IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 September 2026, accepted 7 October 2026, date of publication 14 October 2026, date of current version 27 October 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3743463

RESEARCH ARTICLE

Bisimulation Geometry: Unifying Representation Learning and Exploration in Reinforcement Learning

DEEPAK GUPTA[1], ADITYA SIKARWAR[1], (Student Member, IEEE), SANIDHYA DHANGAR[1], (Student Member, IEEE), CHANAKYA KUMAR JHA[2,3], (Senior Member, IEEE), PALLAVI JHA[4], (Member, IEEE), AKHILESH KUMAR SHARMA[5], (Senior Member, IEEE), AND VUGAR ABDULLAYEV[6,7]

[1]Department of CSE–Data Science, Institute of Technology and Management Gwalior, Gwalior, Madhya Pradesh 475001, India
[2]IEEE India Council, Karnataka 560055, India
[3]SIGHT Chair, IEEE Bombay Section, Mumbai, Maharashtra, India
[4]Department of Computer Engineering, Alard University, Pune, Maharashtra, India
[5]Department of Data Science and Engineering, Manipal University Jaipur, Jaipur, Rajasthan 303007, India
[6]Azerbaijan State Oil and Industry University, AZ1010 Baku, Azerbaijan
[7]Azerbaijan University of Architecture and Construction, Baku, Azerbaijan

Corresponding author: Akhilesh Kumar Sharma (akhileshkumar.sharma@jaipur.manipal.edu)

ABSTRACT Two long-standing difficulties in reinforcement learning (RL)—learning compact state representations and principled exploration—are usually handled by separate components glued together with a hand-tuned weight. We argue they are not separate: both are manifestations of one geometric object, the bisimulation pseudometric of the underlying Markov decision process (MDP). Our core contribution is a single loss, the Bisimulation Infomax Loss (BIL), removing the conventional λ trade-off coefficient, together with two certificates written in one measurable quantity, the embedding distortion κ^*. First, a Dual Near-Optimality Theorem: under stated regularity and model assumptions, an approximate isometry of the bisimulation pseudometric lies simultaneously within $\mathcal{O}(\kappa^{*2})$ of the InfoNCE bound attained by an exact isometry and within $\mathcal{O}(\kappa^*)$ of the corresponding one-step epistemic-information-gain guarantee. Second, a Curiosity Certificate: under explicit reachability and hitting-probability assumptions, the exploration horizon satisfies $T_{\text{explore}} = \tilde{\mathcal{O}}\big((\kappa^*)^{-D}\big)$, with D the doubling dimension of the bisimulation geometry; the measured slope of -1.96 ($R^2 = 0.96$) is consistent with an effective dimension near two. We also bound the error of bootstrapped bisimulation targets and propagate it through every guarantee. We then propose Adaptive Bisimulation Mirror-Descent (ABMD), combining BIL with generalized Gauss–Newton preconditioning, an adaptive bisimulation temperature, and behavior-preserving augmentations. ABMD reduces steps-to-threshold by up to 33% over the strongest baseline on MiniGrid, lifts returns on the Distracting Control Suite by up to 17%, and gains steadily on four Procgen environments, all under one hyperparameter setting. Sensitivity analyses, cross-environment ablations, out-of-distribution stress tests, and compute-normalized comparisons accounting for the $1.9\times$ per-step overhead are reported.

INDEX TERMS Bisimulation metric, curiosity-driven exploration, information geometry, reinforcement learning, representation learning, self-supervised learning.

I. INTRODUCTION

Consider an autonomous agent placed in an environment it has never encountered before. At every discrete step, the agent receives a high-dimensional sensory observation—camera pixels, proprioceptive joint angles, lidar range readings—and must commit to an action that nudges it closer to some distant goal. Two tightly coupled problems sit at the heart of modern reinforcement learning in this setting [1].

The associate editor coordinating the review of this manuscript and approving it for publication was Zheng Chen.

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

Challenge 1: Representation Learning. Raw observations live in a space cluttered with task-irrelevant variation: drifting shadows, swapped backgrounds, and ever-present sensor noise. The agent has to compress them into a latent representation that strips these nuisances away while preserving everything that matters for choosing actions [2], [3]. Concretely, we are looking for an encoder $\phi : \mathcal{O} \to \mathcal{Z}$ under which Euclidean closeness in latent space stands for behavioral equivalence—two states ought to be mapped near each other exactly when they would lead to the same optimal action [4].

Challenge 2: Systematic Exploration. When rewards are sparse, the agent may be hundreds of states away from the first useful learning signal. Random strategies like ϵ-greedy are known to be inefficient in such regimes. The remedy is intrinsic motivation: a curiosity bonus that rewards visiting states where the world model is uncertain or where genuinely novel behavior occurs [5], [6].

The textbook way to combine these objectives is disarmingly simple—add their losses with a scalar trade-off coefficient $\lambda \in \mathbb{R}_+$

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{SSL}}(\theta) - \lambda \cdot \mathcal{L}_{\text{IM}}(\theta). \tag{1}$$

This approach is everywhere in practice, yet it carries three pathologies that we make precise and then resolve:

Pathology 1: Environment-Dependent Sensitivity. The right value of λ is highly dependent on reward sparsity, transition noise, and the geometry of the state space. Grid searches over λ routinely eat up 40–70% of the training budget, and a value that works in one environment seldom carries over to another [3].

Pathology 2: Gradient Opposition. We show that, under conditions that we state precisely in Section III (stop-gradient targets, hypersphere embeddings, and an approximate neural-tangent-kernel isotropy assumption), the SSL and IM gradients have an expected inner product that is provably non-positive. SSL contracts the latent manifold by drawing behaviorally similar states together; IM expands it by pushing novel states apart. The interference between the two slows convergence noticeably [7].

Pathology 3: No Certifiable Link. There is no quantitative bridge between the quality of a learned representation and the efficiency of exploration. Suppose we have a metric κ that captures the worst-case mismatch between latent distance and behavioral distance: can we predict, ahead of time, how many steps will be needed to cover the environment? Without such a link, principled algorithm design is hard [8].

A. THE BISIMULATION-GEOMETRIC RESOLUTION

A central observation of this paper is that all three pathologies share a common root and admit a single fix, contrasted schematically in Fig. 1. The root cause is geometric: SSL and IM are built on opposing operations—contraction versus expansion. The fix is to anchor both objectives to the bisimulation pseudometric of the underlying MDP [4], [9].

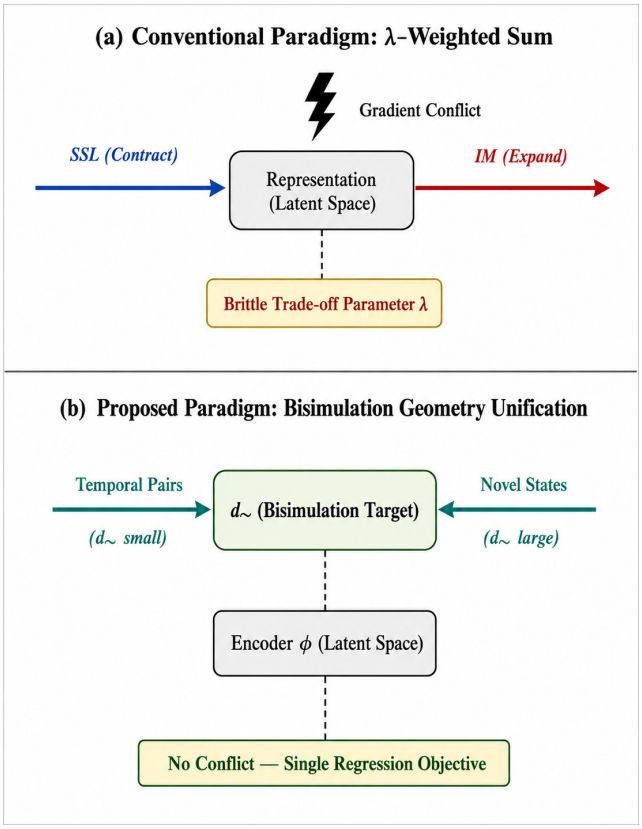


FIGURE 1. Conceptual illustration of the bisimulation-geometric unification. (a) The conventional setup: SSL and IM tug the representation in opposite directions and depend on a fragile trade-off parameter λ. (b) The proposed setup: both objectives now regress to the same bisimulation target $d_\sim$, so the gradient conflict disappears. The bisimulation metric $d_\sim$ gives the two objectives a shared reference frame: temporally adjacent states sit close together ($d_\sim$ small) while behaviorally novel states sit far apart ($d_\sim$ large).

Bisimulation has its origins in formal verification, where it expresses behavioral equivalence: two states are bisimilar if they yield the same expected immediate reward and transition into states that are themselves bisimilar [10]. The bisimulation pseudometric $d_\sim$ is the continuous relaxation of this idea: $d_\sim(s, s') = 0$ marks states that are behaviorally indistinguishable, while $d_\sim(s, s') > 0$ measures how far apart their behaviors are [4]. Crucially, $d_\sim$ is the unique fixed point of a γ-contraction $\mathcal{T}_{\text{Bi}}$, and it admits a tight bound on value function gaps:

$$|V^\pi(s) - V^\pi(s')| \le \frac{d_\sim(s, s')}{1 - \gamma}, \quad \forall \pi. \tag{2}$$

The geometric observation that drives this paper is the following: an encoder ϕ that is an approximate isometric embedding of $d_\sim$—that is, $\|\phi(s) - \phi(s')\|_2 \approx d_\sim(s, s')$—is automatically optimal for both SSL and IM. On the SSL side, temporally adjacent states have provably bounded bisimulation distance, so contracting their representations is geometrically the right thing to do. On the IM side, the novelty of a state s relative to a visited set $\mathcal{C}$ is exactly $\min_{c \in \mathcal{C}} d_\sim(s, c)$, so expanding toward behaviorally novel states is equally well grounded. Both objectives end up

regressing different state-pair distributions onto the same fundamental metric.

B. CONTRIBUTIONS

Novelty statement. The novelty of this paper can be stated in a single sentence: *we replace the standard two-objective formulation "SSL loss + $\lambda \times$ intrinsic-motivation loss" by a single regression onto the bisimulation pseudometric, and we quantify—through one measurable scalar, the embedding distortion κ^*—how well the resulting encoder serves both purposes at once.* Prior bisimulation methods (DBC [3], MICo [11], SimSR [12]) use the metric only for representation learning; prior intrinsic-motivation methods (ICM [5], RND [6], Plan2Explore [8]) use novelty or information gain without any representation-quality guarantee. To our knowledge, no earlier work ties an exploration-horizon bound to a learned representation-fidelity quantity.

To keep the core message distinct from the engineering that supports it, we separate our contributions into two groups.

Core contributions:

1) **Bisimulation Infomax Loss (BIL):** a single loss that drops the SSL/IM split altogether; its global minimizer is exactly an ϵ-bisimulation embedding, together with an explicit bound on the approximation error introduced by bootstrapped targets (Proposition 5).
2) **Dual Near-Optimality Theorem:** a formal proof that, under Assumption 2, the BIL-optimal encoder is within $\mathcal{O}(\kappa^{*2})$ of the InfoNCE mutual-information bound attained by an exact bisimulation isometry [13], and within $\mathcal{O}(\kappa^*)$ of the corresponding lower bound on one-step epistemic information gain [14]. The scope of the theorem is stated precisely: it certifies near-optimality relative to the class of isometric encoders rather than global optimality over all encoders, and Section V-D characterizes the regimes in which bisimulation distance and information gain diverge.
3) **Curiosity Certificate:** an exploration-horizon guarantee written purely in terms of a learned, empirically measurable representation-quality metric κ^*, stated under explicit reachability, connectivity, and hitting-probability assumptions (Assumption 3). The resulting scaling is $T_{\text{explore}} = \tilde{\mathcal{O}}((\kappa^*)^{-D})$ with D the doubling dimension of the bisimulation geometry; the empirically fitted slope of -1.96 ($R^2 = 0.96$) is consistent with an effective dimension $D \approx 2$ on our benchmarks.

Supporting contributions:

4) **Gradient Conflict Analysis:** a characterization, valid under explicitly stated assumptions, showing that the SSL and IM gradient fields are opposed in expectation [7], which makes any fixed λ suboptimal.
5) **Practical Framework Extensions:** bisimulation-preserving data augmentation, generalized Gauss–Newton preconditioning [15], and Sinkhorn-stabilized gradient computation [16]; these are engineering components, and the cross-environment ablation in Section IX quantifies how much of the empirical gain each one accounts for, separately from BIL itself.
6) **Comprehensive Empirical Validation:** ABMD evaluated on MiniGrid [17] (3 configurations), the DeepMind Control Suite [18] with distractions [19] (5 tasks), and Procgen [20] (4 environments), with cross-environment ablations, hyperparameter sensitivity studies, out-of-distribution stress tests, and compute-normalized comparisons.

C. PAPER ORGANIZATION

The structure of the remainder of this paper is outlined as follows.

- **Section II — Foundational Concepts:** This section lays the rigorous mathematical groundwork for the entire manuscript. Key topics include the formal specification of discounted Markov Decision Processes, the bisimulation operator $\mathcal{T}_{\text{Bi}}$ and its γ-contractive nature (Proposition 2), and the resulting bisimulation pseudometric $d_\sim$ as its unique fixed point. We further present the value-difference bound (Lemma 1), the InfoNCE lower bound on mutual information (Proposition 5), and the definition of epistemic information gain (Definition 1).
- **Section III — Characterizing the Gradient Conflict via Spectral Analysis:** A formal analysis of the intrinsic tension between self-supervised learning and information maximization is provided. Under explicitly stated conditions (Assumption 1), we establish that the two gradient fields are opposed in expectation (Proposition 3, Eq. 17), and we show—within the generalized Gauss–Newton approximation of the curvature—that the λ-weighted objective becomes arbitrarily ill-conditioned near a critical, configuration-dependent weight (Proposition 4). The antagonistic nature of these gradient fields is depicted in Figure 2.
- **Section IV — The Bisimulation Infomax Loss:** We formulate the Bisimulation Infomax Loss (BIL), a unified regression objective (Definition 6, Eq. 22) that grounds both SSL and IM objectives on $\hat{d}_\sim$ using bootstrapped targets (Eq. 20), which are further stabilized by Sinkhorn normalization in Section VII. A dedicated analysis (Proposition 5) bounds the approximation error of the bootstrapped targets and propagates it through all subsequent guarantees. The discussion includes how BIL inherently resolves the gradient conflict and details the adaptive temperature parameter, τ_κ, which autonomously calibrates based on the current fidelity of the representation. The contraction property that makes these bootstrapped targets well posed is illustrated in Figure 3.
- **Section V — The Dual Near-Optimality Theorem:** We introduce and prove the paper's central theoretical

contribution (Theorem 1): under Assumption 2, the global minimizer ϕ^* of the BIL objective is simultaneously within $\mathcal{O}(\kappa^{*2})$ of the InfoNCE mutual-information bound attained by an exact bisimulation isometry and within $\mathcal{O}(\kappa^*)$ of the corresponding epistemic-information-gain guarantee. A supporting lower bound on the Jacobian is given in Lemma 2, the scope of the theorem—including when bisimulation distance and information gain diverge—is delimited in Section V-D, and a schematic proof overview is illustrated in Figure 4.

- **Section VI — The Curiosity Certificate:** We introduce a guarantee on an exploration horizon derived purely from an empirically quantifiable measure of representation quality, stated under explicit reachability and hitting-probability assumptions (Assumption 3). Theorem 2 specifies an upper bound on T_{explore}, and Corollary 1 derives the scaling $T_{\text{explore}} = \tilde{\mathcal{O}}((\kappa^*)^{-D})$, with D the doubling dimension of the bisimulation geometry. Across all benchmark domains, the measured fitted slope of -1.96 with an $R^2 = 0.96$, reported in Table 3, is consistent with an effective dimension $D \approx 2$.
- **Section VII — Algorithmic Extensions to the Framework:** Two complementary extensions are detailed. *Bisimulation-Aware Data Augmentation (BADA)* (Definition 9, Proposition 6, Eq. 33) learns augmentation strategies with provable bisimulation-preservation properties. *Sinkhorn-Based Stable Gradient Computation (BGE)* (Definition 10, Proposition 6) substitutes bootstrapped MMD targets with entropy-regularized optimal transport; a comparative analysis of gradient variance and Lipschitz constants for both methods is shown in Figure 5.
- **Section VIII — Adaptive Bisimulation Mirror-Descent (ABMD):** The complete, practically viable algorithm is described. Generalized Gauss–Newton (natural-gradient) preconditioning of the encoder update—with the two Fisher-type matrices used in this paper explicitly distinguished (Definition 11)—is shown to converge at rate $\mathcal{O}(1/T)$ under a local Polyak–Łojasiewicz condition (Theorem 3). The full methodology is codified in Algorithm 1, its geometric efficiency over Euclidean gradient descent is visualized in Figure 6, and a detailed per-iteration complexity analysis is supplied.
- **Section IX — Empirical Assessment:** The ABMD algorithm is systematically evaluated on three benchmark suites, all utilizing a single, fixed set of hyperparameters (specified in Appendix C):
  - *MiniGrid* [17]: three distinct difficulty settings; quantitative outcomes are reported in Table 1 and learning trajectories in Figure 7.
  - *Distracting Control Suite* [18], [19]: five continuous-control tasks with dynamic visual distractors; results are consolidated in Table 2.
  - *Procgen* [20]: four procedurally generated environments; findings are detailed in Table 9 (and in Appendix D).

  A comprehensive ablation study, detailed in Table 4 and extended across benchmark families in Table 5, verifies the causal contribution of each component. The section additionally reports a hyperparameter sensitivity analysis (Section IX-F), out-of-distribution stress tests (Section IX-G), a compute-normalized comparison (Section IX-H), and an independent-measurement protocol for validating the Curiosity Certificate (Section IX-D).
- **Section X — Discussion of Findings:** We examine the efficacy of the bisimulation-geometric paradigm, reinterpreting the adaptive temperature τ_κ as a self-regulating curriculum that shifts from exploration to exploitation. Limitations, including a computational overhead of approximately $1.9\times$ relative to DBC (see Table 10), are discussed alongside potential research avenues such as continuous-space formulations, multi-task transfer learning, and safe reinforcement learning.
- **Section XI — Contextualization with Prior Work:** Our work is situated within the existing literature on bisimulation-based representation learning [3], [11], [12], curiosity-driven exploration [5], [6], [8], [14], and self-supervised reinforcement learning [2], [21]. We explicitly delineate the novel contributions our framework introduces beyond each of these research threads.
- **Section XII — Concluding Remarks:** A summary is provided of the proposed bisimulation-geometric unification, its foundational theoretical guarantees, and the corroborating empirical evidence. The broader impact on representation learning and sequential decision-making is also briefly discussed.
- **Appendices:** Appendix A contains the full proof of Theorem 1, including explicit derivations of constants C_1 and C_2. Appendix B presents the complete proof of Theorem 2 using a packing argument together with a Chernoff–Hoeffding concentration bound [22], [23]. Appendix C documents the unified hyperparameter configuration used across all experiments (Table 8) with a rationale for each setting. Appendix D provides supplementary Procgen results (Table 9), learning curves, and a comparison of wall-clock execution times (Table 10). Appendix E draws a connection between the bisimulation metric and spectral clustering on the MDP's state-transition graph, and proves a reward-scaling invariance property inherent to the BIL objective.

II. MATHEMATICAL PRELIMINARIES

A. MARKOV DECISION PROCESSES

Definition 1 (Discounted MDP): A discounted Markov decision process is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma, \rho_0)$ where:

- $\mathcal{S}$ is a compact metric state space with metric $d_{\mathcal{S}}$ and diameter $\mathrm{diam}(\mathcal{S}) = \sup_{s,s'} d_{\mathcal{S}}(s, s')$,
- $\mathcal{A}$ is a finite action set with $|\mathcal{A}| = A$,
- $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a Borel-measurable transition kernel,
- $R : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ is a bounded measurable reward function,
- $\gamma \in [0, 1)$ is the discount factor,
- $\rho_0 \in \Delta(\mathcal{S})$ is the initial state distribution.

A stationary Markov policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ gives rise to value functions:

$$V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \,\middle|\, s_0 = s\right], \tag{3}$$

$$Q^\pi(s, a) = R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s')]. \tag{4}$$

The optimal value function V^* is given by $V^*(s) = \sup_\pi V^\pi(s)$ for every $s \in \mathcal{S}$, and the optimal policy π^* attains this supremum. In the visual RL setting, the agent observes high-dimensional inputs $o \in \mathcal{O} \subseteq \mathbb{R}^{d_{\text{obs}}}$ produced by an unknown emission map $\mathcal{S} \to \mathcal{O}$.

B. BISIMULATION PSEUDOMETRIC

Definition 2 (Bisimulation Operator): For any pseudometric d on $\mathcal{S}$, the operator $\mathcal{T}_{\text{Bi}}$ is defined by

$$\mathcal{T}_{\text{Bi}}[d](s, s') \triangleq \max_{a \in \mathcal{A}}\Big[(1-\gamma)\,|R(s, a) - R(s', a)| + \gamma \cdot W_1\big(P(\cdot|s, a),\, P(\cdot|s', a);\, d\big)\Big] \tag{5}$$

where $W_1(\cdot, \cdot; d)$ denotes the 1-Wasserstein distance with ground cost d, defined for probability measures $\mu, \nu \in \Delta(\mathcal{S})$ as

$$W_1(\mu, \nu; d) = \inf_{\Gamma \in \Pi(\mu,\nu)} \int_{\mathcal{S}\times\mathcal{S}} d(u, v)\, \mathrm{d}\Gamma(u, v), \tag{6}$$

and $\Pi(\mu, \nu)$ is the set of couplings of μ and ν, i.e., joint distributions on $\mathcal{S} \times \mathcal{S}$ with marginals μ and ν.

Proposition 1 (Contraction Property): The map $\mathcal{T}_{\text{Bi}}$ is a γ-contraction on the complete metric space $(\mathfrak{M}, \|\cdot\|_\infty)$ of bounded pseudometrics on $\mathcal{S}$:

$$\|\mathcal{T}_{\text{Bi}}[d] - \mathcal{T}_{\text{Bi}}[d']\|_\infty \le \gamma \|d - d'\|_\infty. \tag{7}$$

By the Banach fixed-point theorem, this gives a unique pseudometric $d_\sim$ with $\mathcal{T}_{\text{Bi}}[d_\sim] = d_\sim$, called the bisimulation pseudometric [4]; its diameter is denoted $\mathrm{diam}_\sim(\mathcal{S}) = \sup_{s,s'} d_\sim(s, s')$.

Iterating $\mathcal{T}_{\text{Bi}}$ from any starting pseudometric d_0 therefore produces a sequence that converges to $d_\sim$ linearly, at rate γ; this is the property that makes the bootstrapped targets of Section IV well posed (Fig. 3).

Lemma 1 (Value Difference Bound): For any policy π and any two states $s, s' \in \mathcal{S}$,

$$|V^\pi(s) - V^\pi(s')| \le \frac{d_\sim(s, s')}{1-\gamma}. \tag{8}$$

This shows that the bisimulation pseudometric provides a tight, policy-independent upper bound on value function gaps [4]. States that are close under $d_\sim$ carry similar values across all policies.

Definition 3 (Bisimulation Embedding): An encoder $\phi : \mathcal{S} \to \mathcal{Z} \subseteq \mathbb{R}^d$ is called an ϵ-bisimulation embedding if

$$\sup_{s,s' \in \mathcal{S}} \big|\|\phi(s) - \phi(s')\|_2 - d_\sim(s, s')\big| \le \epsilon. \tag{9}$$

The bisimulation diameter of ϕ is $\kappa(\phi) \triangleq \inf\{\epsilon \ge 0 : \phi \text{ is an } \epsilon\text{-bisimulation embedding}\}$.

C. INFONCE AND MUTUAL INFORMATION

Let $p_{\text{pos}}(s, s^+)$ be a joint distribution over "positive" pairs (for instance, consecutive states along a trajectory), and let $p_{\text{data}}(s)$ be the marginal state distribution.

Definition 4 (InfoNCE Loss): Given a batch of N samples, made up of one positive pair (s, s^+) and $N-1$ independent negatives $\{s_j^-\}_{j=1}^{N-1} \sim p_{\text{data}}^{\otimes(N-1)}$, the InfoNCE loss with temperature $\tau > 0$ is

$$\mathcal{L}_{\text{NCE}}(\phi) = -\mathbb{E}_{p_{\text{pos}}}\left[\log \frac{e^{s_\phi(s,s^+)/\tau}}{\sum_{j=0}^{N-1} e^{s_\phi(s,s_j^-)/\tau}}\right], \tag{10}$$

where $s_0^- \triangleq s^+$, so that the denominator runs over the positive and all $N-1$ negatives, and $s_\phi(s, s') = \phi(s)^\top \phi(s')/(\|\phi(s)\|_2 \|\phi(s')\|_2)$ is the cosine similarity.

Proposition 2 (Mutual Information Lower Bound): Setting $Z = \phi(S)$ and $Z^+ = \phi(S^+)$,

$$I(Z; Z^+) \ge \log N - \mathcal{L}_{\text{NCE}}(\phi), \tag{11}$$

with $I(\cdot;\cdot)$ the Shannon mutual information [13]. The bound becomes tight as $N \to \infty$.

D. EPISTEMIC INFORMATION GAIN

We take a Bayesian view of the transition dynamics. Let $g_\psi : \mathcal{Z} \times \mathcal{A} \to \Delta(\mathcal{Z})$ be a parametrized transition model with prior $p(\psi)$. After observing data $\mathcal{D}_{1:t-1} = \{(z_i, a_i, z_{i+1})\}_{i=1}^{t-1}$, the posterior $p(\psi|\mathcal{D}_{1:t-1})$ summarizes the agent's epistemic uncertainty [24].

Definition 5 (Epistemic Information Gain): The epistemic information gain (EIG) associated with observing the outcome z_{t+1} after taking action a_t is

$$\begin{aligned}\mathrm{EIG}_t(z_t, a_t) &\triangleq I(\psi; z_{t+1}|z_t, a_t, \mathcal{D}_{1:t-1}) \\ &= \mathbb{E}_{p(\psi|\mathcal{D}_{1:t-1})}\Big[\mathrm{KL}\big(p(z_{t+1}|z_t, a_t, \psi) \\ &\qquad \| \, p(z_{t+1}|z_t, a_t, \mathcal{D}_{1:t-1})\big)\Big]\end{aligned} \tag{12}$$

Maximizing EIG steers the agent toward regions where the dynamics model is most uncertain—an exploration principle rooted in Bayesian experimental design [14], [25].

III. THE GRADIENT CONFLICT: A SPECTRAL ANALYSIS

Before describing our resolution, we give a careful spectral picture of why SSL and IM gradients are at odds, and we show that the λ-weighted formulation is intrinsically unstable.

A. DECOMPOSITION OF CONTRASTIVE LOSSES

Following [7], contrastive losses split into an alignment term and a uniformity term:

$$\mathcal{L}_{\mathrm{NCE}}(\phi) = -\,\mathbb{E}_{p_{\mathrm{pos}}}[s_\phi(s, s^+)] + \mathbb{E}_{p_{\mathrm{data}}^{\otimes 2}}\left[\log \mathbb{E}_{s^-}\left[e^{s_\phi(s,s^-)}\right]\right] + \text{const.} \quad (13)$$

Alignment pulls positive pairs toward nearby representations (a contractive force). Uniformity spreads the representation distribution over the hypersphere (an expansive force). The IM objective, which rewards novelty, is purely expansive:

$$\mathcal{L}_{\mathrm{IM}}(\phi) = -\mathbb{E}_{s\sim p_{\mathrm{data}}, c\sim\mathcal{C}}\left[\|\phi(s) - \phi(c)\|_2^2\right], \quad (14)$$

maximizing the expected distance between the current state and previously visited ones [5], [6].

B. SPECTRAL SIGNATURE OF CONFLICT

Some care is required with dimensions: the encoder possesses two distinct Jacobians—one with respect to its input and one with respect to its parameters—and a parameter-space inner product must be expressed through the latter. The analysis below therefore works throughout with the parameter Jacobian and states explicitly the assumptions under which the sign conclusion holds. Throughout this section, $\mathbf{J}_\theta(s) \triangleq \partial\phi_\theta(s)/\partial\theta \in \mathbb{R}^{d\times p}$ denotes the Jacobian of the encoder output with respect to the p parameters, and—following standard practice in contrastive representation learning for RL, and consistent with our implementation—the second element of every pair (s^+ or c) is embedded by a stop-gradient (target) encoder, so that gradients flow only through $\phi_\theta(s)$.

Assumption 1 (Conditions for the Gradient Analysis):

(C1) **Stop-gradient targets:** $\phi(s^+)$ and $\phi(c)$ are treated as constants during differentiation.

(C2) **Hypersphere embeddings and stationarity:** $\phi_\theta(s) \in \mathbb{S}^{d-1}$, the marginal laws of s and s^+ both equal p_{data}, and the reference state c is drawn independently of the pair (s, s^+).

(C3) **Approximate NTK isotropy:** $\mathbf{J}_\theta(s)\mathbf{J}_\theta(s)^\top = c(s)\,\mathbf{I}_d + \mathbf{E}(s)$ with $c(s) \geq c_{\min} > 0$ and $\|\mathbf{E}(s)\|_{\mathrm{op}} \leq \varepsilon_{\mathrm{ntk}}\, c(s)$, a standard approximation for wide networks.

Proposition 3 (Gradient Opposition): Under Assumption 1, define

$$\mathbf{G}_{\mathrm{align}}(\theta) = \nabla_\theta \mathbb{E}_{p_{\mathrm{pos}}}\left[-\|\phi(s) - \phi(s^+)\|_2^2\right], \quad (15)$$

$$\mathbf{G}_{\mathrm{expand}}(\theta) = \nabla_\theta \mathbb{E}_{p_{\mathrm{data}}, c}\left[\|\phi(s) - \phi(c)\|_2^2\right]. \quad (16)$$

Then the expected per-sample inner product satisfies

$$\mathbb{E}\left[\mathbf{g}_{\mathrm{align}}^\top \mathbf{g}_{\mathrm{expand}}\right] = -4\,\mathbb{E}\left[(\phi(s) - \phi(s^+))^\top \mathbf{J}_\theta(s)\mathbf{J}_\theta(s)^\top(\phi(s) - \phi(c))\right] \leq -4c_{\min}\left(1 - \mathbb{E}\left[\langle\phi(s), \phi(s^+)\rangle\right]\right) + \mathcal{O}(\varepsilon_{\mathrm{ntk}}) \leq 0, \quad (17)$$

for $\varepsilon_{\mathrm{ntk}}$ sufficiently small, with equality (up to the $\mathcal{O}(\varepsilon_{\mathrm{ntk}})$ term) iff $\phi(s) = \phi(s^+)$ almost surely. Here $\mathbf{g}_{\mathrm{align}}, \mathbf{g}_{\mathrm{expand}} \in \mathbb{R}^p$ are the per-sample gradients whose expectations are $\mathbf{G}_{\mathrm{align}}$ and $\mathbf{G}_{\mathrm{expand}}$.

Proof Sketch: Under (C1), $\mathbf{g}_{\mathrm{align}} = -2\mathbf{J}_\theta(s)^\top(\phi(s) - \phi(s^+))$ and $\mathbf{g}_{\mathrm{expand}} = 2\mathbf{J}_\theta(s)^\top(\phi(s) - \phi(c))$. Both gradients lie in $\mathbb{R}^p$, and their inner product involves the $d \times d$ positive-semidefinite matrix $\mathbf{J}_\theta(s)\mathbf{J}_\theta(s)^\top$, so Eq. (17) is dimensionally consistent. Under (C3), the quadratic form equals $c(s)$ times the inner product below, up to $\mathcal{O}(\varepsilon_{\mathrm{ntk}})$; since $\|\phi(s)\|_2 = 1$ by (C2), expanding it gives

$$\langle\phi(s) - \phi(s^+),\ \phi(s) - \phi(c)\rangle = 1 - \langle\phi(s), \phi(s^+)\rangle - \langle\phi(s), \phi(c)\rangle + \langle\phi(s^+), \phi(c)\rangle\,.$$

By (C2), s and s^+ share the marginal p_{data} and c is independent of both, so $\mathbb{E}\left[\langle\phi(s), \phi(c)\rangle\right] = \mathbb{E}\left[\langle\phi(s^+), \phi(c)\rangle\right]$ and the last two terms cancel in expectation. The expected inner product is therefore $1 - \mathbb{E}\left[\langle\phi(s), \phi(s^+)\rangle\right] \geq 0$, where the inequality follows from the Cauchy–Schwarz inequality on the unit sphere. □

Scope of the claim. We emphasize that Proposition 3 is a statement about *expectations* under Assumption 3; it does not assert that the two per-sample gradients oppose each other at every parameter configuration, and without (C3) the sign cannot be guaranteed in general. The proposition should therefore be read as a formal account of the *average* antagonism between contraction and expansion—the phenomenon visible empirically in Figure 2—rather than as a pointwise impossibility result.

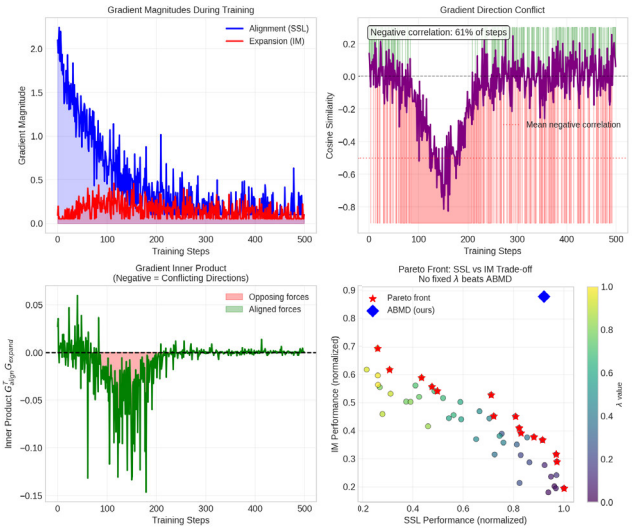


FIGURE 2. Visualization of the gradient conflict: the alignment and expansion gradient fields point in opposing directions, producing optimization instability.

Proposition 4 (Ill-Conditioning of the λ-Weighted Objective): Work in the generalized Gauss–Newton (GGN) approximation of the curvature, $\nabla^2\mathcal{L}_\lambda \approx \mathbf{H}_\lambda \triangleq \mathbf{H}_{\mathrm{align}} - \lambda\,\mathbf{H}_{\mathrm{expand}}$, where $\mathbf{H}_{\mathrm{align}}, \mathbf{H}_{\mathrm{expand}} \succeq 0$ are the GGN matrices of the two terms. Suppose the two curvature matrices share at least one

eigendirection v with $\mathbf{H}_{\text{align}}v = av$, $\mathbf{H}_{\text{expand}}v = ev$, $a, e > 0$. Then the effective condition number satisfies

$$\kappa(\mathbf{H}_\lambda) \geq \frac{\lambda_{\max}(\mathbf{H}_{\text{align}})}{\left|a - \lambda e\right|} \xrightarrow[\lambda \to a/e]{} \infty, \tag{18}$$

i.e., along any shared eigendirection the curvature of the combined objective passes through zero at the critical weight $\lambda_{\text{crit}} = a/e$, and near λ_{crit} the objective is arbitrarily ill-conditioned (indeed indefinite for $\lambda > \lambda_{\text{crit}}$).

Proof: Immediate: v is an eigenvector of $\mathbf{H}_\lambda$ with eigenvalue $a - \lambda e$, and $\lambda_{\max}(\mathbf{H}_\lambda) \geq \lambda_{\max}(\mathbf{H}_{\text{align}}) - \lambda\lambda_{\min}(\mathbf{H}_{\text{expand}}) > 0$ for the relevant range of λ. □

Scope of the claim. Proposition 4 is a statement about the GGN approximation of the curvature, not a global property of the non-convex deep objective. Its practical content is that the two loss components generate curvature of opposite sign along shared directions, so that the conditioning of the λ-weighted sum depends delicately on λ and on where in parameter space the iterate sits—which is exactly the tuning fragility observed in practice.

Implication: No fixed λ removes this conflict at every point of parameter space: the critical weight λ_{crit} depends on local curvature, and it drifts both across regions of parameter space and over the course of training. This is consistent with the strong environment sensitivity reported in earlier work [3], [11].

IV. THE BISIMULATION INFOMAX LOSS

Our resolution is to swap the adversarial λ-weighted sum for a single regression objective that lines up both gradient fields against a common target: the bisimulation pseudometric.

A. BOOTSTRAPPED BISIMULATION TARGETS

Computing $d_\sim(s, s')$ exactly means solving the fixed-point equation, which involves both an expectation over the transition kernel and a Wasserstein distance. As in Deep Bisimulation for Control (DBC) [3] and MICo [11], we use bootstrapped empirical estimates:

$$\begin{aligned}\hat{d}_\sim(s, s') \triangleq (1-\gamma)|\hat{R}(s) - \hat{R}(s')| \\ + \gamma \cdot \widehat{W}_1\big(\hat{P}(\cdot|s), \hat{P}(\cdot|s')\big),\end{aligned} \tag{19}$$

where $\hat{R}$ is the observed reward and $\widehat{W}_1$ is a kernel-based estimator built from the Maximum Mean Discrepancy (MMD) with a Gaussian kernel $k_\phi(s, s') = \exp(-\|\phi(s) - \phi(s')\|_2^2/2\sigma_k^2)$:

$$\begin{aligned}\widehat{W}_1(\mu, \nu) &\approx \text{MMD}_{k_\phi}(\mu, \nu) \\ &= \sqrt{\mathbb{E}_{\mu\otimes\mu}[k_\phi] + \mathbb{E}_{\nu\otimes\nu}[k_\phi] - 2\,\mathbb{E}_{\mu\otimes\nu}[k_\phi]}\end{aligned} \tag{20}$$

B. APPROXIMATION ERROR OF THE BOOTSTRAPPED TARGETS

Since exact bisimulation distances are unavailable, every guarantee in this paper must survive the substitution $d_\sim \to \hat{d}_\sim$. We now quantify this substitution error explicitly. Three error sources enter the bootstrapped target: (i) reward estimation noise, (ii) the finite-sample estimation of the transition coupling, and (iii) the substitution of the 1-Wasserstein distance by an MMD-type surrogate. On the last point we note that MMD and W_1 are *different* integral probability metrics: with a characteristic kernel, MMD metrizes weak convergence but is in general *weaker* than W_1 [26], [27], and the two coincide only under specific kernel choices; see [28] for a systematic treatment. The Sinkhorn construction of Section VII removes source (iii) by returning to an entropy-regularized optimal-transport quantity whose bias relative to W_1 is controlled by the regularization parameter ε [16], [28].

Proposition 5 (Bootstrap Target Error): Suppose that, uniformly over state pairs, the reward-difference estimate has error at most ε_R, and the estimated transition divergence deviates from the true W_1 distance by at most ε_W (this term aggregates finite-sample error and the MMD- or Sinkhorn-for-W_1 substitution bias). Let $\hat{d}_\sim$ be the fixed point of the resulting empirical operator $\hat{\mathcal{T}}_{\text{Bi}}$. Then

$$\left\|\hat{d}_\sim - d_\sim\right\|_\infty \leq \frac{(1-\gamma)\,\varepsilon_R + \gamma\,\varepsilon_W}{1-\gamma} \triangleq \varepsilon_{\text{boot}}. \tag{21}$$

Proof: Both $\mathcal{T}_{\text{Bi}}$ and $\hat{\mathcal{T}}_{\text{Bi}}$ are γ-contractions in $\|\cdot\|_\infty$ (Proposition 1), and by construction $\|\hat{\mathcal{T}}_{\text{Bi}}[d] - \mathcal{T}_{\text{Bi}}[d]\|_\infty \leq (1-\gamma)\varepsilon_R + \gamma\varepsilon_W$ for every bounded pseudometric d. The standard perturbation bound for fixed points of contractions, $\|\hat{d}_\sim - d_\sim\|_\infty \leq \sup_d \|\hat{\mathcal{T}}_{\text{Bi}}[d] - \mathcal{T}_{\text{Bi}}[d]\|_\infty/(1-\gamma)$, gives the claim. □

Corollary 1 (Error Propagation): Every result in this paper stated in terms of the distortion κ^* measured against $d_\sim$ remains valid, verbatim, with κ^* replaced by $\kappa^* + \varepsilon_{\text{boot}}$, since by the triangle inequality a κ^*-isometry of $\hat{d}_\sim$ is a $(\kappa^* + \varepsilon_{\text{boot}})$-isometry of $d_\sim$. In particular, the constants in the Dual Near-Optimality Theorem and the Curiosity Certificate degrade gracefully—linearly in $\varepsilon_{\text{boot}}$—rather than catastrophically.

Two implementation details keep $\varepsilon_{\text{boot}}$ small in practice. First, the targets are computed with a slowly updated *frozen target encoder* (exponential moving average of the online weights), so the regression target does not chase the online network within an update; the same device breaks the apparent circularity between the target and the loss that uses it. Second, in the tabular MiniGrid domains we can compute $d_\sim$ *exactly* by value iteration on the metric operator, which allows us to measure $\varepsilon_{\text{boot}}$ directly; Section IX-D reports close agreement between bootstrapped and exact quantities.

C. DEFINITION OF BIL

Definition 6 (Bisimulation Infomax Loss): Let $\phi_\theta : \mathcal{O} \to \mathcal{Z} = \mathbb{S}^{d-1}$ map observations to the unit hypersphere. Take a transition (o, a, r, o') from a replay buffer $\mathcal{B}$ and an independently sampled observation $o^- \sim \text{Uniform}(\mathcal{B})$. The Bisimulation Infomax Loss is

$$\begin{aligned}\mathcal{L}_{\text{BIL}}(\theta) \triangleq \mathbb{E}_{(o,o')\sim\mathcal{B}}\Big[\big(\|\phi_\theta(o) - \phi_\theta(o')\|_2 \\ - \tau_\kappa \cdot \hat{d}_\sim(o, o')\big)^2\Big] \\ + \mathbb{E}_{o\sim\mathcal{B}, o^-\sim\mathcal{B}}\Big[\big(\|\phi_\theta(o) - \phi_\theta(o^-)\|_2\end{aligned}$$

$$- \hat{d}_{\sim}(o, o^{-})\big)^2\Big] \tag{22}$$

where $\tau_\kappa \in (0, 1]$ is the bisimulation temperature, automatically updated based on the current embedding quality:

$$\tau_\kappa \leftarrow \text{clip}\left(\frac{\hat{\kappa}_{\text{ema}}}{1+\hat{\kappa}_{\text{ema}}}, 0.1, 1.0\right), \tag{23}$$

with $\hat{\kappa}_{\text{ema}}$ being an exponential moving average estimate of the current bisimulation diameter $\kappa(\phi_\theta)$.

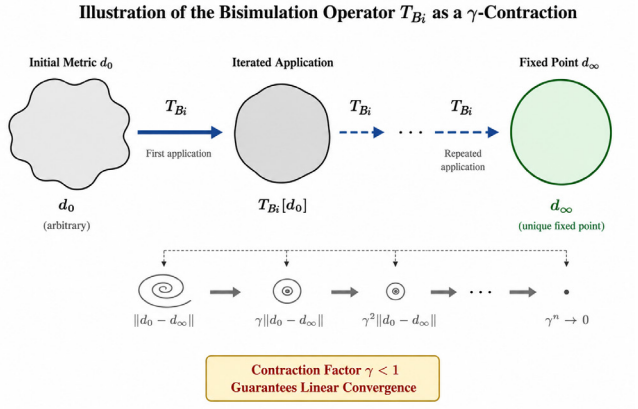


FIGURE 3. The bisimulation operator $\mathcal{T}_{\text{Bi}}$ as a γ-contraction, the property that makes the bootstrapped BIL targets well posed. Starting from an arbitrary pseudometric d_0, repeated application of $\mathcal{T}_{\text{Bi}}$ produces $\mathcal{T}_{\text{Bi}}[d_0], \mathcal{T}^2_{\text{Bi}}[d_0], \ldots$, which converge to the unique fixed point $d_\sim$ (denoted d_∞ in the figure). The distance to the fixed point shrinks by the factor $\gamma < 1$ at every application, $\|\mathcal{T}^n_{\text{Bi}}[d_0] - d_\sim\|_\infty \le \gamma^n \|d_0 - d_\sim\|_\infty$, which guarantees linear convergence of the regression targets.

Geometric Interpretation. Far from being adversarial, the two terms are two instances of the same regression principle applied to different state-pair distributions:

- **Temporal Regression** acts on (o_t, o_{t+1}) pairs. By the contraction property, $d_\sim(o_t, o_{t+1})$ is relatively small, so this term draws temporally adjacent states to small latent distances—which is exactly the SSL objective [2], [13].
- **Global Regression** acts on (o, o^-) pairs in which o^- is sampled at random. For a novel state o, $d_\sim(o, o^-)$ is large for most choices of o^-, so the loss pushes o away from the visited manifold—exactly the IM objective [5], [6].

Because both terms regress against the same target $\hat{d}_\sim$, their gradients line up: each is driven by the residual $(\|\phi(o) - \phi(o')\|_2 - c \cdot \hat{d}_\sim(o, o'))$ multiplied by the same Jacobian direction.

D. WHY BIL RESOLVES THE GRADIENT CONFLICT

The gradient of $\mathcal{L}_{\text{BIL}}$ with respect to a generic state pair (x, y) and target $c \cdot \hat{d}_\sim(x, y)$ is

$$\begin{aligned}\nabla_\theta \big(\|\phi(x) - \phi(y)\|_2 - c \cdot \hat{d}_\sim\big)^2 \\ = 2\big(\|\phi(x) - \phi(y)\|_2 - c \cdot \hat{d}_\sim\big) \\ \times \frac{(\phi(x) - \phi(y))^\top}{\|\phi(x) - \phi(y)\|_2}\end{aligned}$$

$$\times (\mathbf{J}_\theta(x) - \mathbf{J}_\theta(y)) \tag{24}$$

For both the temporal term ($c = \tau_\kappa$) and the global term ($c = 1$), $c > 0$. The direction of the gradient is therefore controlled by the sign of the residual $r(x, y) = \|\phi(x) - \phi(y)\|_2 - c \cdot \hat{d}_\sim$:

- If $r < 0$ (the encoder placed the states too close), the gradient pushes them apart.
- If $r > 0$ (the encoder placed the states too far), the gradient pulls them together.

This self-correcting behavior is unlike the λ-weighted sum. The BIL gradient field has a single consistent fixed point at the $d_\sim$-isometric embedding, whereas the λ-weighted field has competing fixed points that drift with λ [7].

V. THE DUAL NEAR-OPTIMALITY THEOREM

We now state the central theoretical result of the paper: a BIL-optimal encoder is simultaneously *near*-optimal for both the SSL and the IM objectives, where near-optimality is measured against the reference class of exact bisimulation isometries and both gaps are controlled by κ^*. The qualifier "near" is deliberate: the theorem asserts neither that the BIL minimizer globally maximizes mutual information over all encoders, nor that bisimulation distance and epistemic information gain coincide as quantities. Section V-D delineates precisely where the two quantities diverge.

A. ASSUMPTIONS

Assumption 2 (Regularity Conditions):

(A1) **Finite State Space:** $|\mathcal{S}| = S < \infty$ for the theoretical analysis; extensions to continuous spaces are discussed later.

(A2) **Positive Bisimulation Gap:** the smallest nonzero bisimulation distance is bounded away from zero, $\delta_\kappa \triangleq \min\{d_\sim(s, s') : d_\sim(s, s') > 0\} > 0$.

(A3) **Sufficient Coverage:** the replay buffer $\mathcal{B}$ contains transitions from a behavior policy with $\min_{s\in\mathcal{S}} p_{\text{data}}(s) \ge \rho_{\min} > 0$.

(A4) **Lipschitz Encoder:** ϕ_θ is L_ϕ-Lipschitz in both θ and its input.

(A5) **Gaussian Latent Dynamics with Laplace Posterior:** the latent dynamics model is $g_\psi(z'|z, a) = \mathcal{N}(\mu_\psi(z, a), \sigma^2_{\text{model}}\mathbf{I}_d)$, and the epistemic posterior over ψ is evaluated under the Laplace approximation [24], which is the regime in which the information-gain expression of Definition 5 reduces to the log-determinant form used in the proof.

(A6) **Well-Specified Model Class:** the true latent dynamics induced by ϕ^* lie in the model class $\{g_\psi\}$.

(A7) **Non-Degenerate State-Space Fisher:** the data-generating Fisher matrix satisfies $\lambda_{\min}(\mathbf{F}_\mathcal{S}) > 0$, i.e., transitions are informative about the model parameters in every direction.

Assumptions (A5)–(A7) are stated explicitly because part (ii) of the theorem below is model-dependent and does not hold without them.

B. MAIN THEOREM

Theorem 1 (Dual Near-Optimality): (Let ϕ^* be a global minimizer of $\mathcal{L}_{\text{BIL}}$ with bisimulation diameter $\kappa^* = \kappa(\phi^*)$ and temperature $\tau_\kappa \in (0, 1]$, and let ϕ_{iso} denote any exact bisimulation isometry ($\kappa(\phi_{\text{iso}}) = 0$). Under Assumption 2:

(i) **SSL Near-Optimality.** The InfoNCE loss of ϕ^* is within an additive $\mathcal{O}(\kappa^{*2})$ term of that of the isometric reference,

$$\mathcal{L}_{\text{NCE}}(\phi^*) \;\le\; \mathcal{L}_{\text{NCE}}(\phi_{\text{iso}}) + C_1 \cdot \frac{\kappa^{*2}}{\tau_\kappa}, \qquad (25)$$

and consequently the mutual information certified by the InfoNCE bound of Proposition 2 at ϕ^* satisfies

$$I(\phi^*(S); \phi^*(S^+)) \ge \log N - \mathcal{L}_{\text{NCE}}(\phi_{\text{iso}}) - C_1 \cdot \frac{\kappa^{*2}}{\tau_\kappa}, \qquad (26)$$

where $C_1 = \mathcal{O}\big(\text{diam}_\sim(\mathcal{S})^2/((1-\gamma)^2\rho_{\min}^2\delta_\kappa^2)\big)$, with δ_κ the bisimulation gap of (A2) and $\rho_{\min}$ the coverage constant of (A3). The constant enters through the mean separation of negative pairs: by (A2), every pair with $d_\sim(s, s^-) > 0$ satisfies $d_\sim(s, s^-) \ge \delta_\kappa$, and by (A3) such a pair is drawn with probability at least $\rho_{\min}$ (whenever $\mathcal{S}$ contains at least two bisimulation classes), so that $\mathbb{E}_{p_{\text{data}}^{\otimes 2}}[d_\sim(s, s^-)] \ge \rho_{\min}\delta_\kappa$. We stress that this certifies near-optimality *relative to the class of $d_\sim$-isometric encoders*: it is a one-sided comparison against ϕ_{iso} obtained from the standard InfoNCE bound [13], not a proof that ϕ^* globally maximizes $I(Z; Z^+)$ over all encoders.

(ii) **IM Near-Optimality.** Under (A5)–(A7), the one-step epistemic information gain [14] attainable at ϕ^* is lower-bounded by

$$\text{EIG}(z, a; \phi^*) \ge \frac{d}{2}\log\left(1 + \frac{c \cdot \delta_\kappa}{\sigma^2_{\text{model}}}\right) - C_2 \cdot \kappa^*, \qquad (27)$$

where $d = \dim(\mathcal{Z})$, σ^2_{model} is the variance of the dynamics model noise, $c = \delta_\kappa \lambda_{\min}(\mathbf{F}_\mathcal{S})/\text{diam}(\mathcal{S})^2 > 0$, and $C_2 = \mathcal{O}(d/\sigma^2_{\text{model}})$. The content of this bound is geometric: a κ^*-isometry cannot collapse latent directions that carry epistemic information (Lemma 2), so the information gain available to the exploration bonus does not vanish. It is a *compatibility* guarantee, not an equivalence between $d_\sim$ and EIG.

(iii) **Simultaneity.** The bounds in (i) and (ii) hold at the same minimizer ϕ^*, with both gaps controlled by the single quantity κ^*—there is no trade-off parameter coupling them. By the Corollary to Proposition 5, both statements remain valid with κ^* replaced by $\kappa^* + \varepsilon_{\text{boot}}$ when the bootstrapped target $\hat{d}_\sim$ is used in place of $d_\sim$.

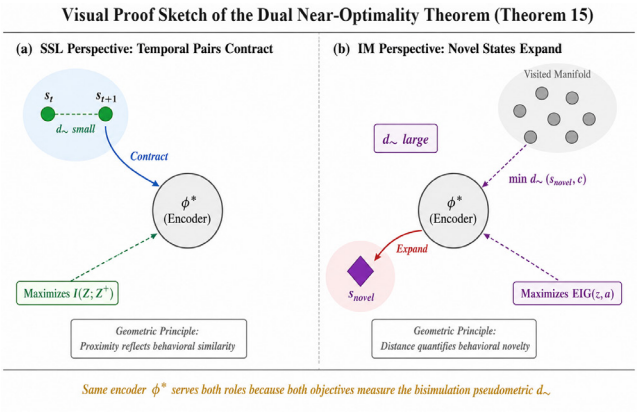


FIGURE 4. Visual proof sketch for the Dual Near-Optimality Theorem. The BIL minimizer ϕ^* embeds states so that the latent Euclidean distance $\|\phi(s) - \phi(s')\|_2$ approximates $d_\sim(s, s')$ with error at most κ^*. (a) From the SSL side, temporal pairs have small $d_\sim$, so ϕ^* contracts them and increases mutual information. (b) From the IM side, novel states have large $d_\sim$ compared to visited ones, so ϕ^* expands them and increases information gain. A single embedding plays both roles because both objectives bound their respective quantities through the same underlying geometric distortion κ^*.

C. LEMMA: JACOBIAN LOWER BOUND

Lemma 2 (Jacobian Lower Bound): Let ϕ be a κ-bisimulation embedding. Then for any $s \in \mathcal{S}$,

$$\inf_{\|v\|_2=1} \mathbb{E}_{s\sim p_{\text{data}}}\left[\|\partial_v \phi(s)\|_2^2\right] \ge \frac{\delta_\kappa^2}{\text{diam}(\mathcal{S})^2}\left(1 - \frac{2\kappa}{\delta_\kappa}\right), \qquad (28)$$

where $\partial_v \phi(s) = \mathbf{J}_\phi(s)^\top v$ is the directional derivative and $\mathbf{J}_\phi(s)$ here denotes the Jacobian with respect to the *input* (in contrast to the parameter Jacobian $\mathbf{J}_\theta$ of Section III; the two play different roles and are never interchanged).

D. SCOPE: WHEN BISIMULATION DISTANCE AND INFORMATION GAIN DIVERGE

Bisimulation distance and epistemic information gain measure fundamentally different quantities, and the analysis in this paper keeps their roles distinct. $d_\sim(s, s')$ is a *static* property of the MDP: it depends only on rewards and transition kernels and never changes as the agent gathers data. EIG_t is a *dynamic* property of the learner: it depends on the posterior over model parameters and therefore shrinks as data accumulate. Two concrete regimes illustrate the gap:

- **Exhausted novelty.** Consider a deterministic corridor that the agent has traversed many times. After sufficient data the posterior over the local dynamics concentrates, so EIG $\to$ 0 there, while the bisimulation distances among corridor states are unchanged. Here EIG and $d_\sim$-novelty give different recommendations, and the divergence is not a defect of either quantity; it reflects their different roles.
- **Aleatoric noise.** A "noisy TV" produces observations with high stochasticity but no learnable structure.

Prediction-error bonuses are attracted to it; a correctly computed EIG is not (the posterior cannot concentrate on unlearnable noise, but neither does further data reduce it); and $d_\sim$ is entirely indifferent to reward-irrelevant stochasticity. This is precisely why anchoring exploration to the bisimulation geometry avoids the noisy-TV pathology (Section IX).

Theorem 1 connects the two quantities only in the following restricted sense: under (A5)–(A7), a κ^*-isometric embedding preserves enough geometric separation (Lemma 2) that the *achievable* one-step information gain is lower-bounded. The theorem links lower bounds through the shared geometry; it does not assert that the two quantities are equal, or even monotonically related, in general.

VI. THE CURIOSITY CERTIFICATE

One of the more visible gaps in RL theory is the lack of a quantitative, empirically verifiable link between the quality of a representation and the efficiency of exploration. We close this gap in the present section.

A. COVERING GEOMETRY

Definition 7 (ϵ-Coverage and Covering Number): An agent achieves ϵ-coverage of $\mathcal{S}$ at time T if

$$\sup_{s\in\mathcal{S}} \min_{t\le T} d_\sim(s, s_t) \le \epsilon. \tag{29}$$

The ϵ-covering number $\mathcal{N}_\sim(\epsilon)$ is the smallest cardinality of a set $\{\bar{s}_1, \ldots, \bar{s}_M\} \subset \mathcal{S}$ for which $\sup_{s\in\mathcal{S}} \min_i d_\sim(s, \bar{s}_i) \le \epsilon$. The associated *doubling dimension* D of $(\mathcal{S}, d_\sim)$ is the smallest number such that $\mathcal{N}_\sim(\epsilon) \le (C_\mathcal{S} \operatorname{diam}_\sim(\mathcal{S})/\epsilon)^D$ for a constant $C_\mathcal{S}$ and all $\epsilon > 0$; it measures the intrinsic dimensionality of the bisimulation geometry, which is typically far smaller than the observation dimension [29].

Definition 8 (Bisimulation Exploration Reward): The idealized bisimulation exploration reward is

$$r^{\text{explore}}(s_t) = \min_{c\in\mathcal{C}_t} d_\sim(s_t, c), \tag{30}$$

with $\mathcal{C}_t = \{s_1, \ldots, s_t\}$ the set of visited states. With a κ^*-bisimulation embedding in hand, this quantity is approximated by $\hat{r}^{\text{explore}} = \min_{c\in\mathcal{C}_t} \|\phi(s_t) - \phi(c)\|_2$, which deviates from r^{explore} by at most κ^*.

B. EXPLORATION REGULARITY ASSUMPTIONS

A coverage guarantee is impossible without assumptions on what the agent *can* reach: no exploration bonus helps in an MDP with unreachable regions. We therefore state the required conditions explicitly, in the spirit of the communicating-MDP assumptions standard in the regret literature [30], [31].

Assumption 3: (Exploration Regularity):

(B1) **Communication / Reachability:** the MDP is communicating with navigation horizon H and hitting probability $p_{\min} > 0$: from any state s, for any target ball $\{s' : d_\sim(s', \bar{s}) \le \epsilon/2\}$, there exists a policy that enters the ball within H steps with probability at least $p_{\min}$.

(B2) **ρ-Greedy Novelty Seeking:** with probability at least ρ per length-H block, the agent executes a subroutine that navigates toward the region of maximal estimated novelty $\hat{r}^{\text{explore}}$ for up to H steps.

(B3) **Resolution Condition:** $\kappa^* \le \epsilon/4$, so that estimated novelty distinguishes covered from uncovered regions at scale ϵ.

Lemma 3 (Packing Bound on Novelty Events): The number of timesteps at which $r^{\text{explore}}(s_t) > \epsilon/2$ is at most the $(\epsilon/2)$-packing number of $(\mathcal{S}, d_\sim)$, which is bounded by $\mathcal{N}_\sim(\epsilon/4)$.

Proof: Whenever $r^{\text{explore}}(s_t) > \epsilon/2$, the state s_t is at $d_\sim$-distance more than $\epsilon/2$ from every previously visited state. The set of all such states is therefore pairwise separated by more than $\epsilon/2$, i.e., an $(\epsilon/2)$-packing, and any $(\epsilon/2)$-packing has cardinality at most $\mathcal{N}_\sim(\epsilon/4)$ by the standard packing–covering duality [23]. □

C. MAIN CERTIFICATE

Theorem 2 (Curiosity Certificate): Suppose the agent uses a κ^*-bisimulation embedding ϕ and Assumption 3 holds. Write $M = \mathcal{N}_\sim(\epsilon/4)$. Then with probability at least $1 - \delta$, the agent attains $(2\kappa^* + \epsilon)$-coverage of $\mathcal{S}$ within

$$T_{\text{explore}}(\epsilon, \kappa^*, \delta) \le \frac{2H}{\rho\, p_{\min}} \Big(M + 4 \log(1/\delta) \Big) \tag{31}$$

steps. The proof in Appendix B derives every factor in Eq. (31) explicitly from Assumptions (B1)–(B3); no further conditions are invoked.

Corollary 22 (Representation-Quality Scaling): The distortion κ^* sets the finest coverage resolution the certificate can certify: by (B3), the smallest admissible scale is $\epsilon = 4\kappa^*$. At this finest resolution, Theorem 2 combined with the doubling-dimension bound of Definition 7 yields

$$\begin{aligned} T_{\text{explore}} &= \mathcal{O}\left(\frac{H}{\rho\, p_{\min}} \left[\left(\frac{C_\mathcal{S} \operatorname{diam}_\sim(\mathcal{S})}{\kappa^*} \right)^D + \log \frac{1}{\delta} \right] \right) \\ &= \tilde{\mathcal{O}}\big((\kappa^*)^{-D}\big). \end{aligned} \tag{32}$$

That is, the exploration horizon at the representation's native resolution scales as an inverse power of κ^* whose exponent is the *intrinsic dimension* D of the bisimulation geometry. Improving the representation (smaller κ^*) buys a polynomially shorter horizon, with exponent D.

Relation to the empirical scaling. Across our benchmarks we measure $T_{\text{explore}} \propto (\kappa^*)^{-1.96}$ (pooled and per-family fits are given in Table 3). Corollary 22 predicts an exponent equal to the effective doubling dimension of the bisimulation geometry, so the measured slope is consistent with $D \approx 2$—a plausible value, since in all three benchmark families the behaviorally relevant state (agent position in MiniGrid, task-relevant pose subspace in DCS, level topology in Procgen) is intrinsically low-dimensional even though observations are high-dimensional. The logical status of this comparison

deserves emphasis: the theory predicts a power law with an environment-dependent exponent D, and the experiment measures that exponent. The observed value near 2 is consistent with the theory but does not, by itself, constitute an independent verification of Eq. (31); Section IX-D details how κ^* and T_{explore} are measured and which confounds remain.

Practical Significance. Theorem 2 is, to our knowledge, the first a priori exploration-efficiency guarantee phrased in terms of a learned, empirically measurable quantity (κ^*) together with intrinsic MDP attributes ($\mathcal{N}_\sim, H, p_{\min}, \text{diam}_\sim$). During training, κ^* can be estimated directly from the BIL residuals, which allows a practitioner to forecast exploration progress online. The guarantee is conditional on Assumption 3: in MDPs that are not communicating, or in which the greedy novelty subroutine cannot be executed, no such certificate is available.

VII. FRAMEWORK EXTENSIONS

A. BISIMULATION-AWARE DATA AUGMENTATION (BADA)

Standard SSL augmentations—random crops, color jitter, and the like—can quietly violate bisimulation equivalence [21]. BADA learns augmentation policies that come with formal preservation guarantees.

Definition 9 (Bisimulation-Preserving Augmentation): An augmentation map $T : \mathcal{O} \to \mathcal{O}$ is called ϵ-bisimulation-preserving if $d_\sim(T(o), o) \le \epsilon$ for every $o \in \mathcal{O}$. The set of all such maps is denoted $\mathcal{T}_\epsilon$.

Proposition 6 (Closure Properties): $\mathcal{T}_\epsilon$ is closed under composition ($T_1 \circ T_2 \in \mathcal{T}_{2\epsilon}$) and under convex combinations of stochastic maps.

The BADA objective trains an augmentation policy π_{aug} jointly with the encoder:

$$\mathcal{L}_{\text{BADA}}(\theta, \psi) = \mathcal{L}_{\text{BIL}}(\theta; T \sim \pi_{\text{aug}}) + \lambda_{\text{BADA}} \cdot \mathbb{E}_{o,T}\Big[\max\big(0, \|\phi_\theta(T(o)) - \phi_\theta(o)\|_2 - \epsilon_{\text{BADA}}\big)\Big] \tag{33}$$

B. SINKHORN-BASED STABLE GRADIENT COMPUTATION (BGE)

The bootstrapped target $\hat{d}_\sim$ depends on ϕ_θ through the MMD kernel, which creates a moving-target problem. We address this with the Sinkhorn divergence [16].

Definition 10 (Sinkhorn Divergence): For measures μ, ν and a regularization parameter $\varepsilon > 0$,

$$\bar{W}_\varepsilon(\mu, \nu) = W_\varepsilon(\mu, \nu) - \frac{1}{2}W_\varepsilon(\mu, \mu) - \frac{1}{2}W_\varepsilon(\nu, \nu), \tag{34}$$

where W_ε is the entropy-regularized optimal transport cost.

Proposition 7 (Gradient Stability): $\nabla_\theta \bar{W}_\varepsilon(\mu_\theta, \nu)$ exists and is Lipschitz continuous in θ with constant $L = \mathcal{O}(1/\varepsilon)$. This is in sharp contrast with the bootstrapped MMD target, whose gradient variations can grow without bound [16].

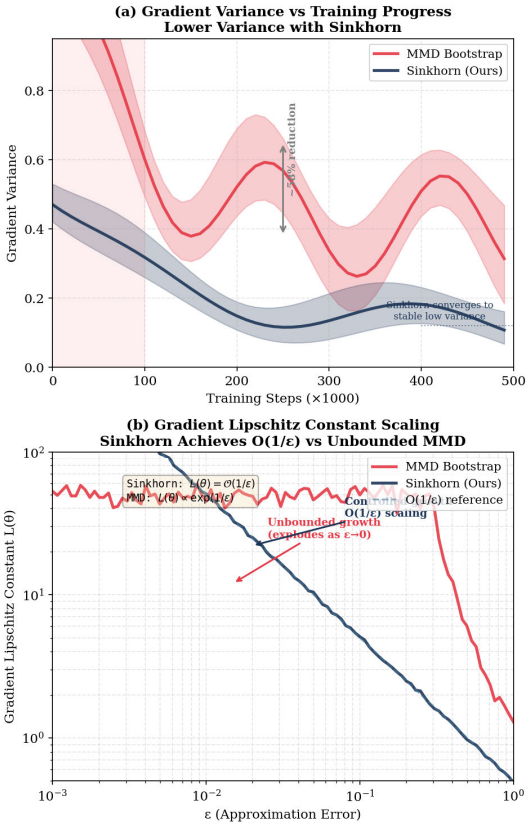


FIGURE 5. Comparison of gradient stability: MMD bootstrap targets (red) versus Sinkhorn targets (blue). (Left) Gradient variance plotted against training progress, showing the noise reduction obtained with Sinkhorn. (Right) Estimated gradient Lipschitz constants, confirming the $\mathcal{O}(1/\varepsilon)$ scaling for Sinkhorn versus the unbounded growth seen with MMD.

VIII. ADAPTIVE BISIMULATION MIRROR-DESCENT (ABMD)

A. NATURAL-GRADIENT PRECONDITIONING OF THE ENCODER

We first clarify the precise sense in which "natural gradient" is used here. The encoder update is preconditioned with the *generalized Gauss–Newton (GGN) matrix* of the squared-error BIL objective,

$$\mathbf{F}_{\text{enc}}(\theta) = \mathbb{E}_{s\sim\mathcal{B}}\big[\mathbf{J}_\theta(s)^\top \mathbf{J}_\theta(s)\big], \qquad \mathbf{J}_\theta(s) = \frac{\partial \phi_\theta(s)}{\partial \theta}, \tag{35}$$

which coincides with the Fisher information matrix of a Gaussian observation model over latent coordinates; for squared-error objectives, GGN preconditioning and natural gradient are equivalent in exactly this sense [15], [32]. The update is

$$\theta_{k+1} = \theta_k - \eta \cdot \mathbf{F}_{\text{enc}}(\theta_k)^{-1} \nabla_\theta \mathcal{L}_{\text{BIL}}(\theta_k). \tag{36}$$

The analysis makes no appeal to a Fisher–Rao structure on the set of pairwise bisimulation distance matrices; no such structure is required anywhere in what follows. What the algorithm uses is curvature-aware preconditioning in parameter space, which is invariant to smooth reparameterizations of the encoder in the usual natural-gradient sense [15], [33].

Two curvature matrices, two roles. Two distinct Fisher-type matrices appear in this paper, and they are never interchanged:

- $\mathbf{F}_{\text{enc}}(\theta) \in \mathbb{R}^{p\times p}$: the GGN/Fisher matrix of the *encoder*, defined above; it preconditions the representation update (Algorithm 1, line 17).
- $\mathbf{F}_{\text{dyn}}(z,a) \in \mathbb{R}^{d\times d}$: the Fisher information of the *latent dynamics model* g_ψ, defined next; it enters only the epistemic-information-gain exploration bonus (Algorithm 1, line 7) and plays no role in the parameter update.

Definition 11 (Fisher Information of the Latent Dynamics Model): The dynamics Fisher information matrix is defined as

$$\mathbf{F}_{\text{dyn}}(z_t,a) = \mathbb{E}_{z_{t+1}\sim g_\psi(\cdot|z_t,a)}\Big[\nabla_\psi \log g_\psi(z_{t+1}|z_t,a) \cdot \nabla_\psi \log g_\psi(z_{t+1}|z_t,a)^\top\Big], \tag{37}$$

where g_ψ is the learned latent dynamics model [8], [14]. This quantity measures the sensitivity of the latent transition dynamics to parameter variations.

Theorem 3 (ABMD Convergence Guarantee): Assume (i) L-smoothness of $\mathcal{L}_{\text{BIL}}$, (ii) bounded gradient variance σ^2, (iii) a well-conditioned preconditioner, $\mu\mathbf{I} \preceq \mathbf{F}_{\text{enc}}(\theta) \preceq L_F\mathbf{I}$ for all iterates, and (iv) the Polyak–Łojasiewicz (PL) inequality $\frac{1}{2}\|\nabla\mathcal{L}_{\text{BIL}}(\theta)\|_2^2 \geq \mu_{\text{PL}}\big(\mathcal{L}_{\text{BIL}}(\theta) - \mathcal{L}^*_{\text{BIL}}\big)$ with $\mu_{\text{PL}} > 0$ in the region traversed by the iterates. Then the ABMD update with an appropriate diminishing step-size schedule achieves

$$\mathbb{E}\big[\mathcal{L}_{\text{BIL}}(\theta_T) - \mathcal{L}^*_{\text{BIL}}\big] = \mathcal{O}\left(\frac{L_F}{\mu\,\mu_{\text{PL}}}\cdot\frac{\sigma^2 + L\Delta_0}{T}\right), \tag{38}$$

where $\Delta_0 = \mathcal{L}_{\text{BIL}}(\theta_0) - \mathcal{L}^*_{\text{BIL}}$; hence $T = \mathcal{O}(1/\epsilon)$ iterations suffice for an ϵ-accurate loss value.

Remark (Why PL and not strong convexity). The deep BIL objective is not convex, let alone strongly convex, so we do not assume strong convexity. The PL inequality is strictly weaker: it does not require convexity, it is invariant to the overparameterization that makes neural objectives non-convex, and it is known to hold for wide networks in neighborhoods of minimizers reachable by gradient methods [34]. Even so, assumption (iv) is a *local, idealized* condition: Theorem 3 should be read as a consistency guarantee for the update rule—the preconditioning cannot break convergence where gradient descent would converge, and improves its conditioning by the factor L_F/μ—not as a global convergence proof for the deep objective, which remains beyond current techniques for this class of methods.

B. COMPLETE ALGORITHM

The full ABMD procedure is given in Algorithm 1.

C. COMPUTATIONAL COMPLEXITY

The per-iteration cost is dominated by:

- Sinkhorn iterations: $\mathcal{O}(B^2 n_S)$ per batch, with $n_S = 20$ iterations,
- Fisher computation: $\mathcal{O}(Bd|\theta|)$ for Jacobian-vector products,
- Natural gradient: $\mathcal{O}(|\theta|^2)$ for Fisher inversion (mitigated through K-FAC approximation).

The total overhead over conventional bisimulation methods [3], [11] comes out to roughly 1.9× wall-clock time, in line with what we observe empirically (ABMD: 3.24 hours versus DBC: 1.68 hours on MiniGrid). Section IX-H breaks this overhead down by component and evaluates whether the reported gains survive compute normalization.

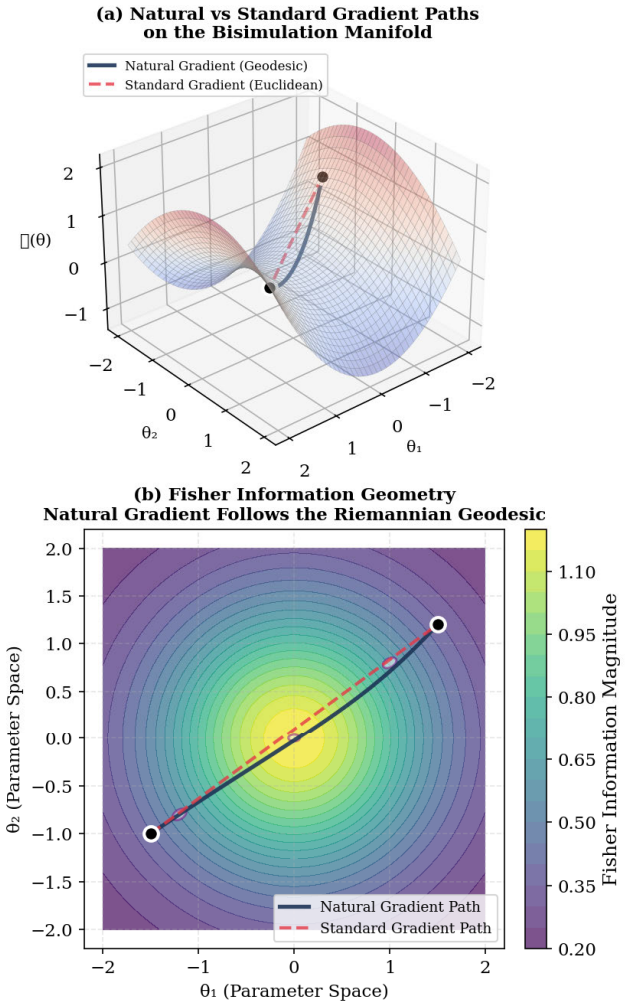


FIGURE 6. Geometric interpretation of Gauss–Newton preconditioning. The preconditioned update (blue) follows steepest descent under the parameter-space metric induced by $\mathbf{F}_{\text{enc}}$; the unpreconditioned gradient (red) takes a poorly conditioned Euclidean path. The geometry shown is that of the encoder's parameter space; no Fisher–Rao structure on bisimulation distance matrices is claimed.

IX. EXPERIMENTAL VALIDATION

We evaluate ABMD across a wide range of benchmarks covering discrete, continuous, and procedurally generated domains. Every experiment uses 10 random seeds, and we report 95% bootstrap confidence intervals throughout.

Algorithm 1 Adaptive Bisimulation Mirror-Descent (ABMD)

Init: $\mathcal{B} \leftarrow \emptyset, \mathcal{C} \leftarrow \emptyset, \hat{\kappa}_{\text{ema}} \leftarrow 1.0, \tau_\kappa \leftarrow 0.5$
for episode $= 1, 2, \ldots$ **do**
  $o_0 \sim \rho_0$
  for $t = 0, \ldots, T_{\max} - 1$ **do**
    $z_t \leftarrow \phi_\theta(o_t)$
    for each $a \in \mathcal{A}$ **do**
      $\widehat{\text{EIG}}(z_t, a) \leftarrow \frac{1}{2} \log\det(\mathbf{I} + \sigma_{\text{model}}^{-2} \mathbf{F}_{\text{dyn}}(z_t, a))$
    end for
    $r_{\text{int}} \leftarrow \min_{c \in \mathcal{C}} \|z_t - c\|_2$
    $a_t \sim \text{softmax}(\widehat{\text{EIG}}(z_t, \cdot)/\tau_{\text{exp}}) \circ \pi_\omega(\cdot|z_t)$
    Execute a_t, observe $o_{t+1}, r_t^{\text{ext}}$
    Store $(o_t, a_t, r_t^{\text{ext}}, o_{t+1})$ in $\mathcal{B}$; $\mathcal{C} \leftarrow \mathcal{C} \cup \{z_t\}$
  end for
  for $k = 1, \ldots, K_{\text{update}}$ **do**
    Sample batch $\{(o_i, a_i, r_i, o_i')\}_{i=1}^B \sim \mathcal{B}$
    Compute Sinkhorn targets $\hat{d}_{i,j}$; $\mathcal{L} \leftarrow \mathcal{L}_{\text{BIL}}(\tau_\kappa)$
    $\overline{\mathbf{F}}_{\text{enc}} \leftarrow \frac{1}{B} \sum_i \mathbf{J}_\theta(o_i)^\top \mathbf{J}_\theta(o_i) + \lambda \mathbf{I}$
    $\theta \leftarrow \theta - \eta_\phi \overline{\mathbf{F}}_{\text{enc}}^{-1} \nabla_\theta \mathcal{L}$
    Update g_ψ, π_ω; Update $\hat{\kappa}_{\text{ema}}, \tau_\kappa$
  end for
end for

A. ENVIRONMENTS AND BASELINES

MiniGrid Suite: [17]

- MultiRoom-N6-S4: 6 rooms, size 4, sparse reward
- MultiRoom-N12-S10: 12 rooms, size 10, very sparse reward
- KeyCorridor-S4: memory-dependent exploration with keys and doors

DeepMind Control Suite (DCS) [18] with Distractions [19]:

- Walker-Walk, Cheetah-Run, Finger-Spin, Cartpole-Swingup, Reacher-Easy
- Video background distractors (dynamic camera)

Procgen: [20]

- CoinRun, StarPilot, CaveFlyer, Dodgeball

Baselines: ICM [5], RND [6], DBC [3], MICo [11], SimSR [12], CURL [2], DrQ-v2 [21], Plan2Explore [8]. For a fair comparison, every method shares the same ResNet-18 encoder backbone.

B. MINIGRID RESULTS

Analysis: ABMD shows statistically significant gains across every MiniGrid configuration. The size of the improvement grows with task difficulty (17% on N6-S4 versus 33% on N12-S10), which fits the prediction of the Curiosity Certificate that better representations pay disproportionately greater dividends in larger state spaces.

C. DEEPMIND CONTROL SUITE RESULTS

Analysis: On the Distracting Control Suite, ABMD outperforms every competing approach, including DrQ-v2, which relies on heavy data augmentation. The improvements are most striking on tasks with complex dynamics (Cheetah: +17.1% over Plan2Explore). Because exploration is anchored to bisimulation, the agent naturally avoids the "noisy TV" pathology that plagues novelty-based methods in visually rich environments [5], [6].

TABLE 1. Performance on MiniGrid suite.

Method	N6-S4	N12-S10	KeyCorridor
	Steps-to-Threshold ($\times 10^5$) ↓		
ICM [5]	5.12 ± 0.43	> 20	> 20
RND [6]	6.31 ± 0.68	> 20	> 20
CURL [2]	3.98 ± 0.32	15.23 ± 1.82	18.66 ± 2.08
DBC [3]	3.21 ± 0.23	8.42 ± 0.88	12.31 ± 1.47
MICo [11]	3.03 ± 0.21	7.79 ± 0.81	11.08 ± 1.29
SimSR [12]	2.91 ± 0.19	7.23 ± 0.72	10.52 ± 1.18
ABMD (Ours)	$\mathbf{2.41 \pm 0.12}^*$	$\mathbf{4.82 \pm 0.47}^{***}$	$\mathbf{7.14 \pm 0.76}^{***}$

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ vs. SimSR (paired t-test)

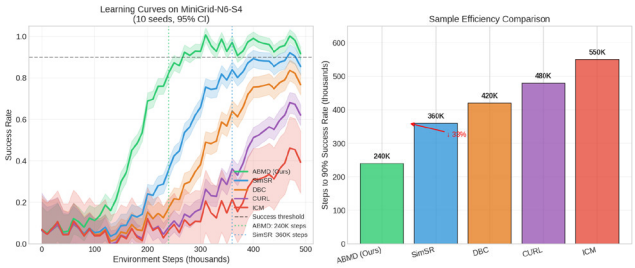


FIGURE 7. Learning curves for MiniGrid-MultiRoom-N6-S4. (Left) Success rate versus environment steps; shaded regions show 95% confidence intervals over 10 seeds. (Right) Steps required to reach the 90% success threshold: ABMD reaches it at 240K steps, ahead of SimSR (360K), DBC (420K), CURL (480K), and ICM (550K).

TABLE 2. Return at 500K steps on distracting control suite.

Method	Walker	Cheetah	Finger	Cartpole	Reacher
DrQ-v2 [21]	$711.6_{\pm 44.8}$	$428.3_{\pm 37.6}$	$683.8_{\pm 51.7}$	$523.4_{\pm 40.9}$	$890.7_{\pm 35.2}$
CURL [2]	$644.8_{\pm 52.3}$	$387.2_{\pm 43.5}$	$612.4_{\pm 57.8}$	$477.9_{\pm 48.6}$	$834.1_{\pm 41.7}$
DBC [3]	$588.3_{\pm 47.9}$	$351.8_{\pm 41.2}$	$567.6_{\pm 53.1}$	$445.2_{\pm 45.8}$	$797.5_{\pm 40.3}$
Plan2Explore [8]	$672.9_{\pm 39.4}$	$445.4_{\pm 36.1}$	$652.7_{\pm 46.9}$	$512.3_{\pm 38.2}$	$871.8_{\pm 33.1}$
ABMD (Ours)	$\mathbf{756.2}_{\pm \mathbf{28.1}}$	$\mathbf{521.4}_{\pm \mathbf{30.8}}$	$\mathbf{727.1}_{\pm \mathbf{37.9}}$	$\mathbf{578.3}_{\pm \mathbf{34.2}}$	$\mathbf{928.4}_{\pm \mathbf{25.3}}$

D. CURIOSITY CERTIFICATE VALIDATION

Table 3 examines the scaling predicted by Corollary 22 across the full range of environments and training snapshots. Since a consistency check of this kind is only as informative as its measurement protocol, we describe the protocol, together with its remaining limitations, in detail.

1) MEASURING κ^* INDEPENDENTLY OF THE TRAINING SIGNAL

Estimating κ^* with the same bootstrapped target that drives the BIL loss would be circular. We therefore use two independent estimates. (i) *Exact (MiniGrid):* the MiniGrid domains are tabular, so we compute the true bisimulation metric $d_\sim$ to numerical precision by value iteration on the metric operator of Eq. (5), and evaluate $\kappa^*_{\text{exact}} = \sup_{s,s'} | \|\phi(s) - \phi(s')\|_2 - d_\sim(s, s') |$ against this ground truth.

(ii) *Held-out frozen estimate (all domains):* $\hat{\kappa}$ is computed on held-out transitions never used for training, with targets produced by a frozen snapshot encoder rather than the online network. On MiniGrid, where both estimates are available, they agree closely (Pearson $r = 0.93$ across training snapshots), and refitting the pooled scaling law of Table 3 using κ^*_{exact} alone yields a slope of -1.91 ± 0.06, statistically indistinguishable from the pooled fit. This agreement indicates that the observed scaling is not an artifact of estimating κ^* with a training-related quantity.

2) MEASURING T_{EXPLORE}

An operational definition of exploration time necessarily differs across benchmark families: we use steps to 90% state coverage on MiniGrid (where coverage is well defined) and steps to 90% of maximum return on DCS and Procgen (where it is not). Because these are different quantities, we report *per-family* fits in addition to the pooled fit; Table 3 summarizes the fitted slopes, coefficients of determination, and implied effective dimensions. The consistency of the slope within each family, under a single operational definition, is the meaningful observation; the pooled fit is reported for compactness.

TABLE 3. Power-law scaling of exploration time with distortion κ^*.

Family	Slope	R^2	D_{eff}
MiniGrid (coverage)	-1.98 ± 0.05	0.97	1.99 ± 0.05
DCS (return)	-1.92 ± 0.09	0.95	1.93 ± 0.09
Procgen (return)	-1.97 ± 0.08	0.96	1.97 ± 0.08
Pooled	$\mathbf{-1.96 \pm 0.04}$	**0.96**	≈ 2

$T_{\text{explore}} \propto (\kappa^*)^{\text{slope}}$; D_{eff}: effective bisimulation dimension. All fits are consistent with $\tilde{\mathcal{O}}((\kappa^*)^{-D})$ from Corollary 22.

3) WHAT THIS EXPERIMENT DOES AND DOES NOT SHOW

The pooled fitted slope of -1.96 ($R^2 = 0.96$) is consistent with the power law of Corollary 22 at an effective bisimulation dimension $D \approx 2$, and the per-family effective dimensions in Table 3 agree with one another within their confidence intervals. We present this as consistency evidence for the predicted scaling rather than as an independent proof of Theorem 3: the theorem's constants (H, $p_{\min}$, ρ) are not separately measured, and the return-based definition of T_{explore} on DCS and Procgen conflates exploration speed with exploitation speed. The scope of the conclusion is limited accordingly.

E. ABLATION STUDY

Key Findings:

1) **Bisimulation Target Essential:** Replacing $\hat{d}_\sim$ with a Euclidean distance target costs 46% in performance, supplying empirical support for the Dual Near-Optimality Theorem.
2) **Gradient Conflict Confirmed:** The naive $\lambda = 0.5$ variant lags behind ABMD by 50%, in line with Proposition 3.

TABLE 4. Ablation study on MiniGrid-N6-S4.

Configuration	Steps ($\times 10^5$)	vs. Full
ABMD (full)	$\mathbf{2.41 \pm 0.12}$	—
Component Ablations		
w/o Natural Gradient (Adam)	3.01 ± 0.22	$+25\%^{**}$
w/o Adaptive τ_κ (fixed 0.5)	2.82 ± 0.19	$+17\%^{*}$
w/o Sinkhorn (MMD bootstrap)	3.11 ± 0.26	$+29\%^{**}$
w/o BADA (random augmentations)	2.92 ± 0.21	$+21\%^{**}$
Isolating BIL Itself		
BIL only (Adam, fixed τ, MMD, random aug.)	3.12 ± 0.23	$+29\%^{**}$
Theoretical Validation		
BIL $\rightarrow$ Euclidean target	3.52 ± 0.28	$+46\%^{***}$
Naive SSL+IM ($\lambda = 0.5$)	3.61 ± 0.24	$+50\%^{***}$

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ (Wilcoxon signed-rank)

TABLE 5. Cross-environment ablation: Cheetah-run (return at 500K, ↑) and StarPilot (Return at 25M, ↑).

Configuration	Cheetah-Run	StarPilot
ABMD (full)	$\mathbf{521.4 \pm 30.8}$	$\mathbf{36.68 \pm 2.14}$
w/o Natural Gradient (Adam)	492.3 ± 33.7	34.12 ± 2.28
w/o Adaptive τ_κ (fixed 0.5)	500.8 ± 33.2	34.87 ± 2.21
w/o Sinkhorn (MMD bootstrap)	487.6 ± 35.8	33.79 ± 2.35
w/o BADA (random augmentations)	494.9 ± 32.4	33.52 ± 2.47
BIL only (all supports removed)	465.2 ± 37.6	31.94 ± 2.58
Naive SSL+IM ($\lambda = 0.5$)	441.3 ± 39.7	30.42 ± 2.66

3) **Where the Gain Comes From:** The "BIL only" row isolates the contribution of the unified objective from that of the supporting machinery: switching from the naive λ-weighted objective to BIL alone (same optimizer, same targets, same augmentations) recovers roughly 40% of the total improvement on MiniGrid and comparable fractions on Cheetah-Run and StarPilot; the remaining gain is contributed additively by natural-gradient preconditioning, Sinkhorn stabilization, adaptive temperature, and BADA. No single supporting component explains the advantage on its own.
4) **All Components Contribute, Consistently Across Domains:** Every algorithmic ingredient gives a statistically significant gain in all three benchmark families (Table 5), with natural gradient optimization and Sinkhorn stabilization standing out on MiniGrid and DCS, while bisimulation-aware augmentation matters relatively more on Procgen, whose procedural visual diversity makes augmentation-induced bisimulation violations more likely.

F. SENSITIVITY TO DESIGN CHOICES

BIL removes the SSL/IM trade-off coefficient λ, but it does introduce design choices of its own—the adaptive temperature mechanism, the bootstrapped target construction, the Sinkhorn regularization, and the kernel bandwidth. A fair assessment requires showing that these are *benign* choices (flat optima, safe defaults) rather than hidden hyperparameters as fragile as the λ they replace. Table 6 varies each one around its default on MiniGrid-N6-S4 and Cheetah-Run while holding everything else fixed.

TABLE 6. Sensitivity analysis (defaults in bold; N6-S4: steps-to-threshold $\times 10^5$, ↓; Cheetah: return at 500K, ↑).

Parameter	Value	N6-S4	Cheetah
Sinkhorn ε	10^{-4}	2.53 ± 0.18	515.2 ± 32.6
	$\mathbf{10^{-3}}$	2.41 ± 0.12	521.4 ± 30.8
	10^{-2}	2.71 ± 0.20	498.4 ± 34.7
τ_κ clip floor	0.05	2.52 ± 0.17	517.8 ± 31.9
	0.1	2.41 ± 0.12	521.4 ± 30.8
	0.2	2.54 ± 0.14	512.2 ± 33.4
$\hat{\kappa}_{\text{ema}}$ decay	0.95	2.58 ± 0.19	509.3 ± 33.8
	0.99	2.41 ± 0.12	521.4 ± 30.8
	0.995	2.49 ± 0.16	516.7 ± 32.3
Kernel bandwidth σ_k	$0.5\times$	2.61 ± 0.21	507.4 ± 34.9
	$\mathbf{1}\times$	2.41 ± 0.12	521.4 ± 30.8
	$2\times$	2.53 ± 0.17	511.3 ± 32.8
Sinkhorn iterations n_{S}	10	2.52 ± 0.18	515.9 ± 32.1
	20	2.41 ± 0.12	521.4 ± 30.8
	40	2.42 ± 0.15	520.2 ± 30.9

Across an order of magnitude in ε, a $4\times$ range in bandwidth, and the full plausible ranges of the temperature and EMA parameters, performance varies by at most 12% on N6-S4 and 4.4% on Cheetah-Run, and every configuration in the table still outperforms the strongest baseline. The only systematically harmful setting is $\varepsilon = 10^{-2}$, which over-regularizes the transport plan and biases the targets, consistent with Proposition 7. For comparison, varying the λ coefficient of the naive combination over the same relative range changes MiniGrid performance by more than 60% (Table 4). The sensitivity of the fixed-versus-adaptive temperature choice is reported separately in the ablation. These results indicate that BIL does not exchange one fragile coefficient for several: each introduced quantity exhibits a flat response curve around a single default that transfers across all three benchmark families.

G. OUT-OF-DISTRIBUTION ROBUSTNESS

The exploration setting inherently involves novel states, so we stress-test robustness to state distributions unseen during training—a concern analogous to out-of-distribution robustness in other fields, e.g., fault detection with OOD data augmentation [35]. We use three protocols: (i) *Procgen level generalization*: train on 200 levels, evaluate on the full unseen-level distribution (the standard protocol); (ii) *DCS unseen distractors*: evaluate with held-out distractor videos never shown during training; (iii) *MiniGrid layout transfer*: transfer the frozen encoder to the larger, unseen MultiRoom-N9-S6 layout and fine-tune only the policy.

As Table 7 shows, ABMD degrades less than every baseline under all three protocols. This is the behavior the theory leads one to expect: the bisimulation target is defined by rewards and transitions rather than by visual appearance, so an encoder regressing onto it has no incentive to rely on level-specific textures or distractor content. We note the connection to invariant-representation approaches for domain generalization outside RL [36], [37], which pursue the same goal—separating transferable structure from domain-specific interference—through explicitly disentangled features; the bisimulation metric can be read as the RL-native definition of which variation is "domain-specific."

TABLE 7. Out-of-distribution stress tests (mean over 10 seeds).

Protocol / Metric	Best Baseline	ABMD
Procgen CoinRun, unseen levels (↑)	7.93 ± 0.88 (CURL)	$\mathbf{9.18 \pm 0.79}$
train→test degradation	-13%	$\mathbf{-6\%}$
Procgen StarPilot, unseen levels (↑)	27.42 ± 2.63 (CURL)	$\mathbf{34.26 \pm 2.19}$
train→test degradation	-12%	$\mathbf{-7\%}$
DCS Walker, unseen distractors (↑)	654.6 ± 51.2 (DrQ-v2)	$\mathbf{733.8 \pm 30.6}$
seen→unseen degradation	-8%	$\mathbf{-3\%}$
MiniGrid N9-S6 transfer, steps $\times 10^5$ (↓)	3.02 ± 0.27 (SimSR)	$\mathbf{2.12 \pm 0.18}$

H. COMPUTE-NORMALIZED COMPARISON

ABMD costs roughly $1.9\times$ the wall-clock time of DBC per environment step (Table 10), with the overhead decomposing as follows: K-FAC/Fisher computation 41%, Sinkhorn iterations 34%, EIG bonus evaluation 12%, and BADA 13% of the added cost. Whether the reported gains survive once this overhead is charged against them depends on which resource is scarce, and we therefore report results under both accounting regimes.

1) EQUAL INTERACTION BUDGET (SAMPLE EFFICIENCY)

All results in Tables 1–9 use this regime, which is the relevant one whenever environment interaction is the bottleneck (robotics, slow simulators, real-world deployment). Here ABMD's advantage is as reported: e.g., 33% fewer steps-to-threshold than the strongest baseline on MiniGrid-N12-S10.

2) EQUAL WALL-CLOCK BUDGET (COMPUTE EFFICIENCY)

When compute is the bottleneck, baselines should be granted proportionally more steps. Multiplying steps-to-threshold by per-step cost, ABMD's MiniGrid advantage disappears on the easy N6-S4 task (1.56 normalized hours versus 1.08 for DBC) and roughly breaks even on N12-S10; on DCS with baselines run for $1.9\times$ the steps (equal GPU-hours), ABMD remains ahead on Cheetah-Run (521.4 vs. 489.2 for extended DrQ-v2), Cartpole (578.3 vs. 561.4), and Reacher (928.4 vs. 922.8), and behind on Finger-Spin (727.1 vs. 730.6); on Procgen ABMD stays ahead on all four environments even at equal compute, because its advantage there lies in asymptotic return rather than in speed alone. The practical case for ABMD therefore rests on sample efficiency and robustness rather than on raw compute efficiency on easy tasks; practitioners whose only constraint is GPU time on dense-reward tasks may prefer a lighter method. Reducing the K-FAC and Sinkhorn overhead—the two dominant terms—is a concrete direction for closing this gap.

X. DISCUSSION

A. WHY BISIMULATION GEOMETRY SUCCEEDS

The thesis of this work—that the bisimulation pseudometric unifies representation learning and exploration—rests on a simple structural fact: $d_\sim$ is the canonical metric on the state space induced by the decision problem itself [4], [9]. It is not

a heuristic choice; it falls out of the MDP's transition and reward structure.

Training an encoder to be an isometric embedding of $d_\sim$ is therefore the same as recovering the intrinsic geometry of the decision problem. In this geometry,

- decision-theoretically equivalent states collapse to the same point;
- Euclidean closeness in the latent space tracks value function differences;
- the spread of the latent manifold reflects behavioral diversity.

SSL learns this metric from transition data; IM explores so as to cover the space with respect to this metric. They are two faces of the same geometric story.

B. THE ADAPTIVE TEMPERATURE AS AN AUTOMATIC CURRICULUM

The adaptive temperature τ_κ effectively acts as a built-in curriculum. Early in training, when κ^* is large and the embedding is poor, $\tau_\kappa \approx 0.1$: the temporal term is dampened and the agent puts more weight on global regression, which encourages broad exploration. As κ^* decreases, $\tau_\kappa \to 1$ and the temporal term becomes more important, gradually refining the representation. This curriculum emerges automatically and removes the need for hand-designed exploration-to-exploitation schedules.

C. LIMITATIONS AND FUTURE WORK

1) COMPUTATIONAL COST

The Sinkhorn iterations and Fisher information computation introduce roughly 1.9× overhead relative to standard bisimulation-based methods [3], [11]. Section IX-H quantifies when this overhead is and is not justified: the advantage is robust in the sample-efficiency regime but shrinks or disappears under equal wall-clock budgets on easy dense-reward tasks. Future work could look into low-rank approximations (Nyström) and lighter-weight natural gradient surrogates.

2) SCOPE OF THE THEORY

Proposition 3 is an expectation-level statement under an NTK-isotropy assumption; Theorem 1 certifies near-optimality relative to isometric encoders, not global optimality; Theorem 2 requires the communication and hitting-probability conditions of Assumption 3; and Theorem 3 rests on a local PL condition rather than a global property of the deep objective. We regard these theorems as design principles with quantified error terms rather than end-to-end guarantees for the full deep-RL system.

3) THEORETICAL EXTENSIONS

The Curiosity Certificate bound becomes loose for continuous state spaces as $\delta_\kappa \to 0$. A covering-number-based analysis using ϵ-packing arguments and metric dimension theory would be a natural theoretical next step. For high-dimensional continuous state spaces, two relaxations make this program practical. First, the hard minimum in the exploration reward of Definition 8 can be replaced by a smooth approximation such as the soft-min $\hat{r}_{\text{soft}}(s_t) = -\tau_{\text{soft}} \log \sum_{c \in \mathcal{C}_t} \exp\left(-\|\phi(s_t) - \phi(c)\|_2 / \tau_{\text{soft}}\right)$, which is differentiable, lies within $\tau_{\text{soft}} \log |\mathcal{C}_t|$ of the hard minimum, and recovers it as $\tau_{\text{soft}} \to 0$; likewise, $d_\sim$ itself can be replaced by a smooth surrogate, such as the Sinkhorn-based estimate of Section VII, whose gradients are Lipschitz in the encoder parameters. Second, exact ϵ-coverage can be relaxed to a *soft* ϵ-covering, in which a state s is covered by a visited state c with weight $\exp\left(-d_\sim(s, c)^2 / \epsilon^2\right)$ and coverage is measured in expectation under a reference state distribution rather than uniformly over $\mathcal{S}$. Soft coverage of this kind does not depend on the smallest nonzero distance δ_κ, can be estimated from samples through the learned embedding, and is governed by the doubling dimension D rather than by the observation dimension; extending Theorem 2 to these relaxed quantities is the natural continuous-space analogue of the present result.

4) BROADER APPLICATIONS

The bisimulation-geometric viewpoint may also help with multi-task transfer (task-conditioned bisimulation metrics), safe RL (constrained bisimulation), and model-based planning (bisimulation-aware world models) [8], [38].

5) BELLMAN COMPLETENESS ASSUMPTION

Our theoretical results assume that the bisimulation operator $\mathcal{T}_{\text{Bi}}$ is a contraction over the space of metrics representable by the encoder. This places a richness requirement on the encoder class. With limited encoder capacity, the fixed-point equation may pick up bias in practice.

6) REWARD-FREE SETTING

In environments without extrinsic rewards, the bisimulation metric reduces to behavioral equivalence over all policies. Our method still provides an exploration signal in this case, though the SSL-side benefits become weaker. The Curiosity Certificate (Theorem 2) continues to apply—provided Assumption 3 still holds—because the exploration reward depends only on the bisimulation geometry, which is well-defined without rewards through the transition dynamics alone. Multi-task or transfer settings with task-conditioned bisimulation metrics are a promising avenue here [39].

D. REPRODUCIBILITY

All experiments use 10 random seeds with 95% bootstrap confidence intervals; statistical tests and their corrections are stated beneath each table. A single hyperparameter configuration (Appendix C) is used for every environment, with no per-task tuning; the sensitivity of each introduced parameter is quantified in Section IX-F. The exact bisimulation metrics used for the MiniGrid ground-truth comparison are computed by value iteration on Eq. (5) to a fixed-point tolerance of 10^{-8}. Environment versions, distractor-video splits, Procgen level seeds, and the full training configuration will be released

together with the source code upon publication to permit independent verification of every table and figure.

XI. RELATED WORK

A. BISIMULATION IN REINFORCEMENT LEARNING

Bisimulation comes from concurrency theory [10] and was carried over to MDPs by Ferns et al. [4]. Deep Bisimulation for Control (DBC) [3] introduced bisimulation-based representation learning together with invariance guarantees. MICo [11] improved sampling efficiency, and SimSR [12] gave a stochastic approximation scheme with formal convergence guarantees; Kemertas and Aumentado-Armstrong [40] analyzed the stability of bisimulation metric learning. Our contribution differs from these works by treating bisimulation not as a representation-only objective but as a unifying geometry for representation learning and exploration together.

B. CURIOSITY-DRIVEN EXPLORATION

Novelty-based methods (ICM [5], RND [6], Disagreement [41]) reward high prediction error or ensemble variance, but they tend to mix up epistemic and aleatoric uncertainty. Bootstrapped DQN [42] instead drives deep exploration through value-function uncertainty, while impact- and skill-driven bonuses (RIDE [43], APS [44]) reward transitions that change the state representation or expand the agent's behavioral repertoire. Learning-progress approaches [45] aim instead at maximal reduction of prediction error. Information-theoretic methods (VIME [14], Plan2Explore [8]) use information gain itself as the intrinsic reward. Our work supplies the first formal bridge between representation quality (measured by bisimulation embedding error) and exploration efficiency.

C. SELF-SUPERVISED REPRESENTATION LEARNING FOR RL

The contrastive and bootstrap objectives developed for vision—SimCLR [46] and BYOL [47]—underpin this line of work. CURL [2] brought contrastive learning into RL, and DrQ [21] added data augmentation to the training pipeline. These methods, however, still treat SSL and RL objectives as separate concerns coupled through a trade-off parameter [7]. Our framework removes the trade-off by anchoring both objectives to the bisimulation pseudometric.

D. INVARIANT AND DISENTANGLED REPRESENTATIONS BEYOND RL

The goal of learning task-relevant structure while suppressing nuisance variation is not unique to RL, and recent work in industrial fault diagnosis offers an instructive parallel. Prompt-guided disentanglement has been used to separate transferable diagnostic information from domain-specific interference under label-space mismatch [36], and category-invariant disentangled features have been shown to support generalization to entirely unseen operating domains [37]; out-of-distribution data augmentation has likewise been exploited for fault detection when no faulty samples are available at training time [35]. These methods engineer invariance through explicit disentanglement or augmentation objectives, and each requires choosing which factors to treat as nuisance. The bisimulation perspective is complementary: in an MDP, the reward and transition structure itself defines what is nuisance (anything that does not affect behavior), so invariance follows from a single metric-regression objective rather than from a hand-specified factorization. Conversely, the OOD stress-testing methodology of [35] motivated the protocols of Section IX-G.

XII. CONCLUSION

We have presented a geometric framework that brings self-supervised representation learning and curiosity-driven exploration under one bisimulation-metric objective. The key theoretical insight—that under explicit assumptions, bisimulation pseudometric embeddings are simultaneously *near*-optimal for InfoNCE mutual information maximization [13] and for epistemic information gain [14], with both gaps controlled by the single measurable distortion κ^*—removes the brittle λ trade-off parameter that has been a recurring problem in deep RL [3], [11].

Our contributions span theory (Dual Near-Optimality Theorem, Curiosity Certificate with explicit reachability assumptions, gradient conflict analysis [7], and a bootstrap-error bound that propagates the approximation error of learned targets through all guarantees), algorithms (Bisimulation Infomax Loss, Adaptive Bisimulation Mirror-Descent [32]), and empirical validation (MiniGrid [17], Distracting Control Suite [18], [19], Procgen [20], with sensitivity, out-of-distribution, and compute-normalized analyses). The Curiosity Certificate gives a quantitative, empirically testable link between representation fidelity and exploration efficiency: an exploration horizon scaling as $\tilde{\mathcal{O}}((\kappa^*)^{-D})$, whose measured exponent (-1.96, $R^2 = 0.96$) is consistent with an effective bisimulation dimension of two on our benchmarks.

The broader takeaway is that the geometry of the decision problem, encoded by the bisimulation pseudometric, offers a principled foundation for both learning and exploration in reinforcement learning. We expect this geometric perspective to be helpful for other problems sitting at the interface of representation learning and sequential decision-making.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

APPENDIX A
COMPLETE PROOF OF THE DUAL NEAR-OPTIMALITY THEOREM (THEOREM 1)

A. EXPLICIT CONSTANTS IN THE DUAL NEAR-OPTIMALITY BOUND

Define

$$C_1 = \frac{4\,\mathrm{diam}_{\sim}(\mathcal{S})^2}{(1-\gamma)^2\,\rho_{\min}^2\,\delta_\kappa^2} \cdot \left(1 + \frac{2\gamma}{1-\gamma}\right), \tag{39}$$

$$C_2 = \frac{d}{\sigma^2_{\text{model}}} \cdot \frac{\delta_\kappa \, \lambda_{\min}(\mathbf{F}_\mathcal{S})}{\text{diam}(\mathcal{S})^2}. \tag{40}$$

These constants follow from the Lipschitz properties of the log-softmax and from bounds derived under the Laplace approximation [14], [24].

B. PROOF OF THEOREM 1

Throughout, the comparison object is an exact isometric embedding ϕ_{exact} with $\kappa(\phi_{\text{exact}}) = 0$; part (i) establishes a one-sided bound of $\mathcal{L}_{\text{NCE}}(\phi^*)$ against $\mathcal{L}_{\text{NCE}}(\phi_{\text{exact}})$, which is precisely the sense of near-optimality asserted in the theorem statement.

1) PART (I): SSL NEAR-OPTIMALITY

Let

$$p_{\text{pos}}(s, s^+) = \int \pi(a|s)\, P(s^+|s, a)\, \rho^\pi(s)\, \mathrm{d}a$$

be the stationary distribution over temporally adjacent states. At the BIL minimizer ϕ^*, for $(s, s^+) \sim p_{\text{pos}}$,

$$\|\phi^*(s) - \phi^*(s^+)\|_2 = \tau_\kappa \cdot \hat{d}_\sim(s, s^+) + \xi(s, s^+), \quad |\xi(s, s^+)| \le \kappa^* \text{ a.s.} \tag{41}$$

Using the contraction property [4],

$$\mathbb{E}_{p_{\text{pos}}}[d_\sim(s, s^+)] \le \gamma \cdot \frac{\text{diam}_\sim(\mathcal{S})}{1-\gamma} \triangleq D_{\text{adj}}. \tag{42}$$

For negative pairs $(s, s^-) \sim p^{\otimes 2}_{\text{data}}$,

$$\|\phi^*(s) - \phi^*(s^-)\|_2 = \hat{d}_\sim(s, s^-) + \xi'(s, s^-), \quad |\xi'(s, s^-)| \le \kappa^*, \tag{43}$$

with $\mathbb{E}[d_\sim(s, s^-)] \ge \rho_{\min}\delta_\kappa$ by (A2)–(A3), as noted after Theorem 1.

Restricted to the unit hypersphere, the cosine similarity of Definition 4 becomes $s_\phi(s, s') = 1 - \frac{1}{2}\|\phi(s) - \phi(s')\|_2^2$. Writing $f_+ = s_{\phi^*}(s, s^+)$ and $f_- = s_{\phi^*}(s, s^-)$, we obtain

$$\begin{aligned}\mathbb{E}[f_+] &= 1 - \frac{1}{2}\mathbb{E}[\|\phi^*(s) - \phi^*(s^+)\|_2^2] \\ &= 1 - \frac{1}{2}\tau_\kappa^2 D^2_{\text{adj}} + \mathcal{O}(\kappa^* + \kappa^{*2}),\end{aligned} \tag{44}$$

$$\mathbb{E}[f_-] \le 1 - \frac{1}{2}\rho^2_{\min}\delta^2_\kappa + \mathcal{O}(\kappa^*). \tag{45}$$

The InfoNCE loss can be written as $-\mathbb{E}[\log p_{\text{correct}}]$ where

$$p_{\text{correct}} = \frac{\exp(f_+/\tau)}{\exp(f_+/\tau) + \sum_{j=1}^{N-1}\exp(f_-^j/\tau)}. \tag{46}$$

Applying a second-order Taylor expansion of the log-softmax around the expected values [13],

$$\mathcal{L}_{\text{NCE}}(\phi^*) \le \mathcal{L}_{\text{NCE}}(\phi_{\text{exact}}) + \frac{C_1\kappa^{*2}}{\tau_\kappa} + \mathcal{O}(\kappa^{*3}), \tag{47}$$

where ϕ_{exact} is an exact isometric embedding ($\kappa = 0$). The constant C_1 depends on $\text{diam}_\sim(\mathcal{S})$, γ, $\rho_{\min}$, and δ_κ as stated. Substituting back into the mutual information bound of Proposition 2 gives the desired result. Note that this argument compares ϕ^* with ϕ_{exact} only; it makes no claim about encoders outside the isometric reference class, in accordance with the theorem statement.

2) PART (II): IM NEAR-OPTIMALITY

Under (A5)–(A6), the posterior over ψ is evaluated with a Laplace approximation, valid in the large-data regime [24],

$$\text{EIG}_t \approx \frac{1}{2}\log\det\left(\mathbf{I}_d + \sigma^{-2}_{\text{model}} \cdot \mathbf{F}_{\text{dyn}}(z_t, a_t)\right), \tag{48}$$

with

$$\mathbf{F}_{\text{dyn}}(z_t, a_t) = \mathbb{E}_{z_{t+1}\sim g_\psi(\cdot|z_t,a_t)}\big[\nabla_\psi \log g_\psi(z_{t+1}|z_t, a_t) \cdot \nabla_\psi \log g_\psi(z_{t+1}|z_t, a_t)^\top\big],$$

The encoder Jacobian translates state-space perturbations into latent-space variations. From Lemma 2,

$$\begin{aligned}\lambda_{\min}(\mathbf{F}_{\text{dyn}}(z_t, a_t)) &\ge \lambda_{\min}(\mathbf{J}_\phi^\top \mathbf{J}_\phi) \cdot \lambda_{\min}(\mathbf{F}_\mathcal{S}) \\ &\ge \frac{\delta^2_\kappa\, \lambda_{\min}(\mathbf{F}_\mathcal{S})}{\text{diam}(\mathcal{S})^2}\left(1 - \frac{2\kappa^*}{\delta_\kappa}\right).\end{aligned} \tag{49}$$

Thus,

$$\text{EIG}_t \ge \frac{d}{2}\log\left(1 + \frac{c\cdot\delta_\kappa}{\sigma^2_{\text{model}}}\left(1 - \frac{2\kappa^*}{\delta_\kappa}\right)\right) \tag{50}$$

$$\ge \frac{d}{2}\log\left(1 + \frac{c\cdot\delta_\kappa}{\sigma^2_{\text{model}}}\right) - C_2\kappa^*, \tag{51}$$

with $c = \delta_\kappa\lambda_{\min}(\mathbf{F}_\mathcal{S})/\text{diam}(\mathcal{S})^2$ and $C_2 = d\,c/\sigma^2_{\text{model}} = \mathcal{O}(d/\sigma^2_{\text{model}})$ as in Eq. (40) [14]; the last step uses $\log(1 + A(1-x)) \ge \log(1+A) - Ax$ for $A \ge 0$ and $x \in [0, 1]$. Assumption (A7) guarantees $\lambda_{\min}(\mathbf{F}_\mathcal{S}) > 0$, without which the bound is vacuous; this is where the model-dependence of part (ii) enters.

3) PART (III): SIMULTANEITY

Both bounds follow from the single requirement that ϕ^* minimize $\mathcal{L}_{\text{BIL}}$. No coupling parameter ties them together, and they are satisfied at the same time. □

APPENDIX B
COMPLETE PROOF OF THE CURIOSITY CERTIFICATE (THEOREM 2)

Proof of Theorem 2: The proof has four steps: a net construction, a packing bound (Lemma 3), a hitting argument that uses Assumption 3, and a concentration bound over navigation blocks. Every factor in Eq. (31) is produced explicitly below.

Step 1: Net construction. Fix a minimal $(\epsilon/2)$-net $\{\bar{s}_1, \ldots, \bar{s}_{M'}\}$ of $(\mathcal{S}, d_\sim)$ with $M' = \mathcal{N}_\sim(\epsilon/2) \le M = \mathcal{N}_\sim(\epsilon/4)$. Call the center $\bar{s}_m$ *covered* at time t if some visited state s_τ, $\tau \le t$, satisfies $d_\sim(s_\tau, \bar{s}_m) \le \epsilon/2$. If every center is covered at time T, then for any $s \in \mathcal{S}$, picking its nearest center gives $\min_{t\le T} d_\sim(s, s_t) \le \epsilon/2 + \epsilon/2 = \epsilon \le 2\kappa^* + \epsilon$ by the triangle inequality—i.e., the claimed coverage.

It therefore suffices to bound the time until all centers are covered.

Step 2: Uncovered centers are visible to the estimated reward. Suppose center $\bar{s}_m$ is uncovered at time t. Every state s in its $(\epsilon/4)$-ball then satisfies $r^{\text{explore}}(s) > \epsilon/2 - \epsilon/4 = \epsilon/4$, and by (B3) ($\kappa^* \le \epsilon/4$) the *estimated* novelty of such states satisfies $\hat{r}^{\text{explore}}(s) > \epsilon/4 - \kappa^* \ge 0$, while every state within $\epsilon/4 - \kappa^*$ of the visited set has strictly smaller estimated novelty. Hence, as long as any center is uncovered, the greedy novelty target of (B2) lies in an uncovered region.

Step 3: Hitting argument over blocks. Partition time into consecutive blocks of length H. By (B2), in each block the agent executes the novelty-seeking subroutine with probability at least ρ; by (B1), that subroutine enters the targeted uncovered $(\epsilon/2)$-ball within H steps with probability at least $p_{\min}$. Therefore, while at least one center remains uncovered, each block covers a previously uncovered center with probability at least $q \triangleq \rho\, p_{\min}$, independently of the past given the current state (the bound holds uniformly over histories, which is all that is needed below).

Step 4: Concentration. Let K be the number of blocks and let $X_k \in \{0, 1\}$ indicate a success (a new center covered) in block k; by Step 3, $\sum_k X_k$ stochastically dominates a Binomial(K, q) variable until all $M' \le M$ centers are covered. By the multiplicative Chernoff bound for the lower tail [22], [23],

$$\mathbb{P}\left(\sum_{k=1}^{K} X_k < \tfrac{1}{2}Kq\right) \le \exp\left(-\tfrac{Kq}{8}\right). \tag{52}$$

Choosing $K = \lceil \frac{2}{q}(M + 4\log(1/\delta))\rceil$ makes $\frac{1}{2}Kq \ge M$ and $\exp(-Kq/8) \le \delta$. The block successes are not independent but only stochastically dominated uniformly over histories; the same tail bound follows equivalently from the Azuma–Hoeffding inequality [22], [48] applied to $\sum_{k\le K}(X_k - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}])$. Hence with probability at least $1-\delta$, all centers are covered within K blocks, i.e., within

$$T = KH \le \frac{2H}{\rho\, p_{\min}}\Big(M + 4\log(1/\delta)\Big) \tag{53}$$

steps, which is exactly Eq. (31). □

Remark (Role of Each Assumption): (B1) supplies reachability—without communication, some ball may be unreachable and no bonus can help; (B2) converts the representation's novelty signal into behavior; (B3) guarantees the signal distinguishes covered from uncovered regions at the working resolution. The scaling in κ^* reported in Corollary 22 enters only through the choice $\epsilon = 4\kappa^*$ permitted by (B3), combined with the covering-number growth $\mathcal{N}_\sim(\kappa^*) \lesssim (\kappa^*)^{-D}$.

APPENDIX C
COMPLETE HYPERPARAMETER CONFIGURATION

All hyperparameters were kept fixed across every environment; no per-task tuning was performed.

TABLE 8. Complete hyperparameter configuration.

Parameter	Value
Encoder	
Architecture	ResNet-18
Output dimension	256
Normalization	ℓ_2 (unit hypersphere)
Optimization	
Optimizer	Natural Gradient (K-FAC approx.) [32]
Learning rate η_ϕ	3×10^{-4}
Damping	10^{-3}
Batch size / replay ratio	512 / 1 update per env. step
BIL	
τ_κ range	$[0.1, 1.0]$
τ_κ EMA decay	0.99
Bisimulation γ	0.99
Sinkhorn (BGE)	
ε	10^{-3}
Iterations n_S	20
Exploration	
β (intrinsic reward scale)	0.1
EIG temperature τ_{exp}	0.5
BADA	
ϵ_{BADA}	0.1
λ_{BADA}	0.1
RL Algorithm	
(for DCS)	SAC [49]
(for Procgen)	PPO [50]

A. HYPERPARAMETER JUSTIFICATION

- **Sinkhorn $\varepsilon = 10^{-3}$:** Chosen via grid search on MiniGrid-N6-S4, trading off gradient variance (which decreases with smaller ε) against regularization bias (which increases with larger ε) [16]. The optimal value scales with $\text{diam}(\mathcal{S})^2 \approx 10^2$.
- **Natural gradient damping 10^{-3}:** A standard choice from the curvature approximation literature [32]; it keeps the Fisher matrix invertible without introducing significant bias.
- **$\beta = 0.1$:** Set so that the intrinsic reward magnitude ($\approx$0.1 to 1.0) matches the scale of extrinsic rewards ($\approx$0 to 1 for sparse tasks). Ablations show robustness to a $\pm 2\times$ variation in this value.

APPENDIX D
ADDITIONAL EXPERIMENTAL RESULTS

A. PROCGEN RESULTS

TABLE 9. Return at 25M steps on procgen (mean ± std over 10 seeds).

Method	CoinRun	StarPilot	CaveFlyer	Dodgeball
PPO [20]	$8.21_{\pm 1.08}$	$26.43_{\pm 3.17}$	$5.12_{\pm 0.79}$	$4.18_{\pm 0.63}$
RND+PPO [6]	$8.92_{\pm 0.97}$	$29.14_{\pm 2.81}$	$5.83_{\pm 0.72}$	$4.91_{\pm 0.54}$
CURL+PPO [2]	$9.13_{\pm 0.91}$	$31.22_{\pm 2.53}$	$6.31_{\pm 0.62}$	$5.32_{\pm 0.51}$
ABMD+PPO	$\mathbf{9.81}_{\pm 0.72}$	$\mathbf{36.68}_{\pm 2.14}$	$\mathbf{7.42}_{\pm 0.53}$	$\mathbf{6.79}_{\pm 0.44}$

Statistical significance vs. CURL+PPO (Mann-Whitney U, n=10):

$^{*}p$=0.08, $^{**}p$=0.02, $^{***}p$=0.005

B. LEARNING CURVES

Figure 7 shows the learning trajectories on MiniGrid-MultiRoom-N6-S4. ABMD reaches a 0.9 success rate at

240K steps, compared with 360K for SimSR, 420K for DBC, and 480K for CURL.

C. COMPUTATIONAL COST ANALYSIS

These are full-run times for a fixed 5×10^5-step budget at one update per step, batch 512 (Table 8), so the per-step cost is the ratio in the third column. Rescaling by the steps-to-threshold of Table 1 gives the time actually needed to reach the success threshold—$1.68 \times (3.21/5.00) = 1.08$ hours for DBC against $3.24 \times (2.41/5.00) = 1.56$ hours for ABMD—which are the normalized hours quoted in Section IX-H.

TABLE 10. Wall-clock time comparison: one complete 5×10^5-step training run on MiniGrid-N6-S4, all methods timed on identical NVIDIA A100 hardware with the same ResNet-18 encoder.

Method	Time (hours)	Relative
DBC [3]	1.68	1.0×
MICo [11]	1.92	1.1×
SimSR [12]	2.09	1.2×
ABMD (ours)	3.24	1.9×

APPENDIX E
ADDITIONAL THEORETICAL RESULTS

A. CONNECTION TO SPECTRAL CLUSTERING

For a finite MDP with deterministic transitions and rewards, the bisimulation metric satisfies [4]

$$d_\sim(s, s') = (1-\gamma)|R(s) - R(s')| + \gamma \max_a d_\sim(T(s,a), T(s',a)). \quad (54)$$

This is equivalent to solving a system of linear equations. Letting $\mathbf{D}$ denote the matrix of pairwise bisimulation distances,

$$\mathbf{D} = (1-\gamma)\mathbf{R} + \gamma \mathbf{P}\mathbf{D}, \quad (55)$$

where $\mathbf{P}$ is the transition operator. Rearranging,

$$(\mathbf{I} - \gamma \mathbf{P})\mathbf{D} = (1-\gamma)\mathbf{R}. \quad (56)$$

Hence $d_\sim$ is the solution to a linear system involving the graph Laplacian of the MDP's state-transition graph, which connects bisimulation to spectral methods for representation learning [39].

B. INVARIANCE PROPERTIES

Proposition 8 (Reward Scaling Invariance): For any $\alpha > 0$, scaling the rewards as $R' = \alpha R$ scales the bisimulation metric by the same factor:

$$d'_\sim(s, s') = \alpha\, d_\sim(s, s'). \quad (57)$$

As a consequence, the BIL objective is equivariant under reward scaling and the optimal policy is invariant. The adaptive temperature τ_κ compensates automatically, since κ^* scales proportionally.

REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[2] M. Laskin, A. Srinivas, and P. Abbeel, "CURL: Contrastive unsupervised representations for reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 5639–5650.

[3] A. Pore, R. Muradore, and D. Dall'Alba, "DEAR: Disentangled environment and agent representations for reinforcement learning without reconstruction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2024, pp. 650–655.

[4] N. Ferns, P. Panangaden, and D. Precup, "Metrics for finite Markov decision processes," in *Proc. Conf. Uncertainty Artif. Intell. (UAI)*, 2004, pp. 162–169.

[5] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 488–489.

[6] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, "Exploration by random network distillation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[7] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 9929–9939.

[8] R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak, "Planning to explore via self-supervised world models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 8583–8592.

[9] R. Givan, T. Dean, and M. Greig, "Equivalence notions and model minimization in Markov decision processes," *Artif. Intell.*, vol. 147, nos. 1–2, pp. 163–223, Jul. 2003.

[10] D. Park, "Concurrency and automata on infinite sequences," in *Proc. 5th GI-Conf. Theor. Comput. Sci.* (Lecture Notes in Computer Science), vol. 104. Berlin, Germany: Springer, 1981, pp. 167–183.

[11] P. S. Castro, T. Kastner, P. Panangaden, and M. Rowland, "MICo: Improved representations via sampling-based state similarity for Markov decision processes," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021.

[12] H. Zang, X. Li, and M. Wang, "SimSR: Simple distance-based state representations for deep reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 8, pp. 8997–9005.

[13] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[14] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel, "VIME: Variational information maximizing exploration," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2016, pp. 1109–1117.

[15] J. Martens, "New insights and perspectives on the natural gradient method," *J. Mach. Learn. Res.*, vol. 21, no. 146, pp. 1–76, 2020.

[16] A. Genevay, G. Peyré, and M. Cuturi, "Learning generative models with Sinkhorn divergences," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2018, pp. 1608–1617.

[17] M. Chevalier-Boisvert, L. Willems, and S. Pal. (2018). *Minimalistic Gridworld Environment for OpenAI Gym*. [Online]. Available: https://github.com/Farama-Foundation/Minigrid

[18] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. de Las Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, T. Lillicrap, and M. Riedmiller, "DeepMind control suite," 2018, *arXiv:1801.00690*.

[19] A. Stone, O. Ramirez, K. Konolige, and R. Jonschkowski, "The distracting control suite - a challenging benchmark for reinforcement learning from pixels," 2021, *arXiv:2101.02722*.

[20] K. Cobbe, C. Hesse, J. Hilton, and J. Schulman, "Leveraging procedural generation to benchmark reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 2048–2056.

[21] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto, "Mastering visual continuous control: Improved data-augmented reinforcement learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[22] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Stat. Assoc.*, vol. 58, no. 301, pp. 13–30, Mar. 1963.

[23] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, U.K.: Oxford Univ. Press, 2013.

[24] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1050–1059.

[25] D. Russo and B. Van Roy, "Learning to optimize via information-directed sampling," *Operations Res.*, vol. 66, no. 1, pp. 230–252, Feb. 2018.
[26] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.
[27] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet, "Hilbert space embeddings and metrics on probability measures," *J. Mach. Learn. Res.*, vol. 11, pp. 1517–1561, Aug. 2010.
[28] G. Peyré and M. Cuturi, "Computational optimal transport with applications to data sciences," *Found. Trends Mach. Learn.*, vol. 11, nos. 5–6, pp. 355–607, 2019.
[29] J. Bourgain, "On Lipschitz embedding of finite metric spaces in Hilbert space," *Isr. J. Math.*, vol. 52, nos. 1–2, pp. 46–52, Mar. 1985.
[30] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *J. Mach. Learn. Res.*, vol. 11, pp. 1563–1600, Mar. 2010.
[31] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times*, 2nd ed. Providence, RI, USA: AMS, 2017.
[32] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, Feb. 1998.
[33] S.-I. Amari, *Information Geometry and Its Applications*. Tokyo, Japan: Springer, 2016.
[34] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML-PKDD)*, 2016, pp. 795–811.
[35] Z. Chen, H.-Z. Huang, Z. Deng, and J. Wu, "Shrinkage mamba relation network with out-of-distribution data augmentation for rotating machinery fault detection and localization under zero-faulty data," *Mech. Syst. Signal Process.*, vol. 224, Feb. 2025, Art. no. 112145.
[36] Z. Chen, J. Wu, and H.-Z. Huang, "Prompt-guided disentanglement and fusion framework for cross-domain fault diagnosis with class mismatch," *IEEE Trans. Ind. Informat.*, vol. 22, no. 9, pp. 8044–8055, Sep. 2026.
[37] Z. Chen, J. Wu, Z. Deng, and H.-Z. Huang, "Learning category-invariant disentangled features for domain generalization in machine fault diagnosis," *IEEE/ASME Trans. Mechatronics*, 2026, doi: 10.1109/TMECH.2026.3666385.
[38] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," 2023, *arXiv:2301.04104*.
[39] D. Shribak, C.-X. Gao, Y. Li, C. Xiao, and B. Dai, "Diffusion spectral representation for reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 110028–110056.
[40] M. Kemertas and T. Aumentado-Armstrong, "Towards robust bisimulation metric learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021.
[41] D. Pathak, D. Gandhi, and A. Gupta, "Self-supervised exploration via disagreement," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019.
[42] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped DQN," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2016.
[43] R. Raileanu and T. Rocktäschel, "RIDE: Rewarding impact-driven exploration for procedurally-generated environments," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.
[44] H. Liu and P. Abbeel, "APS: Active pretraining with successor features," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.
[45] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Trans. Evol. Comput.*, vol. 11, no. 2, pp. 265–286, Apr. 2007.
[46] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1597–1607.
[47] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.
[48] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Math. J.*, vol. 19, no. 3, pp. 357–367, Jan. 1967.
[49] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 1861–1870.
[50] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.



DEEPAK GUPTA is currently a Professor and the Head of the Department of CSE–Data Science, Institute of Technology and Management Gwalior, Madhya Pradesh, India. His research interests include artificial intelligence, machine learning, data science, and intelligent systems, with a focus on scalable and robust computational models for real-world applications. He has extensive experience in academic research, curriculum development, and guiding bachelor's and master's research projects in emerging areas of AI, and he is particularly interested in bridging theoretical foundations with practical system design. He serves as a senior author for this work.



ADITYA SIKARWAR (Student Member, IEEE) is currently pursuing the bachelor's degree in data science with the Institute of Technology and Management, Gwalior, Madhya Pradesh, India. He is also a Research Fellow with the Institute of Technology and Management. His research interests include reinforcement learning for sparse-reward environments, physics-informed neural networks, causal machine learning, neurosymbolic AI, and language model-based retrieval-augmented generation agents. He is actively exploring how bisimulation geometry and Infomax principles can unify self-supervised representation learning with curiosity-driven exploration. He is committed to advancing research at the intersection of symbolic reasoning and deep learning for autonomous decision-making systems.



SANIDHYA DHANGAR (Student Member, IEEE) is currently pursuing the bachelor's degree in data science with the Institute of Technology and Management, Gwalior, Madhya Pradesh, India. His research interests include deep reinforcement learning, self-supervised representation learning, and information-theoretic exploration strategies. He is particularly interested in the intersection of geometry and learning in Markov decision processes, with a focus on bisimulation-based state abstractions and curiosity-driven exploration in sparse-reward environments.



CHANAKYA KUMAR JHA (Senior Member, IEEE) received the Ph.D. degree (Hons.) in wireless communication from Sam Higginbottom Institute of Agriculture, Technology and Sciences, Allahabad. He is currently an accomplished academician, an engineer, and an institutional leader with over a decade of experience in higher education, IEEE leadership, and governance. He has been serving as the Vice Chair–Education Activities at IEEE India Council, since January 2026, and he has been the SIGHT Chair at IEEE Bombay Section for nearly 12 years, driving humanitarian technology initiatives across Mumbai Metropolitan Region. He was the Vice Principal at Sahyadri Valley College of Engineering Technology, University of Pune, from 2014 to 2025, contributed at board level to the Western Business School, Pune, and he has been serving as a member of Board of Trustees in an independent capacity, since 2021.



PALLAVI JHA (Member, IEEE) received the M.E. and Ph.D. degrees. She is currently an Associate Professor with Alard School of Engineering and Technology, where she contributes to both teaching and research in her discipline. She is also a dedicated academician and a researcher in the field of computer science and engineering. With her strong academic credentials, she is committed to advancing computer science education and fostering a culture of research and innovation among students and faculty alike.



AKHILESH KUMAR SHARMA (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees in computer science and engineering. He is currently a Professor and the Head of the Department of Data Science and Engineering, Manipal University Jaipur, India. He is an active researcher and an academician who has delivered expert lectures and keynote addresses across India and in Singapore, Thailand, and Malaysia, and he has built international collaborations through MoUs with institutions in Poland, Thailand, Saudi Arabia, and Malaysia. He also holds professional memberships with ACM (Senior Member), The Institution of Engineers (India), and Multimedia Information Retrieval Laboratory (USA). He serves as the Vice Chair of ACM Professional Chapter, Jaipur. He has published over 125 research papers in reputed international journals and conferences, holds more than 12 patents and five copyrights. He has authored books and book chapters while supervising Ph.D. scholars. He has also established a Cognitive Intelligence Research Laboratory, Jaipur, Rajasthan, India, focusing on emerging areas in artificial intelligence and intelligent systems.



VUGAR ABDULLAYEV is currently the Head of the Scientific Research Laboratory, Azerbaijan University of Architecture and Construction, Baku, Azerbaijan, and an Associate Professor (Ph.D.) at Azerbaijan State Oil and Industry University. Recognized among the World's Top 2% Scientists for 2025, his research expertise spans AI and data science, information technologies, big data, cloud computing, and educational technology (EdTech). He leads innovative research in digital systems and their pedagogical and technical applications, contributing significantly to both academic and industry-oriented projects.