Demystifying Emergent Exploration in Goal-conditioned RL

Mahsa Bastankhah[1*] **Grace Liu**[2*]
Dilip Arumugam[1] **Thomas L. Griffiths**[1] **Benjamin Eysenbach**[1]
[1]Princeton University [2]Carnegie Mellon University
mb6458@princeton.edu, gliu2@andrew.cmu.edu
[*]Equal contribution

Abstract

In this work, we take a first step toward uncovering the underlying dynamics of emergent exploration in unsupervised reinforcement learning. We study Single-Goal Contrastive RL (SGCRL) (Liu et al., 2024), which is capable of solving challenging robotic manipulation tasks without external rewards or curricula. Drawing on methods from cognitive science, we combine theoretical analysis of the algorithm's objective function with controlled experiments to improve understanding of its behavioral drivers. We show that SGCRL implicitly maximizes rewards shaped by its learned representations. The contrastive representations adapt the reward landscape to promote exploration prior to reaching the goal and exploitation thereafter. We also build a simple model of the algorithm without function approximation, isolating the essential components responsible for its exploratory behavior. Finally, we establish connections between SGCRL's exploration dynamics and classical exploration methods, including R-MAX and PSRL.

CODE: mahsa-bastankhah.github.io/demystifying-single-goal-exploration/

1 Introduction

Recent breakthroughs in deep reinforcement learning have revealed emergent behaviors: agents that develop complex skills without explicit rewards (Liu et al., 2024), learn to plan without world models (Bush et al., 2025; Simmons-Edler et al., 2025), and exhibit sophisticated exploration strategies in open-ended environments (Team et al., 2021). While these behaviors suggest promising pathways for scaling agent intelligence, their underlying mechanisms remain poorly understood. The fundamental challenge lies in answering two questions about emergent exploration: (1) What internal variables determine agent behavior? and (2) How do these variables evolve during training to drive exploration? This gap in understanding mirrors long-standing challenges in cognitive science, where researchers study intelligent behavior through theory development and controlled experimentation. Despite the rich toolkit available from cognitive science – including rational analysis, intervention experiments, and cognitive modeling – these approaches remain underutilized in deep RL research.

In this paper, we bridge this gap by characterizing emergent exploration in a way that draws on methods from cognitive science. We apply these principles to analyze Single-Goal Contrastive RL (SGCRL), a self-supervised, goal-reaching algorithm that learns without rewards, demonstrations, or subgoals. These principles include: (1) theoretically analyzing the agent's optimization objective to uncover implicit behavior drivers (Anderson, 1990), (2) conducting controlled intervention experiments on these behavioral drivers, and (3) building a simple model of the exploration mechanism (McClelland, 2009) in a tabular setting. We also ground our findings through comparison with the principled, transparent exploration methods R-MAX and PSRL.

We identify a fundamental interplay between the actor and critic in SGCRL that drives emergent exploration. The actor's objective can be reinterpreted as maximizing a discounted sum of dense rewards that favors states representationally similar to the goal. The critic shapes this implicit reward landscape by decreasing the representational goal similarity of states along unsuccessful trajectories, pruning them from future exploration.

2 Preliminaries

Problem Setup We consider a controlled Markov process (a Markov Decision Process without an explicit reward function) with states s_t and actions a_t. The initial state is sampled $s_0 \sim p_0(s_0)$ and transitions follow $s_{t+1} \sim p(s_{t+1} \mid s_t, a_t)$. The agent is given a single target goal g and must learn a policy $\pi(a \mid s, g)$ by interacting with the environment. Like (Liu et al., 2024). we do not assume any distribution over subgoals nor a predefined reward function.

Single-Goal Contrastive RL (SGCRL). We study SGCRL (Liu et al., 2024), an actor–critic algorithm based on temporal contrastive learning. The critic estimates the likelihood that a state–action pair (s, a) leads to a future state s_f, and is parameterized as $C(s, a, s_f) = \phi(s, a)^\top \psi(s_f)$, where $\phi(s, a)$ and $\psi(s_f)$ are learned embeddings. The critic embeddings are trained with a contrastive backward InfoNCE loss:

$$\mathcal{L}_{\text{critic}}^{\text{backward}} = -\mathbb{E}_{\substack{(s_i,a_i)\sim p(s,a) \\ s_{f,i}\sim p_\gamma^\pi(\cdot|s_i,a_i) \\ i=1,\dots,N}} \left[\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\big(\phi(s_i, a_i)^\top \psi(s_{f,i})\big)}{\sum_{j=1}^{N} \exp\big(\phi(s_j, a_j)^\top \psi(s_{f,i})\big)} \right],$$

where $p_\gamma^\pi(s_f)$ is the discounted state occupancy measure as defined in (Eysenbach et al., 2022).

For each state-action pair (s_t, a_t), positive future states are drawn by looking $\Delta \sim \text{Geom}(1-\gamma)$ steps ahead. Negatives examples are sampled from the marginal distribution $p(s_f)$. We normalize all representations by their ℓ_2 norm. Once trained, the critic encodes a log-Q value, $\phi(s, a)^\top \psi(s_f) = \log p_\gamma^\pi(s_f|s, a) - \log p(s_f)$. The actor is responsible for selecting actions that maximize the likelihood of reaching the goal. Formally, the actor's objective is defined as

$$\max_{\pi(a|s,g)} \; \mathbb{E}_{s\sim p(s),\, a\sim\pi(\cdot|s,g)} \Big[\phi(s, a)^\top \psi(g) + \tau H\big(\pi(\cdot \mid s, g)\big) \Big], \tag{1}$$

where τ is an entropy regularization coefficient.

3 SGCRL representations create an implicit curriculum of rewards

In this section, we theoretically characterize the dynamics that drive exploration in SGCRL. We propose that the actor maximizes a discounted sum of implicit rewards shaped by the critic. Even when the algorithm poorly estimates the probability of reaching the goal, the implicit reward function is well-defined and directs exploration. As the actor optimizes this implicit reward signal, the critic dynamically reshapes the reward landscape. Before the goal is found, the critic decreases the implicit reward for states along unsuccessful trajectories.

3.1 The actor maximizes an implicit, goal-based reward

Our analysis reveals that, although the SGCRL objective is defined with respect to reaching g, it is simultaneously driving the agent toward any state s_f that has high representational similarity to the goal, where similarity is measured by the inner product ψ-*similarity* $= \psi(s_f)^\top \psi(g)$.

Theorem 1 (Maximizing discounted sum of ψ-similarity)**.** *The SGCRL actor objective is equivalent to maximizing the entropy-regularized return*

$$Q(s, a) \;=\; \mathbb{E}_{p_t^\pi(s_t)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \;\Big|\; s_0 = s,\; a_0 = a \right],$$

i.e.

$$\max_{\pi} \; \mathbb{E}_{s\sim p(s),\, a\sim\pi(\cdot|s)} \big[\, Q(s, a) + \tau\, H\big(\pi(\cdot \mid s)\big) \,\big].$$

Where the shaped reward is $r(s, a) \coloneqq \psi(s)^\top \psi(g)$.

Theorem 1 shows that, although SGCRL is an unsupervised algorithm and does not use any external reward from the environment, it implicitly maximizes an internal reward (i.e., ψ-similarity). Refer to Appendix B.2 and B.3 for the proof.

3.2 The critic dynamically shapes implicit rewards to drive exploration

As the actor aims to maximize ψ-similarity, the critic representations update dynamically to adaptively shape this implicit reward during training. We show that contrastive updates reduce the ψ-similarity of frequently visited states along unsuccessful trajectories, driving exploration towards unvisited states.

Theorem 2 (Informal)**.** *Let $\psi(g) \in \mathbb{R}^d$ be a fixed unit vector, with a large dimension, Consider normalized anchor embeddings $\{\phi(s_i, a_i)\}_{i=1}^N$ and future embeddings $\{\psi(s_{f,i})\}_{i=1}^N$, where $s_{f,i}$ denotes the positive example of (s_i, a_i). Assume that the goal $g \notin \{s_{f,i}\}_{i=1}^N$ and that the embeddings are initialized as*

$$\phi^{(0)}(s_i, a_i) \;=\; c\,\psi(g) + \zeta_i, \qquad \psi^{(0)}(s_{f,i}) \;=\; c\,\psi(g) + \kappa_i,$$

where $\zeta_i, \kappa_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\Big(0, \frac{1-c^2}{d}\, I_d\Big)$, and $c < 1$ is a non-zero scalar.

Suppose these embeddings are updated using the InfoNCE gradient descent update rule with a sufficiently large batch size N and sufficiently small learning rate η. Then, with high probability over the random initialization, the system converges to an equilibrium that satisfies:

$$\phi(s_i, a_i)^\top \psi(g) \;=\; \psi(s_{f,i})^\top \psi(g) \;=\; 0 \qquad \textit{for all } i.$$

Theorem 2 shows that if states in a region initially exhibit consistently high ψ-similarity, then repeated InfoNCE updates will progressively reduce their similarity to $\psi(g)$ until they become orthogonal, rendering the region unattractive to the actor. Refer to Appendix B.4 for a formal statement and proof of the theorem.

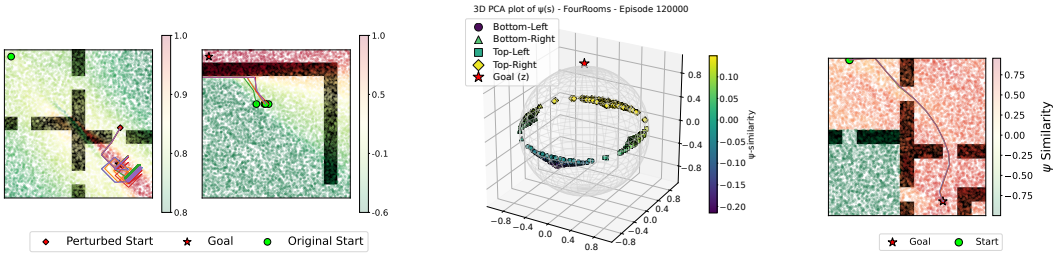
4 Experiments

In this section, we present empirical evidence that the SGCRL actor visits regions of high ψ-similarity and the ψ-similarity decreases for states along unsuccessful trajectories. We study SGCRL on 2D point maze navigation tasks adapted from prior work (Eysenbach et al., 2022; Liu et al., 2024) as well as the Tower of Hanoi goal-reaching task. All training metric curves are averaged over 8 random seeds.

4.1 Verifying behavioral mechanisms with intervention experiments

Agent targets reachable, goal-like states. According to Theorem 1, the agent aims to maximize an implicit reward determined by representational similarity to the goal. To test this theory, we conducted an experiment in which we perturbed the initial position of the agent and mapped its trajectory at training checkpoints. We observed that the agent navigates toward the closest region with high representational goal similarity, even when it could reach the goal directly through a shorter path (Fig. 1a, left side). We can also utilize representational goal similarity to steer agent behavior. We find that fixing the embeddings of the bottom-left room of the FourRooms maze to $-\psi(g)$ forces the agent to reach the goal via the top-right room (Fig. 1c).

Goal similarity decreases along frequently visited trajectories. Theorem 2 proposes that states visited along unsuccessful trajectories should become less representationally similar to the goal, creating an incentive for the agent to target more infrequently visited states. To observe how representations change in a controlled setting, we conducted experiments where we fixed the data collection trajectories to always move in a particular cardinal direction. We find that initially the ψ-similarity of all states is high, but over the course of training states along the frequently traversed paths systematically become less representationally similar to the goal (see Fig 1a, right side). (Fig. 1a, right).



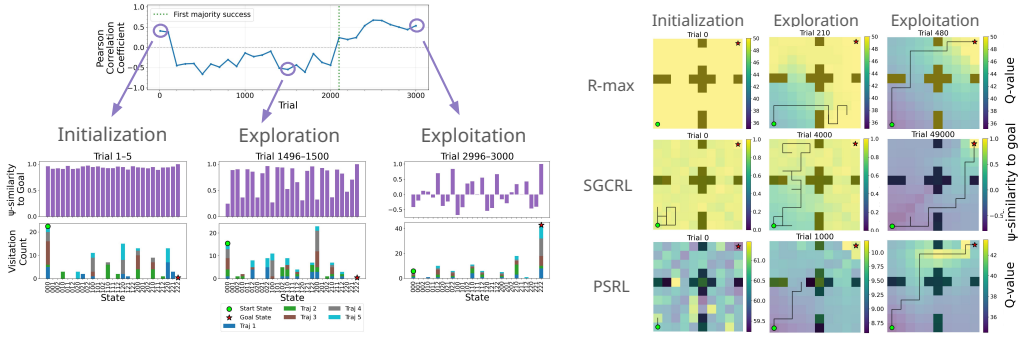
(a) (b) (c)

Figure 1: (a) Intervention experiments (b) Imaginary goal experiment (c) Safety experiment

4.2 Reproducing a simple computation model of SGCRL

To study the key components of SGRCRL in a simplified setting, we design a simple tabular version of the algorithm without function approximation. In this simplified setting, each state s has an embedding $\psi(s)$ stored in a lookup table and updated with InfoNCE gradient descent update rule iteratively. Actions are chosen by $a_t \sim \frac{1}{Z(s_t)} \exp\left(\frac{1}{\tau}\psi(m(s_t, a_t))^\top \psi(g)\right)$, where m is the dynamics model $s_{t+1} \leftarrow m(s_t, a_t)$ which is given to the agent, and $Z(s_t)$ is the normalizer.

Even in this simplified form, SGCRL demonstrates effective exploration and goal reaching, confirming that its exploratory mechanism is algorithmic rather than architectural. Consistent with our theory, training unfolds in three reproducible phases: *initialization* (all states show high ψ-similarity, producing uniform incentives), *exploration* (goal similarity of visited states decreases, pushing the agent to unvisited regions), and *exploitation* (once the goal is reached, states along successful trajectories increase in ψ-similarity, reinforcing the solution) (Fig. 2a). Moreover, state visitation and goal similarity are negatively correlated before reaching the goal and positively correlated afterward. To further test whether representations become orthogonal to the goal for states along unsuccessful trajectories, we conduct an experiment in the FourRooms environment with an imaginary goal representation $\psi(g)$. Initially, all state embeddings align closely with $\psi(g)$, but over training their similarity to $\psi(g)$ diminishes and they drift toward the subspace orthogonal to it (the z-axis)(Fig. 1b).



(a) (b)

Figure 2: (a) phases of SGCRL training (b) Comparison to PSRL and R-MAX

Tabular single goal exploration falls within a class of classical exploration algorithms Classical methods like R-MAX (Brafman and Tennenholtz, 2002) and PSRL (Strens, 2000) share a common principle: they begin with many candidate high-value states and progressively winnow this set until only the true goal remains. R-MAX does so via optimistic value estimates, while PSRL uses epistemic uncertainty (Der Kiureghian and Ditlevsen, 2009). SGCRL follows a similar pattern: goal-conditioned representations initially make many states appear promising, and exploration progressively eliminates them until the goal is found (Fig. 2b), mirroring the optimism-driven behavior of R-MAX.

5 Conclusion

Through theoretical analysis and controlled experiments, we show that SGCRL implicitly maximizes rewards based on representational goal-similarity, enabling effective exploration without explicit rewards. Beyond analyzing SGCRL, we propose general principles for understanding emergent exploration. We draw on methods in cognitive science to build a theoretical model of behavior, conduct intervention experiments, and study the algorithm in a simple setting. In future work, we aim to study whether SGCRL can be proved to obtain an ε-optimal greedy policy in polynomial time in the tabular setting and extend the method empirically to tasks beyond goal-reaching.

Acknowledgments. Thanks to the members of the Princeton RL lab for feedback on preliminary versions of the work. Thanks to Phil Bachman for insightful discussions. DA and TG were supported by ONR MURI N00014-24-1-2748 and ONR grant N00014-23-1-2510. GL acknowledges support by the National Science Foundation Graduate Research Fellowship under Grant No. DGE2140739. BE and MB acknowledge support from the National Science Foundation under Award No. 2441665. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Psychology Press.

Bewley, T. and Lawry, J. (2021). Tripletree: A versatile interpretable representation of black box agents and their environments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11415–11422.

Beyret, B., Shafti, A., and Faisal, A. A. (2019). Dot-to-dot: Explainable hierarchical reinforcement learning for robotic manipulation. In *2019 IEEE/RSJ International Conference on intelligent robots and systems (IROS)*, pages 5014–5019. IEEE.

Brafman, R. I. and Tennenholtz, M. (2002). R-MAX– A General Polynomial Time Algorithm for Near-Optimal Reinforcement Learning. *Journal of Machine Learning Research*, 3(Oct):213–231.

Bush, T., Chung, S., Anwar, U., Garriga-Alonso, A., and Krueger, D. (2025). Interpreting emergent planning in model-free reinforcement learning. *arXiv preprint arXiv:2504.01871*.

Cideron, G., Seurin, M., Strub, F., and Pietquin, O. (2020). Higher: Improving instruction following with hindsight generation for experience replay. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 225–232. IEEE.

Coppens, Y., Efthymiadis, K., Lenaerts, T., Nowé, A., Miller, T., Weber, R., and Magazzeni, D. (2019). Distilling deep reinforcement learning policies in soft decision trees. In *Proceedings of the IJCAI 2019 workshop on explainable artificial intelligence*, pages 1–6.

Der Kiureghian, A. and Ditlevsen, O. (2009). Aleatory or Epistemic? Does it Matter? *Structural Safety*, 31(2):105–112.

Eysenbach, B., Zhang, T., Levine, S., and Salakhutdinov, R. R. (2022). Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620.

Glanois, C., Weng, P., Zimmer, M., Li, D., Yang, T., Hao, J., and Liu, W. (2024). A survey on interpretable reinforcement learning. *Machine Learning*, 113(8):5847–5890.

Greydanus, S., Koul, A., Dodge, J., and Fern, A. (2018). Visualizing and understanding atari agents. In *International conference on machine learning*, pages 1792–1801. PMLR.

Heuillet, A., Couthouis, F., and Díaz-Rodríguez, N. (2021). Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685.

Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., and Doshi-Velez, F. (2019). Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on explainable artificial intelligence*.

Ku, A., Campbell, D., Bai, X., Geng, J., Liu, R., Marjieh, R., McCoy, R. T., Nam, A., Sucholutsky, I., Veselovsky, V., et al. (2025). Using the tools of cognitive science to understand large language models at different levels of analysis. *arXiv preprint arXiv:2503.13401*.

Lesort, T., Díaz-Rodríguez, N., Goudou, J.-F., and Filliat, D. (2018). State representation learning for control: An overview. *Neural Networks*, 108:379–392.

Lesort, T., Seurin, M., Li, X., Díaz-Rodríguez, N., and Filliat, D. (2019). Deep unsupervised state representation learning with robotic priors: a robustness analysis. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Liu, G., Tang, M., and Eysenbach, B. (2024). A single goal is all you need: Skills and exploration emerge from contrastive rl without rewards, demonstrations, or subgoals. *arXiv preprint arXiv:2408.05804*.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, 1(1):11–38.

Onken, L. S. (2015). Cognitive training: targeting cognitive processes in the development of behavioral interventions. *Clinical Psychological Science*, 3(1):39–44.

Raffin, A., Hill, A., Traoré, R., Lesort, T., Díaz-Rodríguez, N., and Filliat, D. (2019). Decoupling feature extraction from policy learning: assessing benefits of state representation learning in goal based robotics. *arXiv preprint arXiv:1901.08651*.

Schooler, L. J. and Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology*, 32(3):219–250.

Sequeira, P. and Gervasio, M. (2020). Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations. *Artificial Intelligence*, 288:103367.

Simmons-Edler, R., Badman, R. P., Berg, F. B., Chua, R., Vastola, J. J., Lunger, J., Qian, W., and Rajan, K. (2025). Deep rl needs deep behavior analysis: Exploring implicit planning by model-free agents in open-ended environments. *arXiv preprint arXiv:2506.06981*.

Steyvers, M. and Griffiths, T. L. (2008). Rational analysis as a link between human memory and information retrieval. *The probabilistic mind: Prospects for Bayesian cognitive science*, pages 329–349.

Strens, M. J. (2000). A Bayesian Framework for Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 943–950.

Team, O. E. L., Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., et al. (2021). Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., and Levine, S. (2020). Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR.

Zahavy, T., Ben-Zrihem, N., and Mannor, S. (2016). Graying the black box: Understanding dqns. In *International conference on machine learning*, pages 1899–1908. PMLR.

Appendix

A Related Work

The study of emergent behaviors in deep RL has gained attention as agents demonstrate increasingly sophisticated skills without explicit programming (Liu et al., 2024; Bush et al., 2025; Simmons-Edler et al., 2025; Team et al., 2021). However, despite this growing attention, the field lacks systematic approaches to understand the mechanisms underlying emergent behaviors.

Traditional approaches to understanding deep RL agents aim to improve algorithm transparency or explain behavior post-hoc (Heuillet et al., 2021; Glanois et al., 2024). Transparency-based methods include learning low-dimensional, meaningful representations of the state space (Lesort et al., 2018, 2019; Raffin et al., 2019), labeling actions based on targeted reward components (Juozapaitis et al., 2019), and maintaining subgoals with hierarchical RL (Beyret et al., 2019; Cideron et al., 2020). Although transparency methods improve algorithmic understanding, they require specific system architectures or auxiliary training tasks. In contrast, post-hoc methods assume that the algorithm is a black-box. Examples include building T-SNE or saliency maps of representations (Greydanus et al., 2018; Zahavy et al., 2016), analyzing interaction data (Sequeira and Gervasio, 2020), and distilling policies into decision trees (Bewley and Lawry, 2021; Coppens et al., 2019). Although these black-box methods can improve understanding of agent behavior, they do not explain behavior from an algorithmic level. To address these limitations, we propose an approach grounded in cognitive interpretability that improves algorithmic understanding of an RL system without the overhead of auxiliary training tasks.

Cognitive science has developed rich methodologies for understanding intelligent behavior through controlled experimentation and modeling. Example methods include modeling human memory as an optimal solution for information retrieval (Anderson, 1990; Schooler and Anderson, 1997; Steyvers and Griffiths, 2008), identifying mechanisms underlying behavior using cognitive interventions (Onken, 2015), and building simple cognitive models as a tool for understanding (McClelland, 2009). Recent work has proposed improving understanding of LLMs using tools from cognitive science (Ku et al., 2025). However, these approaches remain underutilized understanding the behavior of deep RL agents. We address this gap by presenting and instantiating a framework for cognitive interpretability of deep RL agents. We do so by (1) theoretically analyzing the agent's optimization objective to uncover implicit behavior drivers (Anderson, 1990), (2) conducting controlled intervention experiments on these behavioral drivers, and (3) building a simple model of the exploration mechanism in a tabular setting (McClelland, 2009).

B Theoretical results

B.1 Gradient descent updates for optimizing the InfoNCE objective

We derive the gradients of the backward InfoNCE loss with respect to the different representation parameters in order to characterize their update dynamics. Consider the batch consisting of $\{s_i, a_i, sf_i\}_N$ where sf_i is the positive example for state action pair s_i, a_i.

$$L = -\sum_i \log \frac{\exp\left(\phi(s_i, a_i)^\top \psi(sf_i)\right)}{\sum_k \exp\left(\phi(s_k, a_k)^\top \psi(sf_i)\right)}$$

$$p_{ij} \coloneqq \frac{\exp\left(\phi(s_i)^\top \psi(sf_j)\right)}{\sum_k \exp\left(\phi(s_k, a_k)^\top \psi(sf_j)\right)} \tag{2}$$

Note that $\sum_i p_{ij} = 1$ but $\sum_j p_{ij} \neq 1$

$$\nabla_{\phi(s_i,a_i)} L = -\sum_j (\delta_{i,j} - p_{ij}) \cdot \psi(sf_j) \tag{3}$$

$$\nabla_{\psi(sf_j)} = -\sum_i (\delta_{i,j} - p_{ij}) \cdot \phi(s_i, a_i) \tag{4}$$

$$\phi^{(t)}(s_i, a_i) = \phi^{(t-1)}(s_i, a_i) + \eta \sum_j (\delta_{i,j} - p_{ij}) \cdot \psi^{(t-1)}(sf_j) \tag{5}$$

$$\psi^{(t)}(sf_j) = \psi^{(t-1)}(sf_j) + \eta \sum_i (\delta_{i,j} - p_{ij}) \cdot \phi^{(t-1)}(s_i, a_i) \tag{6}$$

where η is the learning rate and $\delta_{i,j} \coloneqq \mathbf{1}[i = j]$. The update rules for the forward loss are analogous; the only difference is that, in the denominator of p_{ij}, the summation is taken over $\psi(sf_k)$.

If the representations are meant to be unit norm, after every update, we should normalize them too.

Equilibrium.

Definition 1 (Equilibrium)**.** *The equilibrium of the system characterized by Equations 5 and 6 is a configuration in which the updates cease to alter any of the representations.*

In the non-normalized *case, this corresponds to vanishing gradients:*

$$\nabla_{\phi(s_i,a_i)} L \;=\; \nabla_{\psi(sf_i)} L \;=\; 0 \quad \forall i.$$

In the normalized *case, equilibrium arises either when the gradients vanish, or when they are parallel to the representations themselves, i.e.,*

$$\nabla_{\phi(s_i,a_i)} L \;=\; c_i' \, \phi(s_i, a_i), \qquad \nabla_{\psi(sf_i)} L \;=\; c_i \, \psi(sf_i), \quad \forall i,$$

for some scalars c_i, c_i'. In this case, normalization preserves the representation since scaling by a constant leaves the direction unchanged:

$$\frac{(c_i' + 1)\phi(s_i, a_i)}{\|(c_i' + 1)\phi(s_i, a_i)\|} \;=\; \phi(s_i, a_i),$$

and similarly for $\psi(sf_i)$.

B.2 Formal version of Lemma ?? and proof

First, we note that under our sampling procedure, any state s_f with positive discounted occupancy measure, i.e. $p_\gamma^\pi(s_f \mid s, a) > 0$, will necessarily be used as a positive example for the state–action pair (s, a). Given this observation, we now proceed to the formal version of the lemma and its proof.

Lemma 1 (Formal version of Lemma **??**)**.** *Let $N \gg 1$ be the batch size and $d \gg 1$ be the representation dimension. Consider anchors (s_i, a_i) and their corresponding positive states sf_i, $i = 1, \ldots, N$, with initializations*

$$\phi^0(s_i, a_i) = \zeta_i, \quad \psi^0(sf_i) = \kappa_i,$$

where $\zeta_i^0, \kappa_i^0 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0, \frac{1}{d} I_d\big)$. Suppose the representations are updated using gradient descent update rule of either the backward or forward InfoNCE loss (with normalization), with update rules as characterized in Appendix B.1. Then, with high probability over the initialization, the following holds at equilibrium:

$$\phi(s_i, a_i) = \psi(sf_i), \quad \forall i.$$

Moreover, if each anchor (s_i, a_i) has more than one positive example $\{sf_i^k\}_k$, then at equilibrium

$$\phi(s_i, a_i) = \psi(sf_i^k), \quad \forall i, k,$$

where k indexes the positive examples.

Proof. If each anchor (s_i, a_i) has only one positive example sf_i, then this lemma is a direct corollary of Theorem 3 with $c = 0$. In particular, Claim 2 of Theorem 3 completes the proof in this case.

The more realistic setting, however, is when each anchor is associated with multiple positive examples. To handle this, we slightly adapt the proof of Theorem 3 to again establish full alignment. We use the same notation as in that proof and restrict attention to the case where each anchor has $K = 2$ positive examples. The extension to any $K > 1$ follows identically.

Concretely, we assume a batch of size N, with anchor representations denoted $\{\mathbf{u}_i\}_{i=1}^N$ (corresponding to $\phi(s_i, a_i)$) and positive representations $\{\mathbf{v}_i\}_{i=1}^N$ (corresponding to $\psi(sf_i)$). Each anchor $\mathbf{u}_i$ is duplicated, i.e., $\mathbf{u}_{2i} = \mathbf{u}_{2i+1}$, while its two positive examples $\mathbf{v}_{2i}, \mathbf{v}_{2i+1}$ are independent. In other words, each $\mathbf{u}_{2i}$ has two distinct positive examples.

We establish the following claims by induction, grouping indices $(2i, 2i+1)$ into bundles:

(a) $\alpha^{(t)} \coloneqq \langle \mathbf{v}_{2i}^{(t)}, \mathbf{u}_{2i}^{(t)} \rangle = \langle \mathbf{v}_{2i+1}^{(t)}, \mathbf{u}_{2i}^{(t)} \rangle$ does not depend on the choice of i, and $\alpha^t > 0, \forall t \geq 1$, and $\alpha^{(\infty)} = 1$.

(b) $\beta^{(t)} \coloneqq \langle \mathbf{v}_{2i}^{(t)}, \mathbf{v}_{2i+1}^{(t)} \rangle$ does not depend on the choice of i, and $\beta^t > 0, \forall t \geq 2$, and $\beta^{(\infty)} = 1$.

(c) $\lambda^{(t)} \coloneqq \langle \mathbf{v}_i^{(t)}, \mathbf{v}_j^{(t)} \rangle = \langle \mathbf{u}_i^{(t)}, \mathbf{v}_j^{(t)} \rangle = \langle \mathbf{u}_i^{(t)}, \mathbf{u}_j^{(t)} \rangle = 0$ whenever $\lfloor i/2 \rfloor \neq \lfloor j/2 \rfloor$; that is, cross-inner products are zero across bundles.

These properties hold at initialization ($t = 0$), since independent Gaussian vectors are almost surely orthogonal in high dimensions. We show that the update dynamics then preserve these index invariant properties We simplify the probability matrix as defined in Equation 2 using the induction assumptions at step $t-1$:

$$p \coloneqq \frac{\exp\langle \mathbf{u}_i^{(t-1)}, \mathbf{v}_j^{(t-1)} \rangle}{\sum_k \exp\langle \mathbf{u}_k^{(t-1)}, \mathbf{v}_j^{(t-1)} \rangle}, \quad \text{if } \lfloor i/2 \rfloor = \lfloor j/2 \rfloor,$$

and

$$q \coloneqq \frac{\exp\langle \mathbf{u}_i^{(t-1)}, \mathbf{v}_j^{(t-1)} \rangle}{\sum_k \exp\langle \mathbf{u}_k^{(t-1)}, \mathbf{v}_j^{(t-1)} \rangle}, \quad \text{if } \lfloor i/2 \rfloor \neq \lfloor j/2 \rfloor.$$

These satisfy the normalization condition $2p + (N-2)q = 1.$, and since N is large, q is small.

Now we write the GD update rule noting that each anchor $\mathbf{u}_i$ receives two gradient updates per iteration (corresponding to its two positives):

$$\begin{aligned}
\hat{\mathbf{u}}_{2i+1}^{(t)} &= \hat{\mathbf{u}}_{2i}^{(t)} \\
&= \mathbf{u}_{2i}^{(t-1)} + \eta \left((1-2p) \left(\mathbf{v}_{2i}^{(t-1)} + \mathbf{v}_{2i+1}^{(t-1)} \right) - 2 \sum_{j \neq 2i, 2i+1} q\, \mathbf{v}_j^{(t-1)} \right), \\
\mathbf{u}_{2i}^{(t)} &= \frac{\hat{\mathbf{u}}_{2i}^{(t)}}{\|\hat{\mathbf{u}}_{2i}^{(t)}\|}, \qquad \mathbf{u}_{2i+1}^{(t)} = \frac{\hat{\mathbf{u}}_{2i+1}^{(t)}}{\|\hat{\mathbf{u}}_{2i+1}^{(t)}\|},
\end{aligned}$$

$$\begin{aligned}
\hat{\mathbf{v}}_{2i}^{(t)} &= \mathbf{v}_{2i}^{(t-1)} + \eta \left((1-2p)\, \mathbf{u}_{2i}^{(t-1)} - \sum_{j \neq 2i, 2i+1} q\, \mathbf{u}_j^{(t-1)} \right), \\
\mathbf{v}_{2i}^{(t)} &= \frac{\hat{\mathbf{v}}_{2i}^{(t)}}{\|\hat{\mathbf{v}}_{2i}^{(t)}\|}.
\end{aligned}$$

We first prove (c). For simplicity, we denote $\alpha^{(t-1)} = \alpha$ and $\beta^{(t-1)} = \beta$. We also note that by the induction hypothesis, the norms $\|\hat{\mathbf{u}}_i^{(t)}\|^2$ and $\|\hat{\mathbf{v}}_i^{(t)}\|^2$ are independent of the index i and we denote these norms at time $t-1$ by r_u , r_v. We write the cross inner product and simplify it using the fact that $\lambda^{(t-1)} = 0, q \approx 0$ and η is small and representations at time $t-1$ are unit norm.

$$\begin{aligned}
\langle \mathbf{u}_i^{(t)}, \mathbf{v}_j^{(t)} \rangle &= \frac{\langle \hat{\mathbf{u}}_i^{(t)}, \hat{\mathbf{v}}_j^{(t)} \rangle}{\|\hat{\mathbf{u}}_i^{(t)}\| \|\hat{\mathbf{v}}_i^{(t)}\|} \\
&= \frac{1}{\sqrt{r_u\, r_v}} \Big[-\eta q \big(3 + \beta\big) + \mathcal{O}(\eta^2) \Big] \underset{q \approx 0}{\approx} 0, \quad \text{whenever } \lfloor i/2 \rfloor \neq \lfloor j/2 \rfloor.
\end{aligned}$$

The proofs for the cross inner products $\langle \mathbf{v}_i^{(t)}, \mathbf{v}_j^{(t)} \rangle$ and $\langle \mathbf{u}_i^{(t)}, \mathbf{u}_j^{(t)} \rangle$ when $\lfloor i/2 \rfloor \neq \lfloor j/2 \rfloor$ are entirely analogous to the argument above. Note that the invariance to the index, carry from timestep $t-1$ to t. Now we analyze the evolution of $\beta^{(t)}$ and $\alpha^{(t)}$, which similarly remain invariant to the choice of index at time t, if being invariant to index at time $t-1$.

$$\begin{aligned} r_u^{(t)} &:= \left\| \mathbf{u}_{2i}^{(t-1)} + \eta \left((1-2p)\big(\mathbf{v}_{2i}^{(t-1)} + \mathbf{v}_{2i+1}^{(t-1)}\big) - 2 \sum_{j \neq 2i, 2i+1} q\, \mathbf{v}_j^{(t-1)} \right) \right\|^2 \\ &= 1 + \mathcal{O}(\eta^2) + 4\eta(1-2p)\alpha, \end{aligned} \tag{7}$$

$$\begin{aligned} r_v^{(t)} &:= \left\| \mathbf{v}_{2i}^{(t-1)} + \eta \left((1-2p)\, \mathbf{u}_{2i}^{(t-1)} - \sum_{j \neq 2i, 2i+1} q\, \mathbf{u}_j^{(t-1)} \right) \right\|^2 \\ &= 1 + \mathcal{O}(\eta^2) + 2\eta(1-2p)\alpha. \end{aligned} \tag{8}$$

$$\begin{aligned} \beta^{(t)} &= \langle \mathbf{v}_{2i}^{(t)}, \mathbf{v}_{2i+1}^{(t)} \rangle \\ &= \frac{1}{r_v^{(t)}} \left(\beta + \eta \cdot 2(1-2p)\alpha \right) \\ &= \frac{\beta + 2\eta(1-2p)\alpha}{1 + 2\eta(1-2p)\alpha} \qquad (9) \\ &= \beta \; + \; \frac{2\eta(1-2p)\alpha\big(1-\beta\big)}{1 + 2\eta(1-2p)\alpha}. \qquad (10) \end{aligned}$$

Using the induction assumption that $\alpha^{(t)} > 0$ for all $t \geq 1$, every gradient descent update increases $\beta^{(t)}$ starting at $t = 2$. At equilibrium, $\beta^{(t)}$ can no longer increase, which implies that $\beta^{(\infty)} = 1$.

$$\begin{aligned} \alpha^{(t)} &= \left\langle \mathbf{u}_{2i}^{(t)}, \mathbf{v}_{2i}^{(t)} \right\rangle \\ &= \frac{1}{\sqrt{r_u\, r_v}} \left(\alpha + \eta(1-2p)\big(2+\beta\big) + \mathcal{O}(\eta^2) \right) \\ &\approx \frac{\alpha + \eta(1-2p)(2+\beta)}{\sqrt{1 + 4\eta(1-2p)\alpha}\,\sqrt{1 + 2\eta(1-2p)\alpha}} \qquad \text{(by the given form of } r_v, r_u) \qquad (11) \\ &\underset{\text{set } x := \eta(1-2p)}{=} \frac{\alpha + x(2+\beta)}{\sqrt{1+4x\alpha}\,\sqrt{1+2x\alpha}} \\ &\approx \frac{\alpha + x(2+\beta)}{(1+2x\alpha)\,(1+x\alpha)} \qquad (12) \\ &\underset{\eta \text{ small}}{\approx} \frac{\alpha + x(2+\beta)}{(1+3x\alpha)} \\ &= \alpha + \frac{x(2+\beta-3\alpha^2)}{(1+3x\alpha)} \qquad (13) \end{aligned}$$

Where Equation 11 follows from the approximation $\sqrt{1+cx} \approx 1 + \frac{1}{2}cx$ for small x (equivalently, small η).

First, note that by Equation 11 and since $\beta \geq 0$ (including at initialization), we have $\alpha^{(t)} > 0$ for all $t \geq 1$. This also completes the induction step to show that $\beta^{(t)} > 0$ for all $t \geq 2$, using the update equation for β (Equation 10).

Moreover, as established earlier, at equilibrium we must have $\beta = 1$. Substituting this into Equation 13 further implies that $\alpha = 1$ at equilibrium. Hence, we have successfully proved properties (a) and (b) too.

□

B.3 Proof of Theorem 1

Here we apply a weaker version of Lemma **??**. While Lemma **??** asserts that, at equilibrium, the representation of an anchor (s_i, a_i) exactly equals that of its positive example $s_{f,i}$, here we instead use the weaker condition that if

$$\phi(s_i, a_i) \;=\; \psi(s_{f,i}) \quad \forall\, s_f \text{ with } p^\pi_\gamma(s_f \mid s, a) > 0,$$

then it must also hold that

$$\phi(s, a) \;=\; \mathbb{E}_{p^\pi_\gamma(s_f \mid s,a)}\big[\psi(s_f)\big].$$

Proof. From Lemma **??**:

$$\phi(s, a) \;=\; \mathbb{E}_{p^\pi_\gamma(s_f \mid s,a)}[\psi(s_f)].$$

Substituting this into the SGCRL actor objective, we obtain

$$\begin{aligned}
&\max_\pi \; \mathbb{E}_{s\sim p(s),\, a\sim\pi(a\mid s,g)}\Big[\phi(s,a)^\top \psi(g) \;+\; \tau H(\pi(\cdot \mid s,a))\Big] \\
&= \max_\pi \; \mathbb{E}_{s\sim p(s),\, a\sim\pi(a\mid s,g)}\Big[\mathbb{E}_{p^\pi_\gamma(s_f\mid s,a)}[\psi(s_f)]^\top \psi(g) \;+\; \tau H(\pi(\cdot \mid s,a))\Big] \\
&= \max_\pi \; \mathbb{E}_{s\sim p(s),\, a\sim\pi(a\mid s,g)}\Big[\mathbb{E}_{p^\pi_\gamma(s_f\mid s,a)}[\psi(s_f)^\top \psi(g)] \;+\; \tau H(\pi(\cdot \mid s,a))\Big].
\end{aligned}$$

Expanding the discounted future state distribution yields

$$= \max_\pi \; \mathbb{E}_{s\sim p(s),\, a\sim\pi(a\mid s,g)}\left[(1-\gamma)\sum_{s_f}\Big(\sum_{t=0}^{\infty}\gamma^t\, p^\pi_t(s_t = s_f \mid s,a)\Big)\,\psi(s_f)^\top\psi(g) \;+\; \tau H(\pi(\cdot \mid s,a))\right]$$

$$= \max_\pi \; \mathbb{E}_{s\sim p(s),\, a\sim\pi(a\mid s,g)}\left[\sum_{t=0}^{\infty}\gamma^t\, \mathbb{E}_{p^\pi_t(s_t)}[\psi(s_t)^\top\psi(g) \mid s_0 = s, a_0 = a] \;+\; \tau H(\pi(\cdot \mid s,a))\right]$$

$$= \max_\pi \; \mathbb{E}_{s\sim p(s),\, a\sim\pi(a\mid s,g)}\mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\,\psi(s_t)^\top\psi(g) \;\Big|\; s_0 = s, a_0 = a\right] \;+\; \tau H(\pi(\cdot \mid s,a)).$$

This is exactly the reinforcement learning objective with reward function

$$r(s,a) \;=\; \psi(s)^\top\psi(g).$$

Therefore, maximizing the SGCRL objective is equivalent to maximizing the Q-value induced by this reward function. □

B.4 A stronger version of Theorem 2 and proof

We now analyze the equilibrium dynamics of the InfoNCE update rule in the following theorem. Claims 1 and 3 of this theorem directly imply the result stated in Theorem 2. For simplicity of notation, we write $\mathbf{u}_i$ in place of $\phi(s_i, a_i)$ and $\mathbf{v}_i$ in place of $\psi(sf_i)$.

Theorem 3 (InfoNCE representations at equilibrium)**.** *Let $\mathbf{z} \in \mathbb{R}^d$ be a fixed unit vector, with $d \gg 1$. Let $\{\mathbf{u}_i\}_{i=1}^n$ and $\{\mathbf{v}_i\}_{i=1}^n \subset \mathbb{R}^d$ be anchor and future embeddings, initialized as:*

$$\mathbf{u}_i^0 = c\,\mathbf{z} + \zeta_i^0, \quad \mathbf{v}_i^0 = c\,\mathbf{z} + \kappa_i^0$$

where $\zeta_i^0, \kappa_i^0 \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \frac{1-c^2}{d} I_d\right)$ and c is a scalar. Suppose these vectors are updated via gradient descent on the backward (or forward) InfoNCE loss with batch size $N \gg 1$ and step size $\eta > 0$, followed by unit-norm normalization. We assume η is sufficiently small.

Then, with high probability over the initialization, the dynamics satisfy the following:

1. *At every step t, each representation decomposes as*
$$\mathbf{u}_i^t = c^t \mathbf{z} + \zeta_i^t, \quad \mathbf{v}_i^t = c^t \mathbf{z} + \kappa_i^t,$$
where $\zeta_i^t, \kappa_i^t \perp \mathbf{z}$ and c^t is the same for all i.

2. *At fixed point (i.e., when all the gradients are zero), $\mathbf{u}_i^{(\infty)} = \mathbf{v}_i^{(\infty)}, \ \forall i$, and $\langle \mathbf{u}_i^{(\infty)}, \mathbf{v}_j^{(\infty)} \rangle = 0, i \neq j$*

3. *At fixed point: $c^{(\infty)} = 0$, i.e., all representations become orthogonal to $\mathbf{z}$ as $t \to \infty$.*

Proof. We establish the above results, together with three additional claims regarding the InfoNCE update dynamics at equilibrium, via induction. At iteration t, we decompose each representation as
$$\mathbf{u}_i^t = c^{(t)}\, \mathbf{z} + \zeta_i^t, \qquad \mathbf{v}_i^t = c^{(t)}\, \mathbf{z} + \kappa_i^t,$$
where the residuals ζ_i^t and κ_i^t are orthogonal to the unit vector $\mathbf{z}$. We define the following quantities and note that they are invariant with respect to the choice of index i (or $i \neq j$ where applicable):

(a) $\langle \mathbf{u}_i^t, \mathbf{z} \rangle = \langle \mathbf{v}_i^t, \mathbf{z} \rangle = c^{(t)}$ for all i.

(b) $\alpha^{(t)} := \langle \zeta_i^t, \kappa_i^t \rangle$.

(c) $\lambda^{(t)} := \langle \zeta_i^t, \kappa_j^t \rangle = \langle \zeta_i^t, \zeta_j^t \rangle = \langle \kappa_i^t, \kappa_j^t \rangle = 0$ for all $i \neq j$ and for all t.

(d) $r^{(t)} := \|\zeta_i^t\|^2 = \|\kappa_i^t\|^2$ for all i.

Base case ($t = 0$): By initialization, $\zeta_i^0, \kappa_i^0 \sim \mathcal{N}(0, \frac{1-c^2}{d} I_d)$ i.i.d., and $\zeta_i^0, \kappa_i^0 \perp \mathbf{z}$. In high dimensions, with probability 1 these vectors are all orthogonal to each other therefore $\lambda^{(0)} = \alpha^{(0)} = 0$ and c^0 is the same for all vectors by construction. And $\|\zeta_i^0\| = \|\kappa_i^0\| = 1 - c^2$ by construction. Hence, all four properties a,b,c,d hold at $t = 0$.

Inductive step: Assume the properties hold at time $t-1$. We now prove they also hold at time t.

We first prove it for the case that representations are normalized and for the backward InfoNCE loss, the proof for the forward InfoNCE loss is exactly the same due to the symmetry at initialization, which, as we will see later, is maintained through all the updates. The backward InfoNCE updates with normalization are given by:
$$\hat{\mathbf{u}}_i^t = \mathbf{u}_i^{t-1} + \eta \left(\mathbf{v}_i^{t-1} - \sum_j p_{ij} \mathbf{v}_j^{t-1} \right) \tag{14}$$
$$\mathbf{u}_i^t = \frac{\hat{\mathbf{u}}_i^t}{\|\hat{\mathbf{u}}_i^t\|}, \qquad \text{similarly for } \mathbf{v}_i^t \tag{15}$$
where $p_{ij}^{(t-1)} = \frac{\exp(\langle \mathbf{u}_i^{t-1}, \mathbf{v}_j^{t-1} \rangle)}{\sum_k \exp(\langle \mathbf{u}_k^{t-1}, \mathbf{v}_j^{t-1} \rangle)}$.

Due to symmetry at time $t-1$, we have:
$$p_{ij}^{(t-1)} = p_{ji}^{(t-1)}, \quad p_{ii}^{(t-1)} =: p^{(t-1)} \quad \text{for all } i, j$$
so the matrix $P^{(t-1)}$ is symmetric, with equal diagonals and exchangeable off-diagonals. For ease of notation we use $p := p_{ii}^{(t-1)}$ and $q := p_{ij}^{(t-1)}$. Note that $p + (N-1)q = 1$

Property (a), (d): projection onto $\mathbf{z}$ and norm From the updates and the fact that all $c^{(t-1)}$ are equal, we get:
$$\left\langle \mathbf{v}_i^{t-1} - \sum_j p_{ij} \mathbf{v}_j^{t-1}, \mathbf{z} \right\rangle = c^{(t-1)} - \sum_j p_{ij} c^{(t-1)} = 0$$
Thus,
$$\langle \hat{\mathbf{u}}_i^t, \mathbf{z} \rangle = \langle \mathbf{u}_i^{t-1}, \mathbf{z} \rangle = c^{(t-1)}, \quad \Rightarrow \quad \langle \mathbf{u}_i^t, \mathbf{z} \rangle = \frac{c^{(t-1)}}{\|\hat{\mathbf{u}}_i^t\|} =: c^{(t)} \tag{16}$$

and the same holds for $\mathbf{v}_i^t$. In order to prove (a), we need to show that $\|\mathbf{u}_i\|$ and $\|\mathbf{v}_i\|$ are equal and invariant to the index for any t. (For ease of notation we use α, λ, r instead of $\alpha^{(t-1)}, \lambda^{(t-1)}, r^{(t-1)}$.)

$$\begin{aligned}\left\|\hat{\mathbf{u}}_i^t\right\|^2 &= \left\| c^{(t-1)}\mathbf{z} + \zeta_i^{t-1} + \eta\left((1-p)\kappa_i^{t-1} - \sum_{j\neq i} q\kappa_j^{t-1}\right)\right\|^2 \\ &= \left(c^{(t-1)}\right)^2 + r + 2\eta(1-p)\alpha - 2\eta(N-1)q\ \lambda \\ &= 1 + 2\eta(1-p)\cdot\alpha + \mathcal{O}(\eta^2) \end{aligned} \tag{17}$$

We used the fact that due to normalization $(c^{(t-1)})^2 + r = 1$, we also used $\lambda = 0$.

It is straightforward to verify that the expression for $\|\hat{\mathbf{u}}_i^t\|^2$ is independent of the choice of index i. Furthermore, if we expand $\|\hat{\mathbf{v}}_i^t\|^2$, we encounter the same number of matched pairwise or cross-term inner products, with only the order of terms being swapped. As a result, we obtain the same expression.

We therefore denote this common quantity by

$$L^{(t-1)} \coloneqq \left\|\hat{\mathbf{u}}_i^t\right\|^2 = \left\|\hat{\mathbf{v}}_i^t\right\|^2.$$

Therefor, it follows from Equation 16 that $\langle \mathbf{u}_i^t, \mathbf{z}\rangle$ is identical for all representations. Since all these representations share the same norm and identical projection onto $\mathbf{z}$ (i.e., the parallel component), it must be that their orthogonal components—namely, ζ_i^t and κ_i^t—also have equal norms. Hence, condition (d) is satisfied too, this also ends the proof for claim 1 of the theorem statement.

Properties (b), (c): Matching and cross inner product We expand:

$$\begin{aligned}\zeta_i^t &= \frac{1}{L^{(t-1)}}\left(\zeta_i^{t-1} + \eta\left((1-p)\kappa_i^{t-1} - \sum_{j\neq i} q\kappa_j^{t-1}\right)\right) \\ \kappa_i^t &= \frac{1}{L^{(t-1)}}\left(\kappa_i^{t-1} + \eta\left((1-p)\zeta_i^{t-1} - \sum_{j\neq i} q\zeta_j^{t-1}\right)\right)\end{aligned}$$

Then:

$$\begin{aligned}\alpha^{(t)} = \langle \zeta_i^t, \kappa_i^t\rangle &= \frac{1}{(L^{(t-1)})^2}\left[\alpha + 2\eta(1-p)r - 2\eta(N-1)\lambda q + \mathcal{O}(\eta^2)\right] \\ &= \frac{1}{(L^{(t-1)})^2}[\alpha + 2\eta r(1-p) + \mathcal{O}(\eta^2)]\end{aligned}$$

This expression is independent of the choice of index i; this proves statement (b). Similarly, we can evaluate $\langle \zeta_i^{(t)}, \kappa_j^{(t)}\rangle$ and $\langle \zeta_i^{(t)}, \zeta_j^{(t)}\rangle$ for $i \neq j$:

$$\begin{aligned}\zeta_i^t &= \frac{1}{L^{(t-1)}}\left(\zeta_i^{t-1} + \eta\left((1-p)\kappa_i^{t-1} - \sum_{l\neq i} q\kappa_l^{t-1}\right)\right) \\ \zeta_j^t &= \frac{1}{L^{(t-1)}}\left(\zeta_j^{t-1} + \eta\left((1-p)\kappa_j^{t-1} - \sum_{l\neq j} q\kappa_l^{t-1}\right)\right) \\ \kappa_j^t &= \frac{1}{L^{(t-1)}}\left(\kappa_j^{t-1} + \eta\left((1-p)\zeta_j^{t-1} - \sum_{l\neq j} q\zeta_l^{t-1}\right)\right)\end{aligned}$$

$$\begin{aligned}\langle \zeta_i^t, \zeta_j^t \rangle &= \frac{1}{(L^{(t-1)})^2}[\lambda + 2\eta(1-p)\lambda - 2\eta q(N-2)\lambda - 2\eta q\alpha + \mathcal{O}(\eta^2)] \\ &\underset{N \text{ Large}}{=} \frac{1}{(L^{(t-1)})^2}[\lambda + 2\eta(1-p)\lambda - 2\eta q(N-1)\lambda + \mathcal{O}(\eta^2)] \\ &\underset{p+(N-1)q=1}{=} \frac{1}{(L^{(t-1)})^2}[\lambda + \mathcal{O}(\eta^2)]\end{aligned}$$

$$\begin{aligned}\langle \zeta_i^t, \kappa_j^t \rangle &= \frac{1}{(L^{(t-1)})^2}[\lambda + 2\eta(1-p)\lambda - 2\eta q(N-2)\lambda - 2\eta q r + \mathcal{O}(\eta^2)] \\ &\underset{N \text{ Large}}{=} \frac{1}{(L^{(t-1)})^2}[\lambda + 2\eta(1-p)\lambda - 2\eta q(N-1)\lambda + \mathcal{O}(\eta^2)] \\ &\underset{p+(N-1)q=1}{=} \frac{1}{(L^{(t-1)})^2}[\lambda + \mathcal{O}(\eta^2)]\end{aligned}$$

Since N is large and $p, q \geq 0$ with $p + (N-1)q = 1$, it follows that $q \approx 0$. We observe that these inner products are also independent of the specific choice of i and j, due to the same underlying symmetry. Moreover $\langle \zeta_i^{(t)}, \kappa_j^{(t)} \rangle$ and $\langle \zeta_i^{(t)}, \zeta_j^{(t)} \rangle$ are all zero given that λ is zero (induction) hence (c) is also proved.

This completes the inductive step, i.e., the proof for a, b, c, d.

Now we assess how $\alpha^{(t)}$ and $r^{(t)}$ change to prove claims 2, 3 of the theorem statement.

$$\begin{aligned}\alpha^{(t)} &= \frac{1}{(L^{(t-1)})^2}[\alpha + 2\eta r(1-p)\cdot r + \mathcal{O}(\eta^2)] && (18)\\ &\approx \frac{\alpha + 2\eta(1-p)\cdot r}{1 + 2\eta(1-p)\cdot\alpha} \\ &= \alpha + \frac{2\eta(1-p)r\cdot\left(1-\alpha\frac{\alpha}{r}\right)}{1+2\eta(1-p)\cdot\alpha} \\ &\underset{\alpha\leq 1}{\geq} \alpha + \frac{2\eta(1-p)r\cdot\left(1-\frac{\alpha}{r}\right)}{1+2\eta(1-p)\cdot\alpha} && (19)\\ &\geq \alpha + \frac{2\eta(1-p)r\cdot(1-\frac{\alpha}{r})}{1+2\eta(1-p)\cdot\alpha} && (20)\\ &\geq \alpha\end{aligned}$$

Since the denominator is positive and $1 - \alpha/r \geq 0$. α only stops growing if the equality holds i.e., $\alpha/r = 1$,

$$\frac{\langle \zeta_i^{(t-1)}, \kappa_i^{(t-1)} \rangle}{\|\zeta_i^{(t-1)}\| \, \|\kappa_i^{(t-1)}\|} = 1 \quad \iff \quad \cos\left(\zeta_i^{(t-1)}, \kappa_i^{(t-1)}\right) = 1,$$

so as long as $\cos\left(\zeta_i^{(t-1)}, \kappa_i^{(t-1)}\right) < 1$, $\alpha^{(t)}$ increases, at equilibrium its growth should be stopped and that means the alignment of each positive pair increases strictly, and positive pairs keep aligning until they are fully aligned. This result, along with the fact that $\lambda^{(t)} = 0$, $\forall t$, which we proved before, completes the proof for claim 2 of the theorem statement.

Finally, we prove Claim 3 of the theorem by showing that, at equilibrium, $r = 1$, i.e., the squared norm of the component orthogonal to $\mathbf{z}$. Since each representation is normalized, this immediately implies that the parallel component must vanish.

$$r^{(t)} = r + 2\eta(1-p)\alpha$$

Since $\alpha^{(0)} = 0$ and, by Equation 19, $\alpha^{(t)}$ increases strictly (Note that due to norm 1 constraint on the representations p can never approach 1) until full alignment, it follows that $\alpha^{(t)} > 0$ for all $t \geq 1$. Consequently,

$$r^{(t)} > r^{(t-1)} \qquad \text{for all } t \geq 2.$$

Thus, the sequence $\{r^{(t)}\}$ is strictly increasing until it saturates at the unit-norm bound.

□

C Experimental Details

Table 1: SGCRL Hyperparameters.

hyperparameter	value
Standard SGCRL (Liu et al., 2024)	
batch size	256
learning rate	3e-4
discount	0.99
actor target entropy	0
hidden layers sizes (policy, critic)	(256, 256)
initial random data collection	10,000 transitions
replay buffer size	1e6
samples per insert[1]	256
representation dimension (dim($\phi(s, a)$), dim($\psi(s_g)$))	64
actor minimum std dev	1e-6
Tabular SGCRL	
batch size	128
learning rate	1e-3
discount	0.99
initial random data collection	False
replay buffer size	1e3
representation dimension (dim($\psi(s_g)$))	16

[1] How many times is each transition used for training before being discarded.

D Additional Experiments

Single-goal exploration in robotic manipulation tasks. We find that the characterization of SGCRL detailed in Section 3 holds in a Sawyer robotic manipulation task where the agent must pick up a block and place it in a bin (Yu et al., 2020). During early stages of training, the agent moves the robotic end-effector towards regions of high representational similarity to the goal. Subsequently, the representational goal similarity of these frequently visited regions decreases, and the agent visits new regions (see Fig. 3). These results suggest that out characterization of single-goal exploration generalizes beyond 2D navigation tasks.

D.1 SGCRL data collection structures the representations optimally

For a goal-conditioned exploration algorithm to be effective, its dynamics must rule out states already visited without containing the goal, preventing repeated exploration of irrelevant regions. Theorem 2 assures that SGCRL's contrastive representations achieve this by driving the corresponding ψ-similarity of non-goal areas close to zero.

In practice, however, the assumptions of Theorem 2 for instance symmetric initialization do not necessarily hold. For instance, in the four-room environment, the agent begins in the bottom-left room and gradually visits new ones. Adding states from newly visited rooms to the replay buffer alters the updates of earlier rooms, breaking the symmetry assumption and eliminating the guarantee of orthogonality between room representations. In this section, we address two key questions: 1) Does the phenomenon of decreasing ψ-similarity persist under in more realistic settings? 2) If so,

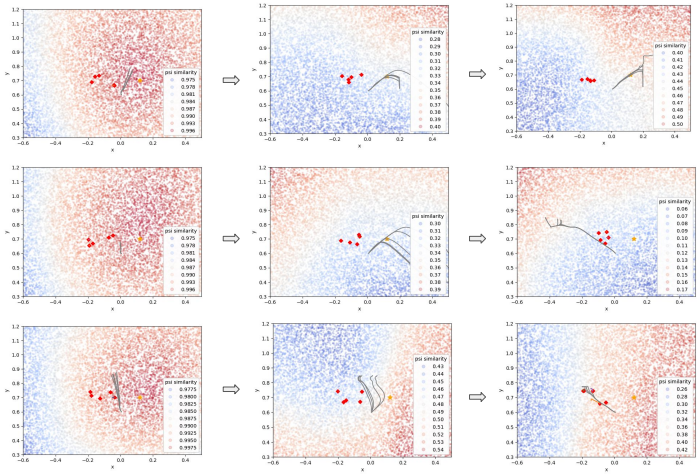


Figure 3: **XY cross section of representational goal similarity for the Sawyer Bin environment**. Each row represents checkpoints throughout training for different training seeds. The gray lines show the trajectory of the end-effector across 5 episodes. The agent moves the end-effector towards regions of high ψ-similarity to the goal, and then those regions subsequently develop low ψ-similarity, driving continued exploration to new areas.

what are the underlying dynamics, and are they unique to SGCRL's goal-directed data collection, or can other exploration strategies achieve the same effect?

The imaginary goal experiment (Figure 1b), demonstrates that even in realistic scenarios the representations of visited rooms drift farther from the goal representation at $(0, 0, 1)$ on the z-axis. This behavior is intuitive under the chosen initialization strategy, where all state representations are set to z plus small Gaussian noise,making every state initially appear close to the goal. Because, first, within a single room, the shared component aligned with $\psi(g) = z$ is essentially wasted energy for contrastive learning. To achieve a stronger contrastive loss among states in the same room, the z-component of their representations is dampened (refer to Theorem 2). Second, as the agent explores more and begins visiting new rooms, the introduction of these new room representations pushes the older room representations even further away from the goal. This happens because newly visited rooms are initialized close to z, so in order to maintain contrast, older rooms are better off shifting downward and away from it.

This observation naturally leads to the following question: *what happens if the representations of some areas in the new rooms are not initialized close to z?* To explore this, we design a new experiment. In this setting, the representations of a small patch in the top-left and bottom-right rooms is initialized orthogonal to z, while the rest of the state representations are initialized as z plus small random noise (see Figure 4). The agent starts in the bottom-left room (marked by the green dot).

Apart from representation initialization, the experimental setup is the same as the imaginary goal experiment. i.e., $\psi(g) = z$ is a random Gaussian vector that does not correspond to the representation of any actual maze state, simulating a scenario where many representation updates occur without the agent ever observing the true goal.

We then compare two strategies for data collection:

1. **SGCRL exploration** – actions are selected according to the greedy policy:

$$a_t \sim \frac{1}{Z(s_t)} \exp\left(\frac{1}{\tau}\psi(m(s_t, a_t))^\top z\right).$$

2. **Random-goal exploration** – actions are chosen based on:

$$a_t \sim \frac{1}{Z(s_t)} \exp\left(\frac{1}{\tau}\psi(m(s_t, a_t))^\top \zeta_t\right), \zeta_t \sim N(0, I_d)$$

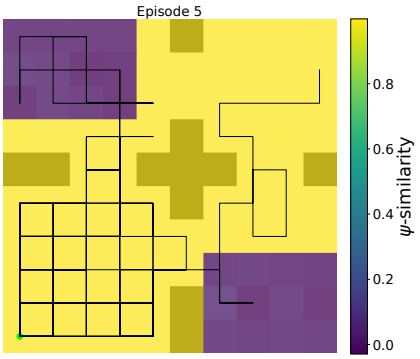


Figure 4: SGCRL data collection vs random goal data collection. At initialization, all state representations are a noisy version of the imaginary goal representation (z) while the two small patches in the top left room and the bottom right room are initialized with an initialization orthogonal to z.

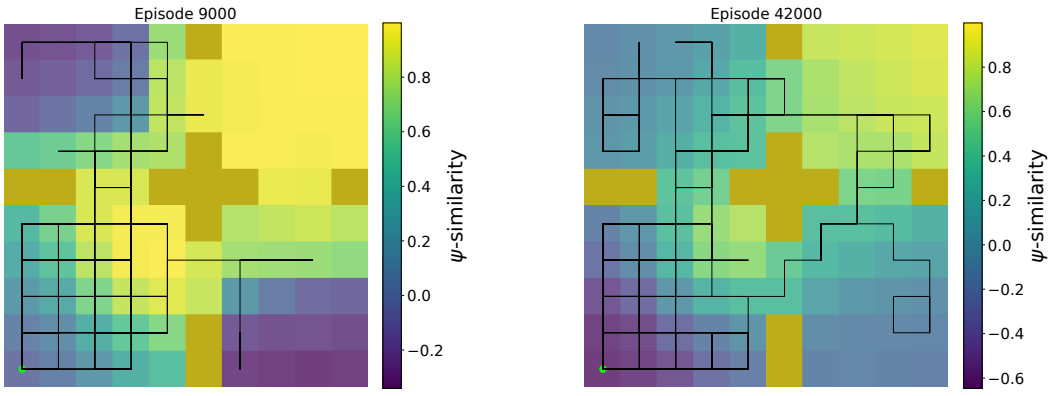
where ζ_t is a Gaussian-sampled goal embedding drawn anew at each action selection. This corresponds to the widely used convention in earlier works where exploration is guided by sampling from a distribution of potential goals to learn more effectively about the environment.

Both data collection strategies start from the same initial representations. We then compare how the representations evolve under these two different data collection strategies.

As demonstrated in Figure 5, at episode 9000—before the representations have fully converged—the SGCRL strategy naturally avoids the small dark patches. Because these patches have lower ψ-similarity compared to their surroundings, the policy does not visit them. Instead, it focuses on areas with higher ψ-similarity. As a result, the representations of the most frequently updated states consistently evolve by moving farther away from the goal, allowing them to form stronger contrasts with the newly visited states that are closer to the goal in representation.

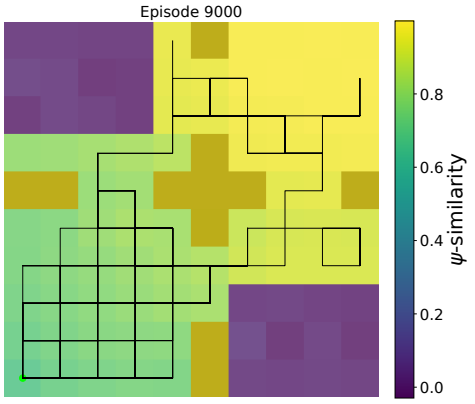
In contrast, the random goal exploration strategy does not avoid the dark patches (Figure 6a). By visiting these areas, it adds their states to the replay buffer, which in turn forces the older room representations to contrast with these new states. This dynamic prevents the older representations from drifting away from the goal. Indeed, even after convergence, the ψ-similarity under random goal exploration fail to approach zero, as shown in Figure 6b.

Figure 6 further visualizes this evolution by projecting the representations into three dimensions using PCA. The key takeaway is that SGCRL's greedy data collection provides an implicit structural benefit: by consistently moving toward regions of higher ψ-similarity, it pushes the representations of previously visited areas farther from the goal. In doing so, it shapes the representations of visited, non-goal regions in a way that naturally encourages further exploration.

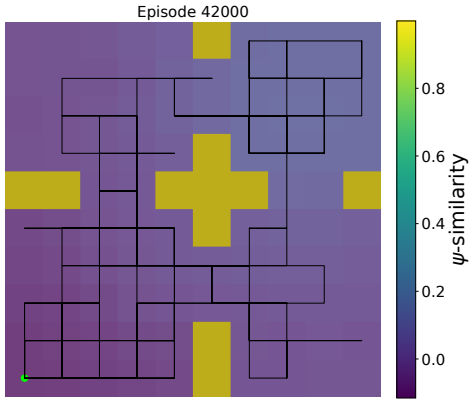


(a) Random goal (before convergence)

(b) Random goal (after convergence)

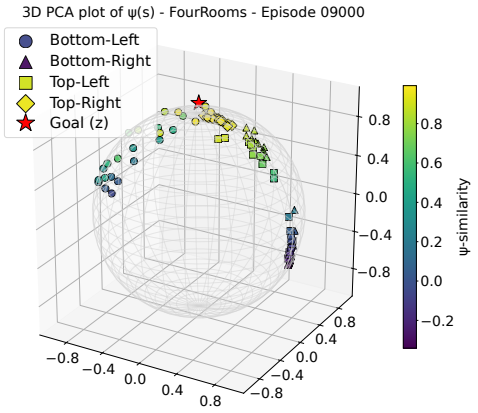


(c) SGCRL (before convergence)

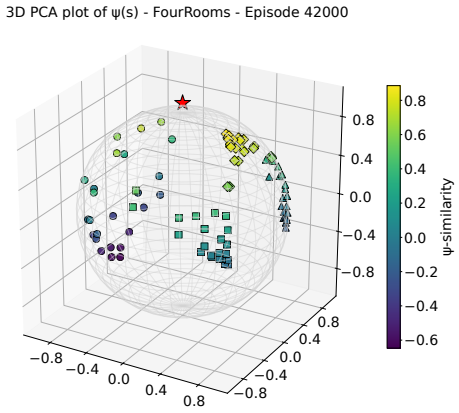


(d) SGCRL (after convergence)

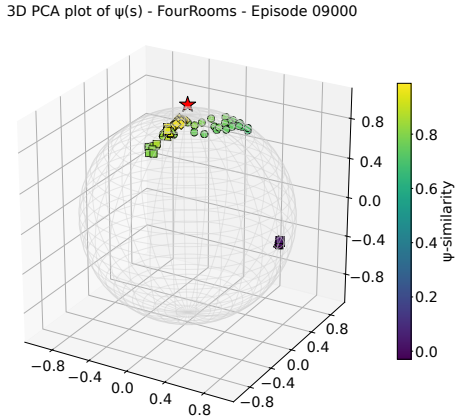
Figure 5: Comparison of representation evolution under two data collection strategies: (Top row) random goal exploration and (Bottom row) SGCRL.



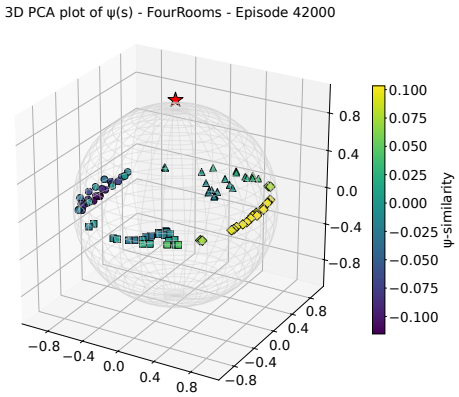
(a) Random goal (before convergence)



(b) Random goal (after convergence)



(c) SGCRL (before convergence)



(d) SGCRL (after convergence)

Figure 6: Comparison of representation evolution under two data collection strategies, with representations projected into 3D using PCA. (Top row) random goal exploration and (Bottom row) SGCRL.