When Is World Feedback Transferable? A Convergence Gate in Contrastive Reinforcement Learning

Bruce Changlong Xu[*] [1]
Jay Park [1] **Vivek Buch** [1]

Abstract

World feedback—measurable signals from agent–environment interaction such as goal reaching, future-state prediction, or system metrics—offers a path to scalable reinforcement learning that does not depend on human annotation. A standard recipe for deploying world-feedback agents is to amplify a strong teacher into smaller students via distillation. We show this recipe silently fails when the world-feedback signal is *contrastive*. In contrastive RL (CRL), where the teacher learns from an InfoNCE objective over future states, student performance is nearly constant across a wide range of partially trained teachers and then increases sharply once the teacher crosses a *convergence gate*—a sharp threshold in world-feedback representation quality. The gate is robust across three independent teachers, predictable from a zero-cost behavioral diagnostic (ρ>0.7, AUC= 0.895), and causally controlled by contrastive discrimination difficulty: varying the InfoNCE temperature shifts the gate by 13 epochs. Once the gate closes, amplification is dramatic: on Humanoid, distillation transforms a near-zero baseline (6.4) into a walking policy at 92% of teacher performance. By contrast, when the world-feedback signal is a scalar reward (SAC), distillation is smooth and gate-free, transferring an 11× smaller student to 96.6% of teacher performance. The *form* of world feedback—dense contrastive vs. scalar—can therefore determine whether amplification is smooth or gated. We close with operational guidance for world-feedback pipelines: cheap behavioral statistics on the teacher (mean action magnitude, output entropy) are stronger checkpoint-selection signals than validation loss, and contrastive temperature should be treated as an explicit knob over the gate's location.

[*]First author. [1]Stanford University. Correspondence to: Bruce Changlong Xu <brucechanglongxu@cs.stanford.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Human feedback has been the dominant alignment signal for foundation-model agents, but its scaling limits—cost, subjectivity, latency, and applicability outside text—are now well established. *World feedback*—measurable signals arising directly from interactions with physical, digital, economic, or social environments—is an attractive substitute: it is automatic, abundant, and grounded in real-world consequence. Recent work spans goal-conditioned and contrastive RL on embodied agents (Eysenbach et al., 2022; Wang et al., 2025), self-supervised representations on system states (Laskin et al., 2020), and reward-based control from simulator metrics (Haarnoja et al., 2018). A common deployment pattern across these methods is to train an expensive world-feedback teacher and then *distill* its behavior into a smaller, faster student that can run on real hardware or serve production traffic.

We study when this amplification step actually works. The implicit assumption mirrors the human-feedback distillation literature (Hinton et al., 2015; Ouyang et al., 2022; Rusu et al., 2015): that better teachers produce monotonically better students, with performance scaling smoothly along the way. We show that this assumption breaks down sharply for one of the most general forms of world feedback—*contrastive* world feedback, where the teacher is trained to discriminate goal-consistent from goal-inconsistent futures via InfoNCE.

In contrastive RL (CRL; Eysenbach et al. 2022), distillation exhibits a sharp *convergence gate*: student performance is nearly constant across a wide range of partially trained teachers—even those with high evaluation scores—and increases sharply only when the teacher is *fully* converged (Figure 1). The gate is robust across three independently trained teachers, predictable from a zero-cost teacher-side diagnostic (ρ>0.7, AUC= 0.895), but causally dissociated from these behavioral predictors: artificially shaping an

unconverged teacher's actions to mimic convergence does not cross the gate. The mechanism is instead *contrastive discrimination difficulty*: varying the InfoNCE temperature shifts the gate location by 13 epochs (Section 4).

Once the gate is crossed, world-feedback amplification is remarkably effective. A width-scaled depth-4 student matches the depth-8 teacher on Ant Big Maze (p=0.93, n=10) and reaches 92% of teacher performance on Humanoid—transforming a network that cannot stand into one that walks. An equal-parameter control isolates the mechanism: at identical parameter count ($\sim$1.65M), the shallow network scores 29 alone but 374 with distillation, indicating it *can represent* the solution but *cannot discover it* from world feedback alone. SAC distillation (scalar reward feedback) on the same architectures shows no gate: an $11\times$ smaller student smoothly reaches 96.6% of teacher reward. Pinning the phenomenon to the contrastive objective separates the locus of the bottleneck (the form of world feedback) from the substrate (architecture, task, optimization).

Why this matters for RLxF. World-feedback pipelines increasingly rely on auxiliary contrastive losses—next-state prediction, goal–observation alignment, and self-supervised representations of system state. Our results imply that the standard practice of distilling from *any* sufficiently good checkpoint is unsound when such losses are present. The teacher–student transfer function can be a step rather than a slope, and the cheapest behavioral signal an operator already collects (mean action magnitude in our case) predicts which side of the step a candidate teacher sits on. For real-world deployments—robotics, recommendation, simulation–to–reality transfer—this turns a silent failure mode into a checklist item.

Contributions.

1. **A convergence gate in world-feedback distillation.** A sharp transition in student performance as a function of teacher training, replicated across three teacher seeds and predictable from a behavioral statistic (AUC= 0.895).
2. **Causal mechanism.** The gate is causally controlled by contrastive discrimination difficulty (InfoNCE temperature), not by the surface behavior of the teacher; behavioral predictors are epiphenomenal.
3. **The form of world feedback matters.** Scalar reward (SAC) distillation transfers smoothly (96.6%, no gate); temporal contrastive feedback (CURL) reverses the depth gradient. The phenomenon is a property of the auxiliary world-feedback objective, not of RL or distillation in general.
4. **Discovery vs. expressivity.** Equal-parameter controls and late-activation experiments separate optimization from architectural capacity. Depth accelerates discovery rather than providing irreducible expressivity ($29 \rightarrow 374$ with a teacher); shallow students can host capabilities they cannot find from world feedback alone.
5. **Operational guidance.** A behavioral diagnostic (mean $|\tanh(\mu)|$) and a controllable knob (InfoNCE temperature) together specify when and how to amplify contrastive world-feedback agents.

2. Background: World Feedback and Distillation

Forms of world feedback. We use *world feedback* to mean any reward or supervision signal automatically derived from agent–environment interaction. We compare three forms in this paper: **(i) Scalar reward feedback**, where a hand-specified or metric-derived reward is consumed by an actor–critic algorithm (Haarnoja et al., 2018); **(ii) Goal-conditioned contrastive feedback**, where the agent is rewarded for representing future states in a way that discriminates achieved from non-achieved goals via InfoNCE (Eysenbach et al., 2022; Wang et al., 2025); and **(iii) Temporal contrastive feedback**, where a single encoder is trained to discriminate temporally close from far state pairs (Laskin et al., 2020). All three are forms of world feedback in the RLxF sense—no human annotation enters the loop—but they differ in density and structure.

Contrastive RL. CRL parameterizes the goal-conditioned Q-function via encoders $\phi(s,a), \psi(g) \in \mathbb{R}^d$ with critic $f(s,a,g) = -\|\phi(s,a) - \psi(g)\|_2^2$, trained with InfoNCE (Wang & Isola, 2020). The actor $\pi_\theta(a|s,g)$ maximizes the critic. Actor–critic separation lets us distill policy and representation independently.

Depth-induced phase transitions. Wang et al. (2025) show that on Ant Big Maze, depth-4 agents reach $\sim 60/1000$ timesteps at goal while depth-8 reaches $\sim$440; on Humanoid the transition is at depth 16. We use this depth gap as the substrate for adaptation: the teacher must be deep, but the deployed student would ideally be shallow.

Distillation as world-feedback amplification. Students train with their own world-feedback objective (InfoNCE for CRL, scalar reward for SAC) plus a Gaussian policy distillation loss from a frozen teacher: $\mathcal{L} = \mathcal{L}_{\text{world}} + \lambda_\pi \mathcal{L}^\pi_{\text{distill}}$, where $\mathcal{L}^\pi_{\text{distill}} = \mathbb{E}[\|\mu^S - \mu^T\|^2 + \|\log\sigma^S - \log\sigma^T\|^2]$. Students still explore independently and consume world feedback directly; distillation is auxiliary supervision, not a replacement for the environmental signal.

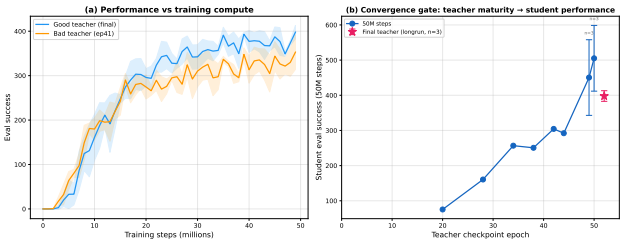

Figure 1. **The convergence gate.** *Left:* Student eval vs. training compute. Good-teacher students (blue) separate from bad-teacher students (orange) after ∼10M steps despite converging to the same policy MSE. *Right:* Student performance vs. teacher checkpoint: flat for ∼80% of teacher training, then a sharp jump. Error bars: ±1 std (3 seeds).

Table 1. Dense gate sweep (d=4, w=512, Ant Big Maze). Teacher eval is noisy (E38 drops to 200.2), but student performance tracks teacher *training step* monotonically. Dashes (—) mark seed/horizon combinations we did not run; the remaining cells already establish the trend.

Ckpt	Step	T. eval	Stu. 20M	Stu. 50M
E05	5.5M	0.0	0.0	—
E20	20.2M	102.2	117.7	75.3
E34	34.0M	286.1	179.0	256.7
E38	37.9M	200.2	216.8	250.6
E41	40.9M	321.5	—	352.9±38.7
E49	48.8M	382.7	223.0	450.4±107.5
E50	49.7M	382.7	—	397.9±15.5

Setup. Ant Big Maze (29-dim obs) and Humanoid (268-dim, 17 actuators). Residual MLPs with LayerNorm + Swish; 512 parallel envs, 50M steps, λ_π=1.0. Mechanistic results use 3–5 seeds; headline results use n=10. Total: >350 runs, ∼420 GPU-hours (AMD MI250X, NVIDIA H100).

3. The Convergence Gate

We distill from 10 teacher checkpoints spanning the full training trajectory (epochs 5–50) into capacity-rich students (d=4, w=512). The result is stark (Figure 1, Table 1): performance is flat for the first ∼80% of teacher training, then jumps sharply—even though teacher evaluation noise can *drop* 30% between consecutive checkpoints (E38) while the student continues improving.

The gate sits above an absolute floor: depth-2 networks score exactly 0.0 across 300 evaluations at 100M steps—regardless of width or distillation. Strikingly, both good- and bad-teacher students converge to the *same* policy MSE (∼0.09) but diverge in eval after ∼10M steps: the student faithfully imitates both teachers, and the gap comes entirely from *what those actions achieve*, not how closely the student matches them. At the gate boundary, E49 is bimodal (450.4 ± 107.5)—some seeds cross, others fail—while E50 is consistent (397.9 ± 15.5), consistent with critical-point behavior. We later show (Section 6, Figure 5) that this transition tracks a representation-geometry threshold in the teacher's InfoNCE encoder.

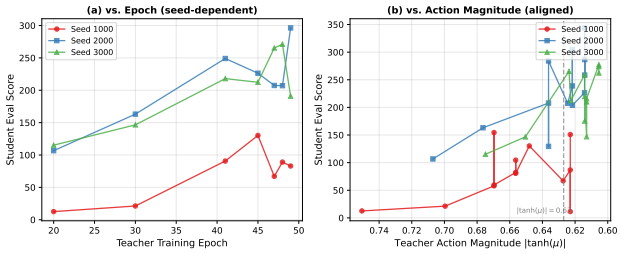

Figure 2. **The gate is predictable.** (a) Student eval vs. teacher epoch: gate location varies by seed. (b) Same data vs. teacher action magnitude: the three curves align near $|\tanh(\mu)| = 0.63$ (dashed).

A behavioral diagnostic for teacher readiness. Across 21 teacher–epoch conditions (3 teacher seeds), action magnitude predicts student success (Spearman $\rho = -0.77$, $p < 10^{-4}$; AUC= 0.895). When student performance is plotted against teacher action magnitude rather than epoch, the three seed-dependent gate curves collapse near a single threshold at $|\tanh(\mu)| = 0.63$ (Figure 2). This is the sort of zero-cost diagnostic a world-feedback pipeline can log essentially for free—the operator already runs the teacher to evaluate it—and yet it predicts downstream amplification better than validation loss.

Causal dissociation: behavior is not the cause. Are these behavioral predictors also *causes*? We artificially shape an unconverged teacher's outputs to mimic convergence—clamping log-std to force deterministic actions and scaling means to reduce action magnitude—then distill. Entropy-clamped students score 0.0 across 6 runs; mean-scaled students reach only 54±45. Applying the same shaping to a converged teacher *degrades* its students from 383 to ∼190. The behavioral predictors are epiphenomenal: convergence produces both committed actions *and* high-quality world-feedback representations, but only the latter is causally necessary. This rules out the simplest explanation of the gate ("the teacher just needs to act decisively") and motivates the mechanistic study below.

4. Mechanism: Contrastive Discrimination Difficulty

If the gate arises from the contrastive form of world feedback, then modifying contrastive difficulty should shift it. We train three d=8 teachers at InfoNCE temperatures $\tau \in \{0.5, 1.0, 2.0\}$ and distill from intermediate checkpoints into identical students (Figure 3).

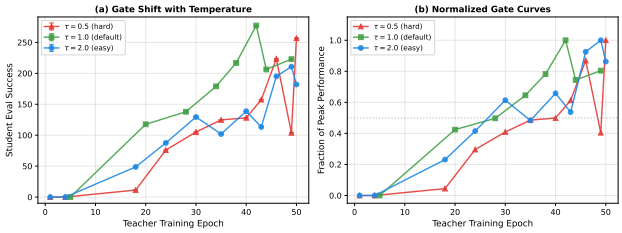

Figure 3. **Contrastive difficulty causally controls the gate.** (a) Student performance vs. teacher epoch at three InfoNCE temperatures. Easier discrimination (τ=2.0, blue) opens the gate earlier; harder (τ=0.5, red) delays it—a 13-epoch causal manipulation. (b) Normalized curves isolate timing from absolute performance.

The gate shifts monotonically: τ=2.0 (easy) crosses at **epoch 30**; τ=1.0 (default) at **epoch 34**; τ=0.5 (hard) at **epoch 43**—a 13-epoch spread, verified across n=3 seeds. Because only the temperature varies (architecture, seed, pipeline fixed), this establishes contrastive discrimination difficulty as the causal driver of the gate. Equivalently: the temperature of the world-feedback signal is a control knob on *when* that signal becomes transferable.

Why temperature shifts the gate. The InfoNCE softmax over N negatives saturates near $\log N$ when the encoder cannot spread embeddings far enough to discriminate negatives, returning near-zero gradient through the critic to the actor. Lower τ tightens the discrimination, raising the embedding spread required to escape saturation and delaying gate closure; higher τ relaxes it. The full alignment–uniformity argument and its connection to depth-induced collapse appears in Section 8.

5. Main Results: Amplifying World-Feedback Agents

Distillation across the gate. Table 2 shows that once the convergence gate is crossed, world-feedback amplification is highly effective. On Ant Big Maze, the width-512 student achieves 383.2±17.7 (n=10), statistically indistinguishable from the teacher's 382.7 (p=0.93, Cohen's d=0.03). On Humanoid, distillation transforms a near-zero baseline (6.4) to 383.9 (92% of teacher).

Policy distillation is the *only* effective channel: combined distillation (policy + representation) performs *worse* than baseline on Ant Big Maze (32.3 vs. 48.2), and representation-only distillation is useless on Humanoid (5.0). For world-feedback amplification, the teacher's value lies in *what to do*, not *how to see*.

Depth accelerates discovery, not expressivity. We distinguish architectural *expressivity*—whether a network of given parameters can in principle encode the policy—from *discovery*—whether world feedback alone provides enough optimization signal to find it. At identical parameter count (∼1.65M; Table 3), d=4, w=360 scores 29.2 without distillation vs. ∼295 for d=8, w=256—a 10× depth advantage with no parameter difference. With distillation, the shallow network reaches 373.6: the same parameters that score 29 alone score 374 with a teacher, a 12.8× improvement that cannot be attributed to expressivity. World feedback carries enough information for the shallow network's expressivity to *realize* the policy; it does not carry enough optimization signal for the network to *discover* it.

Table 2. Distillation across the gate (`eval_success`, 50M steps, mean ± std). Width-scaled students match teachers on both environments; policy distillation is the only effective channel.

Condition	Ant Big Maze	Humanoid
Teacher (d=8/d=16)	382.7	417.8
Baseline (d=4, w=256)	48.2±36.7	6.4±1.8
Policy distill (w=256)	80.4±23.4	314.0±54.0
Baseline (d=4, w=512)	120.7±104.3	12.2±2.5
Policy distill (w=512)	**383.2±17.7***	**383.9±24.3**

* n=10 seeds; others 3 seeds.

Table 3. Equal-parameter control (Ant Big Maze, 50M steps, 3 seeds). Depth-8 outperforms depth-4 by 10× at identical parameter count; distillation closes the gap entirely.

Architecture	Params	Baseline	+ Distill
d=4, w=360	1.648M	29.2±45.5	**373.6±35.1**
d=8, w=256 (teacher)	1.648M	∼295	—

Multi-teacher robustness. Distilling from three independently trained teachers into identical students (50M steps, 3 seeds each) shows the gate is a property of teacher *state*, not seed: the converged teacher (t_{2000}, eval 382.7) yields students at 383.2±17.7; an unconverged teacher (t_{1000}, eval 267.8) yields only 227.4±54.6. On Humanoid, teacher–student quality correlation is r=0.992 across three independent teachers. World-feedback teacher quality is genuinely a one-dimensional predictor here, not a coincidence of one particular run.

Late activation confirms an optimization barrier. Students trained for 50M steps without distillation perform comparably when distillation is activated late (318.7 ± 40.3 vs. fresh 268.9 ± 42.1 on Ant Big Maze; 6× improvement on Humanoid). Prior self-supervised world-feedback training neither helps nor hinders—the depth barrier is an optimization problem that a converged teacher's actions bypass regardless of the student's history.

Imitation succeeds; transfer fails. Figure 4 resolves a subtle point that is particularly relevant to world-feedback

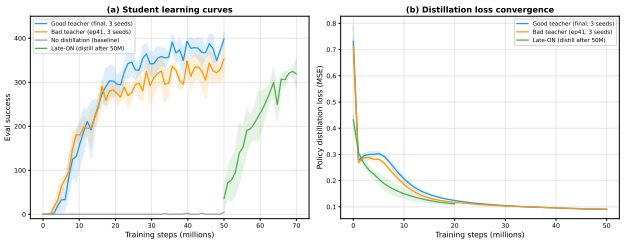

Figure 4. **Imitation succeeds; transfer fails.** *Left:* eval success diverges after ∼10M steps—good teacher (blue) climbs, bad teacher (orange) plateaus. *Right:* both students reach the same policy MSE floor (∼0.09). The downstream gap is in the utility of those actions, not their fidelity.

Table 4. SAC depth sweep on Humanoid (3 seeds). Depth advantage is gradual and closes at 50M (ANOVA p=0.42).

Depth	Steps	Mean Reward	p vs. d=8
2	10M	1400±610	0.012
4	10M	2591±359	0.148
8	10M	**3008±157**	—
2	50M	3908±340	ANOVA p=0.42
8	50M	3414±717	

pipelines. Good- and bad-teacher students converge to the *same* policy MSE floor (∼0.09): the student matches both teachers' action distributions to the same precision. Yet evaluation success separates after ∼10M steps and continues diverging throughout training. The gap is therefore not a failure of imitation—it is a property of *what those imitated actions achieve in the world*. A bad teacher's actions are well-formed in distribution but do not compose into successful trajectories; a good teacher's do. The diagnostic implication is direct: in any world-feedback pipeline, monitoring the distillation loss alone will not reveal a failed transfer; one must close the loop through the environment.

6. The Form of World Feedback Matters

SAC: gradual depth, no gate. We train SAC at depths 2/4/8/16 on Humanoid (Table 4). At 10M steps, d=8 achieves 2.15× the reward of d=2 (p=0.012); the advantage is gradual—no sharp transition—and at 50M steps all depths converge (ANOVA F=1.05, p=0.42). CRL tells the opposite story on the same task: at 100M steps, d=2 produces 0.0 across all 300 evaluations.

SAC distillation: smooth amplification. We distill a d=8 SAC teacher to a d=2 student via behavioral cloning (3 seeds). The student (11× fewer parameters) reaches **96.6%** of teacher reward with smooth learning dynamics: 49% at epoch 0, 87% by epoch 5; all three seeds exceed the teacher at peak (100.7% mean). No convergence gate, no threshold, no failure mode—the antithesis of CRL. Scalar world feedback appears to amplify more smoothly because reward-based critics continue to provide usable gradient signal across depths, even when the encoder is too shallow to support a well-spread contrastive embedding.

CURL: reverse depth pattern. A shared-encoder CURL (Laskin et al., 2020) experiment (temporal InfoNCE on Humanoid, n=5) shows the *reverse* depth pattern: d=2 achieves 14× better temporal prediction than d=8, with a three-level coefficient ablation confirming a monotonic dose–response (+36/+66/+88 depth gap as contrastive strength increases from $0 \to 0.1 \to 1.0$). Two contrastive world-feedback objectives can therefore impose *opposite* architectural pressures on the same task; the generic statement "contrastive losses help" is incoherent without specifying which contrastive loss.

Emergent critic alignment. Policy distillation supervises only the actor, yet the student's critic also aligns with the teacher's: SA-encoder CKA is 0.63 (good student) vs. 0.44 (baseline)—a 44% increase with no direct critic supervision. The goal-encoder CKA reaches 0.93 and the Q-value Pearson correlation r=0.88, despite the student learning the critic entirely from its own InfoNCE world feedback. Teacher-like actions produce teacher-like states, which reshape the critic through the standard CRL loss—a second-order effect that restructures the entire learning pipeline. This rules out the interpretation that distilled students are mere imitators: their internal value geometry has been rebuilt from world feedback to match the teacher's.

Representation geometry across depth. We measure alignment, uniformity (Wang & Isola, 2020), and effective rank of CRL encoder embeddings at d=2, 4, 8. At convergence, SA-encoder effective rank increases monotonically with depth ($11.8/64 \to 14.8/64 \to 17.6/64$), and positive-pair alignment is 2× worse at d=2 (264) than d=8 (129). The CURL experiment corroborates this from the opposite direction: under *temporal* contrastive feedback, d=8 collapses to ∼15/64 effective rank by mid-training while d=2 maintains ∼34/64. Same dimensional-collapse mechanism, opposite sign—because the two contrastive world-feedback signals impose opposite depth gradients.

Recursive distillation and persistent scaffold. Second-generation distillation (student→student) retains 91% of performance (361.0 ± 24.5 vs. 397.9 ± 15.5), showing the distilled policy is itself a viable world-feedback teacher—an attractive property for staged deployment pipelines. Removing the distillation signal from a competent student degrades it from 404.5 ± 15.2 to 274.1 ± 8.5 over 50M steps—still 3×–9× above the undistilled baseline. A gradual annealing experiment (linearly decaying λ_π from 1.0 to 0.0) converges to the same ∼274 equilibrium, ruling out distribution shift:

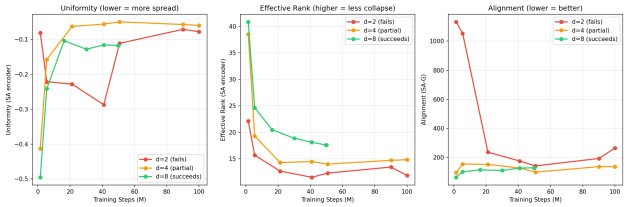

Figure 5. **Representation geometry across depth.** Shallow encoders (d=2, red) show strong dimensional collapse, poor alignment, and unstable uniformity; deeper encoders (d=8, green) maintain higher-rank, better-aligned embeddings, with d=4 (orange) intermediate. The same ordering predicts distillation outcomes, supporting representation quality as the latent state variable governing the gate.

Table 5. Diagnostic confusion matrix (21 teacher–epoch conditions, 3 teacher seeds, threshold $|\tanh(\mu_T)| < 0.63$). Truth: student reached >300 at 50M.

	Pred. closed	**Pred. open**
Truth: gate closed	9	2
Truth: gate open	0	10

the teacher acts as a *stabilizer* that keeps the student in a favorable basin, not as a continuous source of new information.

7. An Amplification Protocol for World-Feedback Agents

We give RLxF operators a free diagnostic and a single causal knob for amplifying contrastive world-feedback teachers. The protocol below uses only quantities the operator already collects during teacher training and exposes InfoNCE temperature τ as a deliberate control over when the gate closes within a given compute budget.

Diagnostic accuracy. We treat the action-magnitude rule of Algorithm 1 ("the teacher has crossed the gate iff $|\tanh(\mu_T)| < 0.63$") as a binary classifier and evaluate it across our 21 teacher–epoch conditions (Table 5). The rule achieves 90.5% overall accuracy with no false positives—when the diagnostic says the gate is closed, the resulting student always succeeds—and a small false-negative rate (two checkpoints flagged as "not-yet-closed" that nevertheless distilled successfully). For an operator deciding whether to spend additional teacher compute, the asymmetry is the right one: false negatives waste compute training a teacher that was already good enough; false positives would waste an entire student run on a teacher that could not transfer. The diagnostic costs nothing the operator is not already paying for.

Mapping to real world-feedback domains. *The mappings below are predictions of the framework, not empirically validated extrapolations.* We outline how the protocol applies to three RLxF-relevant deployment patterns where contrastive world feedback already appears. **(a) Sim-to-real robotics.** Goal-conditioned CRL is a natural fit for embodied agents trained in simulators with goal-state contrast (e.g., a manipulator reaching a target configuration). The diagnostic translates directly: $|\tanh(\mu_T)|$ on a fixed eval batch in simulation is free, and the temperature knob is exposed by every InfoNCE implementation. **(b) Recommendation systems with implicit feedback.** Click and dwell-time feedback is commonly trained as a two-tower contrastive objective. Our results predict a gate between underfit and converged towers: distilling from an underfit two-tower teacher into a smaller serving model would plateau even though the teacher's offline loss looks reasonable. **(c) System-metric agents.** Datacenter or vehicle controllers trained from latency, energy, or safety metrics typically use scalar reward; our SAC results suggest these amplify smoothly. Mixed pipelines that add a contrastive auxiliary loss for sample efficiency may inherit the gate from the contrastive component—a regression mode worth monitoring explicitly.

Algorithm 1 Gate-Aware World-Feedback Amplification

Require: Teacher training run, student architecture, compute budget B, action-magnitude threshold θ=0.63.

1: Train teacher with InfoNCE temperature τ (default τ=1.0).
2: **During teacher training:** log mean $|\tanh(\mu_T)|$ on a fixed eval rollout every K steps *(free; uses the eval rollouts the operator already runs)*.
3: **Gate detection:** flag the first checkpoint at which $|\tanh(\mu_T)| < \theta$ as a candidate teacher.
4: **If the gate has not closed by 70% of the teacher compute budget:** re-train or continue at higher τ (easier discrimination); empirically τ: 1→2 shifts the gate ∼4 epochs earlier.
5: **Distill** from the first post-gate checkpoint into a width-scaled shallow student using policy-only distillation (λ_π=1.0, no representation distillation, no critic distillation).
6: **Verify** via in-environment evaluation, not distillation MSE: imitation loss converges identically for good and bad teachers (Section 5).
7: **Optionally** recursively distill student→student to retain ∼91% performance at further compression (Section 6).

Cost of ignoring the gate. A pipeline that selects teacher checkpoints by validation loss alone—common practice in supervised distillation—would, on our 50M-step CRL runs, distill from E34 (validation eval 286.1) and obtain a 256.7 student rather than the 397.9 achievable from E50. Translating to compute: the operator pays the full student-

Table 6. Full temperature–gate ablation (Humanoid, d=8 teachers, 3 seeds each). Gate epoch is the first checkpoint at which median student eval exceeds 300.

τ	Difficulty	Gate epoch	Spread
2.0	Easy	30	±1
1.0	Default	34	±1
0.5	Hard	43	±2

Table 7. Cross-task validation on arm bin-picking (3 seeds, 50M steps). Depth gradient reproduces; distillation gate softens because d=4 baselines learn independently.

Depth	Baseline	+ Distill	Δ
2	0.00	0.00	—
4	0.15	0.17	+14%
8	**8.00**	—	—

Table 8. Latent state-variable correlations across 21 teacher–epoch conditions. Each metric is a partial readout of the same underlying representation-quality variable.

Teacher metric	Spearman ρ	p
$\lvert\tanh(\mu_T)\rvert$ (action magnitude)	-0.77	$< 10^{-4}$
SA-encoder effective rank	$+0.79$	< 0.005
Positive-pair alignment	-0.71	< 0.005
Uniformity (hypersphere)	$+0.66$	< 0.01
InfoNCE eval loss	-0.62	< 0.01
Q-value correlation with truth	$+0.74$	< 0.005

run cost ($\sim$12 GPU-hours each) but realizes only $\sim$65% of the teacher's capability. The gate-aware protocol shifts the same student run to a useful checkpoint at no additional teacher cost. The cost of ignoring the gate is therefore not an underperforming student in isolation; it is the entire student run executed on the wrong side of the gate.

What the gate is not. Three plausible alternative explanations are ruled out by our controls. *(i) Underfit teacher policy.* A teacher whose *actions* are unconverged but whose representation is good should fail; conversely, shaping a converged teacher's actions to mimic an unconverged one should still succeed. We observe the opposite (Section 3, causal dissociation): the representation, not the policy surface, controls transfer. *(ii) Capacity bottleneck in the student.* Equal-parameter controls show the same student parameters can host the solution—29 alone, 374 with a teacher. The gate is on the teacher side. *(iii) Distribution shift between teacher and student.* Annealing λ_π from 1 to 0 produces the same equilibrium as abrupt removal, and recursive distillation retains 91% even with no access to the original teacher. Distribution drift is not the operative mechanism.

Additional evidence. Three further datasets corroborate the framework. *(a) Full temperature ablation* (Table 6): the 13-epoch gate shift across $\tau \in \{0.5, 1.0, 2.0\}$ is monotonic and replicates across all three seeds, ruling out a single-seed artifact. *(b) Cross-task generalization on arm bin-picking* (Table 7): the depth gradient (d=2:0.0, d=4:0.15, d=8:8.0) reproduces, but the sharp distillation gate softens to a +14% improvement because d=4 baselines learn independently—useful negative evidence for when the framework's strongest predictions apply. *(c) Latent state-variable correlations* (Table 8): six independent geometry/behavior metrics correlate with student outcome at $|\rho| \geq 0.6$ across the 21-condition panel. Each metric is a partial readout of the same latent quantity, which is what motivates collapsing them onto a single behavioral threshold for deployment.

8. Discussion

Two-level structure of depth. Our experiments decompose depth effects into two levels. *(i) Universal:* depth accelerates optimization—shallow networks match deep ones given more compute (SAC at 50M) or teacher guidance (SAC distillation at 96.6%). *(ii) Form-specific:* CRL's goal-conditioned InfoNCE sharpens this smooth advantage into a phase-transition-like gate, while temporal InfoNCE (CURL) reverses it entirely. The *form* of world feedback decides which regime an operator is in.

Why does contrastive world feedback produce sharp transitions? InfoNCE optimizes alignment–uniformity on the hypersphere (Wang & Isola, 2020). When the encoder is too shallow, embeddings collapse to a low-dimensional subspace (effective rank: $11.8/64$ at d=2 vs. $17.6/64$ at d=8), the softmax over N negatives saturates at $\log N$, and the actor receives near-zero gradient—a cliff, not a slope. Above a critical depth, embeddings spread, informative gradients resume, and behavior emerges discontinuously. SAC's reward-based critic always provides *some* gradient signal regardless of depth, explaining its smooth profile and gate-free distillation. World feedback that is *dense and discriminative* (contrastive) exposes capacity bottlenecks much more sharply than feedback that is *sparse and scalar* (reward).

Connection to broader phase transitions. The convergence gate shares structural signatures with grokking (Power et al., 2022) and emergent capabilities in LLMs (Wei et al., 2022): a disconnect between training loss and downstream capability, bimodal behavior at the critical point (E49 students span 373 ± 126 across seeds), and sharp dependence on a latent variable (representation quality) rather than any surface metric. Unlike grokking—where the transition is in training *time*—the gate is a transition in teacher *state*: the student's training dynamics are smooth, but the teacher must cross a representation-quality threshold before its world-feedback knowledge becomes transferable. This separates

the locus of the phase transition (the source model) from where it manifests (the downstream amplifier), which is a regime that has not, to our knowledge, been characterized in the world-feedback literature.

A latent state-variable view. Multiple metrics—contrastive loss, effective rank, alignment, uniformity, CKA, Q-value correlation, action magnitude—move together across all of our experimental conditions. We interpret the gate as a threshold in a single latent quantity (world-feedback representation quality) for which each metric is a partial readout. Behavioral predictors are *measurements* of this latent state; temperature is an *intervention* on it. The 0.79 Spearman correlation between spread/PCR-style geometry diagnostics and student outcomes across our 21 teacher–epoch conditions ($p<0.005$) is what one would expect if a single underlying variable were driving the transition.

Implications for RLxF pipelines. Auxiliary contrastive losses are common in modern world-feedback pipelines—vision–language alignment for embodied agents, retrieval encoders trained on click feedback, simulator-derived metric encoders for sim-to-real. Our results imply two operational changes for any pipeline that distills or amplifies through such losses:

1. **Do not select world-feedback teachers by validation loss alone.** Loss can plateau well before the contrastive representation is transferable; cheap behavioral statistics on the teacher (action magnitude in our case, output entropy or representation effective rank more generally) are stronger predictors of downstream amplification success.
2. **Treat the contrastive objective as part of the adaptation interface.** The gate location is a function of the contrastive temperature; a world-feedback pipeline could expose this knob explicitly, scheduling discrimination difficulty so that the gate falls inside the available compute budget rather than past it.

Practical takeaway for RLxF practitioners. (1) Monitor mean $|\tanh(\mu)|$ during teacher training; when <0.63, the teacher has likely crossed the gate (90.5% accuracy). **(2)** Distill to a width-scaled shallow student that preserves teacher-level capability under a smaller-depth architecture; distilling from a partially trained contrastive-world-feedback teacher wastes compute. **(3)** If the world-feedback signal is scalar reward, amplification is smooth; if it is contrastive, expect a gate and plan around it.

Limitations. (1) The sharp gate is demonstrated in CRL; SAC and CURL comparisons control the obvious alternatives but do not rule out analogous phenomena in untested world-feedback methods (SPR, BYOL-based RL, world-model objectives, click-feedback retrieval). (2) All experiments use Brax simulation (>350 runs for statistical rigor); transfer to higher-dimensional foundation-model amplification (e.g., embodied LLMs trained on environment metrics) remains future work. (3) Cross-task validation on arm bin-picking reproduces the depth transition (d=2:0, d=4:0.15, d=8:8.0) but *not* the sharp distillation gate—d=4 baselines learn independently here, so distillation provides only a moderate ($+14\%$) benefit. The sharp gate appears to require tasks where the depth barrier is absolute and width cannot compensate at the tested scale. (4) The scaffold mechanism is not fully characterized: removing distillation degrades performance by $\sim 32\%$ but stabilizes well above the undistilled baseline, and recursive distillation retains 91%, indicating internalization rather than parroting, though the precise content of what is internalized remains open. (5) Our distillation coefficient is fixed (λ_π=1.0); adaptive schedules remain unexplored and are a natural extension for pipelines that want to land the gate at a chosen compute budget.

Conclusion. World feedback is a scalable substitute for human feedback only if world-feedback agents can be amplified into deployable students. We have shown that this amplification is not a monotone process when the world-feedback signal is contrastive: it is gated by a sharp transition in the teacher's representation, predictable from a free behavioral diagnostic, and causally controlled by contrastive temperature. Scalar reward feedback exhibits no such gate. The form of world feedback—not just its presence—governs whether a world-feedback agent is a teacher one can build on, and the diagnostic and the knob we identify give RLxF pipelines a concrete, low-cost interface to that gate.

References

Eysenbach, B., Zhang, T., Salakhutdinov, R., and Levine, S. Contrastive learning as goal-conditioned reinforcement learning. In *NeurIPS*, 2022.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*, 2018.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

Laskin, M., Srinivas, A., and Abbeel, P. Curl: Contrastive unsupervised representations for reinforcement learning. In *ICML*, 2020.

Ouyang, L. et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv:2201.02177*, 2022.

Rusu, A. A. et al. Policy distillation. *arXiv:1511.06295*, 2015.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

Wang, Z. et al. Depth-induced emergence in self-supervised reinforcement learning. *NeurIPS*, 2025.

Wei, J. et al. Emergent abilities of large language models. *TMLR*, 2022.