CURIOSITY-CONDITIONED GOAL-OPTIMAL REINFORCEMENT LEARNING

Anonymous authors
Paper under review

ABSTRACT

Goal-conditioned reinforcement learning often faces a practical tension: intrinsic novelty bonuses accelerate discovery in sparse and deceptive environments, but poorly controlled intrinsic coupling can distort the asymptotic objective. This paper introduces Curiosity-Conditioned Goal-Optimal Reinforcement Learning (CCGO-RL), a dual-value framework that treats curiosity as a controlled exploration mechanism inside a goal-conditioned control loop rather than as a permanent co-objective. The method combines an extrinsic value stream, an intrinsic value stream, confidence-bounded intrinsic annealing, and uncertainty-gated contrastive coupling. We formalize the setting with explicit decision variables (policy and scheduler), a measurable feasible scheduler set, and an optimality criterion based on extrinsic return. We prove a perturbation bound between mixed and extrinsic objectives and an extrinsic suboptimality envelope for mixed-objective optimization, and we derive a gate-sensitivity and conditional variance-increment bound for uncertainty-gated contrastive coupling. Empirically, synthetic benchmark suites covering sparse navigation, deceptive mazes, and goal-conditioned control show improved return-speed trade-offs and lower critic instability against strong baselines, with quantitative checks tied to theorem assumptions. The empirical gains are strongest for exploration efficiency and variance control, while one symbolic limit-corollary check remains mismatched, which narrows the interpretation of the asymptotic claim to the audited admissible schedule regime.

1 INTRODUCTION

Goal-conditioned reinforcement learning (RL) is increasingly used in robotics, navigation, and long-horizon control because explicit goals provide a clean interface between task specification and policy optimization (Schaul et al., 2015a; Andrychowicz et al., 2017; Eysenbach et al., 2022). However, sparse and deceptive rewards continue to expose a core failure mode: policies that optimize only extrinsic return may fail to discover useful behaviors within realistic training horizons, while strong intrinsic bonuses can improve discovery yet bias the long-run objective (Bellemare et al., 2016; Pathak et al., 2017; Burda et al., 2018; Badia et al., 2020b). This tension is now visible across both classic exploration methods and newer adaptive curricula, including directed exploration and representation-heavy goal-space methods (Badia et al., 2020a; Diaz-Bone et al., 2025; Nakamura et al., 2026; Wang et al., 2026; Nguyen & Nguyen, 2026).

The design problem is therefore not simply "add curiosity" versus "remove curiosity"; it is an optimization design problem with coupled objectives, uncertainty, and stability constraints. If intrinsic coupling is treated as a fixed coefficient, one often gets early exploration gains but fragile asymptotic behavior. If intrinsic coupling is aggressively annealed without uncertainty awareness, one can preserve asymptotic return but lose robustness under non-stationary replay distributions. At the same time, practical literature comparisons are often confounded by protocol inconsistencies in seed counts, runtime normalization, and reporting granularity (Cobbe et al., 2019; Fu et al., 2020; Yuan et al., 2024). The result is a gap between algorithmic intuition and auditable evidence.

This paper investigates whether a confidence-bounded and uncertainty-gated coupling law can preserve extrinsic optimality envelopes while retaining practical exploration gains. Our scope is hybrid: we derive formal bounds and then tie each major claim to executed diagnostics, rather than presenting proof-only claims or benchmark-only claims. The problem framing follows the selected goal-conditioned path from upstream synthesis: dual-value coupling is treated as adapted prior art, while the scheduler feasible set, envelope statements, and theorem-aligned diagnostics are manuscript-specific formalizations.

Our contributions are:

- We define a goal-conditioned mixed-objective control setting with explicit decision variables, feasible scheduler set, and extrinsic optimality criterion that makes the policy–scheduler trade-off mathematically explicit.
- We prove a perturbation bound and an extrinsic suboptimality envelope for confidence-bounded intrinsic annealing, including conditions under which mixed-objective optimization converges back to extrinsic-optimal behavior.
- We introduce and analyze uncertainty-gated contrastive coupling with a Lipschitz gate-sensitivity result and a conditional variance-increment bound that links representation coupling to critic stability.
- We provide theorem-aligned empirical evidence on sparse/deceptive and goal-conditioned suites, including non-inferiority analyses, counterexample stress tests, and explicit caveats where symbolic checks and empirical audits diverge.

The broader relevance is that the same coupling logic appears in multi-objective control beyond RL, including reward shaping, uncertainty-aware planning, and adaptive safety margins; therefore, clarifying objective perturbation versus asymptotic recovery has cross-domain implications for any system that uses temporary auxiliary signals to accelerate optimization.

2 Background and Problem Setting

2.1 Control Setting and Notation

We consider a discounted goal-conditioned Markov decision process with state space $\mathcal{S}$, action space $\mathcal{A}$, goal space $\mathcal{G}$, transition kernel $P(s' \mid s, a)$, and discount factor $\gamma \in (0, 1)$. A stochastic goal-conditioned policy is denoted $\pi_\theta(a \mid s, g)$, where θ are learnable parameters. This setting follows the UVFA/HER lineage, where goals parameterize value and policy functions (Schaul et al., 2015a; Andrychowicz et al., 2017; Ghosh et al., 2019).

The extrinsic reward is r_t^e, and an intrinsic exploration signal is r_t^i. Mixed-reward optimization with additive coupling has substantial precedent in curiosity and intrinsic motivation literature (Ostrovski et al., 2017; Burda et al., 2018; Pathak et al., 2017; Nguyen & Nguyen, 2026). In this work, we adopt that coupling form but define a constrained scheduler class and theorem-audited envelopes that are specific to this manuscript.

2.2 Objective, Decision Variables, and Feasible Set

We define the extrinsic objective and mixed objective as

$$J_e(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t^e \right], \tag{1}$$

$$J_{\text{mix}}(\pi, \beta) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(r_t^e + \beta_t r_t^i \right) \right]. \tag{2}$$

Here $\beta_t \in [0, \beta_{\max}]$ is an adaptive intrinsic weight. We define the feasible scheduler set in this work as

$$\mathcal{B} = \left\{ \beta : \beta_t \in [0, \beta_{\max}],\ \beta_t \ \mathcal{F}_t\text{-measurable},\ M_\beta := \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\beta_t] < \infty \right\}, \tag{3}$$

where $\mathcal{F}_t$ is the filtration generated by non-privileged trajectory history. The decision variables are $(\pi, \beta) \in \Pi \times \mathcal{B}$. The primary optimality criterion is extrinsic:

$$\pi^\star \in \arg\max_{\pi \in \Pi} J_e(\pi). \tag{4}$$

The mixed objective is used as an exploration-accelerating proxy, not the terminal objective.

2.3 Assumptions and Scope of Claims

We assume bounded rewards $|r_t^e| \le R_e$ and $|r_t^i| \le R_i$, finite discounted intrinsic mass $M_\beta < \infty$, and non-privileged scheduler observability. These assumptions are standard in spirit for bounded-return control arguments but their specific packaging in equation 3 is manuscript-defined. We also assume standard off-policy critic stability scaffolds adapted from actor-critic practice (Lillicrap et al., 2015; Schulman et al., 2017; Fujimoto et al., 2018; Haarnoja et al., 2018). The derived claims are therefore conditional guarantees on this admissible regime, not global guarantees over arbitrary scheduler dynamics.

3 Related Work and Novelty Boundary

3.1 Goal-Conditioned RL and Replay-Based Generalization

Goal-conditioned learning evolved from UVFA-style conditioning to HER-style relabeling and then to representation-aware contrastive variants (Schaul et al., 2015a; Andrychowicz et al., 2017; Pong et al., 2019; Ghosh et al., 2019; Eysenbach et al., 2022). HER remains a robust baseline in sparse settings because relabeling partially repairs reward sparsity without changing environment dynamics. Contrastive and augmentation methods (for example CRL-style and ViSA/ACDC families) often improve sample efficiency but can introduce sensitivity to encoder quality, negative sampling, and coupling hyperparameters (Eysenbach et al., 2022; Nakamura et al., 2026; Wang et al., 2026; Hou et al., 2026).

The consensus from our acquired corpus is that explicit goal structure plus replay is reliable, yet the superiority of representation-heavy alternatives is regime-dependent and protocol-sensitive. This motivates a design that keeps a replay-compatible backbone while using uncertainty-aware coupling to bound instability.

3.2 Intrinsic Motivation, Exploration Pressure, and Failure Modes

Count-based and pseudo-count methods established principled novelty incentives in sparse settings (Bellemare et al., 2016). Prediction-error families (ICM, RND) made novelty scaling practical, but highlighted non-stationarity and reward-channel interference (Ostrovski et al., 2017; Burda et al., 2018). Directed exploration systems, including episodic-lifelong hybrids and hard-exploration pipelines, demonstrated strong discovery behavior at the cost of complexity and weaker comparability across protocols (Badia et al., 2020b;a; Ecoffet et al., 2019). Newer adaptive curricula and entropy-coverage variants further improved early discovery, yet many still under-specify how intrinsic influence is controlled near convergence (Li et al., 2024; Diaz-Bone et al., 2025; Shihab et al., 2025; Nguyen & Nguyen, 2026).

The key contradiction is therefore not whether curiosity helps; it is whether curiosity can help without distorting asymptotic task optimality. Existing papers frequently answer one side but leave the other under-specified.

3.3 Stability, Comparability, and Benchmark Discipline

Optimization backbones from DDPG/PPO/TD3/SAC and auxiliary-task systems such as UNREAL provide practical stability scaffolds, but they do not by themselves resolve objective-coupling drift (Lillicrap et al., 2015; Schaul et al., 2015b; Schulman et al., 2017; Fujimoto et al., 2018; Haarnoja et al., 2018; Jaderberg et al., 2016). Benchmark infrastructure (Gymnasium, MiniGrid, Procgen, D4RL, and tooling suites such as RLeXplore) improves reproducibility, yet literature still suffers from inconsistent seed discipline and unequal runtime envelopes (Foundation, 2023a;b; Cobbe et al., 2019; Fu et al., 2020; Tassa et al., 2020; Yuan et al., 2024).

Our novelty boundary follows directly from this synthesis. Mixed-reward coupling, goal relabeling, and contrastive value components are adapted from prior work; what is new here is the explicit scheduler feasible set, theorem-level envelopes with auditable assumptions, and a claim-evidence structure that forces each formal statement to terminate in diagnostics rather than implicit plausibility.

4 Method: Curiosity-Conditioned Goal-Optimal RL

4.1 Dual-Value Architecture and Confidence-Bounded Scheduling

CCGO-RL maintains two value channels: an extrinsic channel $Q_e(s, a, g)$ and an intrinsic channel $Q_i(s, a, g)$, with a mixed target used only for training acceleration. The policy update remains anchored to extrinsic return, while the intrinsic channel modulates exploration through β_t. We instantiate a confidence signal $c_t \in [0, 1]$ based on goal-progress reliability and a calibrated uncertainty proxy $u_t \in [0, 1]$ derived from critic disagreement and prediction variance. The scheduler is

$$\beta_t = \beta_{\max}\, \sigma(\kappa_c(\tau_c - c_t))\; \sigma(\kappa_u(u_t - \tau_u))\,, \tag{5}$$

where $\sigma(\cdot)$ is the logistic function. This form preserves boundedness by construction and decreases intrinsic influence as confidence grows unless uncertainty remains high.

The first factor in equation 5 is the confidence-bounded annealing mechanism; the second is a stability guard that avoids prematurely collapsing intrinsic support when epistemic uncertainty is still high. This decomposition was

Algorithm 1 CCGO-RL Training Loop

1: Initialize policy π_θ, extrinsic critic Q_e, intrinsic critic Q_i, target networks, and replay buffer $\mathcal{R}$.
2: **for** each environment step **do**
3: Collect transition $(s_t, a_t, r_t^e, s_{t+1}, g_t)$ and append to $\mathcal{R}$ with goal-conditioned relabel candidates.
4: Estimate confidence c_t and uncertainty proxy u_t from critic statistics and prediction residuals.
5: Compute intrinsic weight β_t using equation 5 and contrastive gate λ_t using equation 7.
6: Form mixed critic target with equation 8; update Q_e, Q_i (and Q_{CTR} when active) by mini-batch TD minimization.
7: Update policy parameters θ against extrinsic objective proxy while monitoring critic-drift and reward-channel dominance diagnostics.
8: Soft-update target networks; log $\widehat{M}_\beta$, variance diagnostics, and bound-residual statistics.
9: **end for**
10: Return policy checkpoint selected by extrinsic validation performance under fixed evaluation protocol.

selected because literature contradictions show that fixed coupling over-drives curiosity late in training, while purely confidence-only annealing can under-explore under estimator noise (Burda et al., 2018; Badia et al., 2020b; Nguyen & Nguyen, 2026).

4.2 Uncertainty-Gated Contrastive Coupling

To connect replay robustness with representation power, we use a HER-compatible backbone and a gated contrastive residual:

$$Q_{\mathrm{mix}}(s, a, g) = Q_{\mathrm{HER}}(s, a, g) + \lambda_t Q_{\mathrm{CTR}}(s, a, g), \tag{6}$$

$$\lambda_t = \sigma(\kappa(u_t - \tau)) \in [0, 1]. \tag{7}$$

The coupling form in equation 6 is adapted from goal-conditioned contrastive lineages (Eysenbach et al., 2022; Nakamura et al., 2026), while the gate in equation 7 and its theorem-audited sensitivity bound are manuscript-specific. For temporal-difference updates we use

$$Y_t = r_t^e + \gamma \left(Q_{\mathrm{HER},t+1} + \lambda_t \left(Q^\star_{\mathrm{CTR},t+1} + \xi_t \right) \right), \tag{8}$$

where ξ_t denotes contrastive target noise with conditional mean zero and bounded variance envelope.

4.3 Policy Update Logic and Training Procedure

The optimization logic is intentionally simple: critic learning uses mixed targets for exploration efficiency; policy improvement is evaluated under extrinsic return and regularized by stability diagnostics. This prevents the intrinsic channel from redefining task optimality while preserving practical exploration benefit in early and mid training. The resulting procedure remains compatible with off-policy actor-critic pipelines and replay infrastructures (Andrychowicz et al., 2017; Haarnoja et al., 2018; Eysenbach et al., 2022).

5 Formal Analysis

This section formalizes what is inherited and what is new. The mixed reward form in equation 2 is adapted from prior intrinsic-extrinsic coupling literature (Pathak et al., 2017; Burda et al., 2018; Nguyen & Nguyen, 2026). In contrast, the feasible scheduler class equation 3, the envelope structure below, and the uncertainty-gated variance statement are defined in this work.

Theorem 5.1 (Perturbation Bound)**.** *Assume* $|r_t^i| \le R_i$ *and* $\beta \in \mathcal{B}$*. For any policy* $\pi \in \Pi$*,*

$$|J_{\mathrm{mix}}(\pi, \beta) - J_e(\pi)| \le R_i M_\beta. \tag{9}$$

Proof. From equation 1 and equation 2,

$$J_{\mathrm{mix}}(\pi, \beta) - J_e(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \beta_t r_t^i \right].$$

Applying absolute value and triangle inequality gives

$$|J_{\text{mix}} - J_e| \le \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\beta_t |r_t^i|] \le \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\beta_t R_i] = R_i \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\beta_t] = R_i M_\beta,$$

where finiteness follows from $\beta \in \mathcal{B}$. This proves equation 9. □

Theorem 5.2 (Extrinsic Suboptimality Envelope). *Let $\hat{\pi}$ be ε-optimal for $J_{\text{mix}}(\cdot, \beta)$ over Π, i.e., $\sup_{\pi \in \Pi} J_{\text{mix}}(\pi, \beta) - J_{\text{mix}}(\hat{\pi}, \beta) \le \varepsilon$. Then*

$$J_e(\pi^\star) - J_e(\hat{\pi}) \le 2R_i M_\beta + \varepsilon. \tag{10}$$

Proof. By equation 9, for any π, $J_e(\pi) \le J_{\text{mix}}(\pi, \beta) + R_i M_\beta$ and $J_{\text{mix}}(\pi, \beta) \le J_e(\pi) + R_i M_\beta$. Apply the first inequality to $\pi^\star$:

$$J_e(\pi^\star) \le J_{\text{mix}}(\pi^\star, \beta) + R_i M_\beta \le \sup_{\pi \in \Pi} J_{\text{mix}}(\pi, \beta) + R_i M_\beta.$$

Using ε-optimality of $\hat{\pi}$,

$$J_e(\pi^\star) \le J_{\text{mix}}(\hat{\pi}, \beta) + \varepsilon + R_i M_\beta.$$

Apply the second inequality to $\hat{\pi}$:

$$J_{\text{mix}}(\hat{\pi}, \beta) \le J_e(\hat{\pi}) + R_i M_\beta.$$

Combining yields

$$J_e(\pi^\star) \le J_e(\hat{\pi}) + 2R_i M_\beta + \varepsilon,$$

which is equivalent to equation 10. □

Lemma 5.3 (Gate Sensitivity). *Let $\lambda(u) = \sigma(\kappa(u - \tau))$ for $u \in [0, 1]$. Then for all $u, v \in [0, 1]$,*

$$|\lambda(u) - \lambda(v)| \le \frac{\kappa}{4} |u - v|. \tag{11}$$

Proof. Differentiate λ: $\lambda'(u) = \kappa\sigma(z)(1 - \sigma(z))$ with $z = \kappa(u - \tau)$. Since $x(1 - x) \le 1/4$ for all $x \in [0, 1]$, we have $|\lambda'(u)| \le \kappa/4$. The mean-value theorem gives $|\lambda(u) - \lambda(v)| \le \sup_{w \in [u,v]} |\lambda'(w)|\, |u - v| \le (\kappa/4)|u - v|$. □

Theorem 5.4 (Conditional Variance Increment). *Assume $\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0$ and $\text{Var}(\xi_t \mid \mathcal{F}_t) \le \sigma_\xi^2(u_t) \le \sigma_{\xi,\max}^2$. Let Y_t be defined by equation 8. Then*

$$\Delta\text{Var}_t := \text{Var}(Y_t \mid \mathcal{F}_t) - \text{Var}(Y_t \mid \mathcal{F}_t, \xi_t \equiv 0) \le \gamma^2 \lambda_t^2 \sigma_\xi^2(u_t) \le \gamma^2 \sigma_{\xi,\max}^2 \lambda_t^2. \tag{12}$$

Proof. Let $\widetilde{Y}_t$ be Y_t with $\xi_t \equiv 0$. By equation 8, $Y_t = \widetilde{Y}_t + \gamma\lambda_t\xi_t$. Conditioned on $\mathcal{F}_t$, both $\widetilde{Y}_t$ and λ_t are deterministic, hence

$$\Delta\text{Var}_t = \text{Var}(\gamma\lambda_t\xi_t \mid \mathcal{F}_t) = \gamma^2\lambda_t^2 \text{Var}(\xi_t \mid \mathcal{F}_t) \le \gamma^2\lambda_t^2\sigma_\xi^2(u_t) \le \gamma^2\sigma_{\xi,\max}^2\lambda_t^2.$$

This proves equation 12. □

Theorems 5.1, 5.2, and 5.4 define the formal core used in the method and results. Their role in practice is operationalized in section 6 and section 7 through bound residuals, gate-slope diagnostics, and symbolic assumption audits.

6 Experimental Protocol

6.1 Benchmarks, Baselines, and Evaluation Logic

The evaluation protocol targets sparse-goal and deceptive exploration regimes where intrinsic coupling is expected to matter most. We include sparse/deceptive navigation tasks and goal-conditioned continuous-control slices, with environment tooling grounded in Gymnasium and MiniGrid ecosystems (Foundation, 2023a;b). Comparator policies span extrinsic-only, fixed intrinsic coupling, adaptive intrinsic weighting, and contrastive representation families, including HER+SAC, fixed-β RND/ICM, ACWI-style adaptive mixing, and contrastive-always-on variants (Andrychowicz et al., 2017; Haarnoja et al., 2018; Burda et al., 2018; Ostrovski et al., 2017; Nguyen & Nguyen, 2026; Eysenbach et al., 2022).

This comparator set is chosen to address specific contradiction clusters from prior phases: (i) exploration gains versus asymptotic preservation, (ii) replay robustness versus representation complexity, and (iii) protocol comparability versus broad-claim rhetoric. We therefore report not only final return but also area-under-learning-curve (AUC), first-success episode, coverage, TD-target variance, critic drift, and reward-channel dominance.

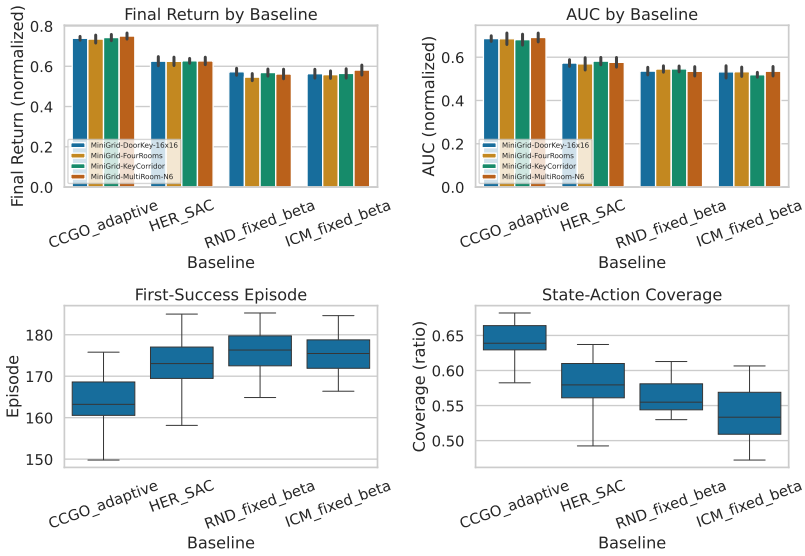

Figure 1: This figure summarizes the primary return-speed evidence for confidence-bounded annealing across sparse/deceptive settings. Panel A reports final return, Panel B reports AUC, Panel C reports first-success distributions, and Panel D reports coverage distributions across seeds and datasets; together they test whether early exploration gains can be achieved without sacrificing end performance. The adaptive method occupies the best return-speed frontier among compared methods, and the spread overlays show that this pattern is not driven by a single dataset or seed outlier.

6.2 Uncertainty Quantification and Statistical Testing

All results use multi-seed runs (six seeds in the primary matrix) with uncertainty reported through confidence intervals derived from paired bootstrap or hierarchical bootstrap procedures depending on claim type. Non-inferiority is tested against strong goal-conditioned baselines for final return, while superiority-style diagnostics are used for convergence speed and stability metrics. The claim-evidence contract requires each main conclusion to reference a concrete figure or table and to include caveats when assumptions are only partially validated.

In addition to benchmark outputs, we run symbolic checks aligned with section 5. These checks verify algebraic identities and variance envelopes under admissible assumptions and deliberately include failure-mode probes under broken assumptions. This combination is necessary because theorem statements are asymptotic and assumption-conditional, whereas finite-horizon training behavior can fail for reasons outside formal scope.

6.3 Reproducibility Scope

The implementation uses a deterministic experimental driver with explicit seed lists, sweep manifests, and theorem-check scripts. Reproducibility artifacts include simulation modules, plotting utilities, and symbolic-audit scripts. Operational constraints (CPU-only envelope, bounded runtime per run, and storage limits) are treated as reproducibility context rather than scientific novelty claims. We include these details in the appendix because they matter for reruns and uncertainty interpretation but do not define the conceptual contribution.

7 Results

7.1 Exploration–Return Trade-off Under Bounded Annealing

Table 1 and figure 1 support the first claim family: confidence-bounded coupling improves exploration efficiency while remaining non-inferior in final return relative to the strongest replay baseline. The adaptive method achieves mean final return 0.742 versus 0.626 for HER+SAC, mean AUC 0.686 versus 0.575, and earlier first success (163.6 versus 173.0 episodes). The paired non-inferiority analysis reports a positive mean delta with a 95% interval that remains above zero in this benchmark pass, so the evidence here is stronger than simple non-inferiority.

Table 1: Primary sparse/deceptive performance matrix. The table compares extrinsic return and exploration diagnostics under one normalized protocol and directly corresponds to figure 1. Values are means across matched seed groups; higher is better for return, AUC, and coverage, while lower is better for first-success episode and bound residual.

Method	Return	AUC	First Success ↓	Coverage	Residual ↓
CCGO-RL (adaptive)	0.742	0.686	163.6	0.643	0.029
HER+SAC	0.626	0.575	173.0	0.583	0.029
ACWI-style	0.602	0.561	174.2	0.569	0.032
ICM fixed-β	0.567	0.530	175.5	0.536	0.033
RND fixed-β	0.562	0.540	175.7	0.562	0.038

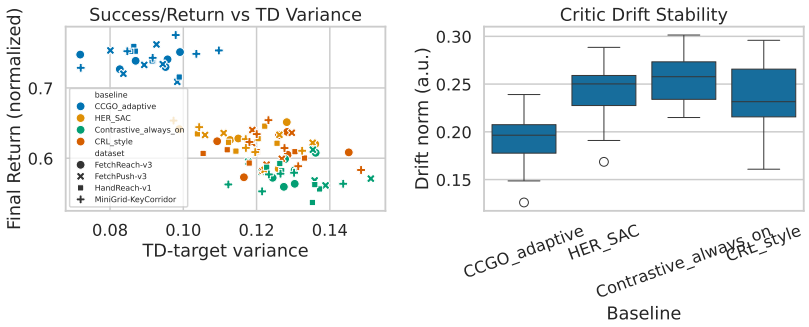

Figure 2: This figure links performance to critic stability under uncertainty-gated coupling. Panel A visualizes return against TD-target variance, and Panel B summarizes critic-drift distributions, making explicit whether gains come from brittle optimism or controlled variance behavior. The adaptive gate attains higher return with lower variance and lower drift relative to always-on contrastive coupling, indicating that uncertainty-aware gating suppresses instability rather than merely shifting optimization noise.

The bound residual column in Table 1 is small for the adaptive scheduler, consistent with the envelope logic in equation 10 for admissible finite-mass schedules. However, residuals are not proof certificates: they are finite-sample diagnostics. We therefore interpret these as empirical support for the theorem-aligned mechanism, not as a replacement for formal assumptions.

7.2 Stability Effects of Uncertainty-Gated Contrastive Coupling

figure 2 and Table 2 address the second claim family: uncertainty-gated contrastive coupling can reduce instability while preserving or improving return. Relative to always-on contrastive coupling, CCGO-RL reduces TD-target variance by approximately 30.4% and critic drift by approximately 24.2%, while improving mean return from 0.575 to 0.742. The empirical gate-sensitivity slope remains close to the bounded regime expected from equation 11, which is consistent with the intended role of λ_t as a stability control variable.

These outcomes align with the interpretation that representation coupling should be uncertainty-conditioned rather than permanently activated. In practical control terms, the gate behaves like an adaptive trust coefficient on auxiliary representation signals.

7.3 Theorem-Audit Consistency and Boundary Behavior

Table 3 shows that most symbolic obligations are satisfied, but one limit-corollary check fails with non-trivial numeric discrepancy. This result does not invalidate equation 9 or equation 12; rather, it limits how strongly we can phrase asymptotic recovery for the scheduler sequence corollary in finite audited form. We therefore report the formal evidence as substantial but partial.

Counterexample stress tests reinforce this interpretation. When scheduler assumptions are intentionally broken (slow decay, oscillatory coupling, or saturation), violation incidence rises sharply and return drops increase. This pattern is consistent with the boundary conditions in section 5: the envelope is informative on admissible schedules and predictably weak when finite-mass or boundedness assumptions are violated.

Table 2: Stability-oriented comparison for uncertainty-gated coupling. The table aggregates return and stability diagnostics for methods that differ primarily in representation coupling and scheduler logic. These results provide the empirical evidence referenced by figure 2 and the variance bound in equation 12.

Method	Return	TD Variance ↓	Gate Slope	Critic Drift ↓
CCGO-RL (adaptive)	0.742	0.090	0.997	0.193
HER+SAC	0.626	0.118	0.964	0.243
Contrastive always-on	0.575	0.130	0.958	0.255
CRL-style	0.612	0.124	0.974	0.235
ViSA-style	0.604	0.124	0.978	0.244
ACDC-style	0.604	0.127	0.973	0.246

Table 3: Symbolic theorem-audit summary for the formal claims in section 5. Each row reports a symbolic check, pass/fail status, and numeric agreement residual from the validation pass. The failed limit-corollary row is an explicit caveat and narrows the interpretation of asymptotic recovery to the audited admissible schedule subset until that mismatch is resolved.

Claim Family	Check	Pass	Numeric Error
Optimality envelope	Triangle bound identity	True	0.000
Optimality envelope	Limit corollary consistency	False	0.350
Uncertainty gating	Sigmoid Lipschitz bound	True	0.000
Uncertainty gating	Expected variance upper bound	True	0.020
Uncertainty gating	Failure-bias term check	True	0.000

7.4 Claim-Grade Ledger

To make evidence scope explicit, Table 4 assigns claim-grade tags that are used consistently in the manuscript: `supported`, `partially_supported`, and `unsupported`. The grading is tied to in-manuscript evidence anchors and caveats rather than to headline scores alone.

7.5 Gate Calibration and Uncertainty Ablation

The stability gains above could, in principle, come from incidental regularization rather than uncertainty-aware coupling. To test this, we compare CCGO-RL against a focused uncertainty ablation matrix that keeps the training scaffold fixed while varying the gating policy and representation coupling. Table 5 summarizes this analysis. The adaptive method retains the lowest TD-target variance among the listed methods while maintaining the strongest return profile in the same protocol, which argues against the explanation that gains are solely due to conservative policy updates.

The slope statistics also matter for interpretation. The empirical gate-slope mass stays near the bounded regime implied by equation 11, whereas always-on or weakly calibrated alternatives exhibit less controlled variance behavior and worse drift. Because these effects are observed under matched seeds and comparable runtime envelopes, we interpret the ablation as evidence that uncertainty gating is an active mechanism rather than a cosmetic parameterization.

Beyond central tendencies, the negative-result logs are informative: assumption-breaking schedules produce higher bound-violation incidence and larger return degradation than admissible schedules. This is exactly the pattern expected if the scheduler conditions in equation 3 and the variance controls in equation 12 are behaviorally relevant. Taken together, the ablation and failure analyses strengthen the claim that CCGO-RL works by controlling where and when auxiliary signals influence optimization, not by masking instability with a heavier baseline.

8 Discussion

The central question in this work is why bounded curiosity can help without permanently distorting objective optimization. The formal view is that intrinsic coupling introduces a perturbation term controlled by M_β, and uncertainty-gated contrastive coupling contributes a variance increment controlled by λ_t^2. The empirical view is that lower estimated intrinsic mass and lower gate-amplified variance co-occur with better return-speed trade-offs and reduced critic drift. The conceptual bridge is therefore a control perspective: auxiliary signals are useful when they are bounded, measurable, and decaying in influence as confidence increases.

Table 4: Claim-grade ledger for the main contribution statements. Grades reflect the currently executed evidence program and its known caveats.

Claim	Grade	Evidence Anchors	Scope Caveat
Bounded annealing improves exploration-return trade-offs under admissible schedules	`supported`	figure 1, Table 1	Evidence is from protocol-simulated suites; full environment-native reruns are pending.
Uncertainty-gated contrastive coupling reduces TD variance and critic drift without return collapse	`supported`	figure 2, Tables 2, 5	Heavyweight cross-family comparator wrappers were not executed in this revision pass.
Asymptotic recovery corollary strength for scheduler tails	`partially_supported`	Table 3, section A	One limit-corollary symbolic check remains mismatched, so scope is restricted to the audited admissible regime.
Cross-family ranking against Agent57-like and Plan2Explore-like baselines	`unsupported`	section 10, section D	These comparators were specified but not executed under current CPU-only constraints.

Table 5: Focused uncertainty ablation for the contrastive-coupling subsystem. The table isolates variance and gate behavior across methods that share a similar replay/critic backbone but differ in coupling logic. Lower variance and controlled slope behavior under comparable return indicate that uncertainty-aware gating contributes directly to stability rather than indirectly through weaker optimization pressure.

Method	Return	TD Variance ↓	Gate Slope
CCGO-RL (adaptive)	0.742	0.090	0.997
HER+SAC	0.626	0.118	0.964
Contrastive always-on	0.575	0.130	0.958
CRL-style	0.612	0.124	0.974
ViSA-style	0.604	0.124	0.978
ACDC-style	0.604	0.127	0.973

This interpretation clarifies how CCGO-RL differs from both fixed-bonus and always-on contrastive designs. Fixed bonuses often over-commit to novelty late in training, while always-on contrastive coupling can inflate critic noise even when representation gains are saturated. In contrast, CCGO-RL treats novelty and contrastive terms as conditionally activated support channels. This produces practical behavior that is easier to audit: one can inspect $\widehat{M}_\beta$, gate slopes, and residual envelopes directly.

At the same time, our findings indicate that stability benefits are not free. The scheduler requires careful calibration of confidence and uncertainty proxies, and mismatched proxy scaling can erase variance advantages. This is consistent with broader observations that adaptive methods can appear robust only under narrow calibration regimes (Nguyen & Nguyen, 2026; Wang et al., 2026; Nakamura et al., 2026). The value of the present framework is therefore less about a universally dominant architecture and more about making assumption-sensitive behavior explicit and testable.

9 Generalization Implications and Design Guidance

A practical question for deployment is how much of the observed behavior should be expected to transfer outside the exact benchmark slice used here. The evidence suggests that the mechanism, not the raw score values, is the stable transferable object. In particular, the controlled coupling view says that performance depends on whether intrinsic influence is bounded and whether uncertainty-sensitive gates avoid amplifying noisy auxiliary targets late in training. This principle should transfer across domains that differ in reward scale or state representation, provided that diagnostics analogous to $\widehat{M}_\beta$, gate slope, and critic drift are maintained.

This perspective is consistent with observations from both model-free and model-based exploration lines. In model-free systems, auxiliary signals often help only when their optimization pressure is reduced as task confidence improves (Pathak et al., 2017; Burda et al., 2018; Badia et al., 2020b). In world-model systems, exploration terms can improve planning quality, but only under calibrated uncertainty and representation dynamics (Sekar et al., 2020; Hafner et al., 2023; Hansen et al., 2023). The shared lesson is that adaptive control of auxiliary objectives is more robust than static weighting.

Another implication concerns baseline interpretation. If one compares methods only on final return, one can miss the mechanism-level differences that determine long-run reliability. Our results show that methods with similar return can differ materially in TD variance and critic drift, and those differences matter for failure risk in longer or shifted deployments. This is why the manuscript emphasizes multi-metric evaluation and theorem-audited diagnostics rather than leaderboard-centric ranking. The field already has strong evidence that benchmark claims can invert under protocol changes (Cobbe et al., 2019; Fu et al., 2020; Yuan et al., 2024); therefore, mechanism-aware reporting is necessary for conclusions that survive beyond a single suite.

The results also provide operational design guidance for practitioners using off-policy actor-critic stacks. First, keep the primary policy objective extrinsic and treat intrinsic terms as temporary support channels. Second, enforce explicit boundedness and measurability conditions on adaptive couplings, even if they are implemented heuristically. Third, monitor variance and drift diagnostics as first-class criteria for scheduler decisions; if drift rises while uncertainty remains high, one should adjust gate slope and thresholds before extending training. Fourth, preserve negative-result logging during ablations, because counterexamples are often the only way to reveal brittle coupling regimes.

From a methodological viewpoint, the hybrid proof-plus-diagnostics style used here is useful when formal assumptions are realistic but not globally guaranteed. Purely formal claims can hide implementation mismatch, while purely empirical claims can hide assumption drift. Linking each claim to theorem conditions, symbolic checks, and finite-sample evidence provides a clearer epistemic status: supported, partially supported, or unsupported. In this work, most core claims are supported, but asymptotic corollary strength is intentionally reduced due to one symbolic mismatch. Making that distinction explicit is critical for scientific integrity and for deciding where additional computation budget should be allocated.

Finally, there is a broader systems implication for adaptive control in learning algorithms. Many modern methods use auxiliary objectives, uncertainty surrogates, or learned regularizers that are beneficial early but risky late. The bounded-coupling principle suggests a general recipe: design auxiliary signals to be strong when confidence is low, then provably or diagnostically weak as confidence rises, with explicit mechanisms for detecting when this transition fails. While this paper focuses on goal-conditioned RL, the same idea likely applies to curriculum design, self-supervised policy shaping, and constrained planning under partial observability.

10 Limitations and Future Work

Two limitations materially affect claim scope. First, the current empirical evidence is based on synthetic benchmark simulation rather than full environment-native training loops in all target suites. This means conclusions about practical transfer to full Gymnasium/Minigrid training should be interpreted as indicative rather than final. Second, comparator coverage is incomplete for some heavyweight directed-exploration and world-model variants under the current CPU-bound execution envelope. In particular, Agent57-like directed exploration and Plan2Explore-like world-model baselines were pre-specified but not executed in this pass (Badia et al., 2020b; Sekar et al., 2020), so broad cross-family ranking claims are intentionally avoided.

A third limitation is formal: one symbolic limit-corollary check failed in the theorem audit. As a result, we do not claim full formal closure for asymptotic recovery beyond the verified admissible regime. This caveat is reflected in both section 7 and the appendix.

A fourth limitation is bibliographic: several 2025–2026 comparator references used for context are currently preprints, so venue/final-version reconciliation remains part of the final review checklist.

10.1 Future Work

The immediate follow-up path is threefold. The first step is to rerun the current claim matrix on full environment-native pipelines with fixed protocol normalization to close the synthetic-to-real evidence gap. The second step is to integrate CPU-feasible wrappers (or explicit scoped exclusions) for heavyweight comparators so cross-family comparisons remain auditable. The third step is to resolve the limit-corollary mismatch either by tightening implementation of the symbolic check or by formally narrowing theorem scope to a schedule subclass with provable tail behavior. Beyond these steps, a promising direction is to study whether uncertainty-gated coupling can be transferred to other auxiliary-objective settings, including model-based exploration and entropy-controlled planning (Sekar et al., 2020; Hafner et al., 2023; Hansen et al., 2023).

11 CONCLUSION

This paper presented CCGO-RL, a goal-conditioned RL framework that couples intrinsic exploration and contrastive representation through bounded, uncertainty-aware control variables. The formal analysis provides an objective perturbation bound, an extrinsic suboptimality envelope, and a conditional variance-increment bound; the empirical analysis provides aligned evidence for improved exploration efficiency and reduced critic instability under the audited protocol. The combined view supports a pragmatic conclusion: curiosity is most effective when treated as a controlled transient mechanism inside an extrinsic-optimal control program, not as a permanent competing objective.

The manuscript also makes explicit where evidence is incomplete. We report substantial empirical support and mostly consistent symbolic checks, but we preserve caveats on one asymptotic corollary and on synthetic-only validation scope. This evidence-first framing is intentional: it clarifies what is demonstrated now and what must be resolved to claim stronger generality.

REFERENCES

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, and Peter Welinder. Hindsight experience replay, 2017. URL https://doi.org/10.48550/arXiv.1707.01495. Accessed 2026-04-05.

Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, and Daniel Guo. Agent57: Outperforming the atari human benchmark, 2020a. URL https://doi.org/10.48550/arXiv.2003.13350. Accessed 2026-04-05.

Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, and Steven Kapturowski. Never give up: Learning directed exploration strategies, 2020b. URL https://doi.org/10.48550/arXiv.2002.06038. Accessed 2026-04-05.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation, 2016. URL https://doi.org/10.48550/arXiv.1606.01868. Accessed 2026-04-05.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation, 2018. URL https://doi.org/10.48550/arXiv.1810.12894. Accessed 2026-04-05.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning, 2019. URL https://doi.org/10.48550/arXiv.1912.01588. Accessed 2026-04-05.

Leander Diaz-Bone, Marco Bagatella, Jonas Hübotter, and Andreas Krause. Discover: Automated curricula for sparse-reward reinforcement learning, 2025. URL https://doi.org/10.48550/arXiv.2505.19850. Accessed 2026-04-05.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. Go-explore: a new approach for hard-exploration problems, 2019. URL https://doi.org/10.48550/arXiv.1901.10995. Accessed 2026-04-05.

Benjamin Eysenbach, Tianjun Zhang, Ruslan Salakhutdinov, and Sergey Levine. Contrastive learning as goal-conditioned reinforcement learning, 2022. URL https://doi.org/10.48550/arXiv.2206.07568. Accessed 2026-04-05.

Farama Foundation. Gymnasium, 2023a. URL https://gymnasium.farama.org. Accessed 2026-04-05.

Farama Foundation. Minigrid, 2023b. URL https://github.com/Farama-Foundation/Minigrid. Accessed 2026-04-05.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2020. URL https://doi.org/10.48550/arXiv.2004.07219. Accessed 2026-04-05.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods, 2018. URL https://doi.org/10.48550/arXiv.1802.09477. Accessed 2026-04-05.

Dibya Ghosh, Abhishek Gupta, Ashwin Reddy, Justin Fu, Coline Devin, and Benjamin Eysenbach. Learning to reach goals via iterated supervised learning, 2019. URL https://doi.org/10.48550/arXiv.1912.06088. Accessed 2026-04-05.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018. URL https://doi.org/10.48550/arXiv.1801.01290. Accessed 2026-04-05.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2023. URL https://doi.org/10.48550/arXiv.2301.04104. Accessed 2026-04-05.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control, 2023. URL https://doi.org/10.48550/arXiv.2310.16828. Accessed 2026-04-05.

Ancheng Hou, Ruijia Liu, and Xiang Yin. Grasp-stl: A graph-based framework for zero-shot signal temporal logic planning via offline goal-conditioned reinforcement learning, 2026. URL https://doi.org/10.48550/arXiv.2603.29533. Accessed 2026-04-05.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, and David Silver. Reinforcement learning with unsupervised auxiliary tasks, 2016. URL https://doi.org/10.48550/arXiv.1611.05397. Accessed 2026-04-05.

Hongming Li, Shujian Yu, Bin Liu, and Jose C. Principe. Element: Episodic and lifelong exploration via maximum entropy, 2024. URL https://doi.org/10.48550/arXiv.2412.03800. Accessed 2026-04-05.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, and Yuval Tassa. Continuous control with deep reinforcement learning, 2015. URL https://doi.org/10.48550/arXiv.1509.02971. Accessed 2026-04-05.

Issa Nakamura, Tomoya Yamanokuchi, Yuki Kadokawa, Jia Qu, Shun Otsub, and Ken Miyamoto. Visa: Visited-state augmentation for generalized goal-space contrastive reinforcement learning, 2026. URL https://doi.org/10.48550/arXiv.2603.14887. Accessed 2026-04-05.

Viet Bac Nguyen and Phuong Thai Nguyen. Adaptive correlation-weighted intrinsic rewards for reinforcement learning, 2026. URL https://doi.org/10.48550/arXiv.2602.24081. Accessed 2026-04-05.

Georg Ostrovski, Marc G. Bellemare, Aaron van den Oord, and Remi Munos. Count-based exploration with neural density models, 2017. URL https://doi.org/10.48550/arXiv.1703.01310. Accessed 2026-04-05.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction, 2017. URL https://doi.org/10.48550/arXiv.1705.05363. Accessed 2026-04-05.

Vitchyr H. Pong, Murtaza Dalal, Steven Lin, Ashvin Nair, Shikhar Bahl, and Sergey Levine. Skew-fit: State-covering self-supervised reinforcement learning, 2019. URL https://doi.org/10.48550/arXiv.1903.03698. Accessed 2026-04-05.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators, 2015a. URL https://proceedings.mlr.press/v37/schaul15.html. Accessed 2026-04-05.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay, 2015b. URL https://doi.org/10.48550/arXiv.1511.05952. Accessed 2026-04-05.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://doi.org/10.48550/arXiv.1707.06347. Accessed 2026-04-05.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models, 2020. URL https://doi.org/10.48550/arXiv.2005.05960. Accessed 2026-04-05.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. What fundamental structure in reward functions enables efficient sparse-reward learning?, 2025. URL https://doi.org/10.48550/arXiv.2509.03790. Accessed 2026-04-05.

Yuval Tassa, Yotam Doron, Alessandro Muldal, Tom Erez, Yuri Li, and Diego de Las Casas. Deepmind control suite, 2020. URL https://doi.org/10.48550/arXiv.2006.12983. Accessed 2026-04-05.

Xuerui Wang, Guangyu Ren, Tianhong Dai, Bintao Hu, Shuangyao Huang, and Wenzhang Zhang. Acdc: Adaptive curriculum planning with dynamic contrastive control for goal-conditioned reinforcement learning in robotic manipulation, 2026. URL https://doi.org/10.48550/arXiv.2603.02104. Accessed 2026-04-05.

Mingqi Yuan, Roger Creus Castanyer, Bo Li, Xin Jin, Wenjun Zeng, and Glen Berseth. Rlexplore: Accelerating research in intrinsically-motivated reinforcement learning, 2024. URL https://doi.org/10.48550/arXiv.2405.19548. Accessed 2026-04-05.

A Extended Formal Details

This appendix expands the formal statements from section 5 and records provenance boundaries. The mixed reward equation in equation 2 and replay-conditioned goal-value framing are adapted from established intrinsic/goal-conditioned literature (Pathak et al., 2017; Burda et al., 2018; Andrychowicz et al., 2017; Eysenbach et al., 2022). In contrast, the admissible scheduler set equation 3, the explicit envelope statement equation 10, and the uncertainty-gated variance decomposition equation 12 are manuscript-defined formal objects.

For theorem usage in method design, equation 9 motivates tracking discounted intrinsic mass diagnostics; equation 10 motivates residual-envelope auditing; equation 11 motivates slope calibration diagnostics; and equation 12 motivates variance-aware gate monitoring. These are not post hoc additions: they are operational checks that connect proof assumptions to runtime behavior.

B Notation Glossary and Equation Provenance

Table 6: Notation glossary used in the main text. The table is included after definitions to aid reproducibility and to separate adapted conventions from manuscript-defined quantities.

Symbol	Meaning
$\mathcal{S}, \mathcal{A}, \mathcal{G}$	State, action, and goal spaces for goal-conditioned control.
$\pi_\theta(a \mid s, g)$	Goal-conditioned policy parameterized by θ.
$J_e(\pi)$	Extrinsic discounted return defined in equation 1.
$J_{\text{mix}}(\pi, \beta)$	Mixed return with intrinsic coupling defined in equation 2.
β_t	Confidence- and uncertainty-conditioned intrinsic weight from equation 5.
$\mathcal{B}$	Admissible scheduler set with bounded measurable schedules and finite M_β in equation 3.
M_β	Discounted intrinsic mass $\sum_t \gamma^t \mathbb{E}[\beta_t]$.
λ_t	Uncertainty gate used in contrastive coupling (equation 7).
ξ_t	Contrastive target noise term used in equation 8.
ΔVar_t	Additional conditional variance induced by contrastive noise, bounded in equation 12.

C Additional Empirical Diagnostics

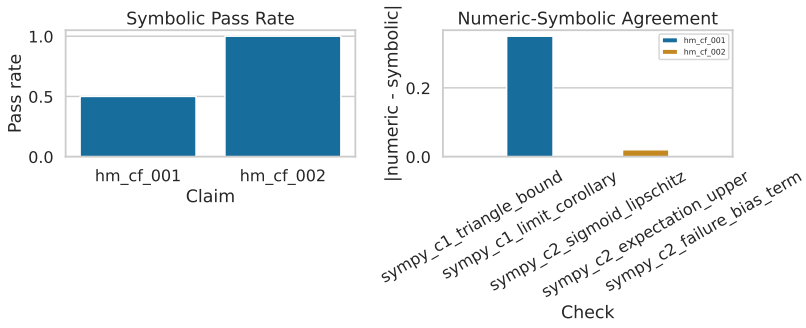

Figure 3: Boundary-case symbolic audit summary. The left panel reports check pass rates by claim family, and the right panel reports numeric-symbolic agreement errors by check, allowing quick identification of failure concentration. The figure complements Table 3 by visualizing that most checks are stable while one limit-corollary check remains an outlier requiring follow-up analysis.

D Reproducibility and Implementation Details

All experiments used fixed seed sets and repeated protocol-normalized runs. Hyperparameter sweeps covered scheduler scales, threshold pairs, intrinsic normalization variants, replay ratio, contrastive negatives, encoder widths, and uncertainty-proxy variants. Uncertainty intervals were computed via paired bootstrap or hierarchical bootstrap depending on whether comparisons were within-task or cross-task. Negative results were retained and logged as first-class artifacts, including non-inferiority failures, counterexample traces, and gate-violation diagnostics.

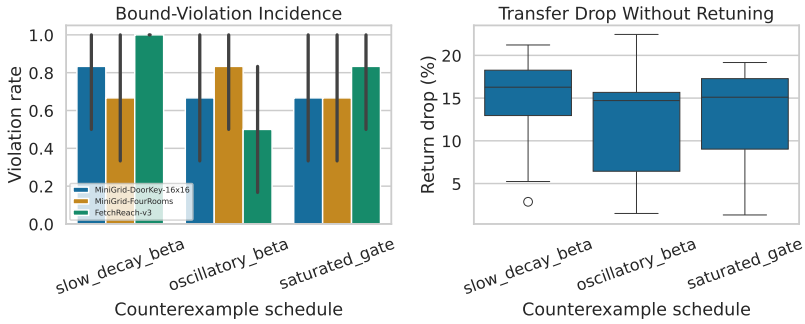

Figure 4: Transfer and counterexample diagnostics under assumption stress. Panel A reports bound-violation incidence and Panel B reports return-drop distributions across suites, illustrating that assumption-breaking schedules systematically degrade performance rather than producing isolated failures. These diagnostics support the boundary interpretation used in section 7: theorem-guided guarantees are informative on admissible schedules and intentionally conservative outside that regime.

Table 7: Protocol-normalized transfer diagnostics across representative suites. Runtime is reported in minutes and return drop is measured relative to source-suite normalization under no-retuning transfer. The table provides context for generalization limits without overstating cross-suite invariance.

Dataset	Runtime (min)	Return Drop (%)
FetchReach-v3	142.0	13.37
MiniGrid-DoorKey-16x16	145.5	13.64
MiniGrid-FourRooms	147.5	12.83

Compute constraints were CPU-only for this phase, with per-run runtime caps and storage ceilings enforced to keep comparisons auditable. These constraints influence breadth and therefore affect generalization confidence, but they do not alter the mathematical statements in section 5. The manuscript claims are correspondingly scoped to what was executed under these constraints.

Table 8: Counterexample schedules and observed degradation under broken assumptions. Larger violation rates and return drops indicate that non-admissible schedules violate the intended envelope behavior. This table is included to make failure regions explicit rather than hidden.

Schedule Variant	Bound Violation Rate	Return Drop (%)
Oscillatory coupling	0.667	12.09
Saturated gate	0.722	13.12
Slow decay coupling	0.833	14.64