Convergence-Gated Distillation for Resource-Adaptive Reinforcement Learning Agents

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

Abstract

Adapting and compressing foundation models through knowledge distillation is typically assumed to be a monotone process: better teachers produce better students. We show this assumption fails sharply for self-supervised reinforcement learning. In contrastive RL (CRL), student performance is nearly constant across a wide range of partially trained teachers and jumps discontinuously only once the teacher crosses a *convergence gate*—a phase boundary in teacher representation quality. The gate is robust across three independent teachers, predictable from a zero-cost behavioral diagnostic (ρ>0.7, AUC= 0.895), but causally controlled by contrastive discrimination difficulty: varying the InfoNCE temperature shifts the gate by 13 epochs. Once crossed, distillation transfers a deep teacher into a shallow student that matches it on Ant Big Maze (p=0.93, n=10) and reaches 92% on Humanoid from a near-zero baseline. An equal-parameter control isolates the mechanism: depth accelerates *discovery*, not *representation* (29 $\rightarrow$ 374 with vs. without a teacher). The gate is objective-specific—SAC distillation transfers smoothly at 96.6% with no gate—suggesting that contrastive self-supervised objectives can introduce sharp adaptation barriers absent from reward-based learning. For practitioners adapting foundation models with auxiliary contrastive losses, this means naive checkpoint selection silently wastes compute, while a single behavioral statistic on the teacher predicts when distillation will pay off.

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

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

Knowledge distillation (Hinton et al., 2015) is one of the most widely used adaptation primitives for foundation models: compressing a large pretrained teacher into a smaller, deployable student that inherits its capabilities. The implicit assumption is monotonicity—a stronger teacher yields a stronger student, with performance scaling smoothly along the way. For supervised classification this assumption is broadly correct; recent work (Cho & Hariharan, 2019) even shows early-stopped teachers sometimes *outperform* fully trained ones. Foundation-model adaptation pipelines—LoRA fine-tuning (Hu et al., 2022), RLHF distillation (Ouyang et al., 2022), policy distillation (Rusu et al., 2015)—inherit this monotone intuition.

We show that the picture inverts under self-supervised reinforcement learning. In contrastive RL (CRL; Eysenbach et al. 2022), where the teacher is trained with an InfoNCE objective, 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 jumps discontinuously 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, adaptation 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* independently. SAC distillation (reward-based Haarnoja et al. 2018) on the same

architectures shows no gate at all—the student smoothly reaches 96.6% of the teacher—pinning the phenomenon to the contrastive objective.

Relevance to adaptive foundation models. Self-supervised contrastive losses are increasingly used as auxiliary signals when adapting foundation models—vision encoders, retrieval-augmented LLMs, embodied agents. 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.

Contributions.

1. **A convergence gate in CRL distillation.** A sharp phase 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. **Objective-specificity.** The gate is CRL-specific: SAC distillation transfers smoothly (96.6%, no gate); CURL exhibits a reverse depth pattern. The phenomenon is a property of the auxiliary objective, not RL in general.
4. **Discovery vs. representation.** Equal-parameter controls and late-activation experiments separate optimization from expressivity. Depth accelerates discovery, not representation ($29 \to 374$ with a teacher); shallow students can host capabilities they cannot find.

2. Background and Method

Contrastive RL. CRL (Eysenbach et al., 2022) 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 ~ 440; on Humanoid the transition is at depth 16. We use this depth gap as the substrate for adaptation.

Distillation. Students train with standard CRL plus a Gaussian 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 $\mathcal{L}^\pi_{\text{distill}} = \mathbb{E}[\|\mu^S - \mu^T\|^2 + \|\log\sigma^S - \log\sigma^T\|^2]$. Students still explore independently and learn from the InfoNCE objective; distillation is auxiliary supervision.

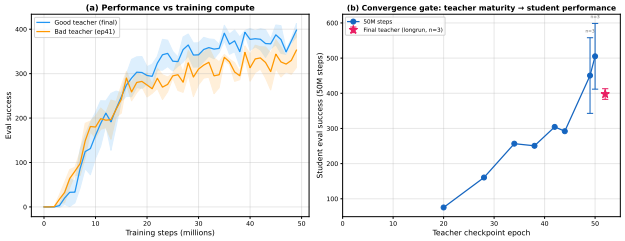

Figure 1. **The convergence gate.** *Left:* Student eval vs. training compute. Good-teacher students (blue) separate from bad-teacher students (orange) after $\sim$10M steps despite converging to the same policy MSE. *Right:* Student performance vs. teacher checkpoint: flat for $\sim 80\%$ of teacher 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, $\sim$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 $\sim 80\%$ 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

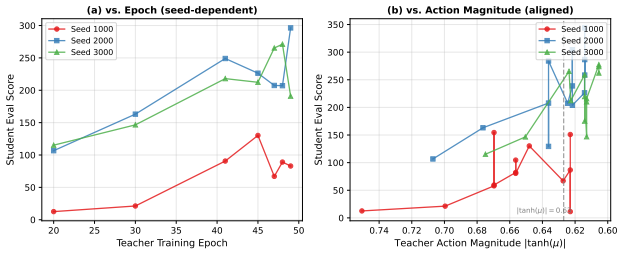

Figure 2. **The gate is predictable.** (a) Student eval vs. teacher epoch: gate location varies by seed. (b) Same data vs. teacher action magnitude: the three curves align near $|\tanh(\mu)| = 0.63$ (dashed).

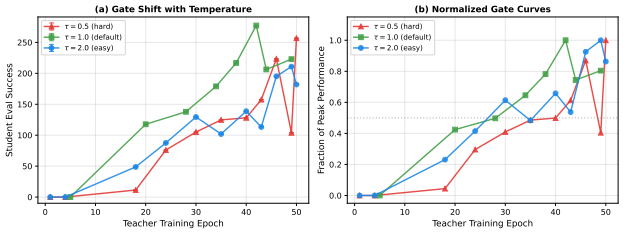

Figure 3. **Contrastive difficulty causally controls the gate.** (a) Student performance vs. teacher epoch at three InfoNCE temperatures. Easier discrimination (τ=2.0, blue) opens the gate earlier; harder (τ=0.5, red) delays it—a 13-epoch causal manipulation. (b) Normalized curves isolate timing from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

($\sim$0.09) but diverge in eval after $\sim$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), a hallmark of phase-transition criticality.

Predictability. 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)—a zero-cost diagnostic for teacher readiness.

Causal dissociation. 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 $\sim$190. 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 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).

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.

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.

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.

5. Main Results and Cross-Algorithm Comparisons

Distillation across the phase boundary. Table 2 shows that once the convergence gate is crossed, distillation 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). The teacher's value lies in *what to do*, not *how to see*.

Depth accelerates discovery, not representation. At identical parameter count ($\sim$1.65M; Table 3), d=4, w=360 scores 29.2 without distillation vs. $\sim$295 for d=8, w=256—a 10$\times$ 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$\times$ improvement that cannot be attributed to capacity.

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	—

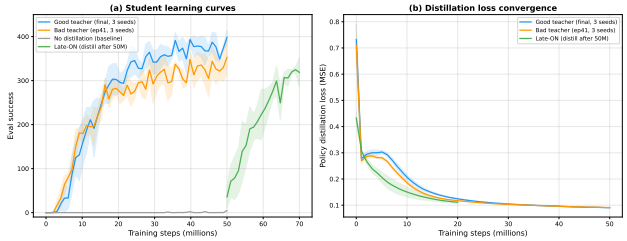

Figure 4. **Imitation succeeds; transfer fails.** *Left:* eval success diverges after ∼10M steps—good teacher (blue) climbs, bad teacher (orange) plateaus. *Right:* both students reach the same policy MSE floor (∼0.09). The downstream gap is in the utility of those actions, not their fidelity.

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, confirming the predictive structure rather than a specific coupling.

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 learning neither helps nor hinders—the depth barrier is an optimization problem that distillation bypasses regardless of the student's history.

Imitation succeeds; transfer fails. Figure 4 resolves a subtle point. 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*. A bad teacher's actions are well-formed in distribution but do not compose into successful trajectories; a good teacher's do. This dissociation is itself diagnostic for adaptive foundation-model pipelines: monitoring the distillation loss alone will not reveal a failed transfer.

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	

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: at 100M steps, d=2 produces 0.0 across all 300 evaluations.

SAC distillation: smooth transfer, no gate. 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.

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 $0 \to 0.1 \to 1.0$).

Emergent critic alignment. Policy distillation supervises only the actor, yet the student's critic also aligns with the teacher's (Table 5): 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 loss. 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 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

Table 5. Emergent critic alignment (Ant Big Maze). Despite no direct critic supervision, good-teacher students develop substantially more teacher-aligned representations.

Comparison	SA CKA	Goal CKA	Q corr.
Good teacher → good student	**0.630**	**0.927**	**0.880**
Bad teacher → bad student	0.608	0.901	0.839
Good teacher → baseline	0.436	0.750	0.686

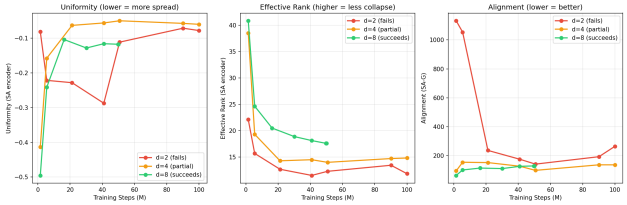

Figure 5. **Representation geometry across depth.** Shallow encoders (d=2, red) show strong dimensional collapse, poor alignment, and unstable uniformity; deeper encoders (d=8, green) maintain higher-rank, better-aligned embeddings, with d=4 (orange) intermediate. The same ordering predicts distillation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

(129). The CURL experiment corroborates this from the opposite direction: under *temporal* contrastive loss, d=8 collapses to $\sim$15/64 effective rank by mid-training while d=2 maintains $\sim$34/64—the same dimensional-collapse mechanism, now *disadvantaging* deeper networks because CRL's and CURL's contrastive objectives 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 teacher. 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 $\sim$274 equilibrium, ruling out distribution shift: the teacher acts as a *stabilizer* that keeps the student in a favorable basin, not as a continuous source of new information.

6. 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) Objective-specific:* CRL's goal-conditioned InfoNCE sharpens this smooth advantage into a phase-transition-like gate, while temporal InfoNCE (CURL) reverses it entirely.

Why does InfoNCE 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.

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, with some crossing the gate and others failing), 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 knowledge becomes transferable. This separates the locus of the phase transition (the source model) from where it manifests (the downstream adapter), which is a regime that has not, to our knowledge, been characterized in the foundation-model 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 (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 adaptive foundation models. Auxiliary contrastive losses are common in modern adaptation pipelines—vision–language alignment, retrieval encoders, embodied agents with intrinsic rewards. Our results imply two operational changes for any pipeline that distills or fine-tunes through such losses:

1. **Do not select teacher checkpoints by validation loss alone.** Loss can plateau well before the contrastive rep-

resentation 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 distillation success.

2. **Treat the contrastive objective as part of the adaptation interface.** The gate location is a function of the contrastive temperature; future adapters could expose this explicitly, scheduling discrimination difficulty to land the gate where the compute budget allows.

Practical takeaway. (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 teacher wastes compute.

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 methods (SPR, BYOL-based RL, non-contrastive self-supervised RL). (2) All experiments use Brax simulation (>350 runs for statistical rigor); transfer to higher-dimensional foundation-model adaptation 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 effect is partially understood: removing distillation degrades performance by $\sim 32\%$ but stabilizes well above the undistilled baseline, and recursive distillation retains 91%, suggesting the knowledge is genuinely internalized rather than parroted. (5) Our distillation coefficient is fixed (λ_π=1.0); adaptive schedules are an obvious next step for adaptive foundation-model pipelines.

Conclusion. Knowledge distillation under a self-supervised contrastive objective is not a smooth interpolation between teacher and student capabilities; it is a phase transition gated by the teacher's representation quality. The transition is causally controlled by contrastive discrimination difficulty, predictable from a single behavioral statistic, and absent under reward-based learning. For the broader adaptive-foundation-models program, the takeaway is that the choice of *which* teacher to distill from—and not just how—can be the dominant factor when contrastive losses are in the pipeline.

References

Cho, J. H. and Hariharan, B. On the efficacy of knowledge distillation. In *ICCV*, 2019.

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.

Hu, E. J. et al. LoRA: Low-rank adaptation of large language models. *ICLR*, 2022.

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.