DATA-DEPENDENT KERNEL-AWARE UNSUPERVISED SKILL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised reinforcement learning (URL) aims to train general agents by learning transferable knowledge through interaction without external rewards, then adapting to downstream tasks with fewer interactions. Skill discovery learns distinguishable, controllable skills with broad coverage through intra-skill consistency and inter-skill diversity. Temporal distance, however, does not explicitly adapt local similarity scales to the empirical state distribution. We propose **I**solation **K**ernel-aware **S**kill **D**iscovery (IKSD), which constructs data-adaptive similarity and geometric scales in a latent space. Online IKSD constrains representation changes with IK dissimilarity; offline IKSD learns representation distances from cumulative IK-weighted transition costs. In DMC Walker and Cheetah, online IKSD achieves the highest final mean valid-posture path length on challenging terrain and competitive position coverage on flat ground. On a fixed MetaWorld dataset, offline IKSD improves downstream performance and solves some tasks excluded from pretraining data collection.

1 INTRODUCTION

Reinforcement learning (RL) achieves strong performance in board games and robot control (Silver et al., 2016; Haarnoja et al., 2018). Training for a specific task, however, often requires a predefined reward and extensive environment interaction. The resulting policy may need further training to adapt to a new objective. Unsupervised reinforcement learning (URL) aims to reduce this repeated learning cost through pretraining without downstream rewards, so that agents can reuse learned knowledge during task adaptation (Laskin et al., 2021).

Skill discovery provides one route to this goal. It learns a policy $\pi(a \mid s, z)$ conditioned on a latent variable z to produce diverse, reusable behaviors without task rewards. Mutual information methods reward states or transitions that help distinguish skills (Eysenbach et al., 2019; Sharma et al., 2020). Distance maximization methods encourage representation displacements in skill-dependent directions and constrain their magnitude using a state distance (Park et al., 2022; 2024b). In these methods, the metric sets the allowed scale of representation changes between adjacent states and thus bounds the corresponding directional rewards. The choice of metric therefore influences which state changes the agent learns to favor.

Existing metrics define state differences through coordinate displacement, conditional transition density (Park et al., 2023b), or the number of interaction steps. We study another source of information: the local density of states in the training data. For two state pairs with the same Euclidean separation, this separation may exceed the typical spacing between nearby samples in a dense region but fall below it in a sparse region. A similarity scale based on absolute separation cannot reflect this difference relative to the local sample distribution. Temporal distance does not explicitly specify how this scale should adapt to empirical density. Calibrating state differences against local density allows the representation constraint and the resulting bounds on directional rewards to adapt to the data distribution. This empirical state density is distinct from the conditional transition density used by CSD and from reachability under the environment dynamics.

We introduce **I**solation **K**ernel-aware **S**kill **D**iscovery (IKSD), which uses the Isolation Kernel (IK) to define density-aware state distances (Ting et al., 2018). IK constructs random partitions from reference data and estimates similarity by the fraction of partitions in which two states fall in the same cell. The local scale of these partitions varies with the data distribution, so the resulting

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

geometry reflects empirical structure without an explicit state density model. The online version of IKSD uses the IK-induced dissimilarity $1 - \kappa_{\text{iso}}(s, s')$ to constrain representation displacements between adjacent states. It retains the objective of maximizing displacement along the skill direction, with the data distribution setting the local constraint scale.

IKSD also supports offline skill pretraining. In this setting, the agent cannot collect new experience to expand coverage and must learn from the structure of a fixed dataset. HILP learns representations in a Hilbert space by fitting temporal distances and uses them to train policies conditioned on directions in that space (Park et al., 2024a). We build the offline version of IKSD within the HILP framework by replacing its unit step cost with an IK-weighted local cost and learning representation distances from cumulative costs. The online and offline versions use density-aware geometry as a representation constraint and a fitting target, respectively, allowing us to examine the design under continued interaction and fixed data.

We evaluate IKSD through local neighborhood retrieval, online locomotion, and offline manipulation. The online experiments compare IKSD, METRA, GenDa (Park et al., 2026a), CSF (Zheng et al., 2025), and CSD. IKSD achieves the highest final mean valid-posture path length on challenging terrain in Walker and Cheetah. With pretrained skills frozen, PPO control tests their composition on downstream tasks. Offline IKSD reaches 67.71% mean success on unseen MetaWorld tasks. Comparisons with fixed and locally adaptive Gaussian kernels, together with window-closing trajectory visualizations, examine kernel choice and skill reuse.

Our main contributions are as follows:

- We examine data-dependent metrics through local neighborhood retrieval and compare fixed Gaussian kernels, locally adaptive Gaussian kernels, and IK within a shared offline learning framework to test the effect of kernel choice on skill learning.
- **I**solation **K**ernel-aware **S**kill **D**iscovery (IKSD) uses dissimilarity induced by IK to adapt the constraint on representation displacements between adjacent states. It achieves higher final mean valid-posture path lengths than METRA, GenDa (Park et al., 2026a), CSF (Zheng et al., 2025), and CSD on challenging terrain in Walker and Cheetah, while maintaining competitive position coverage on flat ground.
- Within the HILP framework, offline IKSD replaces the unit step cost with an IK-based local cost, learns cumulative distances from fixed data, and achieves higher mean success than HILP on seen and unseen MetaWorld tasks.

2 Background and Related Work

2.1 Skill Discovery and State Distances

Skill discovery learns $\pi(a \mid s, z)$ with $z \sim p(z)$ fixed over a trajectory segment. Online pretraining collects experience; offline pretraining uses fixed reward-free transitions $\mathcal{D} = \{(s, a, s')\}$. Downstream tasks use skill selection, composition, or policy fine-tuning. Appendix C.1 gives the full setting.

For $\phi : \mathcal{S} \to \mathbb{R}^d$, distance maximization methods interpret $z \in \mathbb{R}^d$ as a direction and optimize (Park et al., 2022; 2023b; 2024b)

$$\begin{aligned} \max_{\pi,\phi} \quad & \mathbb{E}_{z\sim p(z),\, \tau\sim\pi(\cdot|\cdot,z)} \left[\sum_{t=0}^{T-1} \big(\phi(s_{t+1}) - \phi(s_t)\big)^\top z \right], \\ \text{subject to} \quad & \|\phi(s') - \phi(s)\|_2 \le d(s, s'), \qquad \text{for observed transitions } (s, s'). \end{aligned} \tag{1}$$

The directional reward encourages representation displacement, while d limits its magnitude and determines which state changes permit larger representation differences.

Euclidean, temporal, and conditional transition distances provide alternative scales. IKSD adapts to empirical state density, distinct from CSD's conditional transition density. Appendix C.2 details LSD, CSD, METRA, HILP, and other approaches; Appendix C.3 defines fixed and locally adaptive Gaussian controls.

Offline goal-conditioned RL learns policies from fixed data to reach specified goals. OGBench provides tasks, datasets, and reference implementations for this setting (Park et al., 2025). Offline CRL is an offline implementation of CRL that uses contrastive learning to score compatibility between state-action pairs and goals, then trains a goal-conditioned policy from these scores (Eysenbach et al., 2022). GCIVL adopts HIQL's implicit value learning objective to fit $V(s, g)$ through expectile regression, without a separate Q-function (Park et al., 2023a; 2025). Dual-GCIVL combines GCIVL with dual goal representations based on temporal distances from different starting states to the goal (Park et al., 2026b). Using the triangle inequality for temporal distances, TRL decomposes long-horizon value estimation through intermediate states and combines the resulting estimates (Park et al., 2026c). We use these offline baselines to evaluate goal reaching with the skills learned by IKSD.

2.2 The Isolation Kernel and Its Induced Distance

The Isolation Kernel (IK) defines similarity through random spatial partitions based on reference data (Ting et al., 2018). Given reference states $\mathcal{B}$ and an input representation g, each partition uses ψ randomly sampled reference points. We use complete partitions covering the query space, with $c_\ell(s)$ denoting the cell containing $g(s)$ in partition ℓ. The fraction of m independent partitions in which two states share a cell defines

$$\kappa_{\text{iso}}(s, s'; \mathcal{B}) = \frac{1}{m} \sum_{\ell=1}^{m} \mathbf{1}\{c_\ell(s) = c_\ell(s')\}. \tag{2}$$

Reference samples set cell scales, adapting similarity to the empirical distribution without explicit density estimation.

We use $d_{\text{IK}}(s, s'; \mathcal{B}) = 1 - \kappa_{\text{iso}}(s, s'; \mathcal{B})$. Under complete partitions, this is a normalized Hamming pseudometric in $[0, 1]$ between cell-index sequences. Distinct states can share all cell indices and have zero distance. The Hilbert feature distance is $\sqrt{2d_{\text{IK}}}$. Appendix C.4 gives the feature map and identities; Appendix A proves these properties and bounds finite-partition approximation error.

3 Density-Aware Distances for Skill Discovery

3.1 Local Scale and State Differences

Equal coordinate separations can span several typical sample gaps in dense regions but remain local in sparse regions. A fixed global scale in Equation (1) cannot express this difference. IK's random reference partitions tend to be finer in dense regions and coarser in sparse regions. Equally separated pairs in sparse regions are typically more likely to share a cell, yielding higher similarity and smaller d_{IK} (Ting et al., 2018). IK measures separation relative to local sample spacing.

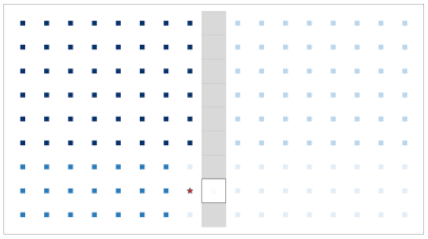
This geometry depends on reference data shaped by the dynamics, behavior policy, and sampling process. Local sample structure can identify neighborhoods relevant to control when it reflects behavioral differences. We examine this connection in two toy experiments: a two-room random walk tests retrieval of states with high arrival probabilities, and PointMaze tests local candidate ordering.

3.2 Neighborhood Retrieval in a Two-Room Random Walk

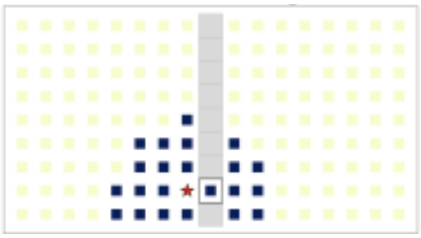
Consider the two-dimensional random walk in Figure 1. Two rooms connect through a narrow doorway, and states have different probabilities of remaining in place. We fix an anchor near the doorway and define the reference set $\mathcal{R}_k$ as the k states with the highest H-step arrival probabilities from that anchor. Each method retrieves k states in order of increasing distance or decreasing similarity, forming a set $\mathcal{N}_k$. We measure set agreement by intersection over union,

$$\text{IoU}(\mathcal{N}_k, \mathcal{R}_k) = \frac{|\mathcal{N}_k \cap \mathcal{R}_k|}{|\mathcal{N}_k \cup \mathcal{R}_k|}. \tag{3}$$

The reference set follows the arrival-probability ranking and differs from the set of all states reachable within H steps. Its definition is independent of the kernel under evaluation. Appendix B.2 gives the parameter settings.



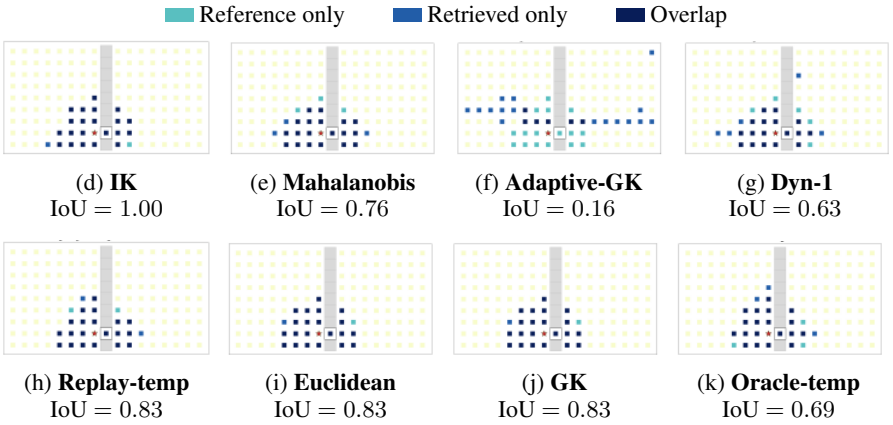
(a) **Environment and stay probabilities.**



(b) **The k states with the highest arrival probabilities.**

Method	IK	Mahalanobis	Adaptive-GK	Dyn-1	Replay-temp	Euclidean	GK	Oracle-temp
IoU ↑	1.00	0.76	0.16	0.63	0.83	0.83	0.83	0.69

(c) **IoU between retrieved and reference sets.** Each method occupies one column.



(d) **IK** IoU = 1.00 (e) **Mahalanobis** IoU = 0.76 (f) **Adaptive-GK** IoU = 0.16 (g) **Dyn-1** IoU = 0.63

(h) **Replay-temp** IoU = 0.83 (i) **Euclidean** IoU = 0.83 (j) **GK** IoU = 0.83 (k) **Oracle-temp** IoU = 0.69

Figure 1: **Local distance diagnostics in the two-room random walk.** The star marks the anchor, and gray regions indicate walls. In (a), darker states have higher stay probabilities. In (b), dark blue states form the reference set. Panel (c) reports IoU, and panels (d) through (k) show the neighborhoods retrieved by all eight methods. Light cyan marks states in the reference set only, blue marks states in the retrieved set only, and dark blue marks their overlap. IK retrieves the reference set exactly.

The baselines include Euclidean distance, Mahalanobis distance, a fixed-bandwidth Gaussian kernel (GK), a locally adaptive Gaussian similarity (Adaptive-GK), and a distance between one-step successor distributions (Dyn-1). We also compare two temporal distances. Replay-temp uses empirical temporal distances in replay trajectories; Oracle-temp uses shortest-path lengths in the true transition graph. The latter provides path-length information, but states at the same shortest-path distance can have different arrival probabilities. It therefore does not give an upper bound for this retrieval task.

In Figure 1(c), IK achieves an IoU of 1.00, compared with 0.83 for Euclidean distance, GK, and Replay-temp, and 0.69 for Oracle-temp. In this setting, the IK neighborhood exactly matches the reference set defined by finite-step arrival probabilities. GK and Euclidean distance give the same result, consistent with their shared ordering. Adaptive-GK achieves an IoU of 0.16, showing that the choice of local scale matters: dependence on data alone does not ensure useful neighborhoods.

This example connects the empirical distribution, a geometric bottleneck, and finite-step arrival probabilities. The stay probabilities are part of the dynamics, so the experiment does not isolate the effect of sampling density. That requires a controlled intervention that changes the empirical sampling distribution while holding the dynamics fixed. IK is symmetric, and its agreement with arrival probabilities depends on the representation and data coverage; it does not recover general directed reachability.

3.3 Local Ranking in PointMaze

Methods that retrieve the same neighborhood can still rank its states differently. To test whether a distance ranks empirically easier-to-reach states first, we construct local ranking tasks from offline data in PointMaze-Umaze and PointMaze-Large (Fu et al., 2020). For a fixed anchor, empirical arrival probabilities distinguish easy from hard candidate states and define the reference ordering. Figure 2 shows the anchors and candidate regions. Each method ranks candidates by distance or similarity. We evaluate the ranking using normalized discounted cumulative gain (NDCG) (Järvelin & Kekäläinen, 2002), where higher values indicate better agreement with the probability-based reference.

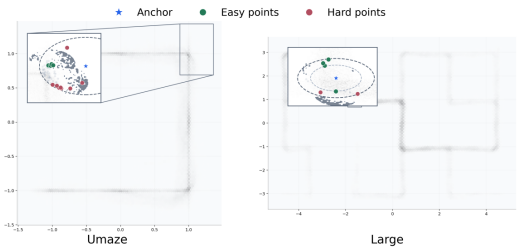


Figure 2: **Local ranking tasks in PointMaze-Umaze (left) and PointMaze-Large (right).** Blue stars mark anchors. Green and red circles mark easy and hard candidates, respectively, as determined by empirical arrival probabilities. Insets show the local spatial relationships among the candidates.

IK achieves the highest NDCG in both tasks, with 0.9864 in Umaze and 1.0000 in Large (Table 1). The strongest baselines are Oracle-temp in Umaze (0.9368) and Euclidean distance in Large (0.9701). Across these two layouts, IK produces an ordering that better matches the empirical arrival probabilities in the corresponding trajectory data. This result concerns the specified anchors and candidate regions; the reference probabilities depend on the behavior policy and data coverage.

3.4 From Local Distances to Skill Learning

The diagnostics connect to skill learning through the objective. With $d = d_{\text{IK}}$ and $\|z\|_2 = 1$, Equation (1) bounds the directional reward $r_z(s, s') = (\phi(s') - \phi(s))^\top z$ by $|r_z(s, s')| \leq d_{\text{IK}}(s, s'; \mathcal{B})$. Similar pairs face tighter constraints; less similar pairs allow larger representation changes.

Appendix C.6 derives this bound and its online and offline uses. Appendix A analyzes constraint violations and their accumulation along paths.

4 Isolation Kernel-Aware Skill Discovery

4.1 Online Unsupervised Pretraining

Online IKSD constructs the kernel in Equation (2) from replay-buffer states $\mathcal{B}$. The kernel input g defines the reference geometry. For fixed reference data and partitions, we maximize Equation (1) with the constraint

$$\|\phi(s_{t+1}) - \phi(s_t)\|_2 \leq 1 - \kappa_{\text{iso}}(s_t, s_{t+1}; \mathcal{B}). \tag{4}$$

The policy receives $r_z(s, s') = (\phi(s') - \phi(s))^\top z$. Observed transitions specify adjacency; IK sets local scales for changes in the reward representation ϕ. We do not interpret kernel similarity as the environment's transition probability. Appendix C.5 gives the expanded objective.

4.2 Offline Unsupervised Pretraining

Given a fixed reward-free dataset $\mathcal{D} = \{(s, a, s')\}$, we use HILP's goal-conditioned representation and skill policy learning framework (Park et al., 2024a). HILP measures temporal distance in interaction steps; offline IKSD defines the local transition cost

$$c_{\text{IK}}(s, s') = \beta + \alpha d_{\text{IK}}(s, s'; \mathcal{B}), \qquad \alpha > 0, \quad \beta \geq 0. \tag{5}$$

Here, β is the base step cost and α weights IK dissimilarity. The goal-conditioned value is the negative representation distance. Temporal-difference learning accumulates local costs, incorporating observed transition structure and local empirical scales.

Table 1: **NDCG in the two local PointMaze ranking tasks.** Each row corresponds to one local task, and each method occupies one column. Higher values indicate better ranking quality.

Task	IK	GK	Adaptive-GK	Euclidean	Mahalanobis	Oracle-temp	Replay-temp	Dyn-1
Umaze ↑	0.9864	0.9047	0.8480	0.7638	0.8059	0.9368	0.8710	0.8004
Large ↑	1.0000	0.8592	0.9277	0.9701	0.9413	0.8592	0.8538	0.9277

Offline RL trains the skill policy from $\mathcal{D}$ with the same directional reward. Appendix B.5 details reference sampling, target-network kernel inputs, stopped cost gradients during representation updates, and offline policy learning.

4.3 Downstream Task Adaptation

We freeze the pretrained low-level skill policy and train a high-level selector with proximal policy optimization (PPO) (Schulman et al., 2017). The selector chooses a skill every 5 environment steps, and the low-level policy executes that skill throughout the interval. We train the selector with downstream external rewards and use the same 500k-step downstream interaction budget for each method. Appendix B.6 gives the task settings.

5 Experiments

We evaluate IKSD through the following questions. (1) **RQ1 (Effectiveness):** Does IKSD improve online exploration, downstream control with frozen skills, and offline success on seen and unseen tasks? (2) **RQ2 (Kernel choice):** How does IK compare with fixed and locally adaptive Gaussian kernels within the same offline framework? (3) **RQ3 (Behavior and representations):** Do the learned skills exhibit inter-skill diversity and intra-skill consistency, and how do unseen-task trajectories relate to pretraining experience in representation space?

5.1 Experimental Setup

We evaluate skill discovery in locomotion and robotic manipulation (Figure 3). Online experiments use Walker and Cheetah from the DeepMind Control Suite (Tassa et al., 2018). We add shallow steps for Walker and smooth hills for Cheetah to test movement as ground contact varies with position. We also evaluate both agents on the original flat ground.

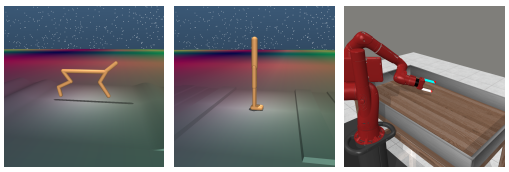


Figure 3: **Environment views.** DMC terrain uses the base scale in Appendix B.3; MetaWorld shows the arm and empty workbench.

These online experiments cover four settings, each with 5×10^6 training steps. The original flat-ground environments serve as standard benchmarks for spatial coverage. Appendix B.3 gives terrain construction and evaluation details.

Offline experiments use the same reward-free MetaWorld (Yu et al., 2020; McLean et al., 2025) data for all methods. Four manipulation tasks supply pretraining data; evaluation covers these and three held-out tasks (Table 2) without downstream fine-tuning. The split holds out task objectives, although related states or local behaviors may occur in pretraining data. Appendix B.4 gives the task split and evaluation protocol.

The online comparison includes IKSD, METRA (Park et al., 2024b), GenDa (Park et al., 2026a), CSF (Zheng et al., 2025), and CSD (Park et al., 2023b). For CSF and CSD, we use implementations adapted to DMC.

Offline baselines include HILP (Park et al., 2024a), Offline CRL (Eysenbach et al., 2022), GCIVL (Park et al., 2023a), Dual-GCIVL (Park et al., 2026b), and TRL (Park et al., 2026c). For the kernel ablation, GKSD and AGKSD replace IK in offline IKSD with a fixed-bandwidth Gaussian kernel and a locally adaptive Gaussian kernel, respectively. These variants retain the data, networks, optimization, and evaluation settings of IKSD. Offline IKSD uses the local cost in Equation (5).

On flat ground, we report the number of position bins visited by the policy to measure spatial coverage. On challenging terrain, we report valid-posture path length per 10-second trajectory to measure the cumulative distance of valid movement. Higher values are better for both metrics, but they measure coverage and valid movement, respectively, so we compare values only within the same setting. Each method uses 8 independent training seeds. Solid lines show the mean, and shaded bands show one standard deviation above and below the mean. Appendix B.3 gives the evaluation details.

5.2 Main Results

At the end of training, IKSD improves mean valid-posture path length over METRA, the strongest baseline, by 11.1% and 27.8% on challenging terrain in Walker and Cheetah, respectively (Figure 4). On the original flat-ground environments, IKSD achieves position coverage comparable to that of strong baselines.

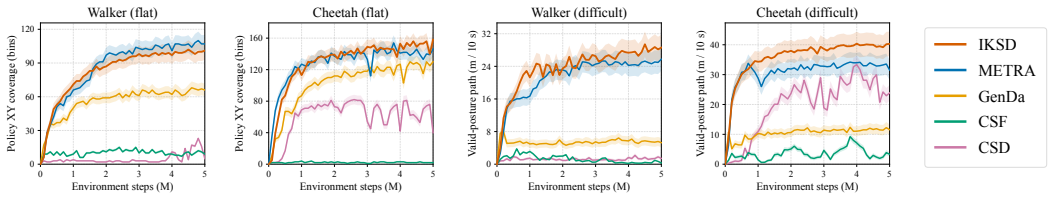


Figure 4: **Online IKSD training performance in DMC.** From left to right: original flat-ground Walker, original flat-ground Cheetah, Walker with challenging terrain, and Cheetah with challenging terrain. The shared legend is on the right. The first two panels report position coverage in bins; the last two report valid-posture path length per 10-second trajectory. The horizontal axis shows environment steps in millions. Solid lines show the mean over 8 independent training seeds, and shaded bands show one standard deviation above and below the mean.

We freeze the pretrained skills on challenging terrain and train a PPO selector that chooses a skill every 5 steps. Figure 5 shows that IKSD maintains the highest speed tracking score on Cheetah Forward and performs comparably to METRA on Walker Forward. It reaches a higher completion fraction late in training on Walker Shuttle, with considerable variation. On Cheetah Backward, its final score falls below METRA, CSD, and GenDa. The frozen IKSD skills thus support forward and shuttle control, while their usefulness varies with the direction of the task.

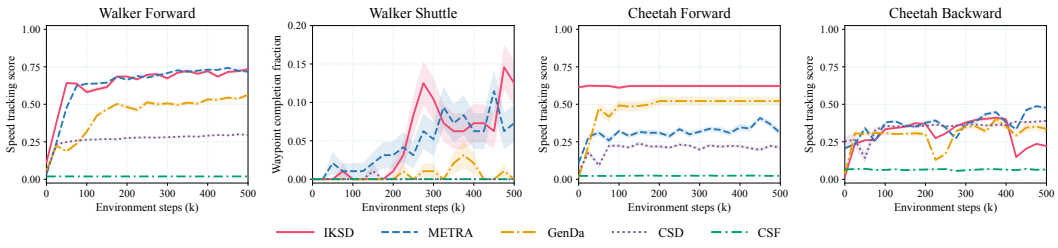


Figure 5: **Downstream PPO control on challenging terrain.** We freeze low-level skills and select a skill every 5 environment steps. From left to right: Walker Forward, Walker Shuttle, Cheetah Forward, and Cheetah Backward. Forward and backward tasks report speed tracking scores; the shuttle task reports the fraction of waypoints completed. The horizontal axis counts downstream environment steps in thousands.

IKSD achieves a mean success rate of 67.71% on the three unseen MetaWorld tasks (Table 2). This exceeds GCIVL, the strongest of the five baselines, by 32.29 percentage points. On the four seen tasks, IKSD also has the highest mean, 93.47%. It leads on window closing and ties for best on button pressing, while GCIVL performs better on faucet closing.

Table 2: **MetaWorld success rates (%).** Means use 8 training seeds and 64 evaluation episodes per seed and task. We evaluate all policies without downstream fine-tuning. The two mean columns average over the four seen tasks and three unseen tasks, respectively. Bold marks the best value in each column, including ties.

	Seen tasks					Unseen tasks			
Method	Door open	Door close	Window open	Dial turn	Mean	Window close	Button press	Faucet close	Mean
HILP	**100.00**	**100.00**	81.25	50.00	82.81	0.00	**62.50**	0.00	20.83
Offline CRL	**100.00**	78.12	37.50	59.38	68.75	0.00	50.00	18.75	22.92
GCIVL	**100.00**	93.75	76.88	84.38	88.75	21.88	34.38	**50.00**	35.42
Dual-GCIVL	**100.00**	**100.00**	65.62	78.12	85.94	0.00	40.62	0.00	13.54
TRL	**100.00**	81.25	**96.88**	78.12	89.06	0.00	59.38	21.88	27.09
GKSD	**100.00**	**100.00**	78.12	62.50	85.16	0.00	**62.50**	0.00	20.83
AGKSD	**100.00**	**100.00**	68.75	78.12	86.72	84.38	34.38	0.00	39.59
IKSD	**100.00**	**100.00**	85.75	**88.12**	**93.47**	**100.00**	**62.50**	40.62	**67.71**

5.3 Offline Kernel Comparison

The offline kernel ablation changes the kernel while keeping all other training and evaluation settings fixed. Local bandwidth adaptation increases mean success on unseen tasks from 20.83% for GKSD to 39.59% for AGKSD; IKSD reaches 67.71% (Table 2). Relative to AGKSD, IKSD gains 6.75 percentage points on seen tasks and 28.12 points on unseen tasks.

5.4 Behavior and Representation Analysis

Figure 6 provides a qualitative view of inter-skill diversity and intra-skill consistency in online IKSD. The four displayed skills differ in posture or movement direction, while the two rollouts of each skill from different initial states produce similar posture patterns and net displacements. Complete examples appear in Appendix D.

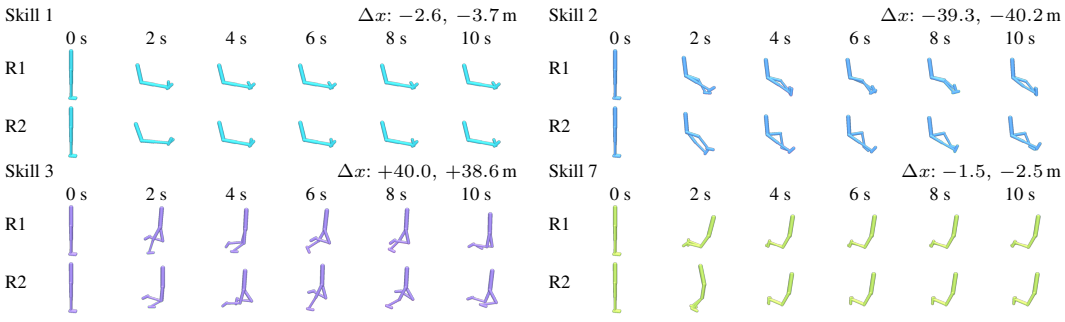


Figure 6: **Online IKSD skills on challenging Walker terrain at 5M steps.** R1 and R2 start from different states. Snapshots follow each agent over 10 seconds; paired Δx values give their signed horizontal displacements. Different skills produce distinct behaviors, with similar outcomes across each pair.

Figure 7 illustrates the window-closing comparison in Table 2. With both pretrained policies frozen, offline IKSD first reaches the success condition at step 48, whereas HILP does not complete the task within the displayed 500-step rollout.

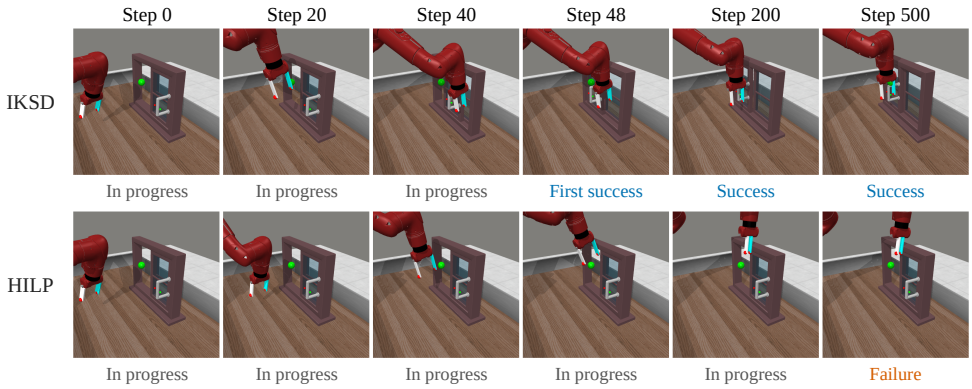


Figure 7: **Frozen policies on held-out window closing.** Columns align evaluation steps. IKSD (top) first succeeds at step 48; HILP (bottom) does not complete the 500-step rollout.

Figure 8 compares a held-out window-closing trajectory with pretraining experience. In the upper PCA projections, the IKSD trajectory lies near window-opening samples and its nearest training trajectory. The displayed pair shows closer local correspondence than in HILP, GKSD, and AGKSD.

The lower row maps pairwise dynamic time warping (DTW) distances with t-SNE (van der Maaten & Hinton, 2008), using one point per trajectory. Window-closing and window-opening trajectories overlap locally in the IKSD panel, supporting reuse of familiar movement segments for a held-out objective. Coordinate scales differ across methods. These projections provide qualitative evidence of shared behavior, but do not establish that a complete window-closing trajectory occurs in the pretraining data.

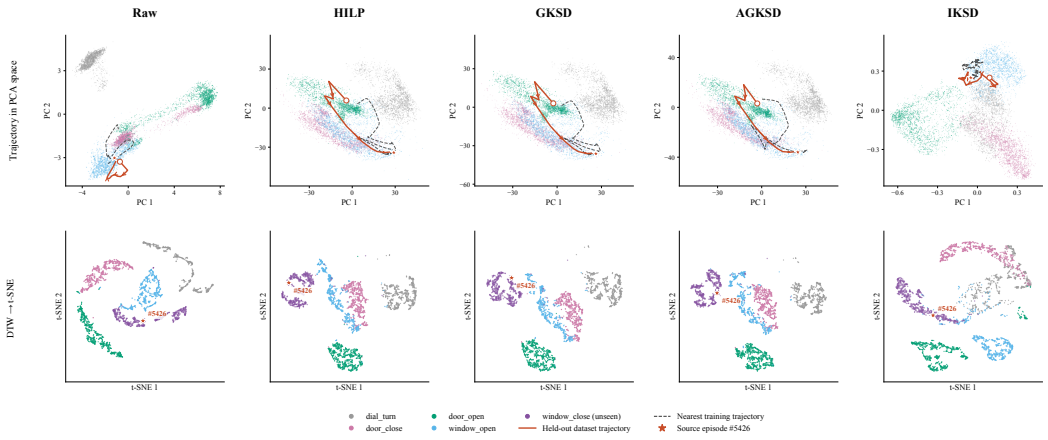


Figure 8: **Window-closing trajectories and pretraining data.** Columns show raw states and four learned representations. Top: PCA, with the held-out trajectory in red and its nearest training trajectory dashed. Bottom: t-SNE of pairwise DTW distances; each point is a trajectory, colors indicate tasks, and the star marks episode #5426.

6 Conclusion

IKSD uses Isolation Kernel dissimilarity to constrain online representations and cumulative IK-weighted costs to learn offline representations within HILP. It improves valid-posture path length on challenging DMC terrain and mean success on seen and unseen MetaWorld tasks. Frozen skills support downstream control, while window-closing visualizations suggest reuse of familiar behaviors. Transfer depends on whether those skills cover the behaviors required by the task.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, including improving wording, grammar, and clarity, and to assist with literature retrieval and discovery, such as identifying potentially relevant related work. All AI-assisted content was reviewed by the authors, and literature identified with AI assistance was checked against the corresponding original sources before inclusion. We take responsibility for the final content of this work, including all text, claims, citations, and artifacts produced with the aid of generative AI.

REFERENCES

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations (ICLR)*, 2019.

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

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv*, abs/2004.07219, 2020.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications. *arXiv*, abs/1812.05905, 2018.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, and Pieter Abbeel. URLB: Unsupervised reinforcement learning benchmark. In *Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, 2021.

Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Hao Liu and Pieter Abbeel. APS: Active pretraining with successor features. In *International Conference on Machine Learning (ICML)*, 2021b.

Reginald McLean, Evangelos Chatzaroulas, Luc McCutcheon, Frank Röder, Tianhe Yu, Zhanpeng He, K.R. Zentner, Ryan Julian, J K Terry, Isaac Woungang, Nariman Farsad, and Pablo Samuel Castro. Meta-World+: An improved, standardized, RL benchmark. In *Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, 2025.

Jongchan Park, Seungjun Oh, Seungho Baek, and Yusung Kim. Learning generalizable skill policy with data-efficient unsupervised RL. In *International Conference on Machine Learning (ICML)*, 2026a.

Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. Lipschitz-constrained unsupervised skill discovery. In *International Conference on Learning Representations (ICLR)*, 2022.

Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. HIQL: Offline goal-conditioned RL with latent states as actions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Seohong Park, Kimin Lee, Youngwoon Lee, and Pieter Abbeel. Controllability-aware unsupervised skill discovery. In *International Conference on Machine Learning (ICML)*, 2023b.

Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with Hilbert representations. In *International Conference on Machine Learning (ICML)*, 2024a.

Seohong Park, Oleh Rybkin, and Sergey Levine. METRA: Scalable unsupervised RL with metric-aware abstraction. In *International Conference on Learning Representations (ICLR)*, 2024b.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking offline goal-conditioned RL. In *International Conference on Learning Representations (ICLR)*, 2025.

Seohong Park, Deepinder Mann, and Sergey Levine. Dual goal representations. In *International Conference on Learning Representations (ICLR)*, 2026b.

Seohong Park, Aditya Oberai, Pranav Atreya, and Sergey Levine. Transitive RL: Value learning via divide and conquer. In *International Conference on Learning Representations (ICLR)*, 2026c.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL https://arxiv.org/abs/1707.06347.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International Conference on Machine Learning (ICML)*, 2020.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. In *International Conference on Learning Representations (ICLR)*, 2020.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind control suite. *arXiv*, abs/1801.00690, 2018.

Kai Ming Ting, Yue Zhu, and Zhi-Hua Zhou. Isolation kernel and its effect on SVM. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2018.

Kai Ming Ting, Bi-Cun Xu, Takashi Washio, and Zhi-Hua Zhou. Isolation distributional kernel: A new tool for point and group anomaly detections. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2697–2710, 2023.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL https://jmlr.org/papers/v9/vandermaaten08a.html.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2020.

Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2004.

Chongyi Zheng, Jens Tuyls, Joanne Peng, and Benjamin Eysenbach. Can a MISL fly? Analysis and ingredients for mutual information skill learning. In *International Conference on Learning Representations (ICLR)*, 2025.

A THEORETICAL PROPERTIES OF IKSD

The Isolation Kernel converts the partition structure of reference data into a constraint on representation changes. We analyze the resulting geometry, the approximation error from finitely many partitions, and the connection to skill returns. We then derive sufficient conditions for terminal-state separation and regional coverage. The analysis covers the scaled distance family

$$\widehat{d}_\alpha(s, s') = \alpha\big(1 - \widehat{\kappa}_m(s, s')\big), \qquad 0 < \alpha < \infty. \tag{A.1}$$

Setting $\alpha = 1$ recovers the distance in the main text. Except in the analysis of objective drift, we fix the reference data, the encoder used to construct the kernel, and the sampled partitions. The scale α is a prescribed constant.

A.1 PARTITION CONSTRUCTION AND KERNEL GEOMETRY

Let $\mathcal{B}$ denote the reference replay data and let $g : \mathcal{S} \to \mathcal{X}$ be the fixed encoder used to construct the Isolation Kernel. If the kernel uses state features directly, g is the identity map. Each random partition H_ℓ divides $\mathcal{X}$ into ψ disjoint cells that cover the query space, with a fixed rule for assigning boundary points. Write $c_\ell(s)$ for the index of the cell containing $g(s)$. Under this complete-partition construction, the empirical co-clustering kernel is

$$\widehat{\kappa}_m(s, s') = \frac{1}{m} \sum_{\ell=1}^{m} \mathbf{1}\{c_\ell(s) = c_\ell(s')\}, \tag{A.2}$$

$$\widehat{d}_0(s, s') := 1 - \widehat{\kappa}_m(s, s'), \qquad \widehat{d}_\alpha = \alpha \widehat{d}_0. \tag{A.3}$$

This co-clustering definition follows the random-partition construction of the Isolation Kernel (Ting et al., 2018). The subscript 0 denotes the unscaled base distance, not the choice $\alpha = 0$. We distinguish the reward representation $\varphi : \mathcal{S} \to \mathbb{R}^d$ from the kernel encoder g: the former is subject to the constraint, whereas the latter defines the reference geometry. Here φ corresponds to the representation ϕ in the main text.

Theorem A.1 (Pseudometric and feature distances)**.** *For any fixed collection of complete partitions, $\widehat{d}_\alpha$ is a pseudometric with values in $[0, \alpha]$. Let $\Phi_m(s) \in \{0, 1\}^{m\psi}$ concatenate the one-hot encodings of the partition cells, and let $\Psi_m(s) = m^{-1/2}\Phi_m(s)$. Then*

$$\widehat{\kappa}_m(s, s') = \langle \Psi_m(s), \Psi_m(s') \rangle, \qquad \|\Psi_m(s)\|_2 = 1, \tag{A.4}$$

$$\widehat{d}_\alpha(s, s') = \frac{\alpha}{2m} \|\Phi_m(s) - \Phi_m(s')\|_1 = \frac{\alpha}{2} \|\Psi_m(s) - \Psi_m(s')\|_2^2. \tag{A.5}$$

On a given set of states, $\widehat{d}_\alpha$ is a metric if and only if the partitions distinguish every pair of distinct states in that set.

Proof. Define $d_\ell(s, s') = \mathbf{1}\{c_\ell(s) \neq c_\ell(s')\}$. If $c_\ell(s) \neq c_\ell(s'')$, then at least one of $c_\ell(s) \neq c_\ell(s')$ and $c_\ell(s') \neq c_\ell(s'')$ holds. Hence

$$d_\ell(s, s'') \leq d_\ell(s, s') + d_\ell(s', s'').$$

Each d_ℓ is nonnegative, symmetric, and zero on the diagonal. Their scaled average $\widehat{d}_\alpha = \alpha m^{-1} \sum_\ell d_\ell$ has the same properties and satisfies the triangle inequality. Within each one-hot block, the inner product equals 1 for matching cells and 0 otherwise. For two different cells, both the ℓ_1 norm and the squared ℓ_2 norm of the block difference equal 2. Summing over blocks proves Eqs. (A.4) and (A.5). The distance vanishes exactly when all cell indices agree, which proves the metric characterization. □

The proof uses the complete-partition assumption. Hypersphere implementations that assign a zero feature block to a query outside all valid balls (Ting et al., 2023) can have $\widehat{\kappa}_m(s, s) < 1$. Theorem A.1 then does not apply to $1 - \widehat{\kappa}_m$. Such an implementation requires an explicit encoding of the residual region or a different distance with its own proof. For finite m and ψ, there are at most ψ^m distinct cell-index combinations, so this construction cannot define a metric on an infinite continuous state space.

Kernel mean embeddings extend the comparison to probability distributions (Gretton et al., 2012; Ting et al., 2023). For a distribution P, define $\mu_P = \mathbb{E}_P \Psi_m(S)$. This expectation exists because the feature vectors have unit norm. The distributional kernel and maximum mean discrepancy are

$$K_m(P,Q) = \langle \mu_P, \mu_Q \rangle = \frac{1}{m} \langle \mathbb{E}_P \Phi_m, \mathbb{E}_Q \Phi_m \rangle , \tag{A.6}$$

$$\mathrm{MMD}^2_{\widehat{\kappa}_m}(P,Q) = K_m(P,P) + K_m(Q,Q) - 2K_m(P,Q). \tag{A.7}$$

Proposition A.2 (Distributional distance identity)**.** *Let $X, X' \sim P$ and $Y, Y' \sim Q$ be mutually independent. Then*

$$\mathrm{MMD}^2_{\widehat{\kappa}_m}(P,Q) = \frac{1}{\alpha}\Big(2\mathbb{E}\widehat{d}_\alpha(X,Y) - \mathbb{E}\widehat{d}_\alpha(X,X') - \mathbb{E}\widehat{d}_\alpha(Y,Y')\Big). \tag{A.8}$$

In particular, $\mathrm{MMD}^2_{\widehat{\kappa}_m}(\delta_s, \delta_{s'}) = 2\widehat{d}_\alpha(s,s')/\alpha$.

Proof. Substitute $\widehat{\kappa}_m = 1 - \widehat{d}_\alpha/\alpha$ into Eq. (A.7). The constant terms cancel. For the point-mass case, set $P = \delta_s$ and $Q = \delta_{s'}$ and use $\widehat{d}_\alpha(s,s) = 0$. □

Equation (A.8) relates expected distances between and within distributions. It does not express the directional reward objective as an MMD between joint distributions. A mean embedding with finitely many features need not distinguish all probability distributions. The quantity $\widehat{d}_\alpha$ is half the squared RKHS distance for the scaled kernel $\alpha\widehat{\kappa}_m$; the corresponding RKHS distance is $\sqrt{2\widehat{d}_\alpha}$.

A.2 Concentration and representation bounds on replay graphs

Conditional on $(\mathcal{B}, g)$, define the population kernel and base distance by averaging over random partitions:

$$\kappa_{\mathcal{B}}(s,s') = \Pr_H\{c_H(s) = c_H(s')\}, \qquad d_{0,\mathcal{B}} = 1 - \kappa_{\mathcal{B}}, \qquad d_{\alpha,\mathcal{B}} = \alpha d_{0,\mathcal{B}}. \tag{A.9}$$

Here the population refers to partition randomness. The reference data $\mathcal{B}$ remain fixed.

Theorem A.3 (Uniform concentration on a finite state set)**.** *Let $\mathcal{V}$ contain $N \geq 1$ states. Conditional on $(\mathcal{B}, g)$, assume that $\mathcal{V}$ does not depend on the randomness of the current partition ensemble and that $H_1, \ldots, H_m$ are independent and identically distributed. For any $\delta \in (0,1)$, with conditional probability at least $1 - \delta$,*

$$\sup_{s,s' \in \mathcal{V}} |\widehat{d}_\alpha(s,s') - d_{\alpha,\mathcal{B}}(s,s')| \leq \alpha\epsilon_m, \qquad \epsilon_m = \min\left\{1, \sqrt{\frac{\log(2N^2/\delta)}{2m}}\right\}. \tag{A.10}$$

Proof. For a fixed state pair, the variables $X_\ell = \mathbf{1}\{c_\ell(s) \neq c_\ell(s')\}$ are conditionally independent Bernoulli variables with mean $d_{0,\mathcal{B}}(s,s')$. Hoeffding's inequality (Hoeffding, 1963) gives

$$\Pr\left\{\left| m^{-1}\sum_{\ell=1}^m X_\ell - d_{0,\mathcal{B}}(s,s')\right| \geq u\right\} \leq 2e^{-2mu^2}.$$

A union bound over at most N^2 state pairs, followed by multiplication by α, gives the square-root bound. Both base distances lie in $[0,1]$, so their difference is also bounded by 1. We can therefore take the smaller of the two bounds. □

The theorem controls the error from approximating a fixed reference kernel with finitely many partitions. It does not bound the error between the replay data and the environment's population distribution. States generated after training need not belong to $\mathcal{V}$. Queries may nevertheless depend on the sampled partitions as long as they remain in the predetermined set $\mathcal{V}$, because the bound holds uniformly on that set.

Let $G = (\mathcal{V}, E)$ be the undirected replay transition graph, with one edge for each observed transition. For an edge $e = (s, s')$, define the constraint violation

$$v_e = \left[\|\varphi(s') - \varphi(s)\|_2 - \widehat{d}_\alpha(s, s')\right]_+ . \tag{A.11}$$

We use undirected edges because the norm constraint is symmetric. The same passage from edge constraints to path bounds applies to distance-maximizing skill discovery with more general constraints (Park et al., 2023b).

Theorem A.4 (Path bounds with constraint violations)**.** *For any path $P = (s_0, \dots, s_L)$ in the graph,*

$$\|\varphi(s_L) - \varphi(s_0)\|_2 \le \sum_{e \in P} \widehat{d}_\alpha(e) + \sum_{e \in P} v_e. \tag{A.12}$$

On the event in Theorem A.3, the right-hand side is at most

$$\alpha \sum_{e \in P} d_{0,\mathcal{B}}(e) + \alpha L \epsilon_m + \sum_{e \in P} v_e. \tag{A.13}$$

If every edge is feasible, so that $v_e = 0$, define the shortest-path pseudodistance within a connected component by

$$D_{G,\alpha}(x, y) = \inf_{P : x \rightsquigarrow y} \sum_{e \in P} \widehat{d}_\alpha(e), \tag{A.14}$$

Then $\|\varphi(x) - \varphi(y)\|_2 \le D_{G,\alpha}(x, y)$. For distributions P, Q supported on the same component,

$$\|\mathbb{E}_P \varphi - \mathbb{E}_Q \varphi\|_2 \le W_{1, D_{G,\alpha}}(P, Q). \tag{A.15}$$

Proof. Applying Eq. (A.11) along the path and using the Euclidean triangle inequality gives Eq. (A.12). Applying Eq. (A.10) to each edge gives Eq. (A.13). If $v_e = 0$, take the infimum over paths from x to y. For any coupling $\Gamma \in \Pi(P, Q)$, Jensen's inequality gives

$$\|\mathbb{E}_\Gamma[\varphi(X) - \varphi(Y)]\|_2 \le \mathbb{E}_\Gamma \|\varphi(X) - \varphi(Y)\|_2 \le \mathbb{E}_\Gamma D_{G,\alpha}(X, Y).$$

Taking the infimum over couplings proves Eq. (A.15). □

Choose a reference state s_{ref} and translate the representation so that $\varphi(s_{\mathrm{ref}}) = 0$. Then $\|\varphi(s)\|_2 \le D_{G,\alpha}(s_{\mathrm{ref}}, s)$ within its connected component. This bounds the representation scale but supplies no positive lower bound on representation distances: a constant representation also satisfies the constraints. The graph distance $D_{G,\alpha}$ depends on the observed paths and generally differs from the kernel distance between two states. A path in the undirected graph need not correspond to an executable trajectory in that direction in the environment.

A.3 Distance scaling and the directional reward objective

For a fixed skill z and a trajectory of length H, the directional reward and objective in the main text take the form

$$r_\varphi(s, z, s') = (\varphi(s') - \varphi(s))^\top z, \tag{A.16}$$

$$J(\pi, \varphi) = \mathbb{E}_{\pi, z} \sum_{t=0}^{H-1} r_\varphi(s_t, z, s_{t+1}). \tag{A.17}$$

If $\|z\|_2 \le 1$ and the relevant transitions satisfy the edge constraints, then $|r_\varphi| \le \widehat{d}_\alpha \le \alpha$. The undiscounted return telescopes:

$$R_z(\tau) = \sum_{t=0}^{H-1} r_\varphi(s_t, z, s_{t+1}) = (\varphi(s_H) - \varphi(s_0))^\top z. \tag{A.18}$$

Proposition A.5 (Scale equivalence for a fixed kernel)**.** *Suppose the constraints apply to a fixed set of state pairs, the representation class $\mathcal{F}$ is closed under positive rescaling, and the kernel does not depend on φ. For the constraints $\|\varphi(s') - \varphi(s)\|_2 \le \widehat{d}_\alpha(s, s')$, the map $\varphi \mapsto \alpha\varphi$ is a bijection from the feasible set at $\alpha = 1$ to the feasible set at scale α, and*

$$J(\pi, \alpha\varphi) = \alpha J(\pi, \varphi). \tag{A.19}$$

If optimal solutions exist, the set of optimal policies for the hard-constrained directional reward objective is independent of α.

Proof. Positive homogeneity gives $\|\alpha\Delta\varphi\|_2 = \alpha\|\Delta\varphi\|_2$, and division by α defines the inverse map. The reward is linear in the representation, which proves Eq. (A.19). For each policy, optimization over feasible representations multiplies the optimal objective value by the same positive constant. The ordering of policies by these values is unchanged. □

Under the same reference sampling distribution, the exact hinge penalty satisfies the related identity

$$-J(\pi, \alpha\varphi) + \lambda\mathbb{E}[\|\alpha\Delta\varphi\| - \alpha\widehat{d}_0]_+ = \alpha\{-J(\pi, \varphi) + \lambda\mathbb{E}[\|\Delta\varphi\| - \widehat{d}_0]_+\}. \tag{A.20}$$

In this case, λ need not change with α. If the objective also includes an entropy term $\beta_{\text{ent}}\mathcal{H}(\pi)$, scaling the reward while keeping the entropy coefficient fixed changes the relative entropy weight to $\beta_{\text{ent}}/\alpha$. Restrictions on network outputs, weight regularization, and the optimizer can also break the ideal scale equivalence. For a squared-norm constraint, $\|\Delta\varphi\|_2^2 \le \alpha\widehat{d}_0$, the corresponding representation scale is $\sqrt{\alpha}$.

A fixed, finite α retains the constraint's protection against unbounded rescaling. If α and the representation are optimized together without an upper bound on α, any feasible solution with positive return can be scaled to make the objective unbounded. For state pairs with $\widehat{\kappa}_m(s, s') = 1$, every positive α still gives zero distance, so rescaling cannot separate their constrained representations.

A.4 Constraint violations under penalization and smoothing

A finite penalty and a hard constraint provide different guarantees. Let x collect the optimization parameters, let ν be the sampling distribution fixed for the current stage, and let $\lambda > 0$. Write

$$u_x(\xi) = \|\Delta\varphi_x(\xi)\|_2 - \widehat{d}_\alpha(\xi), \qquad F(x) = -J(x) + \lambda\mathbb{E}_\nu[u_x(\xi)]_+. \tag{A.21}$$

On the parameter domain Θ, assume $\sup J - \inf J \le B_J < \infty$ and that a feasible point $x_f \in \Theta$ satisfies $u_{x_f} \le 0$ almost surely.

Proposition A.6 (Mean violation for approximate global minimizers)**.** *If $x_\lambda \in \Theta$ and $F(x_\lambda) \le \inf_{x\in\Theta} F(x) + \varepsilon_{\text{pen}}$, then*

$$\mathbb{E}_\nu[u_{x_\lambda}]_+ \le \frac{B_J + \varepsilon_{\text{pen}}}{\lambda}. \tag{A.22}$$

If the state-pair marginal of ν is supported on the set $\mathcal{V} \times \mathcal{V}$ in Theorem A.3, then replacing $\widehat{d}_\alpha$ with $d_{\alpha,\mathcal{B}}$ increases the mean violation by at most $\alpha\epsilon_m$ on the event of that theorem.

Proof. Comparison with the feasible point gives

$$-J(x_\lambda) + \lambda\mathbb{E}_\nu[u_{x_\lambda}]_+ \le -J(x_f) + \varepsilon_{\text{pen}}.$$

Rearranging and using $J(x_\lambda) - J(x_f) \le B_J$ proves Eq. (A.22). The positive-part function is 1-Lipschitz, so replacing the distance changes each violation by at most $\alpha\epsilon_m$. Taking expectations gives the second claim. □

The proposition requires approximate global optimality; a stationary-point condition alone is insufficient. A mean violation bound also does not imply feasibility on every edge. For example, for any $a > 0$, it gives the probability bound

$$\Pr_{\xi\sim\nu}\{u_{x_\lambda}(\xi) > a\} \le \frac{B_J + \varepsilon_{\text{pen}}}{\lambda a}. \tag{A.23}$$

To analyze a smooth objective, we smooth both the norm and the positive-part function. For $\omega, \mu > 0$, define

$$n_\omega(v) = \sqrt{\|v\|_2^2 + \omega^2} - \omega, \qquad h_\mu(u) = \mu \log(1 + e^{u/\mu}), \tag{A.24}$$

$$F_{\omega,\mu}(x) = -J(x) + \lambda \mathbb{E}_\nu h_\mu\big(n_\omega(\Delta\varphi_x) - \widehat{d}_\alpha\big). \tag{A.25}$$

We have $0 \le \|v\|_2 - n_\omega(v) \le \omega$ and $u_+ \le h_\mu(u) \le u_+ + \mu \log 2$. If $x_{\omega,\mu}$ is an $\varepsilon_{\mathrm{pen}}$-approximate global minimizer of the smooth objective, then

$$\mathbb{E}_\nu[u_{x_{\omega,\mu}}]_+ \le \frac{B_J + \varepsilon_{\mathrm{pen}}}{\lambda} + \omega + \mu \log 2. \tag{A.26}$$

At a feasible point, the unweighted smooth penalty is at most $\mu \log 2$. At any point, the exact violation is at most this unweighted penalty plus ω. Substituting these two inequalities into the comparison in Proposition A.6 proves Eq. (A.26). Replacing the empirical distance by the population distance adds $\alpha\epsilon_m$. Smoothness with respect to the parameters still requires the corresponding differentiability of the network and J. Assigning zero probability to nondifferentiable points does not establish this smoothness.

A.5 Optimization with biased gradients and drifting objectives

Changes in the replay distribution or kernel construction can change the objective. Let F_t denote the smooth objective at step t, such as a time-dependent version of Eq. (A.25). Assume F_t and x_t are measurable with respect to the history $\mathcal{F}_t$, each objective has an L_F-Lipschitz gradient on an analysis domain containing the update segments, and all objectives share a lower bound F_{inf}. Suppose

$$\sup_{x \in \Theta} |F_{t+1}(x) - F_t(x)| \le \delta_t, \qquad \mathbb{E}[\widehat{g}_t \mid \mathcal{F}_t] = \nabla F_t(x_t) + b_t, \tag{A.27}$$

$$\mathbb{E}[\|\widehat{g}_t - \mathbb{E}[\widehat{g}_t \mid \mathcal{F}_t]\|_2^2 \mid \mathcal{F}_t] \le \sigma^2. \tag{A.28}$$

Here b_t is the bias relative to the gradient of F_t, and δ_t is a deterministic bound on objective drift. We assume that all iterates remain in the analysis domain.

Theorem A.7 (Average gradient bound for drifting objectives)**.** *For the update $x_{t+1} = x_t - \eta \widehat{g}_t$ with $0 < \eta \le 1/L_F$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F_t(x_t)\|_2^2 \le \frac{2(\mathbb{E}F_0(x_0) - F_{\mathrm{inf}})}{\eta T} + \frac{2\sum_{t=0}^{T-2}\delta_t}{\eta T} + \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|b_t\|_2^2 + L_F \eta \sigma^2. \tag{A.29}$$

Proof. Write $g_t = \nabla F_t(x_t)$. Taking the conditional expectation of the smoothness inequality gives

$$\mathbb{E}_t F_t(x_{t+1}) \le F_t(x_t) - \eta\langle g_t, g_t + b_t\rangle + \frac{L_F\eta^2}{2}(\|g_t + b_t\|_2^2 + \sigma^2).$$

Let $a = L_F\eta \le 1$. Expand the inner product and use $-(1-a)\langle g_t, b_t\rangle \le (1-a)(\|g_t\|_2^2 + \|b_t\|_2^2)/2$ to obtain

$$\mathbb{E}_t F_t(x_{t+1}) \le F_t(x_t) - \frac{\eta}{2}\|g_t\|_2^2 + \frac{\eta}{2}\|b_t\|_2^2 + \frac{L_F\eta^2\sigma^2}{2}. \tag{A.30}$$

Sum over $t = 0, \dots, T-1$. For consecutive objectives, use $F_{t+1}(x_{t+1}) \le F_t(x_{t+1}) + \delta_t$, and for the final step use $F_{T-1}(x_T) \ge F_{\mathrm{inf}}$. Rearranging and dividing by $\eta T/2$ proves the bound. □

For a fixed objective with $b_t = 0$, choosing $\eta = \Theta(T^{-1/2})$ gives an $O(T^{-1/2})$ bound on the average squared gradient norm. For the bound to vanish under objective drift, we also need $\sum_t \delta_t = o(\eta T)$, while the bias must satisfy $T^{-1}\sum_t \mathbb{E}\|b_t\|_2^2 \to 0$. These are first-order guarantees for the stated update. They do not establish convergence of the parameter sequence or global optimality.

We can bound the drift further if the objective is the expectation, under ν_t, of a fixed per-sample loss whose time dependence comes only from ν_t and the kernel distance. Assume the absolute value of this loss is bounded by M, and define

$$\beta_t = \sup_{\xi} |\widehat{d}_{0,t+1}(\xi) - \widehat{d}_{0,t}(\xi)|, \tag{A.31}$$

Since h_μ is 1-Lipschitz,

$$\delta_t \le 2M\,\mathrm{TV}(\nu_{t+1}, \nu_t) + \lambda\alpha\beta_t. \tag{A.32}$$

To prove this bound, first change the sampling distribution and then the distance. The total-variation bound for bounded functions controls the first change, and the Lipschitz property of the smooth penalty controls the second. If α also varies, the kernel-distance change is bounded by $|\alpha_{t+1} - \alpha_t| + \alpha_t\beta_t$.

Neural SAC uses different objectives for the actor, critic, and representation network. Its updates therefore do not directly constitute the joint SGD update in Theorem A.7. Applying the theorem to an implementation requires identifying the objective for each update and controlling the bias b_t from critic estimation, off-policy sampling, and other sources. Fixing the replay data alone need not remove all objective drift.

A separate result holds for the policy subproblem with a fixed representation and reward. Suppose the state and action sets are finite, $0 \le \gamma < 1$, and all relevant transitions satisfy $|r| \le \alpha$. At temperature $\beta_{\mathrm{ent}} > 0$, the soft Bellman operator is

$$(\mathcal{T}_{\mathrm{soft}}Q)(s,a) = \mathbb{E}\left[r + \gamma\beta_{\mathrm{ent}} \log \sum_{a'} \exp\big(Q(s',a')/\beta_{\mathrm{ent}}\big)\right]. \tag{A.33}$$

If $\|Q_1 - Q_2\|_\infty \le c$, the log-sum-exp term including its factor β_{ent} differs by at most c. Each exponential term changes by at most a factor $e^{c/\beta_{\mathrm{ent}}}$, and the reverse comparison gives the other direction. Thus

$$\|\mathcal{T}_{\mathrm{soft}}Q_1 - \mathcal{T}_{\mathrm{soft}}Q_2\|_\infty \le \gamma\|Q_1 - Q_2\|_\infty. \tag{A.34}$$

The finite-dimensional space of bounded functions is complete, so exact soft value iteration converges to the unique fixed point. This result concerns the tabular subproblem with a fixed reward. It does not cover neural function approximation with a changing representation.

A.6 Directional returns and concentration of terminal representations

An upper bound on representation distances permits constant representations. The directional reward encourages displacement along the skill direction, allowing us to relate the attained return to representation concentration. Fix the initial state s_{ref}, the representation used for evaluation, and $K \ge 2$ unit skill vectors z_i, with $\varphi(s_{\mathrm{ref}}) = 0$. Write $R_i = R_{z_i}(\tau)$. Assume every terminal state under the policy class of interest satisfies $\|\varphi(s)\|_2 \le R$, where $0 < R < \infty$. If all such terminal states belong to a finite connected replay graph and the edge constraints hold, any positive $R \ge \max_{s\in\mathcal{V}} D_{G,\alpha}(s_{\mathrm{ref}}, s)$ satisfies this requirement.

Theorem A.8 (Representation concentration from directional returns)**.** *Let $Y_i = \varphi(S_H)$, $J_i = \mathbb{E}[Y_i^\top z_i]$, and $e_i = R - J_i \ge 0$. Then*

$$\mathbb{E}\|Y_i - Rz_i\|_2^2 \le 2Re_i, \tag{A.35}$$

$$\Pr\{\|Y_i - Rz_i\|_2 \ge a\} \le \min\{1, 2Re_i/a^2\}, \qquad a > 0. \tag{A.36}$$

For independent copies Y_i, Y_i' of the terminal representation of one skill, and terminal representations Y_i, Y_j of different skills,

$$\mathbb{E}\|Y_i - Y_i'\|_2 \le 2\sqrt{2Re_i}, \tag{A.37}$$

$$\mathbb{E}\|Y_i - Y_j\|_2 \ge R\|z_i - z_j\|_2 - \sqrt{2Re_i} - \sqrt{2Re_j}. \tag{A.38}$$

Proof. Since $\|Y_i\|_2 \le R$ and $\|z_i\|_2 = 1$,

$$\mathbb{E}\|Y_i - Rz_i\|_2^2 = \mathbb{E}\|Y_i\|_2^2 + R^2 - 2R\mathbb{E}[Y_i^\top z_i] \le 2R(R - J_i).$$

Markov's inequality gives Eq. (A.36). For the same skill, apply the triangle inequality through Rz_i and then the Cauchy-Schwarz inequality. For different skills, use

$$\|Y_i - Y_j\|_2 \geq R\|z_i - z_j\|_2 - \|Y_i - Rz_i\|_2 - \|Y_j - Rz_j\|_2$$

and take expectations to obtain the remaining bounds. □

If $e_i \leq e$ for all skills and the minimum skill separation is $\Delta_z = \min_{i \neq j} \|z_i - z_j\|_2$, then $R\Delta_z > 4\sqrt{2Re}$ suffices for the expected inter-skill representation distance to exceed the intra-skill upper bound. The return gap decomposes as

$$e_i = (R - J_i^\star) + (J_i^\star - J_i), \qquad J_i^\star = \sup_\pi \mathbb{E}[R_i]. \tag{A.39}$$

The first term measures the gap to the directional upper bound under the current geometry and dynamics; the second is the learning error. High directional returns therefore support representation concentration, but the theorem does not ensure a small e_i in every direction. Together with Theorem A.4, the lower bound on inter-skill representation distances also gives a lower bound on inter-skill graph distances. Intra-skill concentration in representation space does not imply concentration of states or trajectories.

Discounting changes the return identity. Writing $f_t = \varphi(s_t)^\top z$, we have

$$\sum_{t=0}^{H-1} \gamma^t (f_{t+1} - f_t) = -f_0 + (1-\gamma) \sum_{t=1}^{H-1} \gamma^{t-1} f_t + \gamma^{H-1} f_H. \tag{A.40}$$

For $\gamma < 1$, the return also depends on intermediate states. The terminal-state interpretation above applies to the undiscounted objective in Eq. (A.18). Approximate optimality of an entropy-regularized objective also does not imply approximate optimality of the unregularized directional return.

A.7 Conditional kernel separation and regional success

Concentration of terminal representations and regional separation in kernel geometry require distinct conditions. Let the K skills correspond to disjoint target regions $A_1, \ldots, A_K \subseteq \mathcal{V}$. Let ρ_i denote the terminal-state distribution of skill i, with $\mathrm{supp}(\rho_i) \subseteq \mathcal{V}$. Assume the population base distance satisfies

$$\sup_{x,y \in A_i} d_{0,\mathcal{B}}(x,y) \leq r_0, \qquad \inf_{\substack{x \in A_i,\ y \in A_j \\ i \neq j}} d_{0,\mathcal{B}}(x,y) \geq \Gamma_0, \qquad 0 \leq r_0 < \Gamma_0 \leq 1. \tag{A.41}$$

The following guarantees assume this regional structure in the environment and representation.

We distinguish a bound on individual trajectory returns from the optimal expected return. Let $M_i < \infty$ be the supremum of the return over admissible trajectories for skill i, and let $J_i^\star = \sup_\pi \mathbb{E}_\pi R_i$. Assume there is a $\Delta_i > 0$ such that every admissible trajectory terminating outside A_i satisfies $R_i \leq M_i - \Delta_i$, and that the learned policy satisfies $J_i^\star - J_i \leq \varepsilon_i^{\mathrm{opt}}$. Define

$$\eta_i = M_i - J_i^\star \geq 0, \qquad q = \min\left\{1, \max_i \frac{\eta_i + \varepsilon_i^{\mathrm{opt}}}{\Delta_i}\right\}. \tag{A.42}$$

The attainability gap η_i accounts for stochastic dynamics and restrictions on the policy class. It can remain positive even when the learning error is zero.

Theorem A.9 (Regional success and kernel separation)**.** *Under the conditions above, $\rho_i(A_i) \geq 1 - q$. On the event in Theorem A.3, let $a = (1-q)^2$ and $u = \min\{1, r_0 + \epsilon_m\}$. For independent terminal-state samples,*

$$\mathbb{E}\widehat{d}_\alpha(X_i, X_j) \geq \alpha a(\Gamma_0 - \epsilon_m)_+, \qquad i \neq j, \tag{A.43}$$

$$\mathbb{E}\widehat{d}_\alpha(X_i, X_i') \leq \alpha\{au + 1 - a\} \leq \alpha(r_0 + \epsilon_m + 2q). \tag{A.44}$$

If $a(\Gamma_0 - \epsilon_m)_+ > au + 1 - a$, the expected inter-skill kernel distance exceeds every intra-skill upper bound, and

$$\mathrm{MMD}^2_{\widehat{\kappa}_m}(\rho_i, \rho_j) \geq 2\left[a(\Gamma_0 - \epsilon_m)_+ - au - (1-a)\right]_+. \tag{A.45}$$

Proof. Let $q_i = \Pr(S_H \notin A_i)$. Each trajectory has return at most M_i, and a trajectory that misses the target region loses at least Δ_i. Hence

$$J_i \le (1-q_i)M_i + q_i(M_i - \Delta_i) = M_i - q_i\Delta_i.$$

The approximate optimality assumption also gives $J_i \ge J_i^\star - \varepsilon_i^{\mathrm{opt}} = M_i - \eta_i - \varepsilon_i^{\mathrm{opt}}$. Combining the two inequalities yields $q_i \le \min\{1, (\eta_i + \varepsilon_i^{\mathrm{opt}})/\Delta_i\} \le q$.

Two independent samples both lie in their respective target regions with probability at least a. For different skills, their empirical base distance on this event is at least $(\Gamma_0 - \epsilon_m)_+$; elsewhere it is nonnegative. This proves Eq. (A.43). For two samples of the same skill, the distance is at most αu when both lie in the target region and at most α otherwise. Since $u \le 1$, substituting the lower bound a on the event probability gives $\mathbb{E}\widehat{d}_\alpha \le \alpha(au + 1 - a)$. The inequality $1 - a = 2q - q^2 \le 2q$ gives the simpler upper bound. Substituting the inter-skill lower bound and both intra-skill upper bounds into Proposition A.2, together with the nonnegativity of squared MMD, proves Eq. (A.45). □

A pure rescaling multiplies $M_i, J_i^\star, J_i, \Delta_i$ by α, leaving the relative gap in Eq. (A.42) unchanged. The common factor α also cancels from the separation condition. Scaling the distance alone therefore does not strengthen this conditional guarantee. Any reduction in the relative gap must come from changes in the attained returns or optimization error.

A.8 Coverage probabilities and replay support

Uniform skill sampling gives the terminal-state mixture $\bar\rho = K^{-1}\sum_i \rho_i$. By Theorem A.9, each target region satisfies

$$\bar\rho(A_i) \ge \frac{1-q}{K}. \tag{A.46}$$

Let C denote the category of the terminal state, with categories $1, \dots, K$ for the target regions and category 0 for their complement.

Proposition A.10 (Regional category entropy and sample complexity)**.** *Under Eq.* (A.46)*, write* $p_c = \Pr(C = c)$*. Then*

$$\sum_{c=0}^{K} p_c^2 \le q^2 + \frac{1-q^2}{K}, \tag{A.47}$$

$$N_{\mathrm{eff}} := \frac{1}{\sum_c p_c^2} \ge \frac{1}{q^2 + (1-q^2)/K}, \qquad H(C) \ge -\log\left(q^2 + \frac{1-q^2}{K}\right). \tag{A.48}$$

If $q < 1$, draw n independent rollouts with terminal-state distribution $\bar\rho$. The condition

$$n \ge \frac{K}{1-q}\log\frac{K}{\delta}, \tag{A.49}$$

ensures that the terminal states include all K target regions with probability at least $1 - \delta$.

Proof. Let $b = (1-q)/K$ and write $p_i = b + r_i$ for $i \ge 1$ and $p_0 = r_0'$. The residual masses are nonnegative and satisfy $\sum_{i=1}^K r_i + r_0' = q$. Thus

$$\begin{aligned}\sum_c p_c^2 &= Kb^2 + 2b\sum_{i=1}^K r_i + \sum_{i=1}^K r_i^2 + (r_0')^2 \\ &\le Kb^2 + 2bq + q^2 = q^2 + \frac{1-q^2}{K}.\end{aligned}$$

Taking reciprocals gives the bound on N_{eff}. Applying Jensen's inequality to $-\log$ gives $H(C) = \mathbb{E}[-\log p_C] \ge -\log \mathbb{E}[p_C] = -\log\sum_c p_c^2$. The probability of missing a particular region in n independent samples is at most $(1-b)^n \le e^{-nb}$. A union bound over the K regions and the condition $Ke^{-nb} \le \delta$ give Eq. (A.49). □

This coverage result concerns the specified target regions and terminal states. It neither assumes that these regions cover the entire reachable state space nor measures their internal physical volume.

Replay support requires a separate exploration condition. States collected along trajectories are usually dependent, so an independent-sample model would impose an additional assumption. We instead state the required condition in terms of conditional region probabilities.

Proposition A.11 (Regional support under dependent sampling)**.** *Let $X_1, \dots, X_n$ be the states added to the reference data, and let $\mathcal{G}_{j-1}$ denote the history before the jth sample. Fix the target regions before sampling. If, for every region and sampling step,*

$$\Pr(X_j \in A_i \mid \mathcal{G}_{j-1}) \geq p_{\min} > 0 \quad \textit{almost surely}, \tag{A.50}$$

then

$$\Pr\{\exists i : \{X_1, \dots, X_n\} \cap A_i = \varnothing\} \leq K(1 - p_{\min})^n \leq K e^{-n p_{\min}}. \tag{A.51}$$

Thus $n \geq p_{\min}^{-1} \log(K/\delta)$ suffices to observe every region with probability at least $1 - \delta$.

Proof. Let $E_{i,j}$ denote the event that the first j samples all miss A_i. Since $E_{i,j-1}$ is measurable with respect to the history,

$$\begin{aligned} \Pr(E_{i,j}) &= \mathbb{E}\left[\mathbf{1}_{E_{i,j-1}} \Pr(X_j \notin A_i \mid \mathcal{G}_{j-1})\right] \\ &\leq (1 - p_{\min}) \Pr(E_{i,j-1}). \end{aligned}$$

Iterating this inequality for n steps and taking a union bound over the regions proves the claim. □

Independent and identically distributed samples with probability at least $p_{\min}$ in each region are a special case. If the condition holds only at sampling times separated by several steps, n counts those sampling times. A general exploration policy need not satisfy Eq. (A.50), and independent minibatch draws cannot recover a region absent from the replay buffer. If $p_{\min}$ decreases exponentially with dimension, the required reference sample size still grows exponentially.

A.9 Trajectory metrics and the scope of the dynamics interpretation

A terminal-state result and a comparison of entire trajectory distributions concern different objects. Represent a trajectory of length H by its empirical occupancy distribution $P_\tau = (H + 1)^{-1} \sum_{t=0}^{H} \delta_{s_t}$. Equation (A.6) then defines the kernel-induced trajectory distance

$$D_m(\tau, \tau') := \|\mu_{P_\tau} - \mu_{P_{\tau'}}\|_2, \tag{A.52}$$

$$D_m^2(\tau, \tau') = K_m(P_\tau, P_\tau) + K_m(P_{\tau'}, P_{\tau'}) - 2K_m(P_\tau, P_{\tau'}). \tag{A.53}$$

The quantity K_m is a similarity, whereas D_m is the corresponding pseudodistance between mean embeddings. In particular, $1 - K_m(P, P)$ need not vanish, so $1 - K_m$ does not in general define a distance between distributions.

For two trajectories of equal length, the kernel distances between states at matching times bound D_m:

$$\begin{aligned} D_m^2(\tau, \tau') &= \left\| \frac{1}{H+1} \sum_{t=0}^{H} \big(\Psi_m(s_t) - \Psi_m(s'_t)\big) \right\|_2^2 \\ &\leq \frac{1}{H+1} \sum_{t=0}^{H} \|\Psi_m(s_t) - \Psi_m(s'_t)\|_2^2 = \frac{2}{\alpha(H+1)} \sum_{t=0}^{H} \widehat{d}_\alpha(s_t, s'_t). \end{aligned} \tag{A.54}$$

The inequality follows from convexity of the squared norm. This bound requires control of state differences along the trajectory; a terminal-state bound alone is insufficient. Kernel means also discard the order of state visits. Evaluation with another trajectory distance requires a separate connection to its definition.

Coverage metrics also require a distinction between category probabilities and kernel similarity scores. The entropy $H(C)$ in Eq. (A.48) uses normalized category probabilities. For state samples $s_1, \ldots, s_n$, the quantity

$$q_m(s) = \frac{1}{n}\sum_{j=1}^{n} \widehat{\kappa}_m(s, s_j) = \left\langle \Psi_m(s), \frac{1}{n}\sum_{j=1}^{n} \Psi_m(s_j) \right\rangle \tag{A.55}$$

is an average similarity score. In general, it lacks density normalization with respect to a reference measure on the state space. Even for a valid density p, samples $s_j \sim p$ give the entropy estimate $-n^{-1}\sum_j \log p(s_j)$, without another factor of $p(s_j)$ inside the sample average. The regional category-entropy guarantee therefore does not directly apply to entropy measures constructed from unnormalized kernel scores. Their probabilistic interpretation requires an explicit density normalization and estimation formula.

A state distribution alone also does not determine transition reachability. Consider the family of two-state transition matrices

$$P_\varepsilon = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}, \qquad 0 < \varepsilon < 1. \tag{A.56}$$

Each matrix has the uniform stationary distribution, but the expected first-passage time from one state to the other is $1/\varepsilon$. The same unordered reference state data can therefore be consistent with different dynamics. Row-normalizing a kernel defines a random walk on a similarity graph; relating this walk to the true MDP requires additional assumptions on the dynamics. The kernel geometry results use a fixed empirical reference distribution, the graph bounds use observed edges, and the regional coverage results depend on return gaps and replay support.

B Additional Experimental Settings

B.1 State Inputs and Hyperparameters

All online and offline experiments use state observations. On challenging DMC terrain, the state input also includes the local height samples described in Appendix B.3. Table 3 lists the online pretraining settings.

We fix the number of IK partitions at $t = 100$, where t corresponds to m in Equation (2). The search space for the number of reference points per partition is $\psi \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$.

Table 3: **Hyperparameters for online skill pretraining.**

Hyperparameter	Value
Input	State
Learning rate	10^{-4}
Optimizer	Adam
Episodes per epoch	8
Gradient steps per epoch	200
Environment steps per episode	250
Minibatch size	256
Discount factor γ	0.99
Replay buffer capacity	3×10^5
Hidden layers	2
Units per hidden layer	1024
Target network smoothing coefficient	0.995
Entropy coefficient	Automatically adjusted
METRA ϵ	10^{-3}
METRA initial λ	30

B.2 Retrieval Settings for the Two-Room Random Walk

The two-room experiment in Section 3.2 uses $k = 22$ and an arrival-probability evaluation horizon of $H = 6$ steps. The reference set contains the k states with the highest six-step arrival probabilities from the anchor. Each method retrieves the same number of states, and we compute IoU using Equation (3). The neighborhood visualizations and IoU values in Figure 1 use this setting.

B.3 DMC Environment Construction and Evaluation

The online experiments use Walker and Cheetah from the DeepMind Control Suite (Tassa et al., 2018), each with original flat ground and challenging terrain. Figure 4 includes five methods across these four settings, giving 20 training curves. DMC simulates these tasks with MuJoCo (Todorov et al., 2012). The flat-ground experiments use the original DMC environments. The challenging environments use terrain that supports physical contact, with the base construction in Table 4.

Table 4: **Base contact terrain for Walker and Cheetah.** Heights correspond to a terrain multiplier of 1; the final multiplier must match the experimental configuration.

Environment	Contact terrain
Cheetah	The center remains flat, with five smooth hills on each side. Their base peak heights are 8, 10, 12, 14, and 16 cm.
Walker	A flat central region spans 2 m. Shallow steps on both sides have widths of 0.5 m and heights of 0, 4, or 8 cm.

The challenging environments allow independent control of terrain height and the initial-position distribution. We sample initial positions from a mixture of a uniform distribution and a centered truncated Gaussian with the same support. A larger mixture coefficient concentrates resets near the center. Observations concatenate the physical state with a scan of local relative heights, using 11 height samples along the forward and backward axis for both Walker and Cheetah. We use the original flat-ground and challenging environments to compare performance under these two conditions.

The flat-ground metric counts position bins visited by the policy (Policy XY coverage in the figure). The challenging-terrain metric measures valid-posture path length per 10-second trajectory (Valid-posture path in the figure). The former counts visited position cells, while the latter accumulates distance traveled when the posture satisfies the validity criterion. A path may revisit the same region, so its length differs from both position coverage and displacement between the trajectory endpoints.

Each curve contains 50 checkpoints from 0.1 M to 5 M environment steps at intervals of 0.1 M. We connect the original means without smoothing and add the origin $(0, 0)$ for plotting only; it does not represent a separate evaluation. Each method uses 8 independent training seeds. Solid lines show the mean across seeds, and shaded bands show one standard deviation above and below the mean. Table 5 reports the final-checkpoint means and standard deviations. We compute the relative improvements in the main text as $(\mu_{\mathrm{IKSD}}/\mu_{\mathrm{METRA}} - 1) \times 100\%$ using unrounded means, without inferring statistical significance from these percentages.

Table 5: **Online evaluation at 5 M environment steps (mean ± standard deviation over 8 independent training seeds).** Flat-ground columns report position coverage in bins; challenging-terrain columns report valid-posture path length (m/10 s). Higher values are better in every column.

	Flat ground		Challenging terrain	
Method	Walker	Cheetah	Walker	Cheetah
IKSD	101 ± 8.25	158 ± 10.38	28.59 ± 3.13	40.31 ± 4.13
METRA	107 ± 7.88	146 ± 9.63	25.74 ± 2.88	31.53 ± 3.38
GenDa	66 ± 5.63	130 ± 8.88	5.29 ± 1.13	11.57 ± 1.75
CSF	9 ± 1.13	2 ± 0.38	0.31 ± 0.13	3.32 ± 0.75
CSD	5 ± 0.88	40 ± 4.63	1.43 ± 0.38	23.86 ± 2.63

CSF and CSD in the figure use implementations adapted to DMC.

B.4 METAWORLD TASK SPLIT AND EVALUATION

The offline dataset contains trajectories collected for `door_open`, `door_close`, `window_open`, and `dial_turn`. These are the four seen tasks in Table 2. We also evaluate `window_close`, `button_press`, and `faucet_close`, none of which contributes trajectories to data collection. This split holds out task objectives, without requiring all associated states or local behaviors to be absent from the dataset. In particular, `window_open` already provides experience with the window.

All methods learn from the same offline data and undergo no downstream fine-tuning. We freeze the trained policies for evaluation. Each method uses 8 independent training seeds, with 64 evaluation episodes per seed and task. Table 2 reports mean success rates across training seeds. We compute the two group means by averaging the four seen-task percentages and three unseen-task percentages, respectively, with equal task weights, then round to two decimal places. We compute differences between group means before rounding.

The kernel comparison keeps the offline framework fixed. GKSD uses a fixed-bandwidth Gaussian kernel, AGKSD uses a locally adaptive Gaussian kernel, and IKSD uses IK. The three variants share the dataset, network architectures, optimization settings, and evaluation procedure.

Figure 7 shows steps 0, 20, 40, 48, 200, and 500 of one window-closing rollout per method. We select skills and execute the frozen policies. IKSD first meets the success condition at step 48, and the snapshots at steps 200 and 500 also indicate success. HILP does not complete the displayed rollout. These examples complement the aggregate success rates in Table 2.

B.5 OFFLINE REPRESENTATION AND POLICY LEARNING

Offline IKSD retains the two representation networks in HILP. For a goal state s_g, network i defines the goal-conditioned value as $V_i(s, s_g) = -\|\phi_i(s) - \phi_i(s_g)\|_2$, with numerical clipping of the norm. We sample reference states from the offline replay buffer, construct the Isolation Kernel in the feature space of the target representation network, and compute the local cost between adjacent states using Equation (5). Kernel similarity is the fraction of hard partitions in which the states share a cell, and the local dissimilarity is $1 - \kappa_{\mathrm{iso}}$.

For samples whose current state has not yet reached the goal, we update the representations with the target

$$y_i = -c_{\mathrm{IK}}(s, s') + \gamma\, V_{\bar{\phi}_i}(s', s_g), \tag{B.1}$$

where γ is the discount factor and $\bar{\phi}_i$ is the target network. We stop gradients through both the kernel cost and the target network output. We determine the expectile weight from the difference between the target and current values and update both representation networks. This objective uses discounted cumulative IK local costs over transitions supported by the data. Fitting the kernel dissimilarity of an individual state pair requires a separate regression objective.

We sample goal states from a mixture of future states in the same trajectory and states from random trajectories. The skill policy uses the directional reward in the main text and learns through IQL (Kostrikov et al., 2022). We fit state values with expectile regression, update action values with temporal differences, and update the policy through advantage-weighted behavior cloning.

B.6 DOWNSTREAM PPO CONTROL

Figure 5 compares frozen skills from IKSD, METRA, GenDa, CSD, and CSF on challenging Walker and Cheetah terrain. For each method, we train a PPO selector that chooses a skill every 5 environment steps. The low-level policy remains fixed throughout downstream training. The budget is 500k environment steps. Walker Forward, Cheetah Forward, and Cheetah Backward report speed tracking scores; Walker Shuttle reports the fraction of waypoints completed. We count downstream interactions separately from reward-free pretraining steps.

Table 6 gives the high-level PPO settings. The controller takes state observations as input and selects skills from the frozen low-level policy.

Table 6: **Hyperparameters for PPO high-level controllers.**

Hyperparameter	Value
Episodes per epoch	64
Gradient steps per episode	10
Minibatch size	256
Entropy coefficient	0.01
GAE λ	0.95
PPO clip threshold ϵ	0.2
Skill selection interval (environment steps)	5

B.7 Window-Closing Trajectory Visualizations

Figure 8 compares raw states with representations from HILP, GKSD, AGKSD, and IKSD. The upper row projects a held-out window-closing data trajectory, its nearest training trajectory, and pretraining samples onto PCA planes. The lower row constructs two-dimensional t-SNE projections from pairwise trajectory DTW distances. Each point represents a trajectory, colors indicate task categories, and a star marks source episode #5426. Window-closing data enter this held-out analysis but do not enter skill pretraining. Coordinate scales differ across panels, so we use the displayed proximity to compare trajectories with pretraining experience qualitatively.

C Additional Background and Method Details

C.1 Problem Setting

Unsupervised reinforcement learning (URL) separates pretraining from task adaptation. An agent first learns representations, policies, or environment models without downstream task rewards, then uses this knowledge to adapt to a task (Laskin et al., 2021). Online pretraining collects experience through continued interaction, whereas offline pretraining learns from a fixed transition dataset $\mathcal{D} = \{(s, a, s')\}$ (Park et al., 2024a). Both settings seek reusable knowledge, but the behaviors available for offline learning depend on the coverage of the dataset.

Consider an environment with state space $\mathcal{S}$, action space $\mathcal{A}$, and transition kernel $P(s' \mid s, a)$. Skill discovery introduces a latent variable $z \sim p(z)$ and learns a policy $\pi(a \mid s, z)$, with z held fixed over a trajectory segment. Intrinsic rewards depend on states, transitions, and skills. Downstream tasks use the resulting behaviors through skill selection, skill composition, or policy fine-tuning. Useful skills therefore need to produce distinct behaviors that support later control.

C.2 Related Approaches to Unsupervised and Skill Learning

LSD uses Euclidean distance to tie representation changes to changes in state coordinates (Park et al., 2022). CSD learns a conditional transition model from data generated by the current skills and uses negative log-likelihood to define a controllability-aware constraint scale. Transitions that are harder for these skills to realize receive larger costs (Park et al., 2023b). METRA constructs a metric-aware abstraction using temporal distance. It imposes a unit constraint on adjacent transitions, so representation distances along a path are bounded by the number of transition steps (Park et al., 2024b). These methods define geometry through coordinate changes, conditional transition distributions, and path lengths. CSD already uses distributional information; our focus is the local scale of the empirical state distribution, which differs from conditional transition density.

Distances can also serve as representation learning targets. HILP learns Hilbert representations from reward-free offline data by fitting temporal distances, then trains policies to move in different directions in that space (Park et al., 2024a). Offline IKSD retains this approach to skill learning and replaces the unit step cost with an IK-weighted local cost, which it accumulates through temporal-difference learning. The online and offline versions thus use local density-aware scales in a representation constraint and a distance-fitting target, respectively.

Other URL methods focus on exploration or model learning. APT estimates state entropy in representation space to encourage visits to new states (Liu & Abbeel, 2021a). Plan2Explore uses a world model to plan exploration and collect experience for subsequent tasks (Sekar et al., 2020). We focus on skill-conditioned policies and study how state distances shape their representations and intrinsic rewards. Useful skills need to produce distinct behaviors that support later control.

Mutual information methods discover skills by making their behaviors distinguishable. DIAYN maximizes mutual information between skills and states, together with policy entropy, using the intrinsic reward $\log q_\eta(z \mid s) - \log p(z)$ (Eysenbach et al., 2019). DADS maximizes conditional mutual information $I(S'; Z \mid S)$ and uses a skill-conditioned transition model to encourage predictable, distinct state changes (Sharma et al., 2020). These objectives identify skills from states or transitions without directly requiring large coordinate displacements or long temporal distances.

Successor features and contrastive learning connect skill discovery to task adaptation. APS combines nonparametric state entropy estimation with variational successor features to explore the environment and learn behaviors that support task inference (Liu & Abbeel, 2021b). CSF analyzes METRA through mutual information skill learning and develops a method based on contrastive successor features (Zheng et al., 2025). These connections show that mutual information and distance maximization are not separate categories: related directional rewards can arise from different representation objectives and normalization choices.

GenDa improves experience reuse and skill generalization by relabeling replay trajectories with the current representation and using a complementary information bottleneck to reduce the policy's dependence on global context (Park et al., 2026a). These changes concern skill supervision and policy inputs, while IKSD studies the state distance that defines representation geometry. Our experiments compare IKSD with METRA, CSF, and GenDa as complete methods and compare kernel choices within a shared offline learning framework.

C.3 Gaussian Similarities and Local Scale Adaptation

Kernel functions provide another way to define state similarity. A fixed-bandwidth Gaussian kernel uses $\kappa_{\mathrm{GK}}(s, s') = \exp\{-\|g(s) - g(s')\|_2^2/(2\sigma^2)\}$, where g is the input representation and σ is a global bandwidth. For a fixed representation, ranking by decreasing kernel similarity gives the same order as ranking by increasing Euclidean distance. A locally adaptive Gaussian similarity instead estimates a neighborhood scale for each state. For example, taking $\sigma(s)$ as the distance to the kth nearest neighbor gives $\exp\{-\|g(s) - g(s')\|_2^2/[\sigma(s)\sigma(s')]\}$ (Zelnik-Manor & Perona, 2004). This provides a direct baseline for local scale adaptation. Comparing these similarities and their induced dissimilarities helps distinguish the effect of density adaptation from the choice of implementation.

C.4 Isolation Kernel Feature Map and Distance Identities

Concatenating the one-hot cell encodings from all partitions gives a feature vector $\Phi_m(s) \in \{0,1\}^{m\psi}$. With $\Psi_m(s) = \Phi_m(s)/\sqrt{m}$, the kernel is the inner product of these normalized features, as illustrated in Figure 9. The dissimilarity in the main text has the equivalent expressions

$$\begin{aligned} d_{\mathrm{IK}}(s, s'; \mathcal{B}) &= 1 - \kappa_{\mathrm{iso}}(s, s'; \mathcal{B}) \\ &= \frac{1}{m} \sum_{\ell=1}^{m} \mathbf{1}\{c_\ell(s) \neq c_\ell(s')\} \\ &= \frac{1}{2} \|\Psi_m(s) - \Psi_m(s')\|_2^2. \end{aligned} \tag{C.1}$$

This is the normalized Hamming distance between cell-index sequences. Under the complete-partition assumption, d_{IK} is a pseudometric with values in $[0, 1]$. Finitely many partitions can assign the same sequence to distinct states, so zero distance does not imply identical states. The corresponding Hilbert feature distance is $\sqrt{2d_{\mathrm{IK}}}$. Appendix A establishes these properties and analyzes finite-partition approximation error.

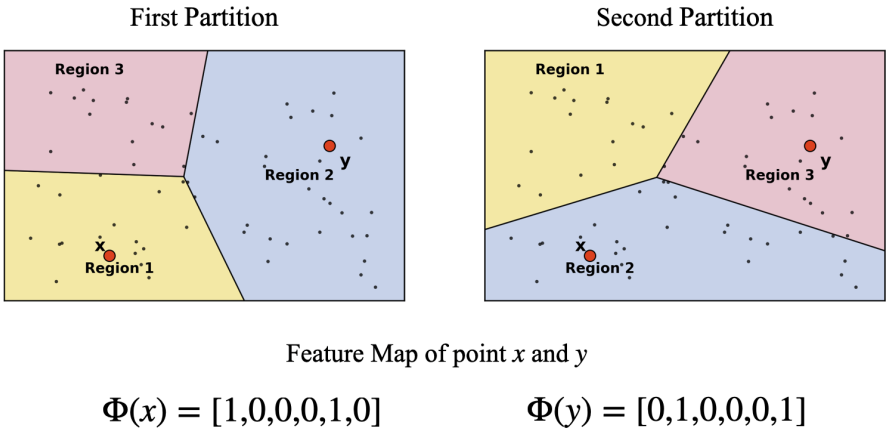


Figure 9: **Finite-dimensional IK feature maps with $m = 2$ partitions and $\psi = 3$ cells per partition.** Each state has a one-hot cell encoding in each partition. Shared nonzero entries determine the kernel similarity.

C.5 Online Objective and Reference Geometry

IKSD shapes skill representations with local empirical scales. Online IKSD constrains representation displacements with IK dissimilarity; offline IKSD learns distances from cumulative IK-weighted transition costs. Both train skill-conditioned policies with the same directional reward.

Online IKSD constructs the kernel in Equation (2) from replay-buffer states $\mathcal{B}$. The kernel input g defines the reference geometry that constrains changes in the reward representation ϕ between adjacent states. For fixed reference data and partitions, the objective is

$$\begin{aligned} \max_{\pi,\phi} \quad & \mathbb{E}_{z\sim p(z),\,\tau\sim\pi(\cdot|\cdot,z)}\left[\sum_{t=0}^{T-1}\big(\phi(s_{t+1})-\phi(s_t)\big)^\top z\right], \\ \text{subject to} \quad & \|\phi(s_{t+1})-\phi(s_t)\|_2 \le 1-\kappa_{\text{iso}}(s_t,s_{t+1};\mathcal{B}). \end{aligned} \tag{C.2}$$

The policy receives $r_z(s,s') = (\phi(s')-\phi(s))^\top z$. Observed transitions specify adjacency, and IK sets local constraint scales from the reference distribution. We do not interpret kernel similarity as the environment's transition probability.

C.6 From Local Distances to Skill Learning

The two-room and PointMaze diagnostics assess neighborhood membership and candidate ordering, respectively. Their implications for skill learning depend on how distance enters the objective. For $\|z\|_2 = 1$, substituting d_{IK} into Equation (1) and applying the Cauchy-Schwarz inequality gives

$$\left|\big(\phi(s')-\phi(s)\big)^\top z\right| \le \|\phi(s')-\phi(s)\|_2 \le d_{\text{IK}}(s,s';\mathcal{B}). \tag{C.3}$$

IK bounds the one-step directional reward when the constraint holds. Similar state pairs face tighter constraints; less similar pairs allow larger representation changes. Appendix A analyzes constraint violations and their accumulation along paths.

Online IKSD constrains representation changes on observed transitions, combining data-dependent geometry with trajectory adjacency. Offline IKSD replaces HILP's unit step cost with an IK-weighted cost and learns cumulative distances. Both condition policies on representation-space directions. The MetaWorld experiments compare IK with fixed-bandwidth and locally adaptive Gaussian kernels within the same offline framework.

D WALKER SKILL EXAMPLES

Figures 10 and 11 show all eight skills in this visualization. The policy is evaluated at the 5M training checkpoint, with two 10-second rollouts per skill from different initial states.

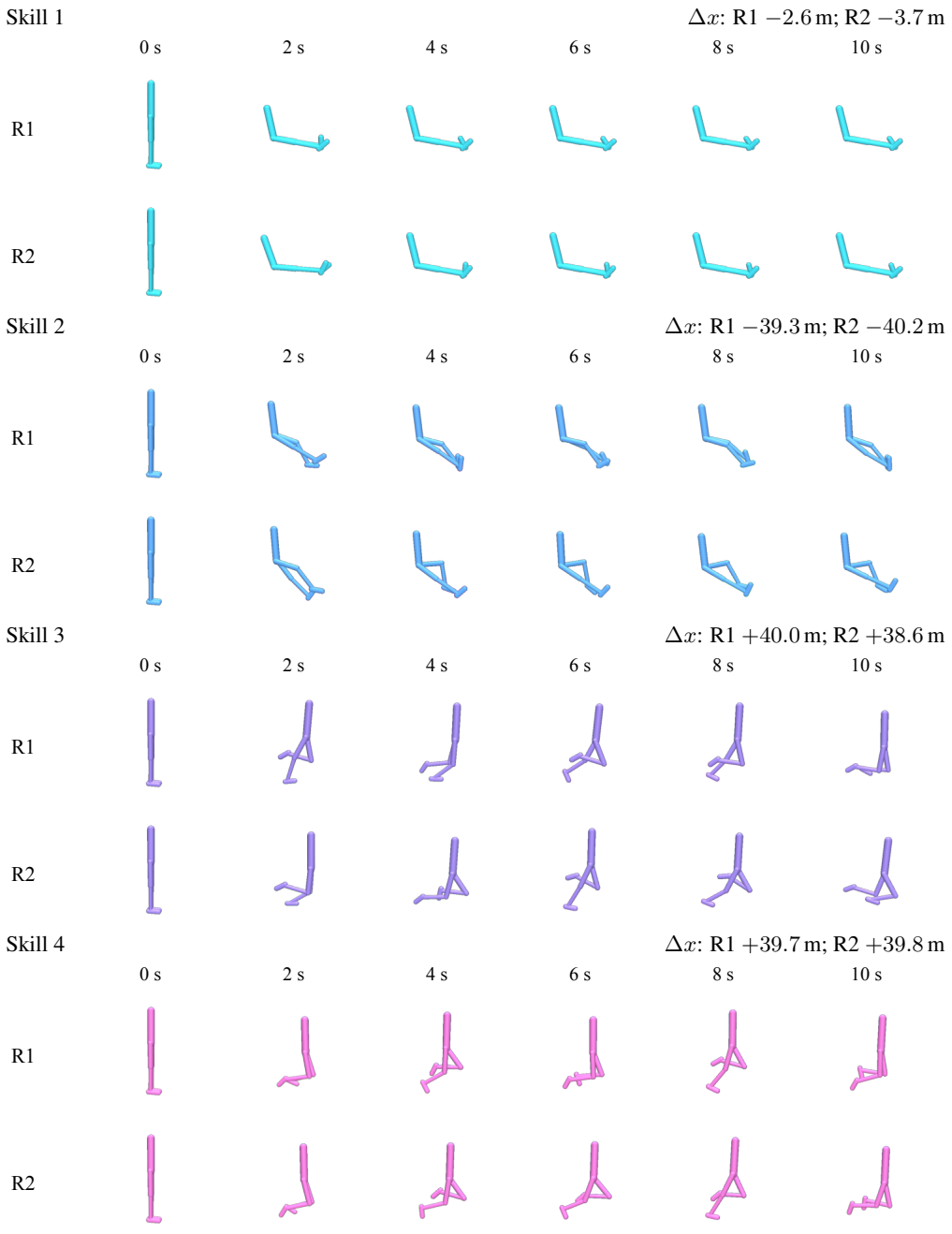


Figure 10: **Online IKSD skills 1–4 on challenging Walker terrain.** Each pair of rows shows two rollouts of one skill. Time progresses from left to right, and snapshots follow the agent. Δx denotes signed net horizontal displacement over 10 seconds.

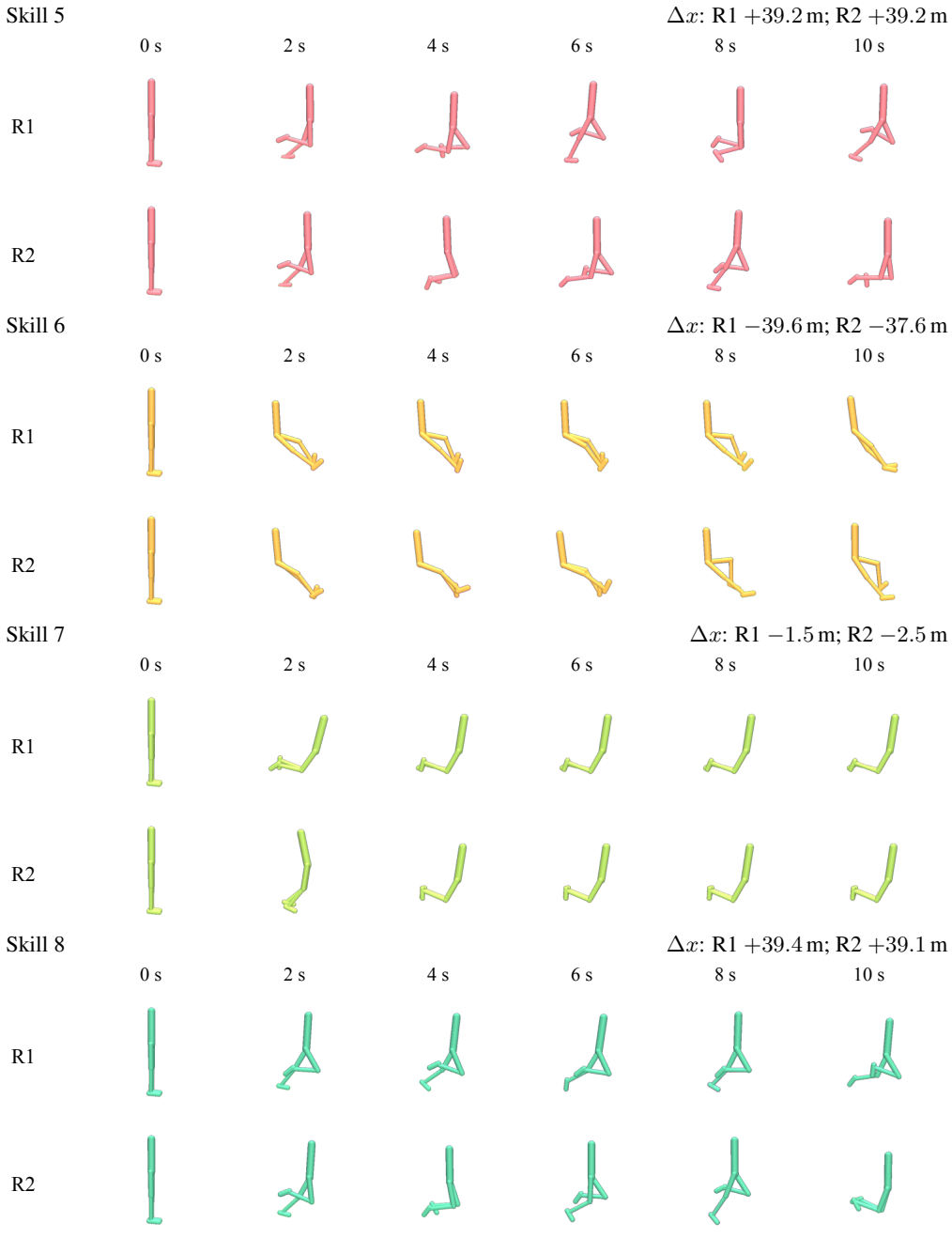


Figure 11: **Online IKSD skills 5–8 on challenging Walker terrain, using the same display conventions as Figure 10.** The examples include both large displacements and little net movement.