples Cond A or Leak with probability $\frac{1}{2}$, the discriminating condition of Sec. 5.3.

In-loop gating recipes. All policies are trained with PPO (Schulman et al., 2017) in MuJoCo/robosuite environments (Todorov et al., 2012; Zhu et al., 2020). S-DVG (selective dynamic visual gating) trains a physics-only reference, burns in a visual encoder, then trains a mixed policy with a per-step binary gate $g_t \in \{0, 1\}$ on the visual pathway ($h = h_P + g \cdot \mathrm{proj}(h_V)$), supervised toward a label ℓ_t derived from a utility signal, plus a sparsity prior; the reference also serves as a distillation anchor. Two supervision families are studied: *teacher-relative* labels $\ell = \mathbf{1}[V_{\mathrm{mix}} - V_{\mathrm{ref}} > \tau_{\mathrm{vg}}]$ (the original design), and *counterfactual-magnitude* labels $\ell = \mathbf{1}[|\Delta_{\mathrm{cf}}| > \hat{\tau}]$ with $\Delta_{\mathrm{cf}} = V_\theta(h_P + \mathrm{proj}(h_V)) - V_\theta(h_P)$ and a scale-anchored threshold $\hat{\tau} = \alpha \cdot \mathrm{EMA}|V_\theta(h_P)|$ (EMA: exponential moving average; the corrected design, with mask-dropout keeping the masked pathway in-distribution). Training details, the loss-normalization fix that all recipes share, and hyperparameters are in Appendix G.

4 POST-HOC COUNTERFACTUAL GATING

4.1 LABEL CORRECTNESS UNDER APPROXIMATE VALUES

We formalize when the gate-supervision label agrees with an ideal notion of visual influence, *without* assuming Bellman-optimal value functions. All statements are at a fixed training iterate: π denotes the current mixed policy, d the state distribution of its training process, and ρ^{μ} the (normalized, discounted) state occupancy of a policy μ; expectations conditional on s_P are taken under d.

Target quantities. Let $V^{\pi}(s)$ be the true value of the current policy under the full observation, and let

$$\bar{V}^{\pi}(s_P) \;:=\; \mathbb{E}_{s\sim d}[\, V^{\pi}(s) \mid s_P \,] \tag{1}$$

be its best prediction from proprioception alone. The *target visual influence* is

$$U^{\circ}(s) \;:=\; \big|\, V^{\pi}(s) - \bar{V}^{\pi}(s_P) \,\big|, \tag{2}$$

the magnitude by which the visual observation moves the value estimate away from its proprio-conditional mean. The learned proxy is $\hat{U}(s) := |\Delta_{\mathrm{cf}}(s)| = |V_\theta(h_P{+}\mathrm{proj}(h_V)) - V_\theta(h_P)|$, and the training label is $\ell(s) = \mathbf{1}[\hat{U}(s) > \hat{\tau}]$ with $\hat{\tau} = \alpha \cdot \widehat{m}$, where $\widehat{m}$ is the running estimate of the value scale $\mathbb{E}_d|V_\theta(h_P)|$.

Assumption 4.1 (Approximation quality)**.** There exist $\varepsilon_F, \varepsilon_P, \varepsilon_\tau \geq 0$ such that, for all s in the support of d (and, for Proposition C.1, of $\rho^{\pi^{\circ}_{g}}$): (i) $|V_\theta(h_P{+}\mathrm{proj}(h_V)) - V^{\pi}(s)| \leq \varepsilon_F$; (ii) $|V_\theta(h_P) - \bar{V}^{\pi}(s_P)| \leq \varepsilon_P$; (iii) $|\hat{\tau} - \tau^{\star}| \leq \varepsilon_\tau$ for the nominal threshold $\tau^{\star} = \alpha\, \mathbb{E}_d|V_\theta(h_P)|$.

Condition (ii) asks the masked head to be calibrated to the proprio-conditional value, precisely the regression that *mask-dropout* training performs (Sec. 4.3), so the mechanism that stabilizes training also grounds the assumption. Condition (iii) is monitorable (Cor. 4.3): $\widehat{m}$ averages $|V_\theta(h_P)|$, whose pooled dispersion is small relative to its mean (0.88 vs 13.1 over 786k samples). Neither value head is assumed optimal.

Theorem 4.2 (Label correctness under approximate values)**.** *Under Assumption 4.1, for every $s \in \mathrm{supp}(d)$ with $\big|U^{\circ}(s) - \tau^{\star}\big| > \bar{\varepsilon}$, where $\bar{\varepsilon} := \varepsilon_F + \varepsilon_P + \varepsilon_\tau$,*

$$\mathbf{1}\Big[\hat{U}(s) > \hat{\tau}\Big] \;=\; \mathbf{1}[U^{\circ}(s) > \tau^{\star}]\,.$$

In words: the label is right wherever the true influence is clearly above or clearly below the threshold; errors can only hide in a band around it whose width is the total approximation error. Hence the mass of states on which the realized label can disagree with the ideal label is at most the d-mass of the margin band $\{\, s : |U^{\circ}(s) - \tau^{\star}| \leq \bar{\varepsilon} \,\}$.

Proof in Appendix D.

Corollary 4.3 (Robustness to scale estimation)**.** *If the running scale estimate is additionally biased by $b \geq 0$ (i.e., ε_τ inflates to $\varepsilon_\tau + \alpha b$), labels agree outside the widened band, and if the law of U° under d admits a density bounded by L on the widened band, the mislabeled mass grows by at most $2L\alpha b$, linearly in the estimation bias, with slope damped by the dimensionless α.*

Scope caveats (snapshot semantics, correlated errors, the value-gap variant) are collected in Appendix B.

4.2 FROM LABELS TO RETURN

A correct label does not make a good gate. Proposition C.1 (Appendix C) writes the return gap between the ideal-label gate $g^{\circ}(s) = \mathbf{1}[U^{\circ}(s) > \tau^{\star}]$ and the executed gate, via the performance difference lemma (Kakade & Langford, 2002; Schulman et al., 2015), as an exact sum of *signed* missed-utility and interference terms over disagreement states, bounded (up to $\delta_{\max}/(1{-}\gamma)$) by the margin-band mass of Theorem 4.2 plus the gate–label mismatch μ_{cls} (0.039 at convergence, measured on d). Because g° thresholds a myopic influence measure rather than the return-optimal gate, those terms need not favor the label gate, which is exactly the gap Sec. 5.5 observes.

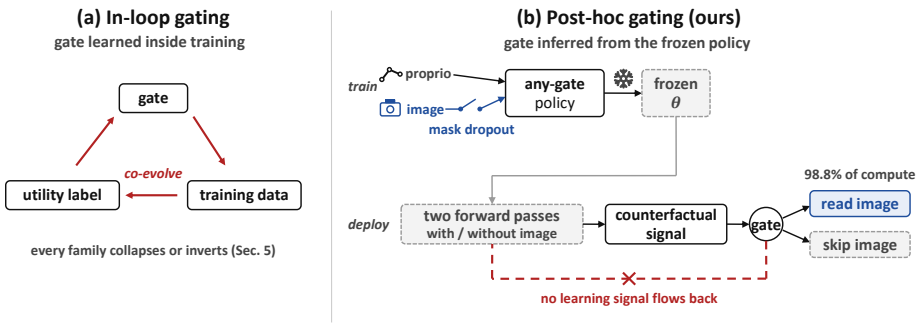

Figure 1: (a) In-loop gating learns the gate inside training: the gate reshapes the data the critic sees and hence the utility label that trains the gate, so the two co-evolve (Sec. 5.3). (b) Post-hoc gating: one any-gate policy is trained with mask dropout and frozen; at deployment two forward passes (with and without the image) yield a counterfactual signal (Sec. 4.3) that sets the gate. No learning signal flows back to θ.

4.3 Decoupled Gating and Its Guarantee

Sections 5.3–5.4 locate the failure of learned gating in a single structural fact: the gating decision lives *inside* the training loop, where it reshapes the critic's data (co-evolution) or the objective's path of least resistance (inversion). We therefore move the decision out (Fig. 1).

Any-gate policies. Train a single policy–critic pair with *mask dropout*: at each update, a fraction p of the batch is processed with the visual pathway masked (m=0, $h = h_P$) and the rest with it admitted (m=1, $h = h_P + \mathrm{proj}(h_V)$). No gate is learned and no utility label is used, so no channel exists through which a decision could shape the data. The result is one network defining two *modes*, $\pi_M(\cdot \mid s) := \pi_\theta(\cdot \mid s_P, s_V)$ and $\pi_P(\cdot \mid s) := \pi_\theta(\cdot \mid s_P, \varnothing)$, both in-distribution by construction; the always-on policy is $\pi_o := \pi_M$.

Post-hoc gating. At execution, a gate $g(s) \in \{0, 1\}$ is computed from the *frozen* network and the executed policy is $\pi_g = g\,\pi_M + (1{-}g)\,\pi_P$. Two counterfactual signals are available at zero training cost: the *value counterfactual* $\hat{U}(s) = |V_\theta(h_P{+}\mathrm{proj}(h_V)) - V_\theta(h_P)|$ of Sec. 4.1, and the *action counterfactual*

$$D(s) \;:=\; D_{\mathrm{TV}}\big(\pi_M(\cdot \mid s),\, \pi_P(\cdot \mid s)\big) \;=\; \tfrac{1}{2}\big\|\pi_M(\cdot \mid s) - \pi_P(\cdot \mid s)\big\|_1, \tag{3}$$

how much admitting vision changes what the policy would *do*.

Remark 4.4 (No co-evolution, by construction). Both signals are deterministic functions of (s, θ) with θ frozen. The loop of Sec. 5.3 (gate $\to$ state distribution $\to$ critic $\to$ label $\to$ gate) has no edge from the gate back to θ, so the separation measured offline is exactly the separation the decision sees.

Proposition 4.5 (Cost of post-hoc gating)**.** *Let ρ^{π_o} be the discounted state occupancy of the always-on policy and $A_{\max} := \sup_{s,a} |A^{\pi_g}(s, a)|$. Then*

$$J(\pi_o) - J(\pi_g) = \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim\rho^{\pi_o}}\Big[\mathbf{1}[g(s){=}0]\,\big(\mathbb{E}_{\pi_M}[A^{\pi_g}(s,\cdot)] - \mathbb{E}_{\pi_P}[A^{\pi_g}(s,\cdot)]\big)\Big] \tag{4}$$

$$\leq \frac{2A_{\max}}{1-\gamma}\,\mathbb{E}_{s\sim\rho^{\pi_o}}\big[\mathbf{1}[g(s){=}0]\,D(s)\big]. \tag{5}$$

Proof in Appendix D.

In words: closing the gate can only cost return at states where admitting vision would change the action, and the cost is at most proportional to how much it would change it.

Corollary 4.6 (A principled frontier)**.** *For the action-counterfactual gate $g_\kappa(s) := \mathbf{1}[D(s) > \kappa]$,*

$$J(\pi_o) - J(\pi_{g_\kappa}) \;\leq\; \frac{2A_{\max}\,\kappa}{1-\gamma}\;\rho^{\pi_o}\big(\{s : D(s) \leq \kappa\}\big),$$

while the fraction of visual computation skipped is the closed mass under ρ^{π_o} (realized savings differ by the self-trap of Sec. 5.6). Sweeping κ traces a frontier with a guaranteed *loss on one axis and savings on the other; for Gaussian policy heads $D(s) \le \sqrt{D_{\mathrm{KL}}(\pi_M \,\|\, \pi_P)/2}$ (Pinsker), so the gate can be computed from the closed-form KL. With observation cost c per unit of open occupancy mass, closing s is certified whenever $D(s) \le \kappa^\star := c(1-\gamma)/(2A_{\max})$.*

Remark 4.7 (Which signal carries the guarantee). The value counterfactual enjoys label correctness (Thm. 4.2) but no return guarantee (it thresholds a myopic quantity; Prop. C.1). The action counterfactual carries the guarantee of Prop. 4.5 because it bounds the very term the performance difference lemma charges: a gate may close where vision changes the value estimate but not the action, never where it changes the action. We test both and their conjunction (Sec. 5.6).

Remark 4.8 (Estimability). Every quantity in equation 5 is measurable on the frozen network: $D(s)$ from the two heads, $A_{\max}$ approximated from the critic's advantage range and observed returns, the closed mass from rollouts of π_o. The bound is therefore reportable alongside the empirical loss, and its tightness is itself a result.

Remark 4.9 (Real savings require deciding without looking). Both counterfactual signals need the visual forward pass, so a per-step gate that computes them saves nothing: the signal *is* the computation. Two decoupled realizations save real compute. (i) *Probe-and-commit*: evaluate the signal on the first k steps of an episode, then commit (optionally re-probing every K steps), saving a $(1-k/T)$ fraction on closed episodes. (ii) *Distilled predictor*: since the frozen network labels its own rollouts with $D(s)$ for free, a small proprioception-only model $\hat{D}(s_P)$ is fit by supervised regression and used at run time without vision; the policy is frozen, so labels do not depend on the predictor and no loop reopens. If $|\hat{D}(s) - D(s)| \le \varepsilon_D$ on the support, closing where $\hat{D} \le \kappa$ implies $D \le \kappa + \varepsilon_D$, and Cor. 4.6 holds with κ inflated by ε_D; where the residual is large the gate should stay open, which the predictor's own uncertainty can enforce.

5 Experiments

Evaluation protocol. All comparisons are internal to one codebase (recipe generations marked where they differ, Appendix G); checkpoints are evaluated deterministically on matched physics-initialization seeds, with paired statistics (Wilcoxon signed-rank) where seeds are matched and rank-sum tests otherwise. We report per-domain retention relative to an always-on (NoGate) control trained identically. Table 3 (Appendix A) collects all symbols.

5.1 The Gradient-Corruption Premise

Gated and asymmetric pipelines are commonly motivated by gradient corruption: spurious visual gradients degrading the proprioceptive representation. We test this with (i) a *naive-joint* variant (physics encoder unfrozen, gate forced open, no reference, no distillation), (ii) a *stop-gradient* control (visual loss path severed from the physics encoder), and (iii) ridge probing of each physics encoder against ground-truth core parameters (Table 6, Appendix F).

Probing differences between naive-joint and clean encoders are systematic but small (seed std 0.001–0.007, n=3), and naive-joint *performance* is statistically indistinguishable from the full pipeline (Cond B: 172.4 ± 62.0 vs 160.9 ± 17.3, p=0.76, n=5v7; direction reversed on Cond D, p=0.31). Sharper still, severing the gradient path in either direction leaves probing R^2 within 0.02 of naive-joint (vis below, phys marginally above): there was nothing to protect. The collapse the premise is invoked to explain is explained by the task needing vision.

5.2 Teacher-Relative Value Gaps

The label $\mathbf{1}[V_{\mathrm{mix}} - V_{\mathrm{ref}} > \tau_{\mathrm{vg}}]$ compares student to teacher, with two observed consequences. *First, scale fragility*: when a physics fix rescaled returns $\sim 70\times$, the unnormalized value loss dominated gate supervision ($28\times$ after weighting) and the fixed threshold fell to 0.5% of the median gap, and the gate silently saturated open in every condition (ablation null at n=7). Loss normalization restores gate–label tracking and is adopted thereafter. *Second, and unfixable by thresholds*: in the Leak domain the trained policy outperforms its physics-only teacher (320.3 vs 267.8) through extra capacity and distillation, not information, so the gap stays positive and the label says "open" exactly where vision

is provably redundant. We verified that no threshold design repairs this (Table 8, Appendix H): an absolute τ_{vg} is scale-fragile, a calibrated relative τ_{vg} permanently closes the gate, a quantile label pins the open rate to its quota, a threshold anchored to the positive part $\Delta^{+}=\max(0, V_{\mathrm{mix}}-V_{\mathrm{ref}})$ normalizes the domain signal away, and a value-anchored one inverts. A signal that measures learning quality cannot be thresholded into measuring utility.

5.3 Apparent Adaptivity Is Gate–Critic Co-evolution

The counterfactual-magnitude label (Sec. 3) produces a striking result: with a single dimensionless threshold (α=0.069), independently trained runs hold the gate open where vision is essential (Cond A 0.890, Cond B 0.755, n=3; deterministic evaluation; on-policy averages under Gumbel sampling run slightly lower, e.g. 0.812) and drive it to exactly 0.000 where vision is redundant (Leak, $3/3$ seeds) and on PiPER (0.000). Offline forensics temper the picture: probing the same label on frozen checkpoints retains only 3.8–$4.4\times$ of the $33\times$ on-policy separation ($92\times$ at the label level). The mechanism is visible in the logs: as the gate closes, the critic specializes to the proprio-only pathway ($\mathrm{frac}(\Delta_{\mathrm{cf}}{>}0)$: $0.563 \to 0.075$ in Leak), amplifying the statistic that drives the gate.

The discriminating experiment is **mixed-domain training**: one policy, episodes sampled from both domains, so genuine discrimination must appear *within a single run*. It does not. The gate collapses to 0.000 in *both* episode types ($3/3$ seeds). Even the early transient is inverted: at iter $\approx$33 the gate is 0.77 open in vision-redundant episodes and 0.22 in vision-essential ones (gap 0.546), the pattern the continuous mechanisms show at convergence (Sec. 5.4). It then decays with intermittent rebounds to sustained zero by iter $\approx$365 on the seed-averaged curve (per seed 310–484; Fig. 4). Each per-domain run had converged to its own co-evolved fixed point rather than learning to discriminate; mixed-domain training is the protocol that exposes this, and one adaptive-fusion claims should pass.

5.4 Continuous Mechanisms Discriminate, but Invert

Is the collapse specific to hard gating? We repeat the mixed-domain protocol on two end-to-end mechanisms with the same backbone and recipe: a *soft attention* weight ($\beta \in [0,1]$, $h = h_P + \beta\,\mathrm{proj}(h_V)$, no label, no sparsity, PPO loss only) and a *two-expert router* (physics-only vs. mixed heads, softmax weights). Both pass the discrimination test the hard gate failed: within a single mixed run, per-episode weights separate cleanly (soft β: gap 0.60 on both seeds, with p_{10}=0.000/p_{90}=1.000, so the weight is genuinely modulated; router: gap 0.33). But the sign is **inverted** relative to utility in every case (Table 1).

Concretely, weight is *low* where vision is essential and *high* where it is provably redundant. The soft variant abandons vision where it is needed (its Cond A return, retention 45.1%, sits at or below the physics-only reference's level), while the router preserves return (92.6%) despite inverted weights, because output-level mixing lets the physics head carry the policy: a separate finding that weight inversion need not cost performance under the right architecture. Two further experiments locate the cause. A *reward-level observation cost* (the most principled in-loop formulation, with the gate as an action and $-c\,g_t$ added to the reward) was swept from 0.4% to 150% of per-step return: the redundant domain never closes (0.999 at 150%), the essential domain is the one that collapses (0.458 open, 47% retention), and mixed training shows the same inversion (A-ep 0.25 vs Leak-ep 1.00); no cost yields correct discrimination.

Intervening on *visual difficulty* at fixed utility (noise sweeps in Leak; an easy-vision Cond A variant) leaves the weights unchanged ($0.96 \to 0.99 \to 0.94$; $0.00 \to 0.02$), ruling out perceptual difficulty as the cause. What remains is cold-start credit lock-in: β=0 is absorbing (a branch that receives no credit gets no gradient and cannot recover), while in the redundant domain the weight drifts unconstrained because either value is costless, the gradient-starvation and shortcut dynamics of Pezeshki et al. (2021) and Geirhos et al. (2020) at the modality level. Together with Sec. 5.3: hard gating fails to discriminate; continuous weighting discriminates, and wrongly.

5.5 Compute–Performance Frontier

Table 2 places the learned gate on the compute–performance frontier (Cond A). One further row deserves a word: attention-privileged learning (APRiL) trains a far better policy (415.7, success 0.78) but uses vision at every step and needs privileged state in training; it answers a different question

Table 1: Vision weight of in-loop mechanisms under the mixed-domain protocol (retention in Table 11). Continuous mechanisms discriminate, with inverted sign; APRiL is always on.

Vision weight	Cond A	Leak	Mixed A-ep	Mixed Leak-ep
Hard gate (cf-mag)	0.890	0.000	0.000	0.000
Soft attention β	0.000	0.964	0.124	0.729
Two-expert router w	0.284	0.583	0.275	0.603
Reward cost (c=0.4%; Mixed at 42%)	0.944	1.000	0.25	1.00
Attention-privileged (APRiL)	1.000	1.000	1.000	1.000

Table 2: Cond A. **Training-time** rows ask how good a policy one can train (return; savings are fixed at training); the **inference-time** rows ask how much computation can be saved on a *given* frozen any-gate policy (retention is relative to that policy's own always-on return; 3 seeds). †privileged state used in training.

Method	Return / Retention	Real MACs saved
Training-time (learned or fixed at training)		
NoGate (always-on; reference)	203.7 ± 10.2	0%
Learned gate (cf-mag)	167.9 ± 20.9	10.9%
Quota q=0.50 / 0.25 (hull)	163.4 / 149.6	40.4% / 64.5%
APRiL† (attention-privileged; gate 1.0)	415.7	0%
Inference-time (post hoc on the frozen NoGate policy)		
Random-episode @ matched budget	79.5% / 65.0%	17.3% / 45.9%
Post-hoc, action-cf / episode	98.3%	17.3%
Post-hoc, action-cf / episode-K50	91.8%	45.9%
Entropy / speed heuristic (cheap signals)	52.2% / 76.7%	–
APRiL† + mask dropout + post-hoc	+15pp vs random	> 0

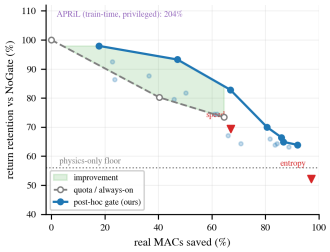

Figure 2: Cond A: retention *relative to the always-on/quota hull* (y=0; shown where the hull is defined, $\leq$ 64.5% savings). Post-hoc gating (blue; dots: 20 configurations) lies above it (+7pp at 17% and +13pp at 46% for the Table 2 cells; best configuration +15pp); the learned in-loop gate (×) lies 12pp below.

(how to train a better policy) than ours (how to allocate computation on a given policy). The two compose: on an APRiL policy trained with mask dropout, post-hoc gating again beats the random control (+15pp on Cond A; Table 2, last row), though mask dropout itself costs this privileged policy 18% on Cond A and 37% on Mixed (nothing on Leak), a domain-dependent cost. The learned gate is not dominated by any single quota, but it sits 12.3pp *below the convex hull* of trivial mixtures: interpolating always-on fusion with quota q=0.50 reaches 94.7% retention at the learned gate's savings level (10.9%), versus its 82.4%; the learned gate occupies the least efficient stretch of the curve. Two further costs are not informational: closing is never free even in Leak (retention 90.2% where vision is uninformative), traced to critic specialization above; and within-domain *state*-level adaptivity is absent (paired spurious-perturbation tests, p=0.42/0.70): the gate never learned *when* within an episode, only *where*, and Sec. 5.3 shows even that was not learned.

5.6 Post-hoc Gating on Frozen Policies

We freeze always-on policies trained with mask dropout (p=0.1; three seeds per domain), i.e. any-gate policies, and gate them post hoc (Sec. 4.3) with either signal under two rules: per episode (probe 5 steps, commit) and with re-probing every K=50 steps. Thresholds are absolute; open rate is an outcome (per-step rules save nothing; Remark 4.9). Two controls isolate what the signal contributes: a *random-episode* rule that closes episodes at the same rate without looking, and, in the mixed domain, a *domain-label reference* that closes every Leak episode.

Tables 2 and 10 report *real* saved computation. Where vision is essential (Cond A/B), the signal selects the right episodes: at matched budgets it beats the random-episode rule by 12–27pp, holding 86–98% retention where random falls to 65–79%; where there is nothing to select (Leak), the two nearly coincide, the signature of a working selector. In return units the signal wins in every domain (A +60.4, B +18.1, Leak +8.5 at the matched-budget cells; Mixed +17.3, Table 10), reported as differences because this control is not seed-paired. Cheap non-visual signals cannot replace it: an

entropy gate mistakes proprioceptive confidence for redundancy and collapses to the physics-only floor (52%), a speed heuristic reaches 77%; the value counterfactual, which needs the visual pass, reaches 89% at 49% savings. The learned gates of Table 2 sat 12pp below the always-on/quota convex hull; post-hoc gating sits 7pp above it at 17% savings and 13pp at 46% (interpolating Table 2), and the full κ sweep stays above the hull wherever a trivial mixture exists (Fig. 2; the value/episode-K50 cell re-evaluated at 100 episodes: 99.2%). A p=0.5 substrate closes more in Leak (0.197 open) at 104.8% retention. In the mixed domain the episode-level signal (231.0 at 53.8% savings) even outperforms the domain-label reference (213.4) and random (213.7): the masked mode π_P is weaker than π_M even in Leak episodes, where vision is uninformative but the two modes are not equally trained, so closing by label is not an upper bound. On the independent PiPER arm, whose controller runs on CPU, the gate removes 40% of visual computation at 99% retention (random: 96%).

Two caveats follow. First, a residual loop survives decoupling: a closed gate changes the states visited, hence the signal, an *execution-time* self-trap (expected minus realized open rate: value signal 0.04, action signal 0.12–0.15, per-step up to 0.34; re-probing helps the action signal, hurts the value signal). Probe length governs commit quality (k=1: 52% retention; k=10: 82%, at K=25), re-probe interval trades retention for savings monotonically, and AND/OR combinations of the two signals move along the same trade-off rather than dominating both single signals (Appendix L). Second, a proprioception-only predictor distilled from the frozen network's own $D(s)$ (Remark 4.9) gates *without looking*: held-out R^2 is 0.64/0.75 on Cond A/B and negative on Leak (where there is no signal to predict), and it saves 39–60% of visual computation with no loss in Leak/Mixed and 84% retention on A/B (Appendix N). Prop. 4.5 holds but is loose: realized loss per unit of closed divergence is 0.03/0.04 on Cond A/B (Appendix M): a certificate, not a prediction.

6 Discussion

Learning versus inference. Read from a frozen network, the utility signal drives a gate that beats trivial mixtures; placed inside the loop, the same signal co-evolves with its critic or locks in at cold start. Gating vision is an inference problem over a trained policy, and mechanisms that decouple estimation from control (Remark 4.9) are the natural family.

Cost model. On our CPU the visual pass is cheap in absolute terms (0.88ms), so a latency-priced return favors always-on (199 vs 156); the case for gating is $7\times$ per-step compute, energy, and headroom for larger encoders.

Reward dependence. Both signals inherit their semantics from the reward: the right dependence for value of information, but validated only under dense-shaped rewards and tied to the task (Cor. 4.6 prices looking through $\kappa^\star$).

Limitations. One algorithm family (PPO; off-policy evidence is representation-level only); domains of our own construction plus one independent robot; four in-loop mechanism families but not transformer cross-attention at scale, and no direct comparison to published information gating (Tomar et al., 2023); no recurrent (POMDP) baseline; limited predictor-input ablations; signed-rank tests for per-step controls only (episode-level controls as differences). Specifications, full tables, and ablations: Appendices E–P.

7 Conclusion

We set out to make a policy learn when to look and found every in-loop way of doing so fails (co-evolution, cold-start lock-in, a cost with no discriminating band), while moving the decision out of the loop works at once: a frozen any-gate policy, gated by its own counterfactuals, beats trivial mixtures without labels. Don't learn when to look; infer it.

AI use statement

The authors conducted the entire research process themselves, including determining the research questions, deciding which hypotheses to pursue and which experiments to conduct, and interpreting

the results. In this process, we used LLM tools as assistants to support theoretical analysis, coordinate experiments, and detect errors. The authors verified all proofs, implementations, and references and take full responsibility for the content.

Reproducibility statement

All experiments run in a single codebase; domain specs, recipes, and hyperparameters are in Appendices E–G; statistics use matched-seed paired evaluation where seeds are matched. Code, post-hoc gating scripts, diagnostics (Appendix J), the mixed-domain protocol, and the orchestration/failure logs will be released.

References

Marcin Andrychowicz et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *CoRL*, 2020.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *ICML*, pp. 1282–1289, 2019.

Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *ICML*, 2020.

Longchao Da et al. A survey of sim-to-real methods in RL: Progress, prospects and challenges with foundation models. *arXiv preprint arXiv:2502.13187*, 2025.

Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous Markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

Matthew Hausknecht and Peter Stone. Deep recurrent Q-learning for partially observable MDPs. In *AAAI Fall Symposium*, 2015.

Zihao He, Hongjie Fang, J. Chen, Hao-Shu Fang, and Cewu Lu. FoAR: Force-aware reactive policy for contact-rich robotic manipulation. *IEEE Robotics and Automation Letters*, 2025.

Irina Higgins et al. DARLA: Improving zero-shot transfer in reinforcement learning. In *ICML*, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ronald A Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. In *ICLR*, 2017.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, 2002.

Ilya Kostrikov, Denis Yarats, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv preprint arXiv:2004.13649*, 2020.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. In *RSS*, 2021.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In *ICML*, 2020.

Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47), 2020.

Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *ICRA*, 2019.

Jiuzhou Lei, Chang Liu, Yu She, Xiao Liang, and Minghui Zheng. Learning when to see and when to feel: Adaptive vision-torque fusion for contact-aware manipulation. *arXiv preprint arXiv:2604.01414*, 2026.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for MDPs. In *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2006.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*, 2017.

Arjun Majumdar et al. Where are we in the search for an artificial visual cortex for embodied intelligence? In *NeurIPS*, 2023.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *CoRL*, 2022.

Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. ModDrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38 (8):1692–1706, 2016.

Tianwei Ni, Benjamin Eysenbach, and Ruslan Salakhutdinov. Recurrent model-free RL can be a strong baseline for many POMDPs. In *ICML*, 2022.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *NeurIPS*, 2021.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. In *RSS*, 2018.

Carlo Rizzardo, Fei Chen, and Darwin Caldwell. Sim-to-real via latent prediction: Transferring visual non-prehensile manipulation policies. *Frontiers in Robotics and AI*, 2023.

Andrei A Rusu et al. Policy distillation. In *ICLR*, 2016.

Sasha Salter, Dushyant Rao, Markus Wulfmeier, Raia Hadsell, and Ingmar Posner. Attention-privileged reinforcement learning. In *CoRL*, 2020.

John Schulman, Sergey Levine, Philipp Moritz, Michael I Jordan, and Pieter Abbeel. Trust region policy optimization. In *ICML*, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Max Schwarzer et al. Data-efficient reinforcement learning with self-predictive representations. *arXiv preprint arXiv:2007.05929*, 2020.

Noam Shazeer et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

Austin Stone, Oscar Ramirez, Kurt Konolige, and Rico Jonschkowski. The distracting control suite – a challenging benchmark for pixel-based reinforcement learning. *arXiv preprint arXiv:2101.02722*, 2021.

Josh Tobin et al. Domain randomization for transferring deep neural networks from simulation to the real world. In *IROS*, pp. 23–30, 2017.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IROS*, pp. 5026–5033, 2012.

Manan Tomar, Riashat Islam, Matthew E Taylor, Sergey Levine, and Philip Bachman. Ignorance is bliss: Robust control via information gating. In *NeurIPS*, 2023.

Joanne Truong et al. Rethinking sim2real: Lower fidelity simulation leads to higher sim2real transfer in navigation. In *CoRL*, pp. 859–870, 2023.

Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E Gonzalez. SkipNet: Learning dynamic routing in convolutional networks. In *ECCV*, 2018.

Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S Davis, Kristen Grauman, and Rogerio Feris. BlockDrop: Dynamic inference paths in residual networks. In *CVPR*, 2018.

Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *ICLR*, 2022.

Jiawen Yu et al. ForceVLA: Enhancing VLA models with a force-aware MoE for contact-rich manipulation. *arXiv preprint arXiv:2505.22159*, 2025.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *IEEE SSCI*, pp. 737–744, 2020.

Yuke Zhu et al. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

A NOTATION

Table 3: Notation.

$s = (s_P, s_V)$	state: proprioceptive / visual part; $h_P,\ h_V$ their encodings
$m \in \{0, 1\}$	visual mask during training; $g(s) \in \{0, 1\}$ gate at execution
$\pi_M,\ \pi_P$	the two *modes* of one network: vision admitted / masked; $\pi_o := \pi_M$ (always-on)
π_g	gated policy $g\,\pi_M + (1{-}g)\,\pi_P$
$\Delta_{\mathrm{cf}}(s)$	value counterfactual $V_\theta(h_P{+}\mathrm{proj}(h_V)) - V_\theta(h_P)$; $\hat{U} = \lvert\Delta_{\mathrm{cf}}\rvert$
$D(s)$	action counterfactual $D_{\mathrm{TV}}(\pi_M(\cdot \mid s), \pi_P(\cdot \mid s))$ (half-L_1); $\le \sqrt{D_{\mathrm{KL}}/2}$ (Pinsker)
$U^\circ(s)$	target visual influence $\lvert V^\pi(s) - \bar{V}^\pi(s_P)\rvert$
$\tau^\star,\ \hat{\tau}$	nominal / estimated value-signal threshold, $\hat{\tau} = \alpha\,\widehat{m}$ (α dimensionless)
κ	post-hoc gate threshold on $D(s)$ (or on $\hat{U}(s)$, value units); $\kappa^\star = c(1{-}\gamma)/(2A_{\max})$ for cost c
$\beta,\ w$	learned soft-attention weight / router weight (in-loop baselines)
q	fixed opening quota (baseline)
$d,\ \rho^\mu$	training state distribution / discounted occupancy of policy μ
$\varepsilon_F, \varepsilon_P, \varepsilon_\tau, \varepsilon_D$	approximation errors: full-input critic, masked critic, threshold, distilled predictor
$\mathrm{proj}(\cdot), T$	projection of h_V into the trunk; episode length
$J(\pi),\ V^\pi,\ A^\pi,\ \gamma$	return, value, advantage, discount factor

B SCOPE REMARKS FOR THEOREM 4.2

Remark B.1 (Scope and honest caveats). (a) The theorem concerns the *supervision label*, not policy optimality. (b) The two forwards in Δ_{cf} share parameters, so their errors are correlated and may partially cancel; the bound uses the worst case. (c) All quantities are snapshots of a nonstationary process (π, d, $\hat{\tau}$ evolve); the statement holds at each iterate, and a clean-reference control (Appendix I) shows the supervision loop is stable rather than self-reinforcing. (d) ε_P absorbs both approximation error and the information loss of conditioning on s_P: the target equation 1 is a function of s_P by construction so this aggregation is part of the target, not an unstated approximation. The value-gap label of our earlier design admits an analogous guarantee (Appendix J), but empirically measures student–teacher quality rather than visual influence (Sec. 5.2).

C PERFORMANCE DECOMPOSITION OF A GATED POLICY

At execution the gate is binary; write $g(s) \in \{0, 1\}$ and let π_M, π_P denote the two *modes* of the trained policy (visual pathway admitted / suppressed, parameters fixed), so $\pi_g = g\,\pi_M + (1{-}g)\,\pi_P$. Let π_g° use the same modes with the ideal gate $g^\circ(s) = \mathbf{1}[U^\circ(s) > \tau^\star]$, and define the mode gap $\delta(s) := \lvert\mathbb{E}_{\pi_M}[A^{\pi_g}(s,\cdot)] - \mathbb{E}_{\pi_P}[A^{\pi_g}(s,\cdot)]\rvert$ with $\delta_{\max} := \sup_{s \in \mathrm{supp}(\rho^{\pi_g^\circ})} \delta(s)$.

Proposition C.1 (Performance decomposition of the gated policy)**.** *Let* $\mathcal{S}_{\neq} := \{s : g(s) \neq g^\circ(s)\}$. *Then*

$$J(\pi_g^\circ) - J(\pi_g) = \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim\rho^{\pi_g^\circ}}\big[\,\ell_{\mathrm{miss}}(s) + \ell_{\mathrm{exp}}(s)\,\big], \tag{6}$$

$$\ell_{\mathrm{miss}}(s) := \mathbf{1}[s \in \mathcal{S}_{\neq},\ g^\circ{=}1]\ \mathbb{E}_{\pi_M}[A^{\pi_g}(s,\cdot)],$$

$$\ell_{\mathrm{exp}}(s) := \mathbf{1}[s \in \mathcal{S}_{\neq},\ g^\circ{=}0]\ \mathbb{E}_{\pi_P}[A^{\pi_g}(s,\cdot)],$$

and consequently

$$J(\pi_g^\circ) - J(\pi_g)\ \le\ \frac{\delta_{\max}}{1-\gamma}\,\rho^{\pi_g^\circ}(\mathcal{S}_{\neq})\ \le\ \frac{\delta_{\max}}{1-\gamma}\,\big(\mu_{\mathrm{band}} + \mu_{\mathrm{cls}}\big), \tag{7}$$

where μ_{band} *is the* $\rho^{\pi_g^\circ}$*-mass of the margin band (Theorem 4.2, with Assumption 4.1 extended to* $\mathrm{supp}(\rho^{\pi_g^\circ})$*) and* μ_{cls} *the mass where the executed gate disagrees with its own label.*

Proof in Appendix D.

Remark C.2 (Reading the decomposition). Equation 6 is exact, and its terms are *signed*: g° thresholds a myopic influence measure, not the return-optimal gate, so a "wrongly closed" state can carry negative advantage; our experiments show the ideal-label gate is not return-optimal, and Equation 6 is the accounting identity that locates the gap. The comparator shares the trained modes; nothing is claimed about mode quality. The measurable handle is μ_{cls} (0.039 at convergence, measured on the training distribution d).

D PROOFS

Proof of Theorem 4.2. By the reverse triangle inequality ($|\cdot|$ is 1-Lipschitz) and Assumption 4.1(i,ii), $|\hat{U}(s) - U^\circ(s)| \le |(V_\theta(h_P + \text{proj}(h_V)) - V_\theta(h_P)) - (V^\pi(s) - \bar{V}^\pi(s_P))| \le \varepsilon_F + \varepsilon_P$. If $U^\circ(s) > \tau^\star + \varepsilon_F + \varepsilon_P + \varepsilon_\tau$, then $\hat{U}(s) \ge U^\circ(s) - \varepsilon_F - \varepsilon_P > \tau^\star + \varepsilon_\tau \ge \hat{\tau}$; the symmetric case is identical. Disagreement is therefore confined to the band.

Proof of Proposition C.1. The performance difference lemma (Kakade & Langford, 2002) gives $J(\pi_g^\circ) - J(\pi_g) = \frac{1}{1-\gamma}\mathbb{E}_{\rho^{\pi_g^\circ}}\mathbb{E}_{a\sim\pi_g^\circ}[A^{\pi_g}(s,a)]$. Off $\mathcal{S}_{\neq}$ the policies share a mode and the inner expectation vanishes ($\mathbb{E}_{a\sim\pi_g}[A^{\pi_g}] = 0$); on $\mathcal{S}_{\neq}$ it equals the cross-mode expectation, whose magnitude equals $\delta(s)$ because the own-mode expectation is zero. For the second inequality, decompose $\mathcal{S}_{\neq} \subseteq \{g \neq \ell\} \cup \{\ell \neq g^\circ\}$: the first event has mass μ_{cls}; the second is contained in the margin band by Theorem 4.2.

Proof of Proposition 4.5. By the performance difference lemma (Kakade & Langford, 2002; Schulman et al., 2015), $J(\pi_o) - J(\pi_g) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\rho^{\pi_o}}\mathbb{E}_{a\sim\pi_o}[A^{\pi_g}(s,a)]$. Where $g(s)$=1, $\pi_g(\cdot \mid s) = \pi_M(\cdot \mid s) = \pi_o(\cdot \mid s)$, so the inner expectation is $\mathbb{E}_{a\sim\pi_g}[A^{\pi_g}] = 0$. Where $g(s)$=0, π_g executes π_P, whose own-mode expectation vanishes, so the inner term equals $\mathbb{E}_{\pi_M}[A^{\pi_g}] - \mathbb{E}_{\pi_P}[A^{\pi_g}] = \int(\pi_M - \pi_P)\, A^{\pi_g}\, \mathrm{d}a \le 2\, D_{\text{TV}}(\pi_M, \pi_P)\, A_{\max}$.

E DOMAIN SPECIFICATIONS

Franka suite (Cond A/B and the Leak/Mixed variants). Table 4 lists the core and spurious parameters, Table 5 the four training conditions, and Fig. 3 shows a representative scene of each.

E.1 FRANKA PANDA SUITE

Table 4: Franka manipulation core and spurious parameter specifications. Core parameters control the underlying physics and reward landscape; spurious parameters affect only appearance or auxiliary dynamics outside the causal path to reward.

Type	Parameter	Nominal	High σ^2	Low σ^2
Core (physics)	Joint stiffness scale	1.00	0.20	0.01
	Action scale	1.00	0.10	0.005
	Target xy range	± 0.15 m	± 0.05 m	± 0.01 m
Spurious (visual / auxiliary)	Table friction	0.95	0.40	0.01
	Joint 7 damping	0.10	0.05	0.001
	Object color (RGB)	128	64	1
	Lighting intensity	1.0	0.4	0.01

The physics observation has $\dim(s_P) = 22$ (joint angles, joint velocities, end-effector pose), excluding the cube and target positions, which the policy must extract visually. The visual observation is a $64 \times 64 \times 3$ RGB frame from a fixed camera. The action is a 7-DoF joint torque command.

Leak variant. Identical to Cond A except that cube and target coordinates are appended to s_P (dimension $22 \to 28$); vision is therefore uninformative beyond proprioception by construction.

Mixed variant. Each episode samples Cond A or Leak with probability $1/2$; s_P is padded to the Leak dimension, so episode identity is inferable from proprioception. Per-episode domain tags are logged.

PiPER. An AgileX PiPER 6-DoF arm in simulation with strong domain randomization ($d_P = 19$), evaluated with a color-temperature perturbation set (PiPER does not read the Franka friction perturbation).

Table 5: Training conditions. Each varies the variance of core (physics) and spurious (visual) signals independently; C/D are earlier-recipe controls used only in Table 6.

Cond.	Label	σ^2_C	σ^2_S	Role in this paper
Reference	Physics only	High	None	physics-only teacher / floor
A	High C, Low S	High	Low	vision-essential, low spurious variation
B	High C, High S	High	High	vision-essential, high spurious variation
C	Low C, Low S	Low	Low	earlier-recipe control (Table 6)
D	Low C, High S	Low	High	earlier-recipe control (Table 6)

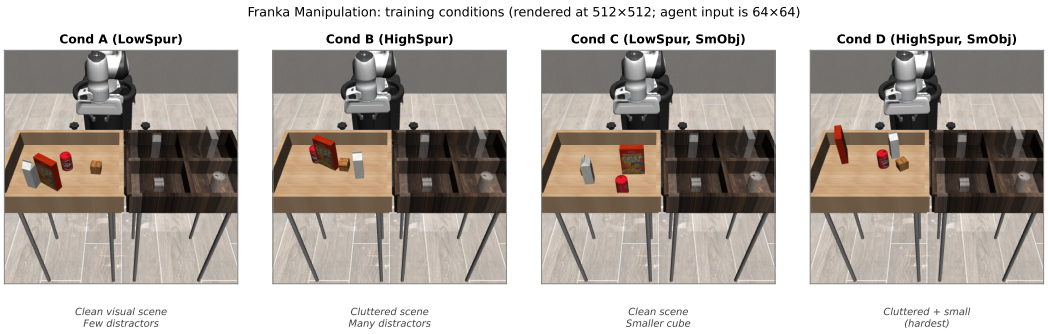

Figure 3: **Visual rendering of the four Franka training conditions.** Each panel shows a representative initial scene from one condition. *LowSpur* conditions (A, C) present cleaner visual workspaces, while *HighSpur* conditions (B, D) add task-irrelevant distractor objects. The *SmObj* suffix marks the earlier-recipe conditions C, D (smaller target), used only in Table 6. See Table 5 for the formal (σ^2_C, σ^2_S) specification.

F The Gradient-Corruption Premise: Full Results

Table 6 reports the naive-joint, stop-gradient, and probing controls of Sec. 5.1 in full.

Table 6: The gradient-corruption premise, tested (earlier recipe generation; internally controlled, not comparable to Tables 2/1). Probing: ridge R^2 on ground-truth core parameters, episode-level GroupKFold. Joint training does not corrupt the physics encoder: naive $\geq$ clean on both conditions.

	Probing R^2		Return	
	Cond B	Cond D	Cond B	Cond D
Clean (Phase-1)	$+0.021$	-0.052	86.5	214.4
Naive-joint	$+0.075 \pm .002$	$-0.016 \pm .001$	172.4 ± 62.0	452.9 ± 41.1
Stop-grad (vis)	$+0.055 \pm .007$	$-0.038 \pm .005$	81.1 ± 24.9	415.3 ± 76.2
Stop-grad (phys)	$+0.086 \pm .001$	$-0.007 \pm .001$	122.6 ± 30.1	368.9 ± 121.0
Full (S-DVG)	–	–	160.9 ± 17.3	488.1 ± 16.6

G Training Details and Recipes

Recipe generations. Tables 6 and 8 (in part) use the earlier recipe; all other results use the current recipe: loss normalization (value targets normalized by a running return std), the counterfactual-magnitude label with α=0.069 ($\tau^\star \approx 0.91$ in value units, calibrated on 786k pooled training-time samples), and mask dropout p=0.1 (batch fraction processed with the visual pathway masked). Any-gate policies for post-hoc gating are NoGate policies trained with the same mask dropout (p=0.1; p=0.5 ablation in Sec. 5.6). Shared PPO settings are listed in Table 7.

Loss-scale failure mode. After a physics fix rescaled returns $\sim$70×, the weighted value loss exceeded the gate loss by 28× and the sparsity term by 376×; loss normalization reduces the ratio to 6–23× and restores gate–label tracking.

G.1 Gumbel-Sigmoid with Straight-Through Estimator

Given the gate logit $z_t = \mathbf{w}_g^\top [h_P; h_V] + b_g$ and logistic noise $\eta \sim \mathrm{Logistic}(0,1)$ (equivalently $\eta = \log u - \log(1-u)$ with $u \sim \mathrm{Uniform}(0,1)$), we compute:

$$g_t^{\mathrm{soft}} \;=\; \sigma\Big(\tfrac{z_t+\eta}{\tau_g}\Big) \in (0,1), \tag{8}$$

$$g_t^{\mathrm{hard}} \;=\; \mathbf{1}\big[g_t^{\mathrm{soft}} > 0.5\big] \in \{0,1\}, \tag{9}$$

$$g_t \;=\; g_t^{\mathrm{hard}} + \big(g_t^{\mathrm{soft}} - \mathrm{stop_grad}(g_t^{\mathrm{soft}})\big), \tag{10}$$

where τ_g is the Gumbel temperature. The forward pass uses g_t^{hard} (binary, decisive at execution); the backward pass flows through g_t^{soft} via the STE identity equation 10, keeping the gate differentiable.

G.2 Loss Term Definitions

Distillation loss.

$$\mathcal{L}_{\mathrm{distill}} \;=\; D_{\mathrm{KL}}\big(\pi_{\mathrm{ref}}(\cdot \mid s_P) \,\|\, \pi_M(\cdot \mid s_P, s_V)\big), \tag{11}$$

anchors the mixed policy to the reference's physics behavior and prevents visual gradients from corrupting the physics-only mode.

Gate supervision loss. We supervise the gate against the binary indicator of whether the reference gap $\Delta_t^{\mathrm{vg}} = V_{\mathrm{mix}} - V_{\mathrm{ref}}$ exceeds a fixed threshold τ_{vg} (earlier recipe; distinct from the nominal counterfactual threshold $\tau^\star$ of Sec. 4.1):

$$\mathcal{L}_{\mathrm{gate}} \;=\; \mathrm{BCE}\big(g_t,\, \mathbf{1}[\Delta_t^{\mathrm{vg}} > \tau_{\mathrm{vg}}]\big), \quad \tau_{\mathrm{vg}} = 0.1. \tag{12}$$

The indicator-and-BCE formulation matches the binary nature of the gate decision and is robust to scale variation in Δ_t^{vg}.

Sparsity regularizer.

$$\mathcal{L}_{\mathrm{sparse}} \;=\; \mathbb{E}\big[\,\sigma(z_t)\,\big] \tag{13}$$

encourages the gate to be closed by default. We regularize the logit's sigmoid rather than the hard sample so that gradients flow smoothly even when the hard gate is fully closed. The Phase-3 objective is $\mathcal{L} = \mathcal{L}_{\mathrm{PPO}} + \lambda_d\,\mathcal{L}_{\mathrm{distill}} + \lambda_g\,\mathcal{L}_{\mathrm{gate}} + \lambda_s\,\mathcal{L}_{\mathrm{sparse}}$, with weights in Table 7.

H Threshold Designs: Full Results

Provenance per design: absolute τ_{vg} (earlier recipe, campaign of $n \leq 7$); offline-calibrated relative τ_{vg} (smoke, 0.3M); quantile (pilot with loss normalization, 0.5M, and Leak 1M); Δ^+- and value-anchored online thresholds (current recipe, 1M, 2 seeds each).

Table 7: Hyperparameters of the three-phase S-DVG pipeline in the earlier recipe (Tables 6 and 8). The Reference column corresponds to Phase 1 (physics-only PPO); Burn-in to Phase 2 (auxiliary visual pre-training); S-DVG to Phase 3 (selective gating in the full environment). The current recipe keeps these values and adds loss normalization, the counterfactual-magnitude label with α=0.069, and mask dropout p=0.1 (*Recipe generations*, Appendix G).

Hyperparameter	**Phase 1 (Reference)**	**Phase 2 (Burn-in)**	**Phase 3 (S-DVG)**
Total timesteps / samples	2×10^6 steps	2×10^4 samples	2×10^6 steps
Epochs	–	20	–
Steps per rollout	4,096	–	4,096
Minibatch size	128	256	128
PPO epochs	8	–	4
Discount γ	0.99	–	0.99
GAE λ	0.95	–	0.95
PPO clip range	0.2	–	0.2
Learning rate (overall)	3×10^{-4}	3×10^{-4}	3×10^{-4}
Visual encoder lr mult.	–	1.0 (×base)	0.1 (×base)
LR schedule	Linear decay	Constant	Linear decay
Gradient clip norm	0.5	1.0	0.5
Entropy coef.	0.005	–	0.005
Value coef.	0.5	–	0.5
Frame stack	3	1	3
Image size	–	64×64 / 32×32	64×64 / 32×32
Visual skip	–	–	4
λ_d (distill)	–	–	0.1
λ_g (gate BCE)	–	–	0.5
λ_s (sparsity)	–	–	0.05
τ_{vg} (value-gap threshold)	–	–	0.1
τ_g (Gumbel temperature)	–	–	1.0
Random seeds	$\{0,1,2\}$ (more in the Table 6 campaign)		

Table 8: Five threshold designs for the teacher-relative label $\mathbf{1}[V_{\text{mix}}-V_{\text{ref}}>\tau_{\text{vg}}]$, each failing instructively (entries are label rates unless marked; measured across recipe generations; provenance per design is listed below). The defect is the label, not the threshold.

Design	Essential (A)	Redundant (Leak)	Failure mode
Absolute τ_{vg}	open (0.85–1.0)	open	scale-fragile
Offline-calibrated rel.	closed (0.000)	–	unreachable τ_{vg}
Quantile q=0.25	0.35, −37% R	0.22	quota pinned
Δ^+-anchored	0.30 (calibration target)	0.32	signal self-erased
Value-anchored	0.16	0.38–0.40	inverted

I Supervision-Loop Stability (Clean-Reference Control)

Under the earlier recipe on Cond B, replacing the self-label with a label from a reference trained in a clean environment changes the gate open-rate (0.083 vs 0.307) but not return: paired difference +0.007 (95% CI $[-0.024, 0.037]$, n=200 matched episodes, p=0.16), with label rate flat over training on one seed and drifting +3.3pp on the other. The supervision loop is stable rather than self-reinforcing; the return-invariance is the signature of disagreement occurring where the two modes barely differ (Prop. C.1).

J Value-Gap Label Theory and Diagnostic Metrics

Value-gap variant of Theorem 4.2. With targets $V^\circ_{\text{mix}} := V^\pi$ and $V^\circ_P(s_P) := \mathbb{E}_d[V^{\pi_{\text{ref}}}(s) \mid s_P]$ for the physics-only reference π_{ref}, and $\Delta^\circ = \max(0, V^\circ_{\text{mix}} - V^\circ_P)$, the same margin-band statement holds with band $\varepsilon_F + \varepsilon_P$, where the two errors now refer to the mixed and reference value heads and

no ε_τ term appears because τ_{vg} is fixed ($\max(0,\cdot)$ is 1-Lipschitz). The label is correct where it is far from τ_{vg}, but, as Sec. 5.2 shows, it is correct about the wrong quantity.

Lemma J.1. *Let $s_V = (s_V^{\mathrm{task}}, s_V^{\mathrm{spur}})$ and suppose the spurious part s_V^{spur} enters neither the transition kernel $P(s' \mid s_P, s_V^{\mathrm{task}}, a)$ nor the reward. Then $\pi^*(\cdot \mid s_P, s_V^{\mathrm{task}}, s_V^{\mathrm{spur}}) = \pi^*(\cdot \mid s_P, s_V^{\mathrm{task}}, s_V^{\mathrm{spur}\prime})$ for all $s_V^{\mathrm{spur}}, s_V^{\mathrm{spur}\prime}$; if in addition s_V^{task} is a function of s_P (the Leak domain), π^* is invariant to all of s_V.*

Proof. By the Bellman optimality equation,

$$\pi^*(s) = \arg\max_a \Big[r(s,a) + \gamma \sum_{s'} P(s' \mid s, a)\, V^*(s') \Big].$$

Since neither P nor r depends on s_V^{spur} by assumption, the Q-function and thus π^* are independent of it; when s_V^{task} is determined by s_P, the same argument removes the dependence on s_V altogether. The spurious-part condition holds by construction in all our domains (object color and lighting enter only the renderer, Table 4); the Leak condition holds in Leak and in the Leak episodes of Mixed. In Cond A/B the target position enters the reward and is observed only through s_V, so vision is essential there. □

This appendix defines the three diagnostic metrics used during development, which we release with the code.

J.1 Spurious Sensitivity Index (SSI)

Let $R_{\mathrm{in}} = J(\pi_\phi, s_{P0}, s_{V0})$ be the policy's nominal performance, and $R_{\mathrm{spur}} = \mathbb{E}_{s_V \sim q_S^{\mathrm{high}}}[J(\pi_\phi, s_{P0}, s_V)]$ be performance when only spurious signals vary (physics fixed at s_{P0}):

$$\mathrm{SSI} = \frac{R_{\mathrm{in}} - R_{\mathrm{spur}}}{|R_{\mathrm{in}}| + \epsilon_0}, \qquad \Delta\mathrm{SSI} = \mathrm{SSI}(\pi_\phi) - \mathrm{SSI}(\pi_{\mathrm{ref}}). \tag{14}$$

$\mathrm{SSI} = 0$ indicates spurious invariance; $\mathrm{SSI} > 0$ indicates interference. $\mathrm{SSI}(\pi_{\mathrm{ref}}) \approx 0$ by construction since π_{ref} has no visual input, so ΔSSI quantifies the cost of visual exposure relative to the reference.

J.2 Core Generalization Index (CGI)

Let $R_{\mathrm{ood}}(\xi)$ denote performance when core parameters are drawn from a distribution ξ times wider than training, with spurious fixed at s_{V0}:

$$\mathrm{CGI}(\xi) = \frac{R_{\mathrm{ood}}(\xi)}{|R_{\mathrm{in}}| + \epsilon_0}, \qquad \Delta\mathrm{CGI}(\xi) = \mathrm{CGI}^{\mathrm{ref}}(\xi) - \mathrm{CGI}(\xi). \tag{15}$$

$\mathrm{CGI}(1) = 1$ by construction; a policy with strong principle learning maintains $\mathrm{CGI}(\xi) \approx 1$ for large ξ. We report CGI at $\xi \in \{1.5, 2.0\}$. Together, SSI and CGI form a 2-D behavioral diagnostic: low SSI & high CGI indicate principle-learned behavior; high SSI & low CGI indicate joint overfitting.

J.3 Visual Relevance Score (VRS)

While SSI and CGI are black-box, VRS is a white-box gradient metric measuring how much the visual branch contributes to action decisions. Let $h_P = f_P(s_P)$ and $h_V = f_V(s_V)$:

$$\mathrm{VRS} = \frac{\mathbb{E}\big[\|\partial \log \pi / \partial h_V\|\big]}{\mathbb{E}\big[\|\partial \log \pi / \partial h_P\|\big] + \mathbb{E}\big[\|\partial \log \pi / \partial h_V\|\big] + \epsilon_0} \in [0,1]. \tag{16}$$

A high VRS in a fully-spurious condition indicates a representation problem invisible to behavior-only metrics.

K MIXED-DOMAIN COLLAPSE TRAJECTORIES

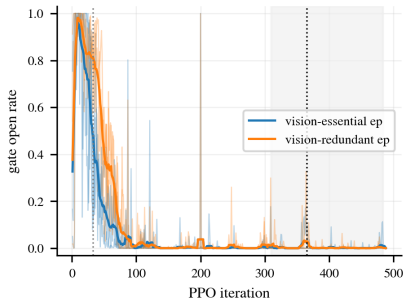

Figure 4: Mixed-domain training collapses the gate in both episode types (3 seeds, thin lines; thick: moving average; on-policy measurements): the early transient is inverted (at iter≈33, first dotted line: 0.77 open in vision-redundant vs. 0.22 in vision-essential episodes, gap 0.546) and decays with intermittent rebounds to sustained zero by iter≈365 (second dotted line; shaded: seed range 310–484); per-domain adaptivity was co-evolution.

L POST-HOC GATING ABLATIONS

Cond A, frozen any-gate policy (p=0.1), entries are retention % (real MACs saved %). Table 9 (left): probe length k versus re-probe interval K (action counterfactual). Right: signal combinations at k=5, K=∞; AND opens only when both signals say open (closes otherwise: maximal savings), OR opens when either does (minimal savings).

Table 9: Post-hoc gating ablations on Cond A: probe length and re-probe interval (left); signal combinations (right).

$k\backslash K$	25	50	100	∞
k=1	52 (64)	54 (71)	54 (74)	55 (81)
k=5	64 (50)	64 (60)	61 (64)	64 (80)
k=10	82 (35)	76 (50)	77 (58)	79 (74)

Combination	Cond A	Mixed
value	67 (80)	106 (52)
action (kl)	74 (51)	87 (52)
value AND action	54 (85)	99 (76)
value OR action	74 (41)	96 (29)

M BOUND TIGHTNESS

Prop. 4.5 bounds the return lost by a post-hoc gate by the divergence mass it closes. Fig. 5 plots the realized loss $J(\pi_o) - J(\pi_g)$ against the closed divergence $\mathbb{E}_\rho[\mathbf{1}[g=0]\, D(s)]$ for the Cond A/B comparison runs. The realized loss per unit of closed divergence is 0.03 (Cond A) and 0.04 (Cond B). The bound holds at every threshold but is loose by roughly four orders of magnitude, as expected of a performance-difference bound whose constant $2A_{\max}/(1-\gamma)$ is worst-case ($1/(1-\gamma) = 100$ alone at $\gamma = 0.99$); realized loss and bound rise together as κ increases, which is the ordering the bound certifies.

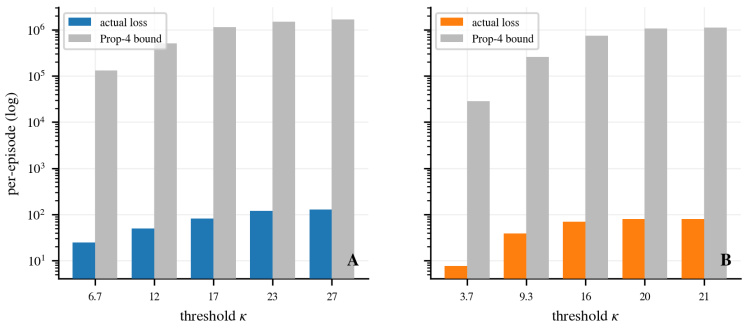

Figure 5: Bound tightness on Cond A (left) and Cond B (right): per-episode realized return loss (colored) against the bound of Prop. 4.5 evaluated on the same episodes (grey), for five thresholds κ of the value-counterfactual gate (value units; the bound applies to any gate); log scale.

N Distilled Predictor and Gate Trajectories

Fig. 6 shows the distilled predictor $\hat{D}(s_P)$ (Remark 4.9) in the Mixed domain, split by episode type. The two distributions overlap heavily, yet the open rate they induce separates the episode types over a range of thresholds (at κ=0.026: Cond A episodes 0.70 open, Leak episodes 0.42), so the predictor discriminates by episode type without reading the image; the retention and savings it achieves are in Sec. 5.6. Fig. 7 shows the per-step open probability within Mixed-domain episodes: after the approach phase (end-effector–target distance settles within the first $\sim$30 steps) the gate stays more open in Cond A episodes ($\approx$0.6) than in Leak episodes ($\approx$0.3) for the rest of the episode, i.e. the separation is sustained rather than confined to the approach.

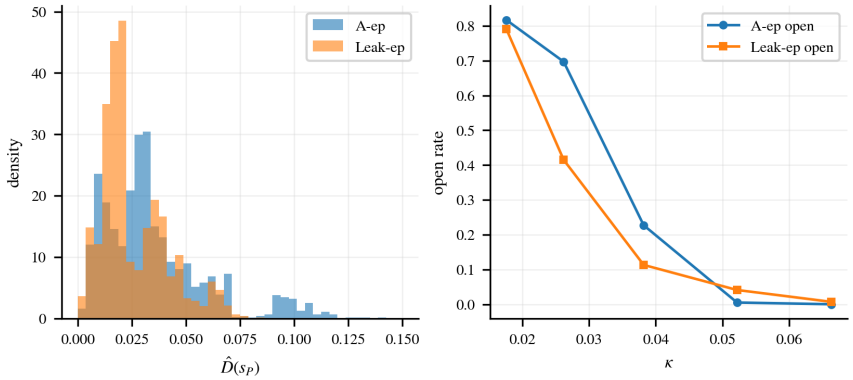

Figure 6: Distilled proprioception-only predictor in the Mixed domain: histogram of $\hat{D}(s_P)$ by episode type (left) and the open rate it induces versus threshold κ (right).

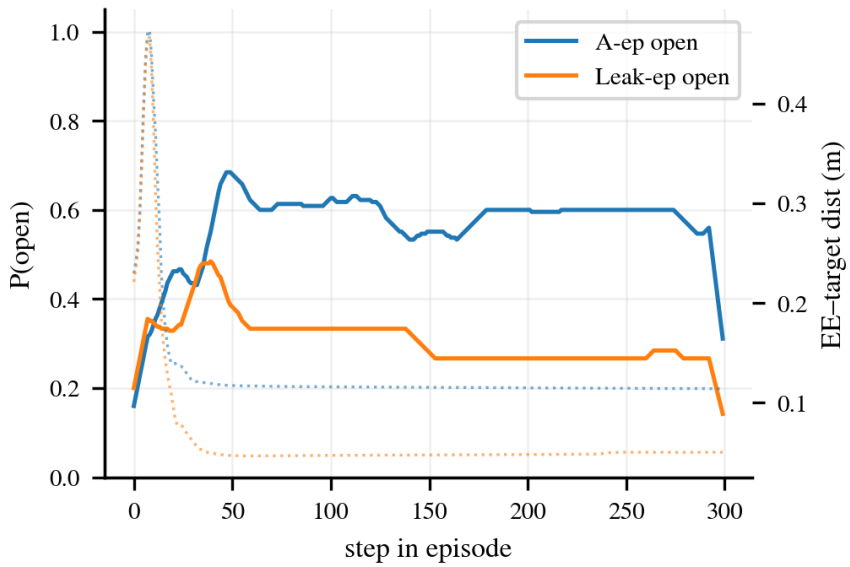

Figure 7: Per-step open probability within Mixed-domain episodes, averaged over episodes and split by episode type (solid, left axis), with the mean end-effector–target distance (dotted, right axis).

O Post-hoc Gating Beyond Cond A

Table 10: Post-hoc gating on frozen any-gate policies (3 seeds, real MACs); "Random" closes episodes at the same rate without a signal.

Domain	Signal / rule	Open	Retention	Real MACs saved	Random @ budget
Cond B	value / episode	0.738	85.8%	25.9%	74.1%
Leak	any	–	90–97%	–	90–92%
Mixed	value / episode-K50	0.455	R=231.0	53.8%	R=213.7; label ref. 213.4
PiPER	value / episode-K50	–	99.4%	39.5%	96.2%

P In-loop Mechanisms: Return Retention

Table 11 reports return retention vs. NoGate for the mechanisms of Table 1.

Table 11: Return retention of the in-loop mechanisms.

	Cond A	Leak	Mixed A-ep	Mixed Leak-ep
Hard gate (cf-mag)	82.4%	90.2%	79.8%	94.5%
Soft attention β	45.1%	87.3%	67.1%	97.1%
Two-expert router w	92.6%	93.0%	95.5%	88.7%
Reward cost (c=5.2%)	93.4%	–	–	–
APRiL	2.04× NoGate	1.12×	–	–