CORE: Uncertainty-Gated Policy Optimization for Reliable Performance Under Distribution Shift

Shawn Azdam

***Abstract*— We present Conservative Online Reliability Enforcement (CORE), an uncertainty-gated policy optimization framework for robust policy updates under distribution shift. The method combines risk-aware trust-region learning with online promote/monitor/rollback control driven by uncertainty-calibrated signals. Across matched-seed benchmark and scenario-model evaluations, our approach improves reliability-floor behavior while preserving fixed-budget update control. We provide algorithm design, theory-guided gate rationale, and diagnostics that clarify where gated updates help and where limitations remain.**

I. Introduction

Reliable policy optimization under distribution shift is still a core failure mode in robotics learning systems. Modest changes in latency, sensing quality, or scene geometry can destabilize updates that appear safe under nominal evaluation. Recent work improves planning-time safety and shift robustness [1]–[4], but update-time rollback control is less explicitly specified.

Our approach addresses this gap with an uncertainty-gated policy optimization loop. It combines a trust-region surrogate update, an ensemble-disagreement penalty, and an online promote/monitor/rollback gate driven by a conservative lower-bound proxy. The novelty is update-time control: unstable candidates are rejected before downstream validation.

On the MetaWorld shifted suite [5], Section IV anchors primary confirmatory claims to PPO-CVaR and reports AdaptManip as an exploratory paper-profile sensitivity lane with explicit fairness caveats. Mean remains secondary/sensitivity-only, and $N = 30$ is sensitivity-only with smaller mean deltas. The unofficial LatencyAware proxy lane remains supplementary-only and excluded from primary claim tables.

Contributions are:

1) An uncertainty-gated update rule that combines risk-aware trust-region optimization with per-iteration promote/monitor/rollback control.
2) A floor-focused theory package with one calibration-conditional deployed-gate design lemma plus diagnostic heuristic envelopes (optional one-sided and false-promote sketches) and matched-seed empirical-floor diagnostics.
3) A reliability-first evidence package with matched-seed diagnostics, targeted seed expansion, ablations, robustness checks, and explicit scope boundaries.

This study focuses on update-time robustness under controlled perturbations. Evidence is simulation-only, and we do not claim deployment-ready safety, hardware validation, or real-robot transfer guarantees.

II. Related Work

Recent work improves contact grounding, long-horizon world-model planning, and distribution-shift robustness [1], [2], [4], [6]–[11]. These advances are complementary to our setting, but most intervene at planning/calibration time or through implicit policy-selection loops rather than explicit update-time commit control.

Sim-to-real and policy-scaling foundations define the current baseline bar [12]–[19]. The unresolved gap for this paper is a conservative, online reject/commit rule for policy updates under fixed evaluation budgets.

Our method is positioned as a policy-optimization method: uncertainty-penalized trust-region updates with an explicit online promote/monitor/rollback gate. Compared with nearby alternatives, the main distinction is update-time rollback control rather than planning-time calibration or model-scale expansion [1], [2], [7]–[11], [20]. In this comparison set, prior lines emphasize planning-time safety proxies or implicit update selection, while our approach uses explicit promote/monitor/rollback control inside the policy-update loop under matched-seed validation.

III. Theory and Design Rationale

This section states one deployed-gate design lemma plus diagnostic heuristic envelopes (optional one-sided false-promote sketch and matched-seed floor remark) that motivate when the gate should promote or reject candidate updates. These are assumption-scoped design-rationale bounds, not claims of new concentration theory.

We consider partially observed control under distribution shift. Let latent state be $s_t \in \mathcal{S}$, observation $o_t \in \mathcal{O}$, action $a_t \in \mathcal{A}$, and history $h_t = (o_{0:t}, a_{0:t-1})$. A policy $\pi_\theta(a_t \mid h_t)$ induces trajectory τ under shift condition $\xi \sim p(\xi)$.

We optimize a risk-aware objective

$$J(\theta) = \mathbb{E}_{\xi\sim p(\xi),\, \tau\sim p(\cdot|\pi_\theta,\xi)}\left[\sum_{t=0}^{T}\gamma^t r_t\right] - \lambda\,\mathrm{CVaR}_{\alpha,\,(\tau,\xi)\sim p_\theta}(-R(\tau,\xi))\,, \tag{1}$$

where $p_\theta(\tau,\xi) := p(\xi)\,p(\tau \mid \pi_\theta, \xi)$, $R(\tau,\xi) = \sum_{t=0}^{T}\gamma^t r_t$, $\lambda \geq 0$ controls robustness, and α sets the tail-risk level. We use $\alpha = 0.4$ (reported as CVaR_{40}). For gate bounds, we separately analyze the return component

$$G(\theta) := \mathbb{E}_{\xi\sim p(\xi),\, \tau\sim p(\cdot|\pi_\theta,\xi)}\left[\sum_{t=0}^{T}\gamma^t r_t\right],$$

while CVaR remains part of the optimization objective in Eq. 1.

At iteration k, our method maintains policy parameters θ_k, world-model parameters ϕ_k, and an uncertainty functional induced by ensemble value disagreement. Let $\{Q_{\phi_k}^{(m)}\}_{m=1}^{M}$ be M bootstrap value heads. For candidate policy π_θ under ensemble snapshot ϕ_k, define

$$\widehat{U}_k^{\mathrm{wm}}(\theta) = \frac{1}{|\widetilde{\mathcal{S}}_k|} \sum_{i \in \widetilde{\mathcal{S}}_k} \frac{1}{|\tau_k^{\theta,(i)}|} \sum_t \mathrm{Var}_m\Big(Q_{\phi_k}^{(m)}(h_t, a_t)\Big), \tag{2}$$

Here $\widetilde{\mathcal{S}}_k := \widetilde{\mathcal{S}}_k(\theta)$ is a short candidate-policy rollout set produced by world-model snapshot ϕ_k from seed-matched incumbent replay starts; gating does not require hardware execution of the candidate policy. $\mathrm{Var}_m(\cdot)$ is the sample variance across the M ensemble heads at fixed (h_t, a_t). Inspired by recent robust planning and world-model work [1], [2], [9], [10], [21], we use the constrained surrogate update

$$\begin{aligned} \theta_k^\star = \ & \arg\max_\theta \ \hat{J}_k(\theta) - \beta \widehat{U}_k^{\mathrm{wm}}(\theta) \\ & \text{s.t.} \quad D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\theta_k}) \le \varepsilon, \end{aligned} \tag{3}$$

Here $\hat{J}_k$ denotes the trust-region surrogate estimate of the risk-aware objective J in Eq. 1; it is distinct from the matched-seed gate statistics used in Eq. 4 and Eq. 7. Eq. 3 uses world-model rollout uncertainty $\widehat{U}_k^{\mathrm{wm}}(\theta)$; gate bounds below use matched-seed uncertainty estimates denoted $\widehat{U}_k^{\mathrm{ms}}(\cdot)$. The trust-region radius is ε and the uncertainty weight is β. The gate in Eq. 7 then maps candidate $\theta_k^\star$ to committed iterate θ_{k+1}.

Assume:

1) local surrogate fidelity: $|G(\theta) - \hat{G}_k(\theta)| \le e_k(\theta)$,
2) uncertainty dominance with empirically calibrated constants: $e_k(\theta) \le c_u \widehat{U}_k^{\mathrm{ms}}(\theta) + c_0$,
3) bounded policy step from the trust region.

Here (c_u, c_0) are fitted from held-out diagnostics (Sec. IV), so Design Lemma 1 is conditional on calibration; Sec. IV reports a disjoint fit/holdout split with holdout Assump.-2 coverage 94.5%. Misspecification sensitivity is summarized by worst gate-agreement 0.055 over 25 perturbations. Define surrogate gain $\hat{\Delta}_k^{\mathrm{sur}} = \hat{G}_k(\theta_k^\star) - \hat{G}_k(\theta_k)$.

Design Lemma 1 (Two-sided improvement bound): Under assumptions (1)–(3),

$$\begin{aligned} G(\theta_k^\star) - G(\theta_k) \ge \ & \hat{\Delta}_k^{\mathrm{sur}} \\ & - c_u\Big(\widehat{U}_k^{\mathrm{ms}}(\theta_k^\star) + \widehat{U}_k^{\mathrm{ms}}(\theta_k)\Big) - 2c_0. \end{aligned} \tag{4}$$

Proof. By assumption (1),

$$\begin{aligned} G(\theta_k^\star) &\ge \hat{G}_k(\theta_k^\star) - e_k(\theta_k^\star), \\ G(\theta_k) &\le \hat{G}_k(\theta_k) + e_k(\theta_k). \end{aligned}$$

Subtracting gives

$$G(\theta_k^\star) - G(\theta_k) \ge \hat{\Delta}_k^{\mathrm{sur}} - e_k(\theta_k^\star) - e_k(\theta_k).$$

Applying assumption (2) to both terms yields Eq. 4. □

If additionally $\widehat{U}_k^{\mathrm{ms}}(\theta_k) \le \widehat{U}_k^{\mathrm{ms}}(\theta_k^\star)$ on the matched-seed estimator used for gate decisions, then

$$G(\theta_k^\star) - G(\theta_k) \ge \hat{\Delta}_k^{\mathrm{sur}} - 2\big(c_u \widehat{U}_k^{\mathrm{ms}}(\theta_k^\star) + c_0\big). \tag{5}$$

Intuition: when $\widehat{U}_k^{\mathrm{ms}}(\theta_k) \le \widehat{U}_k^{\mathrm{ms}}(\theta_k^\star)$, the two-sided correction collapses to the candidate-only term. In deployment, the method treats this one-sided condition as optional and thresholds the two-sided score; the one-sided term is retained only as an auxiliary diagnostic. This follows by direct substitution of $\widehat{U}_k^{\mathrm{ms}}(\theta_k) \le \widehat{U}_k^{\mathrm{ms}}(\theta_k^\star)$ into Eq. 4. If empirical gain exceeds the uncertainty-corrected error term in Eq. 5, expected return improves in the local approximation regime. We use this as design rationale for gating, not as a universal guarantee. Empirical one-sided diagnostic observability gives AdaptManip-paired hold/violation 0.167/0.833 over 84 matched rows, so this simplification is diagnostic-only and excluded from deployed gate decisions and confirmatory claims. In practice, gate bounds apply to the matched-seed return component; CVaR remains explicitly optimized in Eq. 1 and reported as a primary evaluation criterion in experiments. Assumption (4) (surrogate-to-matched-seed bridge): under trust-region local validity and seed-matched evaluation, the empirical gain proxy tracks surrogate gain with bounded error, $|\tilde{\Delta}_k - \hat{\Delta}_k^{\mathrm{sur}}| \le \epsilon_k^{\mathrm{bridge}}$. Objective-bridge identity: with $C(\theta) := \mathrm{CVaR}_\alpha(-R)$ and $J(\theta) = G(\theta) - \lambda C(\theta)$,

$$J(\theta_k^\star) - J(\theta_k) = \big(G(\theta_k^\star) - G(\theta_k)\big) - \lambda\big(C(\theta_k^\star) - C(\theta_k)\big).$$

If $|C(\theta_k^\star) - C(\theta_k)| \le B_k^C$ for a drift budget $B_k^C \ge 0$, then

$$J(\theta_k^\star) - J(\theta_k) \ge \big(G(\theta_k^\star) - G(\theta_k)\big) - \lambda B_k^C.$$

Using Assumption (4), $G(\theta_k^\star) - G(\theta_k) \ge \underline{\Delta}_k^{(2)} - \epsilon_k^{\mathrm{bridge}}$. Combining this with the CVaR drift bound yields the operational bridge

$$J(\theta_k^\star) - J(\theta_k) \ge \underline{\Delta}_k^{(2)} - \epsilon_k^{\mathrm{bridge}} - \lambda B_k^C.$$

Proposition 1: Gate objective non-degradation (assumption-scoped): if Assumptions (1)–(4) hold and a promoted update satisfies

$$\underline{\Delta}_k^{(2)} \ge \tau_{\text{green},k} \quad \text{and} \quad \tau_{\text{green},k} \ge \epsilon_k^{\mathrm{bridge}} + \lambda B_k^C,$$

then $J(\theta_k^\star) - J(\theta_k) \ge 0$.

Proof. From the operational bridge inequality,

$$J(\theta_k^\star) - J(\theta_k) \ge \underline{\Delta}_k^{(2)} - \epsilon_k^{\mathrm{bridge}} - \lambda B_k^C.$$

Substitute $\underline{\Delta}_k^{(2)} \ge \tau_{\text{green},k}$ and $\tau_{\text{green},k} \ge \epsilon_k^{\mathrm{bridge}} + \lambda B_k^C$. □

Hence, Proposition 1 is explicitly assumption-scoped to bridge error and CVaR-drift budget control; we therefore report CVaR as a primary endpoint and do not claim an unconditional CVaR-gated guarantee in this paper. *Implementation interpretation.* In code, $\tilde{\Delta}_k$ is the matched-seed return delta from paired incumbent/candidate evaluations, and $\widehat{U}_k^{\mathrm{ms}}(\theta_k^\star), \widehat{U}_k^{\mathrm{ms}}(\theta_k)$ are estimated on the same pairs. Logs expose these channels (`base_score`, `observed_score`, `total_penalty`); deployed gating uses the two-sided score, with the one-sided score retained for diagnostics. For

robustness to one-sided simplification violations, we compute both diagnostic scores

$$\underline{\Delta}_k^{(1)} = \tilde{\Delta}_k - 2\big(c_u \widehat{U}_k^{\mathrm{ms}}(\theta_k^\star) + c_0\big),$$

$$\underline{\Delta}_k^{(2)} = \tilde{\Delta}_k - c_u\Big(\widehat{U}_k^{\mathrm{ms}}(\theta_k^\star) + \widehat{U}_k^{\mathrm{ms}}(\theta_k)\Big) - 2c_0, \quad (6)$$

and use $\underline{\Delta}_k := \underline{\Delta}_k^{(2)}$ for gate thresholding while logging $\underline{\Delta}_k^{(1)}$ as an auxiliary diagnostic.

Let $\underline{\Delta}_k$ be the uncertainty-corrected lower-bound proxy above. Our approach applies triage thresholds:

$$\begin{aligned} &\text{promote if } \underline{\Delta}_k \geq \tau_{\text{green},k}, \\ &\text{monitor if } \tau_{\text{yellow},k} \leq \underline{\Delta}_k < \tau_{\text{green},k}, \\ &\text{rollback if } \underline{\Delta}_k < \tau_{\text{yellow},k}. \end{aligned} \quad (7)$$

Notation recap for gate decisions: $\hat{\Delta}_k^{\mathrm{sur}}$ denotes surrogate gain in the local bound, $\tilde{\Delta}_k$ denotes raw matched-seed gain used by the gate, and $\underline{\Delta}_k$ is the uncertainty-corrected lower-bound score thresholded by $(\tau_{\text{yellow},k}, \tau_{\text{green},k})$. We use adaptive thresholds $(\tau_{\text{green},k}, \tau_{\text{yellow},k})$ with $\tau_{\text{yellow},k} < \tau_{\text{green},k}$ at each update. The gate is applied online before downstream robustness runs. The monitor and rollback states both keep the incumbent under the same rollout budget; rollback additionally pivots next-cycle hyperparameters.

To reduce threshold miscalibration across tasks/shifts, we use adaptive hysteresis in deployment diagnostics: let $\mathcal{D}_{k-1}$ be a strictly past rolling window of deployed two-sided scores $\underline{\Delta}^{(2)}$, let $S_{k-1} := \mathrm{Spread}(\mathcal{D}_{k-1})$, let $q_{k-1} := Q_q(\mathcal{D}_{k-1})$, and define

$$\tau_{\text{yellow},k} = \max(\tau_{\text{yellow}},\ q_{k-1} + \eta S_{k-1}), \quad (8)$$

$$\tau_{\text{green},k} = \tau_{\text{yellow},k} + \delta S_{k-1}, \quad (9)$$

where Spread is a robust quantile gap (e.g., $Q_{0.9} - Q_{0.1}$) on the score window and we use $q = 0.85$ in experiments. This keeps gate sensitivity high for transient score drops while avoiding over-triggering in regimes with consistently high but non-catastrophic dynamics. In implementation, $\mathcal{D}_{k-1}$ is a strictly past rolling window of matched-seed two-sided scores (excluding the current matched-seed batch), so Eqs. 8–9 are dimensionally matched to Eq. 7.

Using a matched-seed set $\mathcal{S}_k$, we estimate:

$$\tilde{\Delta}_k = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \Big(R(\tau_k^{\star,(i)}, \xi^{(i)}) - R(\tau_k^{(i)}, \xi^{(i)}) \Big),$$

$$\widehat{U}_k^{\mathrm{ms}}(\theta_k^\star) = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \left[\frac{1}{|\tau_k^{\star,(i)}|} \sum_t \mathrm{Var}_m\Big(Q_{\phi_k}^{(m)}(h_t, a_t) \Big) \right].$$

$$\widehat{U}_k^{\mathrm{ms}}(\theta_k) = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \left[\frac{1}{|\tau_k^{(i)}|} \sum_t \mathrm{Var}_m\Big(Q_{\phi_k}^{(m)}(h_t, a_t) \Big) \right].$$

Diagnostic Heuristic 1 (False-promote envelope): Let $n_k = |\mathcal{S}_k|$. For this diagnostic heuristic, $S(\tau, \xi)$ denotes undiscounted simulation benchmark rollout returns from matched-seed evaluations. For each matched seed i, define the one-sided diagnostic variable

$$\begin{aligned} Z_k^{(i)} = \ &\Big(S(\tau_k^{\star,(i)}, \xi^{(i)}) \\ &- S(\tau_k^{(i)}, \xi^{(i)}) \Big) \\ &- 2\Big(c_u \widehat{U}_k^{\mathrm{ms},\star,(i)} + c_0 \Big), \end{aligned}$$

where $\widehat{U}_k^{\mathrm{ms},\star,(i)} := \frac{1}{|\tau_k^{\star,(i)}|} \sum_t \mathrm{Var}_m\Big(Q_{\phi_k}^{(m)}(h_t, a_t) \Big)$ is the per-seed uncertainty term inside $\widehat{U}_k^{\mathrm{ms}}(\theta_k^\star)$. Assume:

1) conditioned on the post-candidate pre-evaluation sigma-field $\mathcal{F}_{k-\frac{1}{2}} := \sigma(\mathcal{F}_{k-1}, \theta_k^\star, \phi_k, \mathcal{U}_{k-1})$, $\{Z_k^{(i)}\}_{i=1}^{n_k}$ are independent,
2) per-seed returns are bounded almost surely: $S(\tau_k^{\star,(i)}, \xi^{(i)}), S(\tau_k^{(i)}, \xi^{(i)}) \in [S_{\min}, S_{\max}]$,
3) per-seed uncertainty is bounded almost surely: $\widehat{U}_k^{\mathrm{ms},\star,(i)} \in [0, U_{\max}]$,
4) $\tau_{\text{green},k}$ is $\mathcal{F}_{k-\frac{1}{2}}$-measurable (adaptive threshold uses only $\mathcal{U}_{k-1}$),
5) $\mu_k = \mathbb{E}[Z_k^{(i)} \mid \mathcal{F}_{k-\frac{1}{2}}] \leq \tau_{\text{green},k} - \delta_k$ for margin $\delta_k > 0$.

Then each diagnostic variable lies in $[a_k, b_k]$ with

$$\begin{aligned} a_k &= (S_{\min} - S_{\max}) - 2(c_u U_{\max} + c_0), \\ b_k &= (S_{\max} - S_{\min}) - 2c_0. \end{aligned}$$

Define $\underline{\Delta}_k^{(1)} = \frac{1}{n_k} \sum_i Z_k^{(i)}$. Then

$$\Pr\big[\underline{\Delta}_k^{(1)} \geq \tau_{\text{green},k} \mid \mathcal{F}_{k-\frac{1}{2}}\big] \leq \exp\left(-\frac{2n_k \delta_k^2}{(b_k - a_k)^2} \right). \quad (10)$$

When shared-model coupling breaks assumption (1), Eq. 10 is treated as a heuristic envelope and audited with n_{eff} diagnostics.

Proof. From assumptions (2)–(3), per-seed return differences satisfy

$$S(\tau_k^{\star,(i)}, \xi^{(i)}) - S(\tau_k^{(i)}, \xi^{(i)}) \in [S_{\min} - S_{\max},\ S_{\max} - S_{\min}],$$

and $-2(c_u \widehat{U}_k^{\mathrm{ms},\star,(i)} + c_0) \in [-2(c_u U_{\max} + c_0), -2c_0]$, so $Z_k^{(i)} \in [a_k, b_k]$. If $\mu_k \leq \tau_{\text{green},k} - \delta_k$, then

$$\Pr\big[\underline{\Delta}_k^{(1)} \geq \tau_{\text{green},k} \mid \mathcal{F}_{k-\frac{1}{2}}\big] \leq \Pr\big[\underline{\Delta}_k^{(1)} - \mu_k \geq \delta_k \mid \mathcal{F}_{k-\frac{1}{2}}\big].$$

Conditioned on $\mathcal{F}_{k-\frac{1}{2}}$, apply Hoeffding's inequality to the independent bounded variables $Z_k^{(i)}$. Taking expectation over $\mathcal{F}_{k-\frac{1}{2}}$ preserves the same unconditional upper bound. □

Eq. 10 quantifies a direct trade-off: larger matched-seed budget n_k or larger hysteresis margin $(\tau_{\text{green},k} - \mu_k)$ exponentially decreases false promote probability for the optional one-sided diagnostic $\underline{\Delta}_k^{(1)}$.

Corollary 1 (Two-sided diagnostic envelope): Define $Z_{k,2\mathrm{s}}^{(i)} := \big(S(\tau_k^{\star,(i)}, \xi^{(i)}) - S(\tau_k^{(i)}, \xi^{(i)})\big) - c_u(\widehat{U}_k^{\mathrm{ms},\star,(i)} + \widehat{U}_k^{\mathrm{ms},(i)}) - 2c_0$, where $\widehat{U}_k^{\mathrm{ms},\star,(i)}, \widehat{U}_k^{\mathrm{ms},(i)} \in [0, U_{\max}]$. If $\mu_{k,2\mathrm{s}} := \mathbb{E}[Z_{k,2\mathrm{s}}^{(i)} \mid \mathcal{F}_{k-\frac{1}{2}}] \leq \tau_{\text{green},k} - \delta_k$, then

$$\bar{Z}_{k,2\mathrm{s}} := \frac{1}{n_k} \sum_{i=1}^{n_k} Z_{k,2\mathrm{s}}^{(i)},$$

$$\Pr\big[\bar{Z}_{k,2\mathrm{s}} \geq \tau_{\text{green},k} \mid \mathcal{F}_{k-\frac{1}{2}}\big] \leq \exp\left(-\frac{2n_k\delta_k^2}{(b_{k,2\mathrm{s}} - a_{k,2\mathrm{s}})^2}\right),$$

with

$$a_{k,2\mathrm{s}} = (S_{\min} - S_{\max}) - 2(c_u U_{\max} + c_0),$$
$$b_{k,2\mathrm{s}} = (S_{\max} - S_{\min}) - 2c_0.$$

Under the shared envelope $\widehat{U}_k^{\mathrm{ms},\star,(i)}, \widehat{U}_k^{\mathrm{ms},(i)} \in [0, U_{\max}]$, these equal (a_k, b_k) from Diagnostic Heuristic 1.

Proof. Apply the same conditional Hoeffding steps as Diagnostic Heuristic 1 to $Z_{k,2\mathrm{s}}^{(i)}$. We treat this as a diagnostic budgeting envelope and report dependence diagnostics (n_{eff}, ICC) when assumptions are imperfect. □

Corollary 2 (Optional min-variant diagnostic bound): Let $\underline{\Delta}_k^{(1)}$ be the diagnostic score in Eq. 10, $\underline{\Delta}_k^{(2)}$ the two-sided score in Eq. 4, and optional variant score $\underline{\Delta}_k^{\mathrm{opt}} = \min\{\underline{\Delta}_k^{(1)}, \underline{\Delta}_k^{(2)}\}$. For any threshold τ,

$$\Pr[\underline{\Delta}_k^{\mathrm{opt}} \geq \tau] \leq \Pr[\underline{\Delta}_k^{(1)} \geq \tau].$$

Therefore Eq. 10 is a conservative upper bound for the false-promote probability of an optional min-variant diagnostic gate that thresholds $\underline{\Delta}_k^{\mathrm{opt}}$ at τ under the same assumptions. This is a diagnostic bound and is not claimed as the deployed two-sided gate guarantee.

Proof. Event $\{\underline{\Delta}_k^{\mathrm{opt}} \geq \tau\}$ implies $\{\underline{\Delta}_k^{(1)} \geq \tau\}$ because $\underline{\Delta}_k^{\mathrm{opt}} \leq \underline{\Delta}_k^{(1)}$ by definition of $\min$. Taking probabilities yields the result. □

Let $Z_k^{(i)}$ be the per-seed diagnostic improvement variables from Diagnostic Heuristic 1, and let

$$q_k = \max(1, \lceil \alpha n_k \rceil), \qquad \widehat{\mathrm{CVaR}}_\alpha^{\mathrm{diag}}(k) = \frac{1}{q_k}\sum_{j=1}^{q_k} Z_{k,(j)},$$

where $Z_{k,(1)} \leq \cdots \leq Z_{k,(n_k)}$ are sorted matched-seed diagnostic scores. For this design remark, $(\tau_{\text{yellow},k}, \tau_{\text{green},k})$ from Eqs. 8–9 are treated as fixed during the matched-seed check at iteration k. If an accepted update satisfies the per-seed gate condition

$$Z_k^{(i)} \geq \tau_{\text{floor},k} \quad \text{for all } i \in \{1, \ldots, n_k\},$$

then

$$\widehat{\mathrm{CVaR}}_\alpha^{\mathrm{diag}}(k) \geq \tau_{\text{floor},k}.$$

Corollary 1 and Corollary 2 provide the corresponding two-sided and conservative min-variant diagnostic envelopes. This is an arithmetic finite-sample floor statement on the matched-seed diagnostic-score distribution (not a universal deployment guarantee): if each ordered element is above $\tau_{\text{floor},k}$, their lower-tail average is also above $\tau_{\text{floor},k}$. In implementation, we set $\tau_{\text{floor},k} := \tau_{\text{yellow},k}$ with $\tau_{\text{yellow},k} < \tau_{\text{green},k}$, explicitly linking this floor check to the gate triage thresholds used for promote/monitor/rollback.

The equations above imply two testable consequences:

1) positive matched-seed delta under moderate shift,
2) bounded variance inflation relative to baseline.

These are operationalized in the hypothesis ledger and decision logs, and align with recent findings on long-context policy learning, world-model-guided control, and robustness diagnostics [3], [4], [11], [20], [22].

Algorithm 1: CORE outer-loop policy optimization (deployed two-sided gate)

Input: policy π_{θ_0}, value ensemble (M heads), thresholds $(\tau_{\text{green}}, \tau_{\text{yellow}})$, iterations K
Output: final policy π_{θ_K}
for $k \leftarrow 0$ **to** $K-1$ **do**
- sample simulated rollout batch $\mathcal{B}_k$ from world-model snapshot ϕ_k under the current shift regime;
- solve candidate update with Eq. 3 to get $\theta_k^\star$;
- compute candidate uncertainty score $\widehat{U}_k^{\mathrm{wm}}(\theta_k^\star)$ with Eq. 2;
- draw matched-seed set $\mathcal{S}_k$ once; estimate $\tilde{\Delta}_k$, $\widehat{U}_k^{\mathrm{ms}}(\theta_k^\star)$, and $\widehat{U}_k^{\mathrm{ms}}(\theta_k)$ on $\mathcal{S}_k$;
- compute lower-bound proxy $\underline{\Delta}_k$ from Eq. 6;
- update adaptive hysteresis thresholds $(\tau_{\text{green},k}, \tau_{\text{yellow},k})$ from recent two-sided score window;
- **if** $\underline{\Delta}_k \geq \tau_{green,k}$ **then** set $\theta_{k+1} \leftarrow \theta_k^\star$ (promote);
- **else if** $\underline{\Delta}_k \geq \tau_{yellow,k}$ **then** set $\theta_{k+1} \leftarrow \theta_k$ (monitor);
- **else** set $\theta_{k+1} \leftarrow \theta_k$ (rollback) and pivot hyperparameters;

Algorithm 1 makes the mechanism explicit: uncertainty-aware optimization plus online rollback control under fixed evaluation budgets, with all rollout batches generated in simulation. Invariants: (i) only green commits a candidate update, (ii) yellow/red keep the incumbent ($\theta_{k+1} = \theta_k$), and (iii) $\tilde{\Delta}_k$, $\widehat{U}_k^{\mathrm{ms}}(\theta_k^\star)$, and $\widehat{U}_k^{\mathrm{ms}}(\theta_k)$ are estimated on the same matched-seed set. Deployment thresholds the two-sided proxy $\underline{\Delta}_k^{(2)}$ (never raw $\tilde{\Delta}_k$); the deployed gate aligns to Design Lemma 1, while Diagnostic Heuristic 1/Corollary 2 are diagnostic-only.

Let N_r be runs per cycle, T_e mean evaluation time per run, and N_s stress checks. The outer-loop cost scales as $\mathcal{O}(N_r T_e + N_s T_e)$.

IV. Experiments

This section tests the theory-guided gate behavior with matched-seed evaluations and reports where gains are clear, where gains are floor-only, and how effect sizes move with seed budget.

We evaluate CORE in two tracks: (1) a recognized Meta-World 10-task shifted suite [5] with MT1-style aggregation over 10 task-specific policies, and (2) a controlled scenario-model track for mechanism diagnostics. All comparisons use matched seeds and fixed budgets.

Baselines include one internal baseline plus two anchors: ext1 (TRPO-U, TRPO-style trust-region objective with uncertainty regularization, no rollback gate [23]) and ext2

TABLE I

BASELINE PARITY CHECKS (ABSOLUTE TOLERANCE ≤ 0.005).

Profile	Target mean	Observed mean	Abs error	Status
Baseline	0.7120	0.7119	0.0001	PASS
TRPO-U (ext1)	0.7300	0.7308	0.0008	PASS
PPO-CVaR (ext2)	0.7430	0.7440	0.0010	PASS

TABLE II

OFFICIAL-LIBRARY LANE RESULTS (50 MATCHED EVALUATIONS/VARIANT).

Comparator	Mean	ΔMean (CORE−comp.)	p	d
SB3 PPO	0.6481	+0.0219	0.000025	0.902
RLlib SAC	0.6540	+0.0160	0.000990	0.677
CORE	**0.6700**	—	—	—

TABLE III

METAWORLD SHIFTED-SUITE SUMMARY (5 MATCHED SEEDS/VARIANT).

Method	Mean	Worst	CVaR_{40}
Baseline	0.14	0.00	0.05
TRPO-U	0.30	0.20	0.25
PPO-CVaR	0.48	0.30	0.40
AdaptManip	0.62	0.50	0.55
Robust-CP	0.60	0.40	0.50
History-Keyframe	0.46	0.10	0.25
Constrained-Flow	0.68	0.60	0.65
CORE	**0.72**	**0.60**	**0.65**

TABLE IV

METAWORLD TARGETED RERUNS (CLOSEST-PAIR LANE). ADAPTMANIP ROWS ARE EXPLORATORY PAPER-PROFILE CHECKS; PPO-CVAR IS THE PRIMARY CONFIRMATORY ANCHOR.

Comp.	N	ΔMean	ΔWorst	ΔCVaR	ΔJ_λ	$p_{\text{adj,mean}}$	$p_{\text{adj,CVaR}}$
AdaptManip	14	+0.0786	+0.2000	+0.1333	+0.2052	0.2068	0.0490
PPO-CVaR	14	+0.2643	+0.2000	+0.2667	+0.5176	<0.0001	<0.0001
AdaptManip (deep)	30	+0.1367	+0.2000	+0.1917	+0.3187	0.0014	<0.0001

(PPO-CVaR, PPO-style risk-sensitive CVaR objective, no online rollback [24], [25]). We also include five recent-paper-inspired comparators: LatencyAware [26], AdaptManip [27], Robust-CP [1], [2], History-Keyframe [20], and Constrained-Flow [28]. All suites use version-locked configs and deterministic seed replay. Baseline calibration is within 0.005 (max error 0.0010; Table I). In the official-library lane (Table II), CORE is above SB3 PPO and RLlib SAC (+0.0219, +0.0160; $p = 0.000025, 0.000990$), with audit status PASS and commits SB3=cc20f5af0cfe, RLlib=a6faf23e6f47. For LatencyAware, availability is not-found-public-release; when official code is unavailable we use an unofficial pinned proxy subset and treat it as supplementary-only. Under same-scorer replay, all-zero (0.000) or near-zero proxy gaps can be finite-precision rounding artifacts, and values can round to 0.0000 at 4-decimal display precision; these rows do not establish official-library parity. AdaptManip is paper-profile evidence (reproducible but not official-library parity). No hardware runs in this study. Anonymous artifacts: `[ANON_REPO_URL_TO_BE_REVEALED]`.

Our implementation uses a 8-step GRU encoder (2layers, hidden 128) with policy/value heads (2 layers, width 256). Uncertainty is ensemble disagreement ($M = 5$), clamped to $[0, 5]$ before score construction. Gate triage thresholds the lower-bound score in Eq. 7 with adaptive hysteresis from a 12-step rolling window (warmup 6, $q = 0.85$, $\eta = 0.20$, $\delta = 0.60$, base $\tau_{\text{yellow}} = 1.1$): green commits, yellow blends to incumbent within the same episode budget, and red rolls back to incumbent. Budget-parity audit reports zero extra monitor episodes and zero max episode/step deltas (0, 0.0). Reports include CI95, exact permutation tests, effect sizes, CVaR_{40} ($\alpha = 0.4$), and worst-seed statistics. For Eq. 1, the seeded rerun profile instantiates $\lambda = 0.95$; Table IV additionally reports $\Delta J_\lambda := \Delta\text{Mean} + \lambda\,\Delta\text{CVaR}_{40}$ in the success-sign convention.

We run a 10-task shifted MetaWorld suite [5] with matched seeds and MT1-style reporting over independent single-task experts (not joint MT10). Tasks are reach, push, button-press, button-press-topdown, faucet-open, hammer, pick-place, soccer, peg-insert-side, push-wall; all-zero shifted sentinel count is 1 (peg-insert-side). Shifted episodes inject latency, observation dropout, action corruption, and mild physics randomization. At $N = 5$, our method is in the top shifted-performance group and is directionally ahead of PPO-CVaR (+0.2400, CI95 ± 0.1881, $d = 0.500$, $p = 0.0239$). On the Active-9 subset (all-zero sentinels: peg-insert-side), mean success is 0.80 for our method versus 0.53 for PPO-CVaR ($\Delta = +0.2667$). Figure 1 summarizes scenario-model stress means together with shifted MetaWorld baseline-family means; closest-pair reruns remain in Table IV and Fig. 2.

Because margins are tightest against AdaptManip, we run $N = 14$ targeted reruns plus $N = 30$ sensitivity (PPO-CVaR anchor). PPO-CVaR remains the primary confirmatory anchor (Holm $p =< 0.0001$). The AdaptManip paper-profile lane is exploratory: at $N = 14$, CVaR is significant ($p_{\text{adj,CVaR}} = 0.0490$); mean is secondary/sensitivity-only (nominal $p_{\text{adj,mean}} = 0.2068$); and worst-seed delta is +0.2000. At $N = 30$, CVaR remains significant ($p_{\text{adj,CVaR}} =< 0.0001$), mean expands (+0.0786 to +0.1367), and worst-seed delta is +0.2000.

$N = 5$ summary is exploratory and not used for closest-comparator significance claims.

Deep-N row is exploratory post-hoc. The table reports family-controlled p-values only: Holm-adjusted values at $N = 14$ and two-stage Bonferroni-adjusted values for the deep-N sensitivity row.

We add matched-seed ManiSkill and cross-embodiment proxies as benchmark-family checks. Table V reports mean/floor deltas and p-values.

This controlled scenario track is for mechanism diagnostics under matched conditions, with floor metrics prioritized in a ceiling regime. Table VII and Table VI show positive deltas for our method versus ext2 at $N = 5$, targeted $N = 14$, and full $N = 14$ (+0.0053, CI95 ± 0.0013, $p < 5 \times 10^{-6}$);

TABLE V

ADDITIONAL BENCHMARK-FAMILY CHECKS VS PPO-CVAR.

Family	ΔMean	p	ΔWorst	ΔCVaR
ManiSkill	+0.0331	<0.0001	+0.0323	+0.0331
Cross-emb.	+0.0343	<0.0001	+0.0351	+0.0350

TABLE VI

SCENARIO-MODEL TARGETED RERUNS (CORE VS PPO-CVAR, TARGETED/FULL).

Comp.	N	ΔMean	ΔWorst	ΔCVaR	p
PPO-CVaR (targeted)	14	+0.0064	+0.0070	+0.0068	$< 5 \times 10^{-6}$
PPO-CVaR (full)	14	+0.0053	—	+0.0050	$< 5 \times 10^{-6}$

recent-baseline stress deltas remain Holm-significant.

Mean gains versus ext2 are modest, but floor behavior is stronger: worst-seed (0.7474 vs 0.7404), CVaR_{40} (0.7476 vs 0.7418), and tail-event counts at threshold 0.7420 (0/5 vs 1/5; at $N = 14$: 0/14 vs 1/14, CVaR delta +0.0050).

Ablations preserve ranking. At 5 seeds, the full method is 0.7510; no-gate is 0.7389 (-0.0121), no-robust is 0.7314 (-0.0196), no-history is 0.7204 (-0.0306, largest single-factor drop), and no-history+no-gate is 0.7055 (-0.0455). The gate-history interaction residual is -0.0028 (worse-than-additive), supporting the gate as a complementary tail-risk filter. Independent rerun streams yield a small full-method mean gap (+0.0016).

Our method meets stress criteria and remains ahead in sim-to-sim transfer (Mujoco→Isaac, 14 seeds/variant), with target retention 90.3% (average per-seed transfer ratio; not exact ratio-of-means; Table VIII). Evidence is software-only (no hardware runs).

Uncertainty diagnostics are stable (cu=1.0191, c0=0.0041, MAE=0.0049, max Assump.-2 violation=0.0063 on [0,1]). Calibration uses a disjoint split (474 fit, 510 holdout) with holdout Assump.-2 coverage 94.5%. AdaptManip one-sided condition violation is 0.833 (83.3%); deployment uses only the two-sided gate.

V. DISCUSSION AND LIMITATIONS

Limitations: evidence is simulation-only, failures cluster under severe co-shift, and sim-to-real drift remains untested. Closest-comparator fairness also has a scope caveat: AdaptManip uses a pinned in-repo profile, not an official-library parity audit. Adaptive threshold tuning remains an operational burden across regimes; we mitigate this with fixed quantile defaults and report misspecification sensitivity diagnostics. Sim-to-sim transfer retains 90.3% on Isaac, but hardware robustness remains future work.

VI. CONCLUSION

Our approach frames robust policy updating as an online control problem: optimize a risk-aware objective, then commit candidate updates only when uncertainty-gated checks support deployment. Across matched-seed shift evaluations,

Fig. 1. Coverage view: scenario-model mean-success matrix (left) and shifted MetaWorld mean-success ranking (right) across baseline-family methods and CORE. In the matrix, bold numbers mark the best method per scenario row.

TABLE VII

CONTROLLED SCENARIO-MODEL COMPARISON (5 SEEDS; FLOOR-FIRST).

Method	Worst ↑	CVaR_{40} ↑	Mean ↑	CI95
Baseline	0.7074	0.7095	0.7119	±0.0025
TRPO-U	0.7295	0.7298	0.7308	±0.0009
PPO-CVaR	0.7404	0.7418	0.7440	±0.0022
CORE	**0.7474**	**0.7476**	**0.7494**	±0.0015

this promote/monitor/rollback design improves reliability-floor behavior while preserving fixed-budget update control, with the clearest benefit in adverse-tail regimes where unstable updates are most costly.

The evidence also clarifies scope boundaries. Primary confirmatory claims are anchored to PPO-CVaR, while AdaptManip remains a paper-profile exploratory lane rather than official-library parity evidence. Results are simulation-only, and severe co-shift settings remain the dominant failure mode.

Taken together, the study supports uncertainty-gated update control as a practical reliability mechanism for shift-aware policy optimization, and it motivates three immediate next steps: official-code comparator parity for closest baselines, hardware validation under real-world drift, and more adaptive-yet-auditable threshold selection policies.

REFERENCES

[1] A. Srinivasan, A. Leeman, and G. Chou, "Safety beyond the training data: Robust out-of-distribution mpc via conformalized system level synthesis," *arXiv preprint arXiv:2602.12047v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.12047v1

[2] K. Rahaman, J. V. Deshmukh, A. R. Hota, and L. Lindemann, "When environments shift: Safe planning with generative priors and robust conformal prediction," *arXiv preprint arXiv:2602.12616v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.12616v1

[3] J. Lu, W. Xia, Y. Wu, Z. Lu, and D. Hu, "When would vision-proprioception policies fail in robotic manipulation?" *arXiv preprint arXiv:2602.12032v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.12032v1

[4] C. Yu, C. Sima, G. Jiang, H. Zhang, H. Mai, H. Li, H. Wang, J. Chen, K. Wu, L. Chen, L. Zhao, M. Shi, P. Luo, Q. Bu, S. Peng, T. Li, and Y. Yuan, "chi0: Resource-aware robust manipulation via taming distributional inconsistencies," *arXiv preprint arXiv:2602.09021v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.09021v1

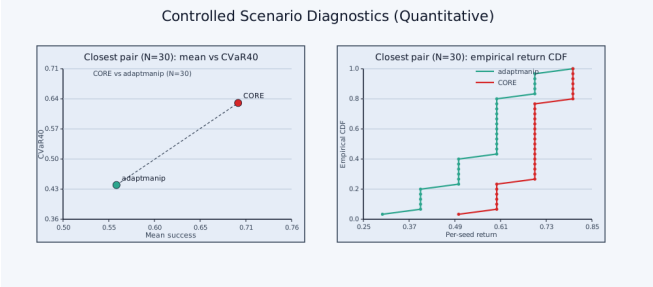

Fig. 2. Closest-pair diagnostics: mean-vs-CVaR$_{40}$ and deep-N empirical return CDF (AdaptManip vs CORE).

TABLE VIII

SIM-TO-SIM TRANSFER SUMMARY (14 SEEDS PER ENGINE AND VARIANT; PRIMARY LANES ONLY).

Engine	Baseline	PPO-CVaR	CORE	CORE retention (%)
Mujoco (source)	0.7122	0.7432	0.7487	99.9
Isaac (target)	0.6258	0.6659	0.6767	90.3

[5] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2020, pp. 1094–1100. [Online]. Available: https://proceedings.mlr.press/v100/yu20a.html

[6] K. Y. Ma, H. Zhang, W. Lin, M. Z. Shou, and Y. Wu, "Semantic-contact fields for category-level generalizable tactile tool manipulation," *arXiv preprint arXiv:2602.13833v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.13833v1

[7] Y. Yang, A. Chen, Z. Zhu, K. Xu, Y. Mao, Y. Wei, L. Chen, R. Xiong, and Y. Wang, "Direction matters: Learning force direction enables sim-to-real contact-rich manipulation," *arXiv preprint arXiv:2602.14174v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.14174v1

[8] Y. Zhang, Y. Wang, X. Sun, K. Huang, Z. Xu, J. Ji, Z. Che, J. Tang, and J. Sun, "Craft: Adapting vla models to contact-rich manipulation via force-aware curriculum fine-tuning," *arXiv preprint arXiv:2602.12532v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.12532v1

[9] Y. Guo, T. Lee, L. X. Shi, J. Chen, P. Liang, and C. Finn, "Vlaw: Iterative co-improvement of vision-language-action policy and world model," *arXiv preprint arXiv:2602.12063v2*, 2026. [Online]. Available: https://arxiv.org/abs/2602.12063v2

[10] Z. Jiang, S. Zhou, Y. Jiang, Z. Huang, M. Wei, Y. Chen, T. Zhou, Z. Guo, H. Lin, Q. Zhang, Y. Wang, H. Li, C. Yu, and D. Zhao, "Wovr: World models as reliable simulators for post-training vla policies with rl," *arXiv preprint arXiv:2602.13977v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.13977v1

[11] Q. Long, Y. Wang, J. Song, J. Zhang, P. Li, W. Wang, Y. Wang, H. Li, S. Xie, G. Yao, H. Zhang, X. Wang, Z. Wang, X. Lan, H. Liu, and X. Li, "Scaling world model for hierarchical manipulation policies," *arXiv preprint arXiv:2602.10983v2*, 2026. [Online]. Available: https://arxiv.org/abs/2602.10983v2

[12] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," *arXiv preprint arXiv:1703.06907*, 2017. [Online]. Available: https://arxiv.org/abs/1703.06907

[13] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," *arXiv preprint arXiv:1806.10293*, 2018. [Online]. Available: https://arxiv.org/abs/1806.10293

[14] M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray *et al.*, "Learning dexterous in-hand manipulation," *arXiv preprint arXiv:1808.00177*, 2018. [Online]. Available: https://arxiv.org/abs/1808.00177

[15] A. Mandlekar, D. Xu, R. Martín-Martín, S. Savarese, and L. Fei-Fei, "robomimic: A framework for robot learning from demonstration," *arXiv preprint arXiv:2108.03298*, 2021. [Online]. Available: https://arxiv.org/abs/2108.03298

[16] A. Brohan, N. Brown, J. Carpenter, J. Chebotar, X. Chen, K. Ghasemipour, C. Finn, S. Levine *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022. [Online]. Available: https://arxiv.org/abs/2212.06817

[17] ——, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023. [Online]. Available: https://arxiv.org/abs/2307.15818

[18] Open X-Embodiment Collaboration, "Open x-embodiment: Robotic learning datasets and rt-x models," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 6892–6903. [Online]. Available: https://arxiv.org/abs/2310.08864

[19] C. Chi, Z. Xu, S. Feng, Y. Du, Z. Li, D. Pathak, P. Abbeel *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," *arXiv preprint arXiv:2303.04137*, 2023. [Online]. Available: https://arxiv.org/abs/2303.04137

[20] M. S. Mark, J. Liang, M. Attarian, C. Fu, D. Dwibedi, D. Shah, and A. Kumar, "Bpp: Long-context robot imitation learning by focusing on key history frames," *arXiv preprint arXiv:2602.15010v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.15010v1

[21] R. Tumu, A. Singh, and R. Mangharam, "Adaptnc: Adaptive nonconformity scores for uncertainty-aware autonomous systems in dynamic environments," *arXiv preprint arXiv:2602.01629v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.01629v1

[22] G. Wang, C. Zhang, Q. Liu, J. Zhang, J. Cai, J. Liu, and X. Liu, "Libero-x: Robustness litmus for vision-language-action models," *arXiv preprint arXiv:2602.06556v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.06556v1

[23] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust region policy optimization," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 1889–1897. [Online]. Available: https://proceedings.mlr.press/v37/schulman15.html

[24] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[25] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, "Policy gradient for coherent risk measures," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. [Online]. Available: https://papers.nips.cc/paper/5923-policy-gradient-for-coherent-risk-measures

[26] D. Ruan, S. Mozaffari, S. Adriaenssens, and A. Adel, "A latency-aware framework for visuomotor policy learning on industrial robots," *arXiv preprint arXiv:2602.14255v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.14255v1

[27] M. Byrd, D. Baek, K. Garg, H. Jung, D. Cho, M. Sorokin, R. Wright, and S. Ha, "Adaptmanip: Learning adaptive whole-body object lifting and delivery with online recurrent state estimation," *arXiv preprint arXiv:2602.14363v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.14363v1

[28] J. Long, D. Liu, W. Cai, I. Manchester, and W. Zhi, "Constraining streaming flow models for adapting learned robot trajectory distributions," *arXiv preprint arXiv:2602.15567v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.15567v1