

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 DANCE: DIFFICULTY AND NOVELTY CO-DRIVEN EX- PLORATION

Anonymous authors

Paper under double-blind review

ABSTRACT

Reinforcement learning in environments with sparse rewards poses a significant challenge. Numerous exploration techniques strive to surmount this challenge by inciting agents to explore novel states. However, as familiarity with the environment burgeons, the novelty of states wanes, yielding an unguided exploration trajectory during later phases of learning. To surmount this quandary, this study posits that the difficulty of attaining a state functions as a more potent intrinsic motivational beacon, guiding the agent throughout the learning process. This difficulty signal encapsulates pivotal insights into the environment’s underlying structure and the task’s trajectory, a facet transcending the exclusive purview of state novelty. Subsequently, we introduce a reward prediction network to acquire a hybrid reward sourced from both state difficulty and novelty. Initially elevated for novel states, this reward progressively converges toward the state’s inherent difficulty as visitations accumulate. This dynamic formulation assuages the scourge of catastrophic forgetting, shepherding the agent precisely across the learning odyssey. We establish the theoretical underpinnings of this reward mechanism as a distinct manifestation of reward shaping. It ensures the consistency between the learned policy and the original policy and additionally transforms the sparse reward problem into a dense reward problem, consequently accelerating the entire learning process. We evaluate the proposed Difficulty and Novelty Co-driven Exploration agent on several tasks with sparse rewards, and it consistently achieves satisfactory results.

1 INTRODUCTION

A fundamental challenge in reinforcement learning (RL) is how to trade-off between exploration and exploitation. An agent has to decide whether to greedily exploit what is already known to maximize the expected cumulative reward or explore the unknown environment to gather more information that helps find a potentially better policy. Although simple exploration strategies such as ϵ -greedy action selection Mozer et al. (2018) and correlated Gaussian noise Lillicrap et al. (2015) work well on a wide range of tasks, they are inefficient in hard exploration tasks with sparse rewards such as the Atari game *Montezuma’s Revenge*. Standard RL algorithms often fail to obtain even a single positive reward.

Recently, various exploration methods Bellemare et al. (2016); Ostrovski et al. (2017); Andrychowicz et al. (2017); Liu et al. (2019); Trott et al. (2019); Warde-Farley et al. (2019); Zhao et al. (2020); Badia et al. (2020); Hartikainen et al. (2020); Zhang et al. (2021a); Henaff et al. (2022); Mutti et al. (2022); Mu et al. (2022); Hou et al. (2025); Bagaria et al. (2025) try to address the sparse-reward problems. Among them, the novelty-based rewarding approaches in intrinsic motivation Ryan & Deci (2000) make remarkable progress by giving agents intrinsic rewards whenever they visit an unexplored or unexpected state, *i.e.*, novel states. Measuring the ‘novelty’ usually requires an additional model to evaluate the environmental states’ statistical distribution, *e.g.*, the count-based exploration methods Bellemare et al. (2016); Ostrovski et al. (2017); Tang et al. (2017) count each visit to a state as a way to quantify its novelty.

One typical class of methods leverages an environment-related transition model’s prediction error to measure state novelty. The prediction error is small if the agent encounters a familiar state. However, most methods of this type suffer from the “Noisy TV” problem Savinov et al. (2019)

054 in a stochastic or partially observable environment. Among works addressing this problem, the
 055 notable ones are episodic curiosity through reachability Savinov et al. (2019) and Random Network
 056 Distillation (RND) Burda et al. (2019). RND employs a fixed randomly initialized neural network's
 057 prediction error as an exploration bonus and achieves competitive performances on several hard
 058 exploration Atari games.

059 However, the intrinsic rewards mentioned above face the problem of gradually losing efficacy dur-
 060 ing the learning process. With continuous exploration, the agent becomes more and more familiar
 061 with the environment, and the novelty of each state gradually vanishes. In this process, the intrinsic
 062 reward slowly loses its guiding significance; thus, learning is only driven by extrinsic rewards, lead-
 063 ing to the agent's undirected exploration in the later learning stage. The agent may forget valuable
 064 states after quickly getting familiar with the environment, which hinders itself from having a deeper
 065 understanding of the tasks in the environment, thus affecting the overall progress of task completion.

066 In this work, we propose a new intrinsic motivation mechanism to make up for the shortcomings of
 067 the existing novelty-based exploration methods while fully using its advantages in the early stage
 068 of learning. Besides state novelty, we utilize the difficulty of reaching a state as another evalua-
 069 tion criterion to guide the exploration. The difficulty level of a state is measured by the estimated
 070 difficulty signal an agent takes from the initial state to the target one. Unlike the state novelty, the
 071 difficulty levels of states will not vanish after repeated visits. More importantly, the state difficulty
 072 is continuously self-optimized during the learning process, automatically generating a learning cur-
 073 riculum according to the agent's current ability. In other words, the state difficulty acts as a stepping
 074 stone toward the states that are more difficult to reach, which is beneficial for designing an effective
 075 exploration strategy.

076 We further build a Reward Prediction Network (RPN) to provide intrinsic rewards for the agent.
 077 The network architecture takes inspiration from the RND network, including a fixed and randomly
 078 initialized target network and a predictor network. Unlike RND, the prediction error between the
 079 target network and the predictor network is fitted to the difficulty signal when reaching the current
 080 state instead of reducing to zero. The prediction error is high for novel states and decreases to the
 081 states' difficulty that the predictor has trained on. In this way, the novelty and difficulty of states are
 082 integrated into the prediction error, helping the agent not only find novel states but also better learn
 083 from the experience. The proposed intrinsic reward provides the agent with a lifelong exploration
 084 and exploitation guidance, which can work even without access to extrinsic rewards.

085 This work makes the following three main contributions:

- 086 • We design an intrinsic exploration mechanism DANCE considering both the difficulty
 087 and novelty of the states to learn exploration strategies that maintain efficient exploration
 088 throughout the learning process.
- 089 • We propose a reward prediction network to provide intrinsic rewards leveraging the state's
 090 difficulty and novelty. The intrinsic reward is high for novel states but decays with repeated
 091 visits, modulated by the state's difficulty signal.
- 092 • We theoretically demonstrate that the DANCE policy is equivalent to the policy that stan-
 093 dard RL learns. We further illustrate that DANCE provides dense rewards and a curriculum
 094 to accelerate the learning process, explaining our mechanism's effectiveness in a principled
 095 way.

097 We evaluate DANCE in several hard-exploration environments. Extensive experimental analyses
 098 and comparisons demonstrate the effectiveness of the learning algorithm. Compared with the state-
 099 of-the-art methods, DANCE significantly improves the learning efficiency in the Maze environment
 100 with vector-based state space and achieves competing results in hard Atari games with image-based
 101 state spaces. To facilitate further studies on reinforcement learning in hard exploration tasks, the
 102 source code and all the experimental results of this work are attached in the supplementary.

103 2 RELATED WORK

104 The sparse reward phenomenon is common in many hard exploration tasks. Most previous
 105 works focus on improving the effectiveness of exploration. Simple heuristic methods such as ϵ -
 106 greedy Volodymyr et al. (2015) or Gaussian control noise Mnih et al. (2016) work well on a wide

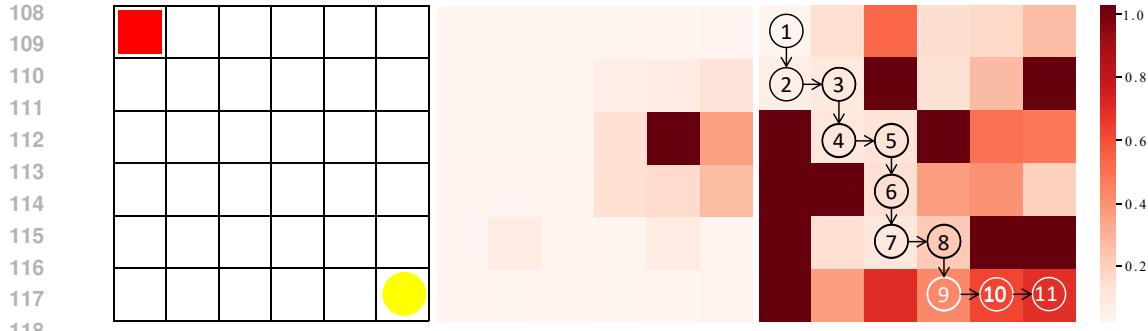


Figure 1: A simple maze environment to exemplify the differences between state difficulty and novelty: (Left) the environment, where the red square state is the initial state, the yellow circle one is the terminal state, and the agent can only move in the horizontal and vertical directions; (Middle) the heatmap of the novelty signal after training for 10K timesteps. and (Right) the difficulty signal’s heatmap after training for 10K timesteps (the numbered arrows denote the final learned policy). The darker color indicates larger intrinsic rewards. Compared with the novelty signal, the difficulty signal greatly densities the rewards and automatically forms a path (the numbered arrows) to the target from easy to hard.

range of tasks, but they are inefficient in sparse reward environments such as those found in the Atari games like *Montezuma’s Revenge* and *Venture*. In this section, we introduce some of the most related methods to our work on exploration in RL, including count-based exploration, curiosity-based exploration, and empowerment-based exploration.

Count-based Exploration. Count-based exploration Oudeyer et al. (2007); Strehl & Littman (2008); Silvia (2012) is a classical algorithm to solve hard exploration problems. These methods achieve good performance in tabular environments. To solve the hard exploration problems with image input, some works Bellemare et al. (2016); Ostrovski et al. (2017); Tang et al. (2017) adopt pseudo counts, density models, hash counts, and so on to measure the count for each visited state.

Curiosity-based Exploration. Recently, Curiosity-based exploration methods Pathak et al. (2017); Burda et al. (2019); Ghosh et al. (2019); Trott et al. (2019); Warde-Farley et al. (2019); Azizsoltani et al. (2019); Badia et al. (2020); Raileanu & Rocktäschel (2020); Zhang et al. (2021a;b); Henaff et al. (2022); Chen et al. (2022); Wang et al. (2023); Wan et al. (2023); Mahankali et al. (2024); Hou et al. (2025); Iten & Krause (2025); Bagaria et al. (2025) have yielded promising results for the hard exploration problems. Different kinds of curiosity models have been proposed, however, the basic idea behind these models is to give novel observations a bonus to encourage the agent to explore. Among them, RND Burda et al. (2019) is a simple-to-implement and easy-to-parallelize curiosity-driven exploration method. It employs the prediction error of a fixed randomly initialized neural network as an exploration bonus and outperforms average human performance in some games.

Empowerment-based Exploration. Empowerment-based exploration methods Mohamed & Rezende (2015); Liu et al. (2019); Leibfried et al. (2019); Kim et al. (2019) explore the environment by enhancing the agent’s control over the environment. Go-Explore Ecoffet et al. (2021) combines the idea of depth-first search and empowerment and obtains amazing scores in the *Montezuma’s Revenge*.

3 INTRINSIC MOTIVATION WITH DIFFICULTY

We introduce the difficulty of reaching a state as intrinsic rewards to guide efficient exploration even when the agent becomes familiar with the environment. The difficulty of a state adapts automatically as learning progress. As the policy converges, the state difficulty is stable no matter when or how many times the agent visits the state.

To illustrate the advantages of state difficulty over state novelty, we take a simple maze environment shown in Figure 1 (Left) as an example. In this environment, an agent at the initial state (red square in the upper left corner) tries to reach the terminal state (yellow circle in the lower right corner). We

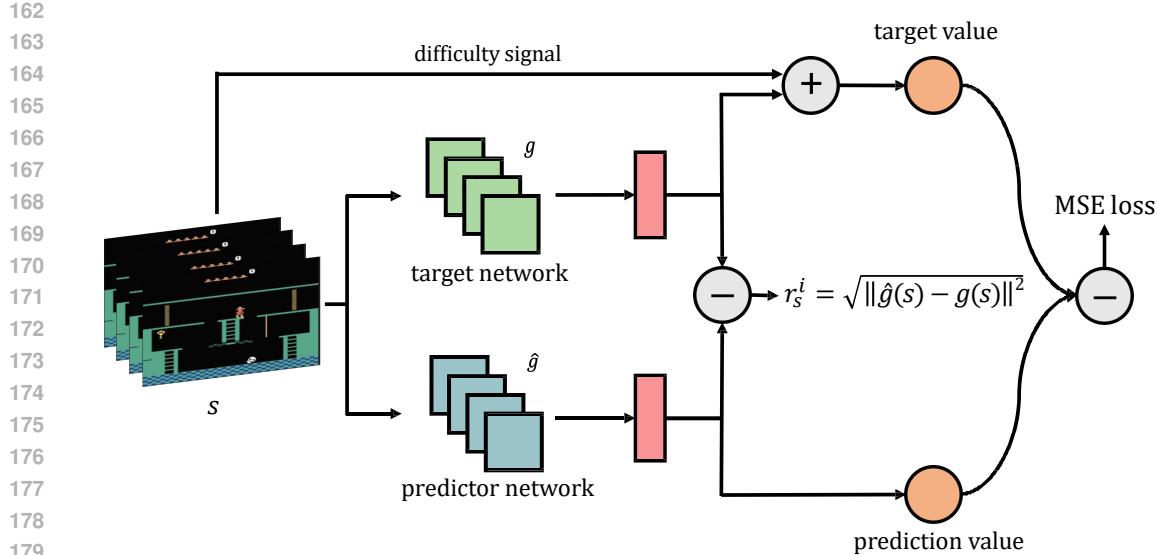


Figure 2: The RPN architecture consists of a target network g and a predictor network \hat{g} . The current step number of state s is normalized and then added to the target network’s output to form the target value. The predictor network is trained on the data collected by the agent to predict the target value. The prediction error between the target and predictor networks is used as the intrinsic reward r_s^i .

dig into state novelty with heatmaps after training for 10K frames. As shown in Figure 1 (Right), the novelty of most states has become the same and decayed to zero. Unlike state novelty, the difficulty of each state converges to different values to densify the intrinsic rewards, revealing the environment structure and task direction. It helps the agent understand the environment and provides lifelong guidance to enable further exploration and a faster learning process. We further combine difficulty and novelty to drive the agent to explore the environment. Essentially, the novel or rewarded states can be seen as learning sub-goals, and the agent needs to transfer to a new state by choosing actions until reaching the sub-goal. The state difficulty can be seen as a type of depth-first search as Go-Explore Ecoffet et al. (2021). With the state difficulty, the agent learns faster to reach sub-goals. It forms environment structures and task directions to guide deeper exploration.

4 DANCE DRIVEN RL AGENT

The previous section introduces the state difficulty and proves its effectiveness in a simple motivating maze environment. The most common way to encourage exploration is to provide the agent with intrinsic rewards proportional to some quantitative measurements. In this section, we introduce DANCE intrinsic rewards which contains state difficulty and novelty.

Although difficulty signals can effectively help agents understand the environment, identifying such signals in complex environments poses an even more challenging problem. To solve this problem, we propose Reward Prediction Network (RPN) to model the difficulty and uncertainty of states. As shown in Figure 2, RPN consists of two neural networks: a fixed and randomly initialized target network $g : \mathcal{O} \rightarrow \mathcal{R}^k$, which takes observation to an embedding, and a predictor network $\hat{g} : \mathcal{O} \rightarrow \mathcal{R}^k$ trained on data collected by the agent. The difficulty signal of visiting a state s is integrated into the target network’s output as the target value. Thus, the prediction problem is set by the target value, and the predictor neural network \hat{g} is then trained by gradient descent to minimize the expected mean-square error:

$$\mathcal{L} = \|\hat{g}(s) - (g(s) + ds(s))\|^2, \quad (1)$$

where $ds(s)$ is the difficulty signal used for reaching this state. In the implementation, we initialize a vector with the same dimension as $g(s)$ and assign all elements in the vector the value of $ds(s)$.

The predictor network \hat{g} learns to predict the sum of the output of the target network g and the difficulty signal of the current state. We prove that the difference between output values of the well-

216 trained predictor network \hat{g} and the target network g on state s , i.e. $\|\hat{g}(s) - g(s)\|_2$, is equal to the
 217 normalized running average step number of state s . We thus define the intrinsic reward of reaching
 218 state s as:

$$219 \quad r_s^i = \|\hat{g}(s) - g(s)\|_2. \quad (2)$$

221 Our method derives the difficulty signal and enables the functionality of novelty-based methods in
 222 the early stages of the exploration. The prediction error is high for novel states while converging to
 223 the states' difficulty level after repeated visits. When the agent is unfamiliar with the environment,
 224 the prediction error for an unknown state is much higher than the normalized pseudo-step; thus, the
 225 agent is encouraged to explore the environment instead of staying still. In the later learning stage,
 226 the explored area's novelty has vanished, while the difficulty of state $ds(s)$ still guides the agent to
 227 enable effective exploration and efficient exploitation.

228 **Theorem 1.** *Suppose there are no two identical states in the state spaces, for a fixed policy π , i.e.,
 229 using argmax to sample action, the intrinsic reward r_s^i for state s modeled by the RPN network is
 230 equal to the difficulty of state $ds(s)$.*

231 **Proof Sketch.** When the policy converges, the gradient of the Eq.(1) loss function vanishes in the
 232 ideal case. The intrinsic reward r_s^i aligns with the difficulty of state $ds(s)$. See Appendix for detail
 233 proof.

235 However, defining the difficulty of states constitutes a fundamental challenge. On the one hand, ac-
 236 curately specifying the difficulty level for each state is highly non-trivial and necessitates substantial
 237 expert knowledge. On the other hand, due to the non-zero property of state difficulty, incorporating
 238 such a non-zero reward function may lead to a deviation of the learning objective from the original
 239 reinforcement learning objective. To address this challenge, we propose that using proportional coef-
 240 ficient to the number of steps to reach a state as the difficulty signal is well-suited (e.g., $ds(s_t) \propto t$).
 241 When the policy converges, this simple yet strong signal exhibits the following properties: (1) For
 242 each state on the converged policy, the step count is fixed, and the intrinsic reward function is corre-
 243 spondingly fixed (align with the state's step number); (2) for states not on the converged policy, after
 244 sufficient exploration, their expected step count strictly exceeds that of states on the converged pol-
 245 icy. The aforementioned properties effectively facilitate the agent's more efficient utilization of past
 246 experience and further encourages exploration of the environment. In the following section, we rig-
 247 orously establish that the learning policy remains unbiased when the state difficulty is proportional
 248 to the number of steps required to reach the state under the converged policy.

249 5 THEORETICAL JUSTIFICATION

251 In this section, we first review a formulation of the intrinsic reward-based RL methods. Next, we
 252 demonstrate that the learning goal of the DANCE mechanism is the same as the original one with
 253 common RL algorithms, even though the intrinsic rewards in DANCE will not decrease to zero.
 254 Then, we argue that the dense rewards provided by DANCE help agents form a curriculum to ac-
 255 celerate learning. Finally, we provide the reward shaping form of DANCE, which further prove that
 256 the learned DANCE policy is consistent with the learned common RL policy.

257 5.1 FORMULATION OF INTRINSIC REWARD BASED RL

259 We consider the standard RL formalism that consists of an agent interacting with an environment to
 260 fulfill some tasks. An environment is described by a set of states $\mathcal{S} = \{s_0, s_1, s_2, \dots, s_t, \dots\}$, a set
 261 of actions $\mathcal{A} = \{a_0, a_1, a_2, \dots, a_t, \dots\}$, and a reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{R}$. After executing
 262 an action a_t at a state s_t , the agent will transfer to a new state s_{t+1} and get an extrinsic reward
 263 $r_t^e(s_t, a_t)$. For hard exploration tasks with sparse rewards, $r_t^e(s_t, a_t)$ is defined to represent the task
 264 goal only in a few states and set to zero for all other states. The goal of the RL agent is to learn a
 265 policy π^* which maximizes the cumulative extrinsic rewards:

$$266 \quad \pi^* = \text{argmax}_{\pi} \mathbf{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_t^e(s_t, a_t) \right], \quad (3)$$

267 where $\gamma \in [0, 1]$ is the discount factor that trades off the importance of immediate and later rewards.

270 **Theorem 2.** *The cumulative extrinsic rewards received by policy π^* should be no less than any*
 271 *policies π^{any} .*
 272

273 *Proof.* Since the policy π^* maximizes the cumulative extrinsic rewards, the cumulative extrinsic
 274 rewards obtained by any other policies π^{any} should not exceed those of policy π^* . \square
 275

276 For the intrinsic reward-based methods, the agent will get two kinds of rewards: extrinsic reward
 277 $r_t^e(s_t, a_t)$ and intrinsic reward $r_{s_t}^i(s_t)$. The goal of the intrinsic reward-based agent is to learn a
 278 policy π_i^* which maximizes the cumulative extrinsic and intrinsic rewards:
 279

$$\pi_i^* = \operatorname{argmax}_{\pi} \mathbf{E}_{\pi} \left\{ \sum_{t=0}^{\infty} \gamma^t [r_t^e(s_t, a_t) + \beta r_{s_t}^i(s_t)] \right\}, \quad (4)$$

282 where $\beta > 0$ represents the weight of the intrinsic reward, and the intrinsic reward is far less than
 283 the extrinsic reward $r_{s_t}^i(s_t) \ll r_t^e(s_t, a_t)$.
 284

285 **Theorem 3.** *The cumulative extrinsic rewards and intrinsic rewards received by policy π_i^* should*
 286 *be no less than any policies π^{any} .*
 287

288 *Proof.* Since the policy π_i^* maximizes the cumulative extrinsic rewards, the cumulative extrinsic
 289 rewards obtained by any other policies π^{any} should not exceed those of policy π_i^* . \square
 290

291 5.2 CONNECTION BETWEEN DANCE AND COMMON RL

292 In this subsection, we further demonstrate that policy π_i^* is at least a subset of policy π^* when the
 293 intrinsic reward $r_{s_t}^i(s_t)$ is proportionate to the step number t reaching state s_t (e.g., $r_{s_t}^i(s_t) = \omega t$,
 294 and ω is a constant that satisfies $0 < \omega \ll 1$).
 295

296 **Theorem 4.** *The policy π_i^* DANCE agent learns is at least a subset of policy π^* common RL method*
 297 *learns.*
 298

299 **Proof Sketch.** According to Bellman Equation Sutton & Barto (1998); Sutton et al. (1999), for any
 300 RL problem, the policy π^* is exist. See Appendix for detail proof.
 301

302 Above all, we prove that policy π_i^* is at least a subset of policy π^* . Here we assume that the MDP
 303 has an infinite horizon, since each finite MDP can be transformed into an equivalent infinite MDP,
 304 *i.e.*, an MDP which does not end at the terminal state but falls into a loop state that can never come
 305 out, so this proposition also applies to finite MDPs.
 306

307 DANCE provides each state with a hybrid intrinsic reward mechanism including the difficulty and
 308 novelty signals using RPN. From RPN we intuitively find that this rewarded signal is non-zero for
 309 each visited state. It thus provides more dense rewards than methods with only the novelty signal.
 310 In addition, the reward signal does not disappear when the number of exploration times increases.
 311 These dense rewards help the agents guide the exploration more efficiently. Therefore, the DANCE
 312 mechanism provides the agent with more dense rewards to guide the exploration efficiently than
 313 standard novelty-based methods.
 314

315 **Theorem 5.** *The DANCE mechanism helps to form a curriculum to accelerate learning.*
 316

317 *Proof.* It is well known that humans learn better when training tasks are organized in a meaning-
 318 ful order, *e.g.*, by starting with easy ones and gradually progressing to more complex ones. The
 319 curriculum learning Bengio et al. (2009) paradigm inspired by this objective fact has shown good
 320 performances in supervised learning. In reinforcement learning, the intrinsic rewards of common
 321 novelty-based methods gradually vanish when the agent becomes more and more familiar with the
 322 environment, which can not provide more guidance for agents in the phase of learning policies to
 323 extrinsic rewards. DANCE helps the agent to explore the environment from easy to difficult. In the
 324 initial stage of exploration, DANCE gives intrinsic rewards according to the novelty of each state.
 325 During the phase of learning policies to extrinsic rewards, the intrinsic rewards given by DANCE
 326 for each state s_t gradually fit the timestamps t . The intrinsic rewards of the states in a trajectory
 327 satisfy the following relationship:
 328

$$r_{s_0}^i(s_0) \leq r_{s_1}^i(s_1) \leq \dots \leq r_{s_t}^i(s_t) \leq \dots \quad (5)$$

324 According to the definition of curriculum learning Bengio et al. (2009), the state sequence
 325 $\{s_0, s_1, \dots, s_t, \dots\}$ of the DANCE mechanism forms a curriculum. \square
 326

327 Figure 1 (Right) provides an illustration of Theorem 5. The heatmap shows that the difficulty signals
 328 of the states are still distinct after the agent gets familiar with the environment, which demonstrates
 329 that DANCE indeed provides dense rewards and more guidance for better exploration and learning
 330 acceleration. Besides, the color of squares on the learned policy (denoted as the numbered arrows)
 331 are getting darker, indicating a series of curriculum from easy to difficult. Notably, Figure 1 (Middle
 332 and Right) validates our hypothesis: compared with the novelty signal, which gradually diminishes
 333 in the environment, the difficulty signal consistently maintains a high level for samples outside the
 334 reward trajectory, persistently facilitating the agent’s exploration.

335 5.3 REWARD SHAPING FORMULATIONS

336 It is worth noting that most existing works Zhang et al. (2021b) suggest that intrinsic rewards need
 337 to satisfy *asymptotic consistency*, i.e., intrinsic rewards should vanish after sufficient exploration.
 338 They argue that if the intrinsic rewards are asymptotically inconsistent, the final policy does not
 339 maximize the cumulative extrinsic reward, deviating from the goal of RL. Interestingly, our DANCE
 340 intrinsic reward is asymptotically inconsistent, but according to Proposition 4, the optimal policy the
 341 DANCE agent learns is equivalent to the optimal policy in the original MDP. This property can also
 342 be explained from the perspective of reward shaping.
 343

344 The Policy Invariance Theorem Ng et al. (1999) states that a *potential-based reward shaping function*
 345 F can guarantee the optimal policy learned by reward shaping will also be an optimal policy in
 346 the original MDP. Let F be a potential-based reward shaping function if there exists a real-valued
 347 function $\phi : \mathcal{S} \rightarrow \mathcal{R}$, $F(s, a, s') = \gamma\phi(s') - \phi(s)$. Our DANCE intrinsic reward function is a
 348 potential-based reward shaping function, the potential function $\phi(s_t)$ is:
 349

$$\phi(s_t) = \sum_{i=0}^t \frac{\omega_i}{\gamma^{t+1-i}}, \quad (6)$$

350 when γ is 1, $\phi(s_t)$ is equal to $\frac{t(t+1)}{2}$. The DANCE intrinsic reward for state s_t is:
 351

$$r_{s_t}^i = \gamma\phi(s_t) - \phi(s_{t-1}) = \omega t. \quad (7)$$

352 These analyses validate that DANCE not only accelerates agents’ learning but also obtains the opti-
 353 mal policy in the original sparse reward problem.
 354

355 6 EXPERIMENTS

356 To verify DANCE’s effectiveness and generalization ability, we evaluate it on two distinct domains:
 357 1) a simple maze; 2) seven hard exploration Atari games, including *Montezuma’s Revenge*, *Pitfall!*,
 358 *Gravitar*, *Solaris*, *Venture*, *Frostbite*, and *Freeway*. We then present the performance results of
 359 various algorithms in the aforementioned environments devoid of extrinsic rewards to assess the
 360 pure exploration capability of the DANCE algorithm.

361 6.1 MAZE

362 To verify the validity of DANCE, we explore the pure exploration capability of DANCE. To verify
 363 this point, we compare different methods in a 50×50 Empty Maze environment. In this Maze
 364 environment, the maximum step number for one episode is set as 100. The implemented methods
 365 are the Deep Q Network (DQN) Volodymyr et al. (2015), DQN with Random Network Distillation
 366 (DQN+RND), DQN with NovelID Zhang et al. (2021b) (DQN+NovelID) and DQN with DANCE
 367 (DQN+DANCE). The hyper-parameter β in RND, NovelID and DANCE is set to 0.01 and each al-
 368 gorithm is run for 100,000 steps, and the explored area for each algorithm is depicted in Figure 3.
 369 We observe that the baseline DQN explores only a limited portion of the environment, covering only
 370 314 different states (12.56% of the environment state coverage). The RND and NovelID algorithms
 371 perform similarly, each exploring 467, 470 different states (18.68% and 18.80% of the environment
 372 coverage). The DANCE algorithm explores the most states, covering 650 different states (26.00% of
 373 the environment coverage). The results show that DANCE is able to explore the environment more
 374 effectively than the other methods. The explored area for each algorithm is shown in Figure 3.
 375

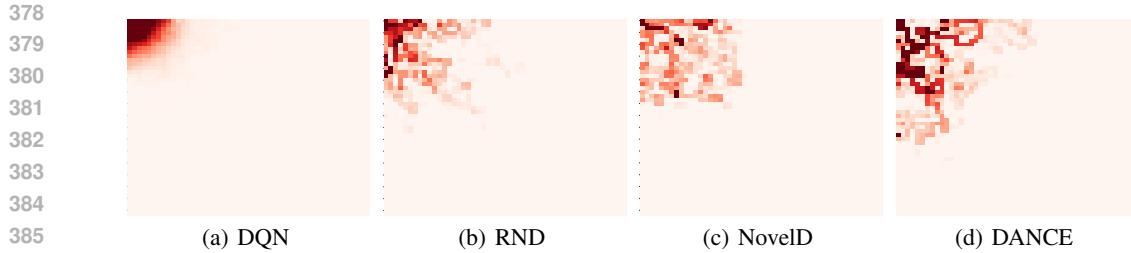


Figure 3: The heatmap of DQN, RND, NovelD and DANCE on pure exploration maze in 100K steps. The initial state is located in the upper left corner.

state coverage), respectively. In contrast, DANCE, which incorporates a difficulty mechanism, significantly expands the exploration range compared to the other algorithms, exploring 617 different states (24.68% of the environment state coverage). These results confirm that DANCE’s persistent difficulty signal effectively mitigates premature convergence, enabling systematic coverage of the state space. It is a critical advantage in pure exploration scenarios where sustained exploration drive is paramount.

6.2 HARD EXPLORATION ATARI GAMES

We convert the 210×160 input RGB frames to 84×84 gray-scale images for hard-exploration Atari games and feed the last 4 stacked gray-scale images into RPN. The hyper-parameter settings of DANCE are the same as those of RND Burda et al. (2019) (β is set to 0.5), using the Proximal Policy Optimization (PPO) algorithm Schulman et al. (2017) with the RNN policy. The step number in DANCE is normalized by dividing the maximum step number in an episode set to 4,500 in our experiment. Most experiments are run for 12.2K rollouts of length 128 per environment with 32 parallel environments, *i.e.*, for a total of 50 million steps (200 million frames) of experience.

To demonstrate DANCE’s effectiveness, we compare its performance on Atari games with sparse rewards *Montezuma’s Revenge*, *Pitfall!*, *Gravitar*, *Solaris*, *Venture*, and *Freeway* Bellemare et al. (2016) with RND and PPO. From Figure 6, DANCE learns much faster and better than RND in games *Montezuma’s Revenge*, *Gravitar*, *Venture*, and *Frostbite*, while achieves a slight improvement in *Pitfall!*. It is a typical result in *Pitfall!* to fail to find any positive rewards, as the extrinsic positive reward is very sparse. DANCE outperforms RND and does not have the decline and instability as RND. These experimental results demonstrate that DANCE’s intrinsic rewards encourage the agent to explore and exploit the environment more effectively than RND.

To further demonstrate DANCE’s effectiveness, we then compare our DANCE with other five competitive exploration methods: AE-SimHash Tang et al. (2017), Exemplar Models Exploration (EX2) Fu et al. (2017), Curiosity-Driven Exploration (ICM) Pathak et al. (2017), Self-Imitation Learning Oh et al. (2018), and Exploration with Mutual Information (EMI) Kim et al. (2019) in seven Atari games. The final training performance of using 5 random seeds in 50M timesteps for

Table 1: Comparisons of representative algorithms with excellent performance on seven hard exploration Atari games. The results of AE-SimHash Tang et al. (2017), EX2 Fu et al. (2017), and EMI Kim et al. (2019) are from the original papers. All results are reported by using 5 random seeds in 50M timesteps.

	DANCE	PPO	AE-SimHash	EX2	ICM	SIL	RND	EMI
Montezuma’s Revenge	5016	2461	75	0	1011	1259	3216	387
Pitfall!	-1.8	-4.0	-	-	-	-	-4.4	-
Gravitar	1631	1516	482	550	427	1629	1314	558
Solaris	1304	1054	4467	2276	2453	2854	1171	2688
Venture	1346	0	445	900	418	0	1150	646
Frostbite	3518	2067	5214	4901	4456	5873	2886	7002
Freeway	33.8	33.6	33.5	33.3	33.6	33.2	33.7	33.8

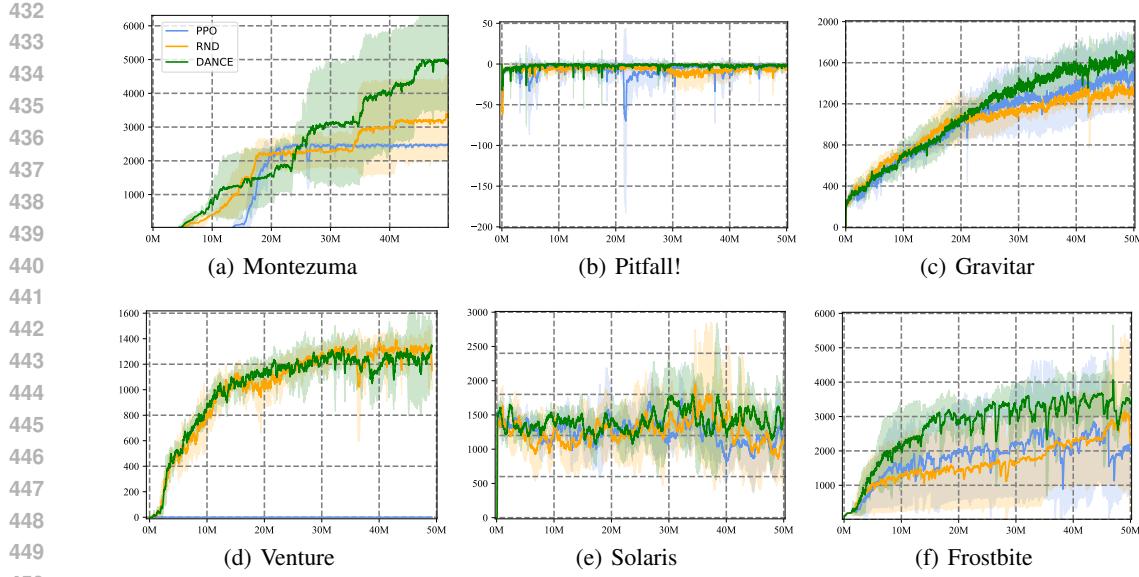


Figure 4: DANCE’s performances on six hard exploration tasks compared to RND and PPO. The curves show the mean episodic return (y -axis) of five different seeds at each iteration (x -axis). The results are reported by using 5 random seeds in 50M timesteps.

each algorithm is listed in Table 1. From Table 1, we observe that DANCE achieves the best results in five of the seven environments. Taking *Montezuma’s Revenge* as a canonical benchmark, we quantify the intrinsic rewards across the first and second rooms (0.047 for DANCE while 0.018 for RND), observing a statistically significant disparity in their mean values. DANCE fails to achieve competitive performance in *Solaris* and *Frostbite*. The highly similar visual backgrounds across these environments introduce significant interference to the intrinsic reward signal—specifically, the step number learning objective based on the Reward Prediction Network. This degradation stems from high background similarity between the environments, which substantially disrupts the intrinsic reward mechanism—specifically, the RPN-based step-wise learning process. Even so, DANCE made a significant improvement compared with other exploration methods. DANCE represents a significant advance in exploration methods, allowing the agents to navigate further and explore more diverse areas. This ultimately improves exploration efficiency, making DANCE a promising strategy for future research.

7 LIMITATION

Just as human beings easily get distracted by specific things and leave behind the task they are doing, we realize that the DANCE mechanism may face the same problem. For example, suppose the environment contains an observable clock. In that case, the predictor network may learn to look solely at the clock to predict the step number, thus may no longer yield any meaningful measure of progress influenced by the agent. This clock example poses a significant challenge for all the intrinsic reward based methods, and few of them can deal with this challenge well.

8 CONCLUSION

In this paper, we propose DANCE, an exploration mechanism which learns the difficulty signal of each state. By considering the difficulty and novelty of the state simultaneously, the intrinsic rewards provided by DANCE can continuously guide the agent to learn efficiently during the whole training process. Theoretical analysis shows DANCE constitutes a reward shaping framework for curriculum reinforcement learning. The detailed experimental results on the maze environment and some hard exploration games show that DANCE is beneficial for a wide range of RL algorithms. DANCE provides a simple yet strong baseline for solving hard exploration tasks.

486 9 REPRODUCIBILITY STATEMENT
487488 To enhance the reproducibility of our work, we have made comprehensive efforts to document and
489 share all relevant resources. The source code for our experiments is provided as supplementary
490 material. Detailed experimental configurations, model hyperparameters, and procedural descriptions
491 are thoroughly documented in the Appendix. Furthermore, complete proofs for theoretical claims
492 and theorems presented in the paper are included in the Appendix to ensure verifiability.
493494 REFERENCES
495

496 Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob
497 McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience re-
498 play. In *Advances in Neural Information Processing Systems*, pp. 5048–5058, 2017.

500 Hamoon Azizsoltani, Yeo Jin Kim, Markel Sanz Ausin, Tiffany Barnes, and Min Chi. Unobserved
501 is not equal to non-existent: Using gaussian processes to infer immediate rewards across contexts.
502 In *International Joint Conference on Artificial Intelligence*, pp. 1974–1980, 2019.

503 Adria Puigdomenech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven
504 Kapturowski, Olivier Tieleman, Martin Arjovsky, Alexander Pritzel, Andrew Bolt, et al. Never
505 give up: Learning directed exploration strategies. In *International Conference on Learning Rep-
506 resentations*, pp. 1–28, 2020.

507 Akhil Bagaria, Anita De Mello Koch, Rafael Rodriguez-Sanchez, Sam Lobel, and George Konidaris.
508 Intrinsically motivated discovery of temporally abstract graph-based models of the world. In
509 *Reinforcement Learning Conference*, 2025.

511 Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi
512 Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Infor-
513 mation Processing Systems*, pp. 1471–1479, 2016.

515 Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In
516 *International Conference on Machine Learning*, pp. 41–48, 2009.

517 Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network
518 distillation. In *International Conference on Learning Representations*, pp. 1–17, 2019.

520 Eric Chen, Zhang-Wei Hong, Joni Pajarinen, and Pulkit Agrawal. Redeeming intrinsic rewards via
521 constrained optimization. *Advances in Neural Information Processing Systems*, 35:4996–5008,
522 2022.

524 Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. First return, then
525 explore. *Nature*, 590(7847):580–586, 2021.

526 Justin Fu, John Co-Reyes, and Sergey Levine. EX2: Exploration with exemplar models for deep
527 reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2577–2587,
528 2017.

530 Dibya Ghosh, Abhishek Gupta, and Sergey Levine. Learning actionable representations with goal
531 conditioned policies. In *International Conference on Learning Representations*, pp. 1–18, 2019.

533 Kristian Hartikainen, Xinyang Geng, Tuomas Haarnoja, and Sergey Levine. Dynamical Distance
534 Learning for Semi-Supervised and Unsupervised Skill Discovery. Technical report, 2020.

535 Mikael Henaff, Roberta Raileanu, Minqi Jiang, and Tim Rocktäschel. Exploration via elliptical
536 episodic bonuses. In *Advances in Neural Information Processing Systems*, pp. 1–10, 2022.

538 Yaqing Hou, Jie Kang, Haiyin Piao, Yifeng Zeng, Yew-Soon Ong, Yaochu Jin, and Qiang Zhang.
539 Cooperative multiagent learning and exploration with min–max intrinsic motivation. *IEEE Trans-
actions on Cybernetics*, 2025.

540 Klemens Iten and Andreas Krause. Scalable and efficient exploration via intrinsic rewards in
 541 continuous-time dynamical systems. In *The Exploration in AI Today Workshop at ICML 2025*,
 542 2025.

543 Hyoungseok Kim, Jaekyeom Kim, Yeonwoo Jeong, Sergey Levine, and Hyun Oh Song. Emi: Ex-
 544 ploration with mutual information. In *International Conference on Machine Learning*, pp. 3360–
 545 3369, 2019.

546 Felix Leibfried, Sergio Pascualdiaz, and Jordi Graumoya. A unified bellman optimality principle
 547 combining reward maximization and empowerment. In *Advances in Neural Information Process-
 548 ing Systems*, pp. 7867–7878, 2019.

549 Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa,
 550 David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv
 551 preprint arXiv:1509.02971*, 2015.

552 Hao Liu, Alexander Trott, Richard Socher, and Caiming Xiong. Competitive experience replay. In
 553 *International Conference on Learning Representations*, pp. 1–16, 2019.

554 Srinath Mahankali, Zhang-Wei Hong, Ayush Sekhari, Alexander Rakhlin, and Pulkit Agrawal. Ran-
 555 dom latent exploration for deep reinforcement learning. In *International Conference on Machine
 556 Learning*, pp. 1–13, 2024.

557 Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Tim Harley, Timothy
 558 Lillicrap, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement
 559 learning. In *International Conference on Machine Learning*, pp. 1928–1937, 2016.

560 Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsi-
 561 cally motivated reinforcement learning. In *Advances in Neural Information Processing Systems*,
 562 pp. 2125–2133, 2015.

563 S. Mozer, M C, and M. Hasselmo. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

564 Jesse Mu, Victor Zhong, Roberta Raileanu, Minqi Jiang, Noah Goodman, Tim Rocktäschel, and
 565 Edward Grefenstette. Improving intrinsic exploration with language abstractions. In *Advances in
 566 Neural Information Processing Systems*, pp. 1–14, 2022.

567 Mirco Mutti, Riccardo De Santi, and Marcello Restelli. The importance of non-markovianity in
 568 maximum state entropy exploration. In *International Conference on Machine Learning*, pp.
 569 16223–16239, 2022.

570 Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations:
 571 Theory and application to reward shaping. In *International Conference on Machine Learning*, pp.
 572 278–287, 1999.

573 Junhyuk Oh, Yijie Guo, Satinder Singh, and Honglak Lee. Self-imitation learning. In *International
 574 Conference on Machine Learning*, pp. 3878–3887, 2018.

575 Georg Ostrovski, Marc G Bellemare, Aaron Van Den Oord, and Remi Munos. Count-based ex-
 576 ploration with neural density models. In *International Conference on Machine Learning*, pp.
 577 2721–2730, 2017.

578 Pierre Yves Oudeyer, F. Kaplan, and V. V. Hafner. Intrinsic motivation systems for autonomous
 579 mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286, 2007.

580 Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration
 581 by self-supervised prediction. In *International Conference on Machine Learning*, pp. 2778–2787,
 582 2017.

583 Roberta Raileanu and Tim Rocktäschel. RIDE: Rewarding Impact-Driven Exploration for
 584 Procedurally-Generated Environments. In *International Conference on Learning Representa-
 585 tions*, pp. 1–10, 2020.

594 Richard M Ryan and Edward L Deci. Intrinsic and extrinsic motivations: Classic definitions and
 595 new directions. *Contemporary Educational Psychology*, 25(1):54–67, 2000.
 596

597 Nikolay Savinov, Anton Raichuk, Damien Vincent, Raphael Marinier, Marc Pollefeyns, Timothy
 598 Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. In *International Conference
 599 on Learning Representations*, pp. 1–10, 2019.

600 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
 601 optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

603 Paul J Silvia. Curiosity and motivation. *The Oxford handbook of human motivation*, pp. 157–166,
 604 2012.

606 Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for
 607 markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.
 608

609 Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press,
 610 1998.

611 Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A frame-
 612 work for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–
 613 211, 1999.

615 Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman,
 616 Filip DeTurck, and Pieter Abbeel. Exploration: A study of count-based exploration for deep
 617 reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2753–2762,
 618 2017.

619 Alexander Trott, Stephan Zheng, Caiming Xiong, and Richard Socher. Keeping your distance: Solv-
 620 ing sparse reward tasks using self-balancing shaped rewards. In *Advances in Neural Information
 621 Processing Systems*, pp. 10376–10386, 2019.

623 Mnih Volodymyr, Kavukcuoglu Koray, Silver David, Rusu Andrei A, Veness Joel, Bellemare
 624 Marc G, Graves Alex, Riedmiller Martin, Fidjeland Andreas K, and Ostrovski Georg. Human-
 625 level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

627 Shanchuan Wan, Yujin Tang, Yingtao Tian, and Tomoyuki Kaneko. Deir: efficient and robust ex-
 628 ploration through discriminative-model-based episodic intrinsic rewards. In *International Joint
 629 Conferences on Artificial Intelligence*, pp. 4289–4298, 2023.

631 Zhihai Wang, Taoxing Pan, Qi Zhou, and Jie Wang. Efficient exploration in resource-restricted
 632 reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 10279–
 633 10287, 2023.

634 David Warde-Farley, Tom Van de Wiele, Tejas Kulkarni, Catalin Ionescu, Steven Hansen, and
 635 Volodymyr Mnih. Unsupervised control through non-parametric discriminative rewards. In *In-
 636 ternational Conference on Learning Representations*, pp. 1–17, 2019.

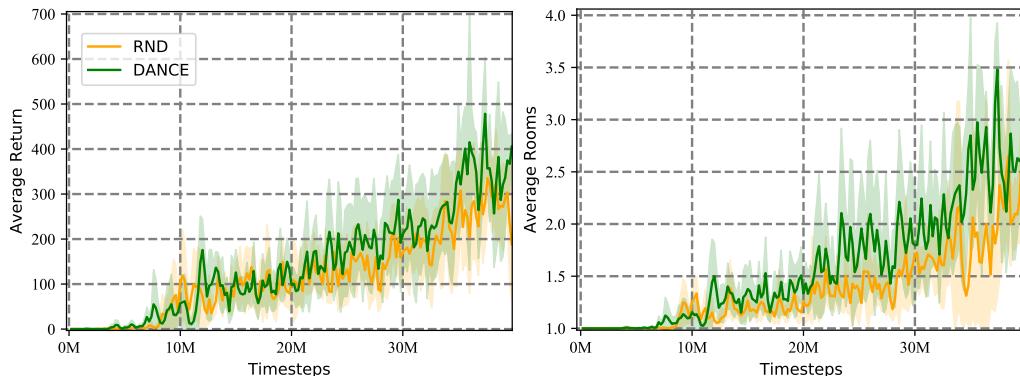
638 Chuheng Zhang, Yuanying Cai, Longbo Huang, and Jian Li. Exploration by Maximizing Renyi
 639 Entropy for Reward-Free RL Framework. In *AAAI Conference on Artificial Intelligence*, pp.
 640 10859–10867, 2021a.

641 Tianjun Zhang, Huazhe Xu, Xiaolong Wang, Yi Wu, Kurt Keutzer, Joseph E Gonzalez, and Yuan-
 642 dong Tian. Noveld: A simple yet effective exploration criterion. In *Advances in Neural Informa-
 643 tion Processing Systems*, pp. 1–10, 2021b.

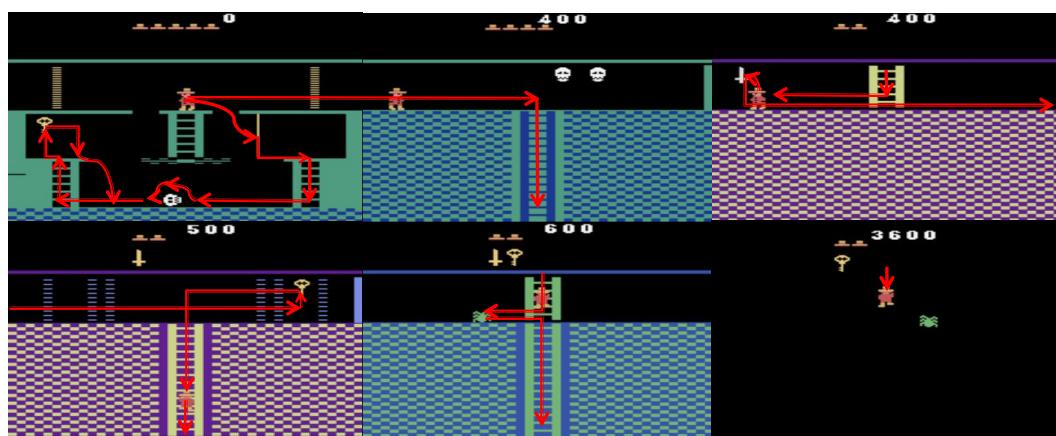
645 Enmin Zhao, Shihong Deng, Yifan Zang, Yongxin Kang, Kai Li, and Junliang Xing. Potential driven
 646 reinforcement learning for hard exploration tasks. In *International Joint Conferences on Artificial
 647 Intelligence*, pp. 2096–2102, 2020.

648 **A APPENDIX**649 **A.1 ADDITIONAL EXPERIMENT**

650 To further demonstrate the validity of DANCE, we evaluate DANCE and RND’s performance in
 651 the absence of any extrinsic reward to see whether the proposed DANCE mechanism improves the
 652 exploration performance. We use *Montezuma’s Revenge* as the experimental environment and set
 653 all the extrinsic rewards as 0. Figure 5 shows the mean episodic return and the mean number of
 654 rooms over the training process. From Figure 5 we see that the DANCE agent explores more rooms
 655 and receives a higher episodic return than the RND agent, which demonstrates the superiority of
 656 DANCE. In many practical applications, agents need to explore environments with extremely sparse
 657 rewards (or even no rewards), and the experimental results show that DANCE outperforms many
 658 competing algorithms. The furthest exploration of DANCE explores 6 rooms and receives 3,600
 659 points, as shown in Figure 6. During the pure exploration training process, the DANCE agent
 660 successfully visit 11 rooms in total without any extrinsic rewards.
 661



662
 663 Figure 5: (Left) Mean episodic return and (Right) mean rooms visited received by the pure explo-
 664 ration agent in *Montezuma’s Revenge* trained without access to extrinsic rewards using RND and
 665 DANCE. The results are reported by using 5 random seeds in 40M timesteps.
 666



667
 668 Figure 6: An example policy the DANCE agent learned. DANCE explores 6 rooms and receives
 669 3,600 points without extrinsic rewards.
 670

671 **A.2 ADDITIONAL PROOF**

672 In this section, we extend the assumptions of Theorem 1 and Theorem 4.
 673

674 **Theorem 1.** *Even when there are n identical states s in a fixed policy π , the timesteps for these
 675 states are t_0, t_1, \dots, t_{n-1} . The difficulty signal $ds(s)$ of the state is the mean timestep of these states*

702 $ds(s) = \frac{t_0 + t_1 + \dots + t_{n-1}}{n}$. Under n accesses, the sum of intrinsic rewards for n identical states s
 703 remains $t_0 + t_1 + \dots + t_{n-1}$.
 704

705 *Proof.* The loss function \mathcal{L}_s for state s is:

$$\begin{aligned} 706 \quad \mathcal{L}_s &= \sum_{i=0}^{n-1} \|\hat{g}(s) - (g(s) + t_i)\|^2 \\ 707 \quad &= n(\hat{g}(s) - (g(s) + \frac{t_0 + t_1 + \dots + t_{n-1}}{n}))^2 + C, \\ 708 \quad & \end{aligned} \quad (8)$$

709 where C is a constant. When the RPN converge, the optimal predictor network $\hat{g}(s)$ should minimize
 710 the loss \mathcal{L}_s for the state s and satisfy:
 711

$$714 \quad \hat{g}(s) - (g(s) + \frac{t_0 + t_1 + \dots + t_{n-1}}{n}) = 0. \quad (9)$$

715 The intrinsic reward of state s is:

$$717 \quad r_s^i = \|\hat{g}(s) - g(s)\|_2 = \frac{t_0 + t_1 + \dots + t_{n-1}}{n}. \quad (10)$$

719 Under n accesses, the sum of intrinsic rewards for n identical states s remain $t_0 + t_1 + \dots + t_{n-1}$.
 720

$$721 \quad n \cdot r_s^i = t_0 + t_1 + \dots + t_{n-1}. \quad (11)$$

722 \square

723 It implies that for a fixed policy π^{fix} , the cumulative intrinsic rewards are the sum of the steps (e.g.,
 724 $\mathbf{E}_{\pi^{fix}}[\sum_{t=0}^{\infty} \gamma^t \beta \omega r_{s_t}^i(s_t)] = \sum_{t=0}^{\infty} \gamma^t \beta \omega t$, when $\gamma = 1$. This ensures that Theorem 4 in the original
 725 paper remains valid even when multiple identical states in the fixed policy.

726 **Theorem 4.** The policy π_i^* DANCE agent learns is at least a subset of policy π^* common RL method
 727 learns.

728 *Proof.* According to Theorem 2, the cumulative extrinsic rewards received by optimal policy π^*
 729 should be no less than π_i^* , we have :

$$732 \quad \mathbf{E}_{\pi^*} \left[\sum_{t=0}^{\infty} \gamma^t r_t^e(s_t, a_t) \right] \geq \mathbf{E}_{\pi_i^*} \left[\sum_{t=0}^{\infty} \gamma^t r_t^e(s_t, a_t) \right], \quad (12)$$

733 and thus:

$$\begin{aligned} 734 \quad & \mathbf{E}_{\pi^*} \left[\sum_{t=0}^{\infty} \gamma^t r_t^e(s_t, a_t) \right] + \sum_{t=0}^{\infty} \gamma^t \beta \omega t \\ 735 \quad & \geq \mathbf{E}_{\pi_i^*} \left[\sum_{t=0}^{\infty} \gamma^t r_t^e(s_t, a_t) \right] + \sum_{t=0}^{\infty} \gamma^t \beta \omega t. \end{aligned} \quad (13)$$

736 According to Theorem 1, for fixed policies π_i^* , the cumulative intrinsic rewards
 737 $\mathbf{E}_{\pi_i^*}[\sum_{t=0}^{\infty} \gamma^t \beta \omega r_{s_t}^i(s_t)]$ is equal to $\sum_{t=0}^{\infty} \gamma^t \beta \omega t$, hence we obtain:
 738

$$\begin{aligned} 744 \quad & \mathbf{E}_{\pi_i^*} \left[\sum_{t=0}^{\infty} \gamma^t r_t^e(s_t, a_t) \right] + \sum_{t=0}^{\infty} \gamma^t \beta \omega t \\ 745 \quad & = \mathbf{E}_{\pi_i^*} \left\{ \sum_{t=0}^{\infty} \gamma^t [r_t^e(s_t, a_t) + \beta r_{s_t}^i(s_t)] \right\}. \end{aligned} \quad (14)$$

750 From Eqns. (13) and (14), we have:

$$\begin{aligned} 752 \quad & \mathbf{E}_{\pi^*} \left[\sum_{t=0}^{\infty} \gamma^t r_t^e(s_t, a_t) \right] + \sum_{t=0}^{\infty} \gamma^t \beta \omega t \\ 753 \quad & \geq \mathbf{E}_{\pi_i^*} \left\{ \sum_{t=0}^{\infty} \gamma^t [r_t^e(s_t, a_t) + \beta r_{s_t}^i(s_t)] \right\}. \end{aligned} \quad (15)$$

756 In addition, according to Theorem 3, the expected cumulative total reward π_i^* gets should be no less
 757 than that of π^* . Thus policy π_i^* also satisfies:
 758

$$\begin{aligned} 759 & \mathbf{E}_{\pi_i^*} \left\{ \sum_{t=0}^{\infty} \gamma^t [r_t^e(s_t, a_t) + \beta r_{s_t}^i(s_t)] \right\} \\ 760 & \geq \mathbf{E}_{\pi^*} \left\{ \sum_{t=0}^{\infty} \gamma^t [r_t^e(s_t, a_t) + \beta r_{s_t}^i(s_t)] \right\}, \end{aligned} \quad (16)$$

765 and thus:

$$\begin{aligned} 766 & \mathbf{E}_{\pi_i^*} \left\{ \sum_{t=0}^{\infty} \gamma^t [r_t^e(s_t, a_t) + \beta r_{s_t}^i(s_t)] \right\} \\ 767 & \geq \mathbf{E}_{\pi^*} \left[\sum_{t=0}^{\infty} \gamma^t r_t^e(s_t, a_t) \right] + \sum_{t=0}^{\infty} \gamma^t \beta \omega t. \end{aligned} \quad (17)$$

772 From Eqns. (15) and (17), we can see that the equality sign in both equations hold. Since
 773

$$\mathbf{E}_{\pi_i^*} \left[\sum_{t=0}^{\infty} \gamma^t r_t^e(s_t, a_k) \right] = \mathbf{E}_{\pi^*} \left[\sum_{t=0}^{\infty} \gamma^t r_t^e(s_t, a_k) \right], \quad (18)$$

774 which proves that policy π_i^* is at least a subset of policy π^* .
 775

776 **Proof of existence for π^* and π_i^* :** According to Bellman Equation Sutton & Barto (1998); Sutton
 777 et al. (1999), for any RL problem, the policy π^* is exist. In addition, from Eqns. (16) and (18), we
 778 see that the policy π^* can maximize the learning goal in Eqn. (4), so the policy π_i^* is exist. \square
 779

780
 781
 782
 783
 784
 785
 786
 787
 788
 789
 790
 791
 792
 793
 794
 795
 796
 797
 798
 799
 800
 801
 802
 803
 804
 805
 806
 807
 808
 809