Capability Transfer Has a Sharp Gate: Objective-Induced Phase Transitions in Distilling Self-Supervised RL Agents

Bruce Changlong Xu*, Jay Park, Vivek Buch
Stanford University
***Corresponding author**

Keywords: contrastive RL, knowledge distillation, phase transitions, capability transfer, big worlds, depth scaling

Summary

When an RL agent in a big world must inherit capability from a larger or longer-trained predecessor, how smooth is that transfer? We study this question in self-supervised contrastive RL (CRL), where a deep teacher is distilled into a shallow student that fits a deployment budget. Across two continuous-control environments and three independently trained teachers, the transfer is not smooth: a capacity-rich student learns almost nothing from partially trained teachers and jumps to near-teacher performance only after the teacher crosses a representation-quality threshold. We give a zero-cost behavioral diagnostic that predicts the threshold (AUC 0.895), and we show that varying the InfoNCE temperature shifts the threshold location by thirteen epochs, identifying contrastive discrimination difficulty as the underlying mechanism. The same comparison run on SAC produces no threshold: a depth-2 student inherits 96.6% of a depth-8 teacher's reward smoothly on the same environment. An equal-parameter control (score 29 alone vs. 374 with a teacher) shows that depth accelerates discovery rather than representation. We read this as evidence that auxiliary objectives shaping internal state can introduce discontinuities in capability transfer that are absent from the reward landscape, with practical consequences for compute-bounded teacher–student pipelines in big worlds.

Contribution(s)

1. A convergence gate in CRL distillation: across three independently trained teachers on two environments, students fail uniformly across partially trained teachers and then jump to near-teacher performance only after the teacher has crossed a representation-quality threshold.
   Context: In supervised distillation, early-stopped teachers can outperform fully trained ones (Cho & Hariharan, 2019). The CRL setting reverses this trend sharply.
2. A causal isolation of the gate's mechanism via a single-knob manipulation of the InfoNCE temperature, producing a thirteen-epoch monotonic shift in the gate location across three temperatures (n=3 seeds), and ruling out a strong behavioral predictor of the gate (Spearman $\rho=-0.77$, AUC 0.895) as its cause.
   Context: The behavioral predictor is itself useful as a teacher-readiness diagnostic. The causal claim concerns the underlying contrastive objective, not the predictor.
3. Evidence that the gate is objective-specific rather than a property of RL or of the environment: on matched environments, SAC distillation transfers smoothly (96.6% at $11\times$ fewer parameters, no threshold), and a CURL ablation shows the depth pattern can reverse under temporal InfoNCE.
   Context: SAC is distilled by behavioral cloning, while CRL uses policy distillation, reflecting the natural target of each objective; the contrast we draw is between the presence and absence of a threshold, not the absolute magnitudes.
4. A discussion of implications for big-world RL pipelines: compute-aware teacher checkpointing, equal-parameter capacity tests that separate representation from discovery, and risks of relying on contrastive self-supervision in capacity-bounded continual settings.
   Context: None.

Capability Transfer Has a Sharp Gate: Objective-Induced Phase Transitions in Distilling Self-Supervised RL Agents

Bruce Changlong Xu*, **Jay Park**, **Vivek Buch**
brucechanglongxu@cs.stanford.edu, jaypark@stanford.edu, vpbuch@stanford.edu

Stanford University

Abstract

An agent in a big world cannot afford to learn every capability from scratch, so capability must transfer between agents of different scales and training histories. We study how this transfer behaves in self-supervised contrastive RL (CRL), where a deep teacher is distilled into a shallow student that fits a deployment budget. We find that the transfer is governed by a sharp threshold—a *convergence gate*—where a capacity-rich student learns little from partially trained teachers and jumps to near-teacher performance only after the teacher has crossed a representation-quality threshold. We provide a zero-cost behavioral diagnostic that predicts the threshold (AUC 0.895), and we show that varying the InfoNCE temperature shifts the threshold location by thirteen epochs, identifying contrastive discrimination difficulty as a key mechanism. By contrast, the same comparison on SAC shows no threshold: a depth-2 student inherits 96.6% of its depth-8 teacher's reward smoothly. An equal-parameter control (score 29 alone vs. 374 with a teacher) indicates that depth accelerates discovery rather than representation. The results suggest that auxiliary objectives shaping internal state can introduce discontinuities in capability transfer that are absent from the reward landscape, which may be relevant for big-world RL pipelines that employ such objectives.

1 Introduction

Learning in a big world forces a concession: the agent's representational capacity is small relative to the environment, and the optimal policy and value function are out of reach of any computation the agent can actually run (Javed & Sutton, 2024; Lewandowski et al., 2025). Practical agents in this regime do not learn each capability from scratch. They inherit it from a larger model, a longer-trained predecessor, or a curriculum of progressively simpler tasks, and commit only the capacity-bounded artifact to deployment. A central question for big-world RL is therefore not just how to learn, but how capability transfers between agents of different scales and training histories.

We study this question in the cleanest available proxy: knowledge distillation (Hinton et al., 2015; Rusu et al., 2015) from a deep, fully trained self-supervised RL teacher into a shallow student that has the same parameter count as the deployment budget allows. The teacher is trained with contrastive RL (CRL; Eysenbach et al. 2022), representative of the broader class of self-supervised RL methods that have become common in long-horizon and goal-conditioned settings. The student trains with the same online RL loop plus an auxiliary policy distillation term from a frozen teacher.

The transfer turns out not to be smooth. Across a dense sweep of teacher checkpoints (Figure 1), student performance is nearly constant for the first eighty percent of the teacher's training, including checkpoints whose own evaluation scores are already high, and jumps sharply only once the teacher

is fully converged. We call this a convergence gate. It is robust across three independent teachers and predictable from a teacher-side behavioral statistic ($\rho > 0.7$, AUC 0.895), but causally dissociated from that predictor: artificially shaping an unconverged teacher's actions to mimic convergence does not cross the gate. Varying the InfoNCE temperature, by contrast, shifts the gate location by thirteen epochs (Section 4), implicating contrastive discrimination difficulty as a key factor.

Three observations make this phenomenon relevant to learning in big worlds. First, compute spent training a teacher that has not yet crossed the gate may be inefficiently used from the perspective of a downstream budget-constrained student, even if the teacher itself appears to be improving. For curricula and teacher–student pipelines, checkpoint selection based on teacher validation loss alone can therefore be misleading. Second, at matched parameter count ($\sim$1.65M), the shallow CRL student scores 29 alone and 374 with a teacher, suggesting the binding constraint is discovery compute rather than representational capacity. Third, the barrier is a property of the objective: on the same environments, SAC distillation transfers smoothly with no threshold. As self-supervised auxiliary losses become more common in big-world RL, auxiliary-objective-induced barriers like the convergence gate merit attention.

Following the workshop's guidance that a submission may study a single aspect of learning in big worlds under simplifying assumptions about others, we use fully observable continuous-control environments (Brax Ant Big Maze and Humanoid) so that capability transfer between agents can be measured cleanly. The agent-side bounds we impose, on parameter count, depth, and training compute, are genuine: the shallow student is the deployment artifact, and the teacher is the auxiliary scaffold amortized across deployments. In the big-worlds taxonomy of Javed & Sutton (2024), this corresponds to a setting where the environment is large enough that the deployable agent cannot learn its policy from raw reward within budget, and capability must be inherited from a separate, compute-rich training process. The gate phenomenon concerns the algorithm's behavior under these bounds and does not depend on the environment itself being small.

2 Background and Method

Contrastive RL (Eysenbach et al., 2022) parameterizes the goal-conditioned Q-function through 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; the actor $\pi_\theta(a|s, g)$ maximizes the critic. The actor–critic separation lets us distill policy and representation independently and ask which channel carries transferable capability. Wang et al. (2025) report that scaling the depth of CRL networks unlocks sharp, discontinuous performance gains: on Ant Big Maze, depth-4 agents reach about 61 of 1000 timesteps at goal while depth-64 agents reach 441 ($7.3\times$), with the critical jump occurring at roughly depth 8; on Humanoid, performance rises from 12.6 at depth 4 to 562 at depth 8 and 649 at depth 64 ($52\times$). Our question is whether the capability locked behind these depth thresholds is representable by a shallow agent that fits the deployment budget, or only discoverable by the deep one.

Students train with the standard CRL objective plus a policy distillation loss from a frozen teacher: $\mathcal{L} = \mathcal{L}_{\text{critic}} + \mathcal{L}_{\text{actor}} + \lambda_\pi \mathcal{L}^\pi_{\text{distill}}$, where $\mathcal{L}^\pi_{\text{distill}} = \mathbb{E}[\|\mu^S - \mu^T\|^2 + \|\log \sigma^S - \log \sigma^T\|^2]$ matches Gaussian policy parameters. Students explore independently, collect their own data, and learn from the InfoNCE objective; distillation provides auxiliary supervision. We evaluate on Brax Ant Big Maze (29-dimensional observation) and Humanoid (268-dimensional observation, 17 actuators), using residual MLPs with LayerNorm and Swish, 512 parallel environments, 50M steps, and $\lambda_\pi = 1.0$. Mechanistic results use 3–5 seeds and headline results use n=10. The total experimental budget is more than 350 runs and roughly 420 GPU-hours on AMD MI250X and NVIDIA H100.

3 The Convergence Gate

We distill from ten teacher checkpoints spanning epochs 5–50 into capacity-rich students (d=4, w=512). The pattern shown in Figure 1 and Table 1 is consistent across teacher seeds: student performance is flat for roughly the first eighty percent of teacher training and then jumps sharply.

The teacher's own evaluation score is a poor proxy for whether the gate has been crossed. The score at E38, for instance, drops by thirty percent relative to E34 due to evaluation noise, yet the student trained from E38 continues to improve as if nothing had happened.

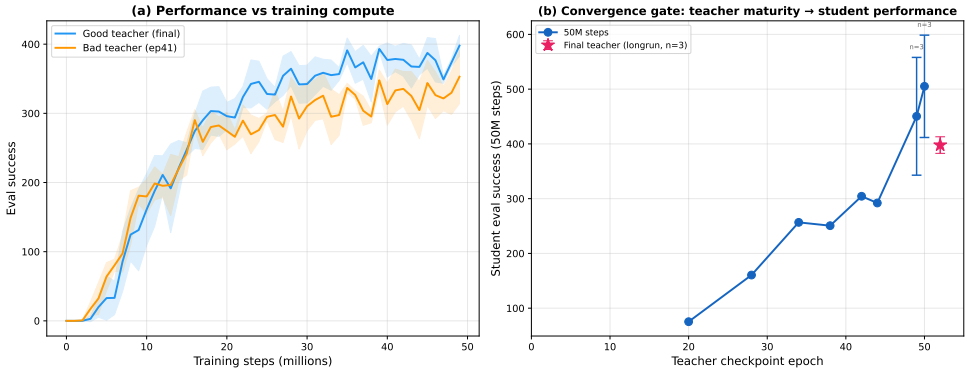

Figure 1: Student evaluation against training compute (left) and against teacher checkpoint (right). On the left, good-teacher students (blue) separate from bad-teacher students (orange) after roughly 10M steps, even though both converge to the same policy MSE. On the right, student performance is flat for about eighty percent of teacher training and then jumps. Error bars are ± 1 std over 3 seeds.

Table 1: Dense gate sweep (d=4, w=512, Ant Big Maze). Teacher evaluation is noisy (E38 drops to 200.2), but student performance tracks the teacher's training steps monotonically. Underlying policy structure, rather than instantaneous evaluation, determines distillation quality.

Ckpt	Step	Teacher eval	Student 20M	Student 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	—	450.4 ± 107.5
E50	49.7M	382.7	—	397.9 ± 15.5

The gate sits above a hard architectural floor: depth-2 networks score zero across 300 evaluations at 100M steps regardless of width or distillation, indicating a minimum representational requirement below which no amount of teacher signal helps. Above the floor, good- and bad-teacher students converge to the same policy MSE (about 0.09), but their evaluation scores diverge after roughly 10M environment steps. The student imitates both teachers equally well; the gap comes from what the imitated actions achieve, not from how closely they match the teacher. At the gate boundary itself, E49 produces bimodal outcomes (450.4 ± 107.5, with some seeds crossing and others failing), while E50 is consistent (397.9 ± 15.5), which is the kind of behavior one expects near a critical point.

The teacher's mean action magnitude turns out to predict student success well across 21 teacher–epoch conditions (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 onto a common threshold at $|\tanh(\mu)| = 0.63$ (Figure 2). For pipelines that cannot afford to evaluate every checkpoint through a full distillation run, this provides a cheap teacher-readiness test.

Predictiveness, however, does not imply causation. To test whether the behavioral signature itself is what matters, we artificially shape an unconverged teacher's outputs to mimic convergence by clamping the policy's log-standard-deviation to force deterministic actions and by scaling the policy means to reduce action magnitude, and then distill. Entropy-clamped students score zero across six

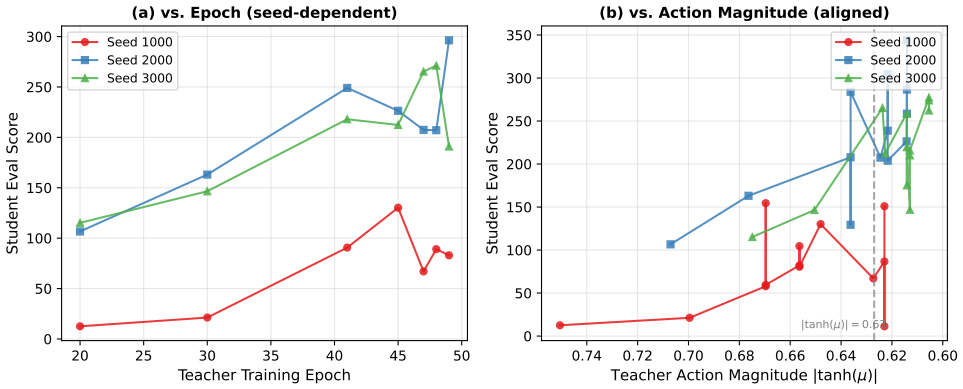

Figure 2: The gate is predictable but causally dissociated from its predictors. (a) Student evaluation against teacher epoch: the gate location varies by seed. (b) The same data plotted against teacher action magnitude: the three curves align near $|\tanh(\mu)| = 0.63$ (dashed).

runs and mean-scaled students reach only 54 ± 45, compared with zero for the unmodified teacher. Applying the same shaping to a converged teacher degrades its students from 383 to about 190. We conclude that the behavioral predictors are epiphenomenal: convergence produces both committed actions and high-quality representations, but only the latter is causally necessary.

4 Mechanism: Contrastive Discrimination Difficulty

If the gate arises from the contrastive objective, then modifying the difficulty of the contrastive task should shift it. We train three depth-8 teachers at InfoNCE temperatures $\tau \in \{0.5, 1.0, 2.0\}$ and distill from intermediate checkpoints into identical students (Figure 3). The gate shifts monotonically with temperature: $\tau = 2.0$ (easier discrimination) crosses at epoch 30, $\tau = 1.0$ at epoch 34, and $\tau = 0.5$ (harder discrimination) at epoch 43, a spread of thirteen epochs. The ordering is consistent across n=3 seeds. Because only the temperature varies between conditions (architecture, seed, and pipeline are held fixed), this identifies contrastive discrimination difficulty as the causal driver of the gate, and singles out a single hyperparameter that controls when a teacher's capability becomes transferable.

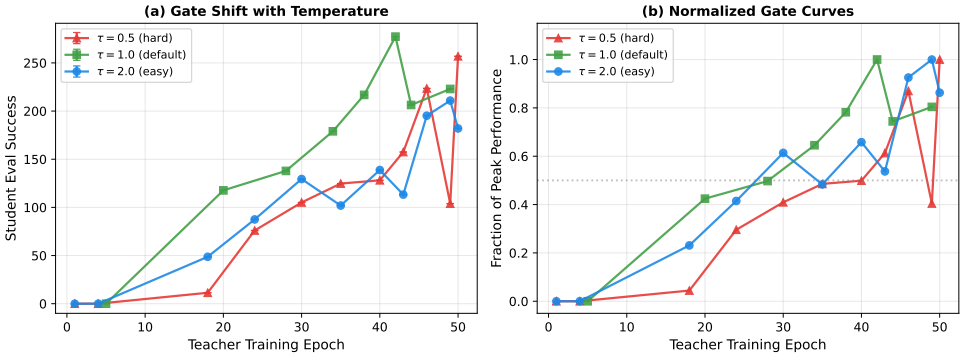

Figure 3: Contrastive difficulty causally controls the gate. (a) Student performance against teacher epoch at three InfoNCE temperatures: easier discrimination (τ=2.0, blue) opens the gate earlier, harder discrimination (τ=0.5, red) delays it, a 13-epoch effect. (b) Normalized curves isolate timing from absolute performance.

5 Capability Transfer: Representation vs. Discovery

Once the gate has been crossed, distillation is highly effective (Table 2). 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 lifts a near-zero baseline (6.4) to 383.9, ninety-two percent of the teacher's score.

Table 2: Distillation across the phase boundary (`eval_success`, 50M steps, mean $\pm$ std). Width-scaled students match teachers on both environments. Policy distillation is the only effective channel: combined policy and representation distillation gives 32.3, and representation-only distillation gives 5.0 (see text).

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 (d=4, 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 (d=4, w=512)	$\mathbf{383.2 \pm 17.7}$ *	$\mathbf{383.9 \pm 24.3}$

* n=10 seeds; all other conditions 3 seeds unless noted.

Policy distillation is the only effective transfer channel. Combined policy and representation distillation performs worse than the baseline on Ant Big Maze (32.3 vs. 48.2), and representation-only distillation is useless on Humanoid (5.0). What the teacher usefully transfers is what to do, not how to see.

The contrast between architecture and optimization is sharpest at matched parameter count. With $\sim$1.65M parameters in both, the shallow d=4, w=360 agent scores 29.2 alone but reaches 373.6 with distillation, while the deep d=8, w=256 agent scores about 295 on its own. The same parameters that score 29 in pure online learning therefore score 374 once a teacher provides a target. The capacity to support the converged policy is already present in the shallow agent; what is missing is the optimization signal needed to find it.

A related late-activation experiment supports the same reading. Students trained for 50M steps without distillation reach 318.7 ± 40.3 on Ant Big Maze when distillation is switched on late, compared with 268.9 for a fresh distilled run, and on Humanoid late activation produces a sixfold improvement over baseline (983.0 ± 489.8, three seeds). Prior self-supervised learning neither helps nor hurts. The depth barrier is an optimization problem that distillation bypasses regardless of the student's history.

To test whether the gate is specific to CRL, we train SAC (Haarnoja et al., 2018) at depths 2, 4, 8, and 16 on Humanoid (Table 3). At 10M steps the d=8 agent earns 2.15 times the reward of d=2 (p=0.012), but the advantage is gradual rather than threshold-shaped, and by 50M steps all depths converge (ANOVA F=1.05, p=0.42). The CRL behavior on the same benchmark is qualitatively different: at 100M steps, the depth-2 agent produces a score of exactly zero across 300 evaluations. We caution that SAC and CRL differ in more than the auxiliary objective (reward shape, replay use, and target-network usage all differ), so this contrast establishes that the gate is not a generic property of off-policy RL on Humanoid, rather than uniquely attributing it to InfoNCE; the temperature manipulation in Section 4 provides the within-CRL causal evidence.

Distillation from SAC to a much smaller SAC student is similarly smooth. We distill a depth-8 SAC teacher into a depth-2 student via behavioral cloning across three seeds. The student, which has eleven times fewer parameters than the teacher, reaches 96.6% of the teacher's reward, starting at 49% at epoch zero, rising to 87% by epoch five, and ultimately exceeding the teacher at peak in all three seeds (100.7% mean). There is no threshold and no failure mode, in direct contrast to CRL on the same environment.

Table 3: SAC depth sweep on Humanoid (3 seeds). Depth provides a $2.15\times$ advantage at 10M steps, but unlike in CRL the effect is gradual, and the gap closes by 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	$\mathbf{3008 \pm 157}$	—
2	50M	3908 ± 340	ANOVA p=0.42
8	50M	3414 ± 717	

A shared-encoder CURL experiment (Laskin et al., 2020) on Humanoid (n=5) shows that the depth pattern can also reverse: at depth 2 the temporal InfoNCE prediction is fourteen times better than at depth 8, and a three-level coefficient ablation gives a monotonic dose-response, with the depth-2-favoring gap growing by 36, 66, and 88 units of score as the contrastive strength rises from 0 to 0.1 to 1.0. Taken together, the SAC and CURL comparisons indicate that the depth–capability map is a property of the auxiliary objective coupling the agent to its environment, not of either in isolation.

6 Discussion

It is useful to separate two layers of depth effects in our results. The first is universal: depth accelerates optimization, and shallow networks can match deep ones given enough additional compute (SAC at 50M steps) or enough external guidance (SAC distillation at 96.6%). The second is objective-specific: in CRL, the goal-conditioned InfoNCE sharpens the smooth optimization advantage into a phase-transition-like gate, while temporal InfoNCE in CURL reverses the depth ordering altogether. For RL in big worlds, where bounded depth is forced on the deployment agent and capability must therefore be transferred from elsewhere, the objective-specific layer is what determines whether the transfer is smooth or gated.

A plausible explanation for the discontinuity is that InfoNCE optimizes alignment and uniformity on the hypersphere (Wang & Isola, 2020), and that when the encoder is too shallow, embeddings collapse to a low-dimensional subspace (effective rank $11.8/64$ at d=2, $17.6/64$ at d=8; Appendix Figure 4). The softmax over N negatives then saturates near $\log N$ and the actor receives almost no gradient, producing a cliff rather than a slope. Above a critical depth, embeddings spread (Jing et al., 2022), informative gradients return, and behavior emerges discontinuously. SAC's reward-based critic always provides some gradient signal regardless of depth, which would explain why its profile is smooth and its distillation gate-free.

We emphasize that the rank and uniformity numbers are observational correlates rather than direct causal mechanisms: they correlate with depth in our experiments but we do not intervene on representation geometry, so this explanation remains mechanistically hypothesized. The causal claim is limited to what the temperature manipulation establishes: contrastive discrimination difficulty causally controls gate location (Section 4). A future study that directly manipulates representation rank or uniformity independent of depth would provide stronger mechanistic evidence.

The gate has structural similarities to grokking (Power et al., 2022) and to emergent capabilities in language models (Wei et al., 2022): a disconnect between training loss and downstream capability, bimodal behavior at the critical point (E49), and sharp dependence on a latent variable rather than on any directly measurable surface statistic. The difference is that grokking is a transition in training time, while the convergence gate is a transition in teacher state. The student's own dynamics are smooth; what is discontinuous is the teacher's transferability. This is consistent with a broader hypothesis that sharp capability boundaries are a generic consequence of auxiliary objectives that shape representation geometry, rather than a peculiarity of CRL.

For big-world RL pipelines, four practical consequences follow. In compute-bounded curricula, the relevant quantity is not the teacher's instantaneous evaluation but whether it has crossed the gate, so intermediate checkpoints can usually be discarded. Before adding depth or width to fit a hard task, an equal-parameter distillation control can determine whether the bottleneck is discovery or capacity: if a shallow student matches the deep agent under distillation, the task is discovery-bounded. In continual settings, auxiliary objectives that gate capability transfer can also gate recovery after distribution shift, so methods that combat catastrophic forgetting via teacher–student replay may inherit the gate. Finally, because the gate is controlled by a single hyperparameter (InfoNCE τ), an online scheduler could in principle anneal contrastive difficulty during training to widen its basin, which we leave to future work.

We also note a partial scaffold dependence. Removing distillation from a competent student degrades its performance by about a third (from $\sim$404 to $\sim$274) but stabilizes well above the undistilled baseline (29–121). Gradual annealing of λ_π from 1.0 to 0.0 over ten epochs converges to the same $\sim$274 equilibrium as abrupt removal, which rules out a simple distribution-shift explanation and suggests instead that the teacher acts as a stabilizer that keeps the policy in a favorable basin. Recursive distillation, from a competent student into a fresh student, retains 91% of teacher performance (361.0 ± 24.5), indicating that the knowledge is genuinely internalized rather than dependent on the teacher's continued presence.

The primary limitations of our study are scope. We demonstrate the gate in CRL and contrast it with SAC and CURL on the same benchmarks, but we cannot rule out analogous phenomena in other untested methods, nor can we rule out that with longer training the CRL gate would dissolve in the same way the SAC depth gap closes by 50M steps; our 50M-step horizon was chosen to match the deployment-compute regime that motivates the question.

Our environments are fully observable continuous-control benchmarks (Ant Big Maze, Humanoid), which was a deliberate choice to measure capability transfer cleanly in a controlled setting. The gate mechanism depends on contrastive representation geometry rather than observability assumptions, so we expect it to persist in observable subproblems of partially observable domains, but this is not directly demonstrated. Future work should validate whether the gate appears in high-dimensional observations (vision) and whether it persists in exploration challenges typical of real big-world deployments.

The mechanistic story relies on a representation-geometry correlation rather than a direct intervention on rank or uniformity (we measure effective rank and uniformity but do not directly intervene on them). The scaffold-dependence finding suggests adaptive λ schedules and gate-aware annealing as natural next steps. Taken together, the results indicate that contrastive auxiliaries can introduce sharp barriers in capability transfer that are absent from the underlying reward landscape, and that a single hyperparameter, together with a cheap behavioral diagnostic, suffices to detect and control them.

References

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *International Conference on Computer Vision*, 2019.

Benjamin Eysenbach, Tianjun Zhang, Ruslan Salakhutdinov, and Sergey Levine. Contrastive learning as goal-conditioned reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, 2018.

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

Khurram Javed and Richard S. Sutton. The big world hypothesis and its ramifications for artificial intelligence. In *Finding the Frame Workshop, Reinforcement Learning Conference*, 2024.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, 2020.

Alex Lewandowski, Aditya A. Ramesh, Edan Meyer, Dale Schuurmans, and Marlos C. Machado. The world is bigger: A computationally-embedded perspective on the big world hypothesis. In *Advances in Neural Information Processing Systems*, 2025.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

Kevin Wang, Ishaan Javali, Michał Bortkiewicz, Tomasz Trzciński, and Benjamin Eysenbach. 1000 layer networks for self-supervised RL: Scaling depth can enable new goal-reaching capabilities. *arXiv preprint arXiv:2503.14858*, 2025.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 2020.

Jason Wei et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Supplementary Materials

The following content was not necessarily subject to peer review.

7 Key Supporting Results

Width sweep (d=4, policy distillation, 50M steps, Ant Big Maze): w=256: 80.4 ± 23.4; w=384: 365.0 ± 2.4; w=512: 383.2 ± 17.7 (n=10); w=768: 394.7 ± 11.6. Performance scales with width, with diminishing returns above w=384.

Equal-parameter control ($\sim$1.65M params each): d=4, w=360 baseline: 29.2 ± 45.5; d=4, w=360 + distillation: 373.6 ± 35.1; d=8, w=256 (teacher): $\sim$295.

Multi-teacher Humanoid (d=4, w=512, 50M steps): Teacher–student correlation r=0.992. Strongest teacher (eval 483.2) produces students at 498.2 ± 25.0, exceeding the teacher. Weakest teacher (eval 123.1) still provides $10\times$ boost over baseline.

Representation geometry at convergence (CRL encoders, Ant Big Maze):

Table 4: Effective rank, alignment, and uniformity vs. depth.

Depth	Eff. Rank (SA)	Alignment	Uniformity (G)
d=2	11.8 / 64	264.4	-0.483
d=4	14.8 / 64	136.4	-0.699
d=8	17.6 / 64	128.9	-0.822

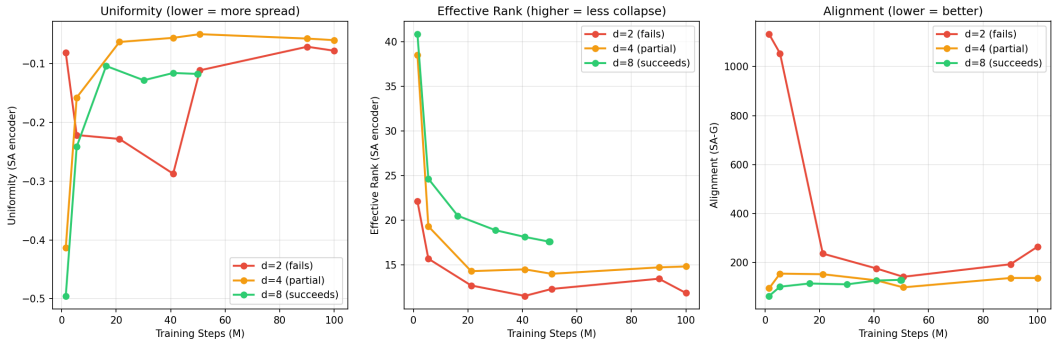

Figure 4: Representation geometry across depth (CRL, Ant Big Maze). Shallow encoders (d=2, red) show strong dimensional collapse, with low effective rank, poor alignment, and unstable uniformity. Deeper encoders (d=8, green) maintain higher-rank, better-aligned representations.

Late-ON distillation (d=4, w=256, 3 seeds): Ant Big Maze late-ON 318.7 ± 40.3 vs. fresh 268.9 ± 42.1; Humanoid late-ON 983.0 ± 489.8 vs. baseline 159.4 ± 28.3 ($6\times$).

Full hyperparameters. 50M env steps (50 epochs $\times$ 1M), 512 parallel envs, buffer 10K, batch 256, Adam lr 3e-4, γ=0.99, InfoNCE τ=1.0 (default), representation dim 64, residual MLP with LayerNorm + Swish, λ_π=1.0. Student seeds: 100/200/300; teacher seeds: 1000/2000/3000.