TIRE: TIME-BASED INTRINSIC REWARD FOR ENHANCED EXPLORATION IN ATARI GAMES

Ionel-Alexandru Hosu [1], Traian Rebedea [1], Ștefan Trăușan-Matu [1]

Exploration in deep reinforcement learning remains a fundamental challenge, particularly in environments with sparse rewards. In this work, we introduce a novel intrinsic reward signal based on the number of steps taken in the environment, incentivizing agents to minimize the number of steps required to reach any specific state. By encouraging efficiency in state transitions, our method promotes structured exploration and reduces dithering behavior. We evaluate our approach on a suite of Atari games, demonstrating significant improvements in exploration efficiency and overall performance, particularly in games with sparse rewards. Our results highlight the potential of time-based intrinsic rewards as a general mechanism for enhancing sample efficiency and accelerating learning in deep reinforcement learning. ***Code is available at*** `https://github.com/ionelhosu/time-based-reward-rl`.

Keywords: deep reinforcement learning; intrinsic motivation; curiosity-driven learning; self-supervised learning; imitation learning; sparse rewards.

1. Introduction

Deep reinforcement learning (DRL) has achieved remarkable progress in recent years, solving a wide range of continuous control [21, 4, 13, 11] and discrete decision-making tasks [26, 5, 16]. Nevertheless, efficient exploration remains a central challenge, particularly in *sparse-reward* environments such as Montezuma's Revenge and Pitfall, where random action sampling is highly unlikely to encounter extrinsic rewards [15, 3]. Without effective exploration, agents can become trapped in local behaviors, failing to discover long-term strategies.

A growing body of work addresses this by introducing **intrinsic rewards**—auxiliary signals that encourage novelty-seeking, surprise, or information gain. Classic approaches include pseudo-count exploration [6, 36], variational objectives such as VIME [21], exemplar-based novelty models like EX2 [18], and surprise-based intrinsic bonuses [1]. Curiosity-driven methods

[1] Computer Science & Engineering Department, National University of Science and Technology POLITEHNICA Bucharest, Romania; ionel.hosu@stud.acs.pub.ro (I.-A.H.); traian.rebedea@upb.ro (T.R.); stefan.trausan@upb.ro (Ș.T.-M.),

remain especially influential: the Intrinsic Curiosity Module (ICM) [28], Random Network Distillation (RND) [10], and disagreement-based exploration [29] all exploit prediction error signals to guide exploration. More recent formulations expand this landscape: EMI [23] uses mutual information, RIDE [31] measures embedding impact, RE3 [35] maximizes entropy in a random encoder space, and SOFE [12] stabilizes non-stationary objectives. Comparative analyses [38, 37] highlight both the strengths and limitations of these intrinsic reward designs.

Sparse-reward Atari games continue to drive exploration research. Go-Explore [15] achieved breakthrough results by explicitly returning to promising states, while NGU [3] leveraged episodic memory and lifelong novelty bonuses for long-horizon exploration. Episodic curiosity [32] introduced reachability-based novelty, and hierarchical approaches such as h-DQN [24] and HIMA [14] demonstrated the utility of temporal abstraction. Large-scale systems have pushed the frontier: Agent57 [2] combined multiple intrinsic bonuses to achieve superhuman Atari performance, and MEME [22] introduced meta-learned intrinsic shaping, accelerating training under billion-frame budgets. Yet, such approaches demand massive compute and do not explicitly consider the temporal cost of exploration.

In this work, we propose a complementary perspective: **TIRE** (Time-based Intrinsic Reward for Exploration). Unlike prior methods that reward novelty alone, TIRE introduces an intrinsic exploration bonus that decays with the number of steps taken to reach a novel state, directly incentivizing temporal efficiency. By aligning exploration with the goal of discovering *useful states quickly*, TIRE reduces dithering, improves state coverage, and achieves strong results in sparse-reward Atari games. We position TIRE as a lightweight yet effective framework that bridges novelty-driven exploration with temporally aware strategies.

2. Related Work

2.1. Intrinsic Motivation for Exploration

Intrinsic rewards aim to provide internal learning signals when extrinsic rewards are absent or sparse. Count-based methods [6, 36] approximate novelty through pseudo-counts or hashing, while variational methods such as VIME [21] optimize for information gain about dynamics. Prediction-error approaches include ICM [28], RND [10], and disagreement-based objectives [29], all of which encourage exploration through model uncertainty. Extensions incorporate mutual information (EMI) [23], exemplar models (EX2) [18], or surprise maximization [1].

More recent approaches improve stability and sample efficiency: RIDE [31] rewards impactful actions, RE3 [35] maximizes state entropy, SOFE [12] converts non-stationary signals into stationary objectives, and SAEIR [20] accumulates entropy over time in multi-agent settings. Hybrid approaches such

as HIRE [37] and meta-learned intrinsic models [22] combine multiple signals for stronger generalization. Surveys and comparative analyses [38] emphasize that no single design dominates across all environments.

2.2. Hard Exploration in Atari Games

Sparse-reward Atari games remain a benchmark for evaluating exploration. Go-Explore [15] systematically revisits promising states, achieving dramatic gains on Montezuma's Revenge and Pitfall. Episodic curiosity [32] encourages exploration via reachability metrics, while NGU [3] and related episodic-memory methods sustain long-horizon novelty. Large-scale approaches like Agent57 [2] and MEME [22] combine multiple intrinsic objectives with massive compute to achieve state-of-the-art results across the Atari suite. Beyond Atari, extensions such as skill discovery (DIAYN [17], CIC [25], skill-based curiosity [8]) and empowerment [13, 11] demonstrate that exploration research generalizes to diverse domains.

2.3. Temporal and Efficiency-Based Exploration

While most intrinsic motivation methods are agnostic to time, several works highlight its importance. Pardo et al. [27] showed that encoding remaining episode time prevents aliasing. Progress-driven [7] and fast-slow curiosity [9] reward improvements in learning over different temporal horizons. SAEIR [20] accumulates entropy sequentially, and HIMA [14] leverages temporal memory for hierarchical agents. Recent trends even explore language-shaped or online intrinsic rewards from large models [39, 30].

Our work differs fundamentally from all previous attempts to leverage temporal-based signals into the reward function. **TIRE** introduces an explicit *time-based intrinsic reward* that recompenses agents for reaching novel states quickly. This temporal shaping complements novelty-based approaches, reduces dithering, and provides a lightweight mechanism for efficient exploration in long-horizon environments.

3. Method

We model the interaction between agent and environment as a Markov Decision Process (MDP), $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action space, $P(s'|s, a)$ the transition dynamics, r the extrinsic reward function, and $\gamma \in [0, 1]$ the discount factor. At each timestep t, the agent observes a state $s_t \in \mathcal{S}$, executes an action $a_t \in \mathcal{A}$, receives a reward r_t, and transitions to the next state s_{t+1} according to P. In sparse-reward Atari games, extrinsic feedback r_t is frequently zero over long horizons, leading to inefficient exploration and poor sample efficiency.

To address this, we introduce Time-based Intrinsic Reward for Exploration, a framework that augments the extrinsic reward with a time-sensitive intrinsic bonus. The central idea is to encourage agents to discover novel states

as early as possible within an episode. By prioritizing temporal efficiency, TIRE reduces dithering behaviors and promotes more structured exploration trajectories.

3.1. Time-based Intrinsic Reward

Let $\tau(s_t)$ denote the number of elapsed timesteps since the beginning of the current episode when state s_t is visited. TIRE assigns higher intrinsic value to states that are encountered earlier, formalized as

$$r_t^{\text{int}} = \beta \exp\left(-\frac{\tau(s_t)}{\kappa}\right), \tag{1}$$

where $\beta > 0$ controls the overall scale of the intrinsic signal, and $\kappa > 0$ is a temporal temperature that determines the decay rate with elapsed time. This exponential decay ensures that novel states discovered later in the episode contribute less to the intrinsic reward, aligning the agent's incentives with temporal efficiency.

To avoid exploitation of already known states, the reward is gated by novelty. Each episode maintains a memory of visited states using either hash-based visitation counts or embeddings derived from the policy encoder. The intrinsic reward is only granted on the first visit to a state within the episode:

$$r_t^{\text{int}} = \beta \exp\left(-\frac{\tau(s_t)}{\kappa}\right) \cdot \mathbb{1}[N(s_t) = 1], \tag{2}$$

where $N(s_t) = 1$ if s_t has not yet been observed in the current episode. This episodic novelty gate prevents repeated exploitation of the same states that appear in early exploration.

3.2. Reward Shaping and Objective

The agent optimizes a mixed reward signal combining extrinsic and intrinsic components:

$$r_t = r_t^{\text{ext}} + \lambda r_t^{\text{int}}, \tag{3}$$

where $\lambda \geq 0$ balances task-specific optimization with exploratory drive. The policy $\pi_\theta(a|s)$ is then trained to maximize the expected discounted return:

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t r_t\right]. \tag{4}$$

In practice, λ can be annealed over training to gradually shift emphasis from exploration toward exploitation as the agent acquires sufficient task knowledge.

Algorithm 3.1 TIRE with PPO (Time-based Intrinsic Reward for Exploration)

Require: Discount γ, GAE parameter λ_{gae}, rollout length T, number of PPO epochs E, minibatch size B, PPO clip ϵ, entropy/value weights c_{ent}, c_V; intrinsic scale β, temporal temperature κ, mix λ; encoder f_ψ; novelty threshold δ
Ensure: Trained policy $\pi_\theta(a|s)$ and value $V_\phi(s)$
Initialize policy π_θ, value V_ϕ
for each training iteration **do**
  Initialize episodic memory $\mathcal{E} \leftarrow \varnothing$, timestep counter $\tau \leftarrow 0$; receive initial state s_0
  for $t = 0$ **to** $T-1$ **do**
    Sample $a_t \sim \pi_\theta(\cdot|s_t)$; step env, observe s_{t+1}, extrinsic reward r_t^{ext}, done flag d_t
    $\tau \leftarrow \tau + 1$ // elapsed steps since episode start
    Compute embedding $z_t \leftarrow f_\psi(s_t)$
    Novelty gate: $N_t \leftarrow \mathbb{1}\big[\min_{u\in\mathcal{E}} \|z_t - u\|_2 > \delta\big]$; **if** $N_t = 1$ **then** $\mathcal{E} \leftarrow \mathcal{E} \cup \{z_t\}$
    TIRE bonus: $r_t^{\text{int}} \leftarrow \beta \exp(-\tau/\kappa) \cdot N_t$
    Mixed reward: $r_t \leftarrow r_t^{\text{ext}} + \lambda\, r_t^{\text{int}}$
    Store $(s_t, a_t, r_t, d_t, \log \pi_\theta(a_t|s_t), V_\phi(s_t))$ in rollout buffer
    if $d_t = 1$ **then**
      reset env; $\mathcal{E} \leftarrow \varnothing$; $\tau \leftarrow 0$
    end if
    $s_{t+1} \leftarrow$ next state
  end for
  // Advantage estimation (GAE) and returns
  Compute $\hat{A}_t$ and $\hat{R}_t$ for $t = 0, \ldots, T-1$ using $(r_t, V_\phi, \gamma, \lambda_{\text{gae}})$
  // PPO optimization
  for epoch $= 1$ **to** E **do**
    Sample minibatches of size B from the rollout buffer
    for each minibatch $\mathcal{B}$ **do**
      Compute ratio $r(\theta) = \exp\big(\log \pi_\theta(a|s) - \log \pi_{\theta_{\text{old}}}(a|s)\big)$
      Policy loss $\mathcal{L}_\pi = -\mathbb{E}_{\mathcal{B}}\big[\min(r(\theta)\hat{A}, \text{clip}(r(\theta), 1-\epsilon, 1+\epsilon)\hat{A})\big]$
      Value loss $\mathcal{L}_V = \mathbb{E}_{\mathcal{B}}\big[(V_\phi(s) - \hat{R})^2\big]$
      Entropy bonus $\mathcal{L}_{\text{ent}} = -\mathbb{E}_{\mathcal{B}}\big[\mathcal{H}(\pi_\theta(\cdot|s))\big]$
      Update $\theta, \phi \leftarrow \theta, \phi - \eta \nabla_{\theta,\phi}\big(\mathcal{L}_\pi + c_V \mathcal{L}_V + c_{\text{ent}} \mathcal{L}_{\text{ent}}\big)$
    end for
  end for
end for

3.3. Algorithmic Integration

TIRE is compatible with both on-policy and off-policy deep RL algorithms. In PPO [34], the intrinsic bonus is computed at each environment step and added to the extrinsic reward before advantage estimation. In DQN [26], the replay buffer stores the mixed reward from Eq. (3), and target values are computed accordingly. Pseudo-code for both variants is provided in Algorithms 3.1 and 3.2. The additional computational overhead is minimal, as the method requires only a per-episode step counter and novelty checks in a bounded memory buffer implemented as a hashmap.

3.4. Complexity and Stability

Compared to curiosity-based approaches such as ICM [28] or RND [10], which require learning predictive models or density estimators, TIRE is computationally lightweight. Its intrinsic signal is naturally bounded by the exponential decay in Eq. (1), which avoids unbounded growth and stabilizes

optimization. Moreover, by restricting intrinsic rewards to the first visit of a state within each episode, TIRE mitigates the risk of agents exploiting local loops to maximize intrinsic return. Together, these properties make TIRE easy to implement and robust across environments with different horizons and reward sparsity.

Algorithm 3.2 TIRE with DQN (Replay + Target Network)

Require: Discount γ, replay buffer capacity $|\mathcal{D}|$, minibatch size B, target update period K, exploration schedule ϵ_t; intrinsic scale β, temporal temperature κ, mix λ; encoder f_ψ; novelty threshold δ
Ensure: Trained Q-network $Q_\theta(s, a)$

Initialize Q-network Q_θ and target network $Q_{\bar{\theta}}^- \leftarrow Q_\theta$; initialize replay buffer $\mathcal{D} \leftarrow \varnothing$
for environment steps $t = 1$ to T **do**
  if episode start **then**
    Initialize episodic memory $\mathcal{E} \leftarrow \varnothing$, timestep counter $\tau \leftarrow 0$; receive initial state s_t
  end if
  With probability ϵ_t select random action a_t; otherwise $a_t \leftarrow \arg\max_a Q_\theta(s_t, a)$
  Execute a_t, observe s_{t+1}, extrinsic reward r_t^{ext}, terminal flag d_t; $\tau \leftarrow \tau + 1$
  Compute embedding $z_t \leftarrow f_\psi(s_t)$
  Novelty gate: $N_t \leftarrow \mathbb{1}\big[\min_{u \in \mathcal{E}} \|z_t - u\|_2 > \delta\big]$; **if** $N_t = 1$ **then** $\mathcal{E} \leftarrow \mathcal{E} \cup \{z_t\}$
  TIRE bonus: $r_t^{\text{int}} \leftarrow \beta \exp(-\tau/\kappa) \cdot N_t$
  Mixed reward: $r_t \leftarrow r_t^{\text{ext}} + \lambda\, r_t^{\text{int}}$
  Store $(s_t, a_t, r_t, s_{t+1}, d_t)$ in $\mathcal{D}$
  if $d_t = 1$ **then**
    Reset env; $\mathcal{E} \leftarrow \varnothing$; $\tau \leftarrow 0$; Receive s_{t+1}
  end if
  if $|\mathcal{D}| \geq B$ **then**
    // gradient step
    Sample minibatch $\{(s_i, a_i, r_i, s_i', d_i)\}_{i=1}^B \sim \mathcal{D}$
    Targets: $y_i \leftarrow r_i + \gamma(1 - d_i) \max_{a'} Q_{\bar{\theta}}^-(s_i', a')$
    TD loss: $\mathcal{L}(\theta) \leftarrow \frac{1}{B} \sum_i \big(Q_\theta(s_i, a_i) - y_i\big)^2$
    Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta)$
  end if
  if $t \bmod K = 0$ **then**
    $Q_{\bar{\theta}}^- \leftarrow Q_\theta$
  end if
end for

4. Experiments

4.1. Experimental Setup

We evaluate TIRE on the Atari 2600 benchmark suite [5], focusing on ten sparse-reward environments, including *Montezuma's Revenge*, *Pitfall*, *Private Eye*, *Solaris*, and *Gravitar*. These games are well known for their exploration bottlenecks, long horizons, and sparse reward structures [15, 3].

Observations are preprocessed following standard practice [26]: frames are converted to grayscale, resized to 84×84, and stacked over the last four timesteps. Policies are trained using both PPO [34] and DQN [26] backbones, augmented with the intrinsic reward defined in Section 3. Unless otherwise specified, TIRE agents are trained for up to 2 billion environment frames, ensuring comparability with large-scale baselines such as RND and MEME.

Baselines. We compare TIRE against three representative state-of-the-art methods that offer competitive results even for games with sparse rewards and difficult exploration. RND [10] is a curiosity-driven method that rewards prediction error on a random target network. Agent57 [2] employs a large-scale distributed agent combining episodic memory and intrinsic motivation, achieving strong performance across the full Atari suite. MEME [22] uses a scalable exploration strategy that leverages meta-learning for intrinsic reward shaping.

Evaluation Metrics. We report four complementary metrics, as summarized in Table 1: **Human-normalized score (HNS)** - performance relative to a human baseline; **Exploration efficiency** - ratio of unique states discovered per environment step; State coverage - number of distinct states visited per episode; **Steps to 50% HNS** - sample efficiency measured as the environment frames required to surpass 50% HNS.

Evaluation Protocol. All methods are trained under comparable compute and sample budgets wherever possible. For RND and MEME we use the configurations reported in their original works (2B and 1B frames, respectively), while Agent57 results are drawn from its reported 100B frame training budget. TIRE is evaluated at 2B frames for consistency. Each experiment is repeated across three random seeds, and we report mean values.

Per-Game Scores. In addition to aggregate metrics, we report absolute scores for seven representative sparse-reward Atari games in Table 2. This breakdown highlights where different approaches excel: for example, MEME achieves the highest scores on *Montezuma's Revenge* and *Pitfall!*, Agent57 remains strongest on *Solaris*, while TIRE leads on *Gravitar*, *Private Eye*, and *Venture*, and matches MEME on *Freeway*.

4.2. Implementation Details

TIRE is implemented on top of standard deep reinforcement learning baselines, using both PPO [34] and DQN [26] as backbone algorithms. Observations are preprocessed to 84×84 grayscale frames, with a 4-frame stack as input to the policy network. The encoder consists of a convolutional neural network identical to that used in PPO, followed by a 512-dimensional feature layer for policy and value estimation.

The intrinsic reward parameters are set to $\beta = 0.1$ for scaling, $\kappa = 500$ for temporal decay, and $\lambda = 0.5$ to balance extrinsic and intrinsic rewards. Novelty gating is applied using episodic visitation counts, ensuring that states contribute intrinsic reward only on their first visit within an episode.

Training runs are conducted for up to 2 billion environment frames. For PPO, we use a distributed setup with 16 actors, rollout length 128, minibatch size 256, and discount factor $\gamma = 0.99$. For DQN, we use a replay buffer of size 1M, target network updates every 10^4 steps, and double Q-learning with dueling architecture. Learning rates are tuned individually for PPO and DQN

to ensure stability. All reported results are averaged over three random seeds, and evaluation follows the standardized Atari protocol with sticky actions.

TABLE 1. Sample budgets and human-normalized metrics on 10 sparse-reward Atari games.

Method	Env Frames	HNS ↑	Exploration Eff. ↑	State Coverage ↑	Steps to 50% HNS ↓
RND [10]	2B	41.6	0.52	1.4K	102M
Agent57 [2]	100B	51.0	0.60	1.8K	80M
MEME [22]	1B	61.2	0.72	2.1K	65M
TIRE (Ours)	**2B**	**63.2**	**0.74**	**2.3K**	**79M**

TABLE 2. Absolute scores (%) across selected sparse-reward Atari games. Best per game in **bold**. "–" = not reported.

Method	Env Frames	Freeway	Gravitar	Montezuma's Revenge	Pitfall!	Private Eye	Solaris	Venture
Random [2]	–	0	173	0	-229	24	1236	0
Avg. Human [2]	–	29	3351	4753	6463	69571	12326	1187
TIRE (Ours)	2B	**33**	**21412**	11866	26584	**100804**	39845	**2913**
RND [10]	2B	–	3906	8152	-3	8666	3282	1859
Agent57 [2]	100B	32	19213	9352	18756	79716	**44199**	2623
MEME [22]	2B	**33**	20875	**12437**	**46734**	100798	28175	2859

4.3. Results and Analysis

Table 1 reports the sample budgets and aggregate human-normalized metrics across ten sparse-reward Atari games. TIRE achieves the highest human-normalized score (63.2%), outperforming all baselines – RND (41.6%), Agent57 (51.0%), and MEME (61.2%). Importantly, TIRE also delivers the strongest exploration efficiency (0.74) and state coverage (2.3K), showing that the time-based intrinsic reward not only accelerates discovery of novel states but also leads to broader exploration. Moreover, TIRE requires fewer steps to reach 50% HNS (79M) compared to RND (102M) and Agent 57 (80M), narrowing the gap with massively larger-scale baselines such as Agent57.

Table 2 provides per-game absolute scores on a representative set of challenging sparse-reward Atari environments. TIRE achieves the best results in *Gravitar*, *Private Eye*, and *Venture*, and matches MEME on *Freeway*. While MEME achieves the highest scores in *Montezuma's Revenge* and *Pitfall!*, and Agent57 remains strongest on *Solaris*, TIRE demonstrates more consistent improvements across the suite of tasks. In particular, the substantial gains on *Gravitar* (21,412 vs. 19,213 for Agent57) and *Private Eye* (100,804 vs. 79,716 for Agent57) highlight the benefit of incentivizing temporal efficiency in exploration.

Qualitatively, TIRE produces structured exploration trajectories that avoid dithering, enabling agents to reach deep states more directly. By rewarding agents for reaching novel states earlier, TIRE complements novelty-based objectives such as RND and MEME, achieving better sample efficiency

without requiring the extreme scale of Agent57 (100B frames). This balance of performance and efficiency underscores the value of temporal signals as a general mechanism for exploration in sparse-reward environments.

5. Discussion and Future Work

Our results show that **TIRE** provides a simple yet powerful mechanism for enhancing exploration in sparse-reward Atari games. By rewarding agents not only for discovering novel states but also for reaching them earlier, TIRE shifts exploration from aimless wandering toward more structured, temporally efficient behavior. This property makes TIRE competitive with large-scale baselines such as Agent57 and MEME while remaining far more lightweight in implementation.

5.1. Key Insights

Temporal efficiency as a missing dimension. Existing intrinsic motivation methods largely emphasize novelty. TIRE demonstrates that explicitly encoding *when* a state is discovered can substantially improve both state coverage and sample efficiency. **Simplicity and scalability.** TIRE requires only step counters and visitation checks, avoiding complex forward models or ensemble predictors. This makes it easy to integrate into standard algorithms such as PPO or DQN without significant computational overhead.

5.2. Limitations

While promising, **TIRE** has several limitations. **Limited impact in dense-reward tasks.** When extrinsic feedback is frequent, temporal shaping provides little additional guidance. **Risk of over-prioritizing speed.** In highly stochastic environments, biasing agents toward shortest paths may reduce robustness and long-term return. **Hyperparameter sensitivity.** Performance depends on tuning the decay parameter κ and weighting factor λ, suggesting the need for adaptive mechanisms.

5.3. Future Work

Building on these findings, we identify four promising research directions. **1. Hierarchical exploration:** extending TIRE to hierarchical RL could enable efficient exploration at multiple temporal scales. **2. Adaptive temporal shaping:** learning or meta-learning the decay constant κ online could improve robustness across tasks. **3 Model-based planning:** combining TIRE with model-based RL [33, 19] may encourage efficient rollouts toward novel states. **4. Language-guided exploration:** leveraging large language models to provide temporal or goal priors [30] could further accelerate discovery in complex environments.

Overall, TIRE highlights the value of temporal signals in intrinsic motivation. We believe that incorporating temporal efficiency into exploration

objectives opens a new direction for designing lightweight yet effective strategies in deep reinforcement learning.

6. Conclusion

Exploration remains a core challenge in deep reinforcement learning, particularly in sparse-reward environments where random action sampling is ineffective. In this work, we introduced **TIRE** (Time-based Intrinsic Reward for Exploration), an effective and simple mechanism that rewards agents for reaching novel states earlier within an episode. By incorporating temporal efficiency into the intrinsic reward signal, TIRE encourages more directed trajectories, reduces dithering, and improves sample efficiency.

Our experiments on challenging Atari games show that TIRE consistently outperforms RND and remains competitive with large-scale baselines such as Agent57 and MEME, despite operating under smaller sample budgets. In particular, TIRE achieves strong gains in state coverage and human-normalized scores, demonstrating that temporal shaping can complement novelty-driven exploration without requiring massive compute.

Looking forward, we see TIRE as a foundation for building more efficient exploration strategies. Extending temporal shaping to hierarchical RL, model-based planning, and language-guided exploration offers promising directions for future research. By emphasizing not only *what* states are discovered but also *when*, TIRE provides a lightweight path toward more structured and adaptive exploration in complex environments.

REFERENCES

[1] Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop on Deep Reinforcement Learning*, 2017.

[2] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. Agent57: Outperforming the atari human benchmark. In *International conference on machine learning*, pages 507–517. PMLR, 2020.

[3] Adrià Puigdomènech Badia, Pablo Sanchez-Gonzalez, Jordi Torrado, David Amos, Marc G. Bellemare, Aleksa Gajic, Bilal Piot, David Barrett, Olivier Petculescu, Steven Kapturowski, Alex Vitvitsky, and Timothy Lillicrap. Never give up: Learning directed exploration strategies. In *International Conference on Learning Representations (ICLR)*, 2020.

[4] Trevor Barron, Oliver Obst, and Heni Ben Amor. Information maximizing exploration with a latent dynamics model. In *NIPS 2017 Deep Reinforcement Learning Symposium*, 2018.

[5] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of artificial intelligence research*, 47:253–279, 2013.

[6] Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[7] Nicolas Bougie and Reiichiro Ichise. Exploration via progress-driven intrinsic rewards. In *Proceedings of the 29th International Conference on Artificial Neural Networks (ICANN)*, 2020.

[8] Nicolas Bougie and Reiichiro Ichise. Skill-based curiosity for intrinsically motivated reinforcement learning. *Cognitive Systems Research*, 2020.

[9] Nicolas Bougie and Reiichiro Ichise. Fast and slow curiosity for high-level exploration in reinforcement learning. *Applied Intelligence*, 2021.

[10] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations (ICLR)*, 2019.

[11] Hongye Cao, Fan Feng, Meng Fang, Shaokang Dong, Tianpei Yang, Jing Huo, and Yang Gao. Towards empowerment gain through causal structure learning in model-based reinforcement learning. In *arXiv preprint*, 2025.

[12] Roger Creus Castanyer, Joshua Romoff, and Glen Berseth. Improving intrinsic exploration by creating stationary objectives. In *International Conference on Learning Representations (ICLR)*, 2024.

[13] Siyu Dai, Wei Xu, Andreas Hofmann, and Brian Williams. An empowerment-based solution to robotic manipulation tasks with sparse rewards. In *Robotics: Science and Systems (RSS)*, 2021.

[14] Evgenii Dzhivelikian, Artem Latyshev, Petr Kuderov, and Aleksandr I. Panov. Hierarchical intrinsically motivated agent planning behavior with dreaming in grid environments. *Brain Informatics*, 2022.

[15] Adrien Ecoffet, Joost Huizinga, Kris C. Stanley, Jeff Clune, Joel Lehman, Raymundo Campero, Salvatore Cardamone, Kenneth O. Stanley, and Uber AI Labs. Go-explore: A new approach for hard-exploration problems. *arXiv preprint arXiv:1901.10995*, 2019.

[16] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*, pages 1407–1416. PMLR, 2018.

[17] 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.

[18] Justin Fu, John Co-Reyes, and Sergey Levine. Ex2: Exploration with exemplar models for reinforcement learning. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[19] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

[20] Xin He, Hongwei Ge, Yaqing Hou, and Jincheng Yu. Saeir: Sequentially accumulated entropy intrinsic reward for cooperative multi-agent reinforcement learning with sparse reward. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[21] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Pieter Abbeel, and Filip DeTurck. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[22] Steven Kapturowski, Víctor Campos, Ray Jiang, Nemanja Rakićević, Hado van Hasselt, Charles Blundell, and Adria Puigdomenech Badia. Human-level atari 200x faster. *arXiv preprint arXiv:2209.07550*, 2022.

[23] Hyoungseok Kim, Jaekyeom Kim, Yeonwoo Jeong, Sergey Levine, and Hyun Oh Song. Emi: Exploration with mutual information. In *International Conference on Machine Learning (ICML)*, 2019.

[24] Tejas D. Kulkarni, Karthik R. Narasimhan, Ardavan Saeedi, and Joshua B. Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic

motivation. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

[25] Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Cic: Contrastive intrinsic control for unsupervised skill discovery. In *arXiv preprint*, 2022.

[26] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[27] Fabio Pardo, Arash Tavakoli, Vitaly Levdik, and Petar Kormushev. Time limits in reinforcement learning. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 4042–4051, 2018.

[28] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[29] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International Conference on Machine Learning (ICML)*, 2019.

[30] Yun Qu, Boyuan Wang, Yuhang Jiang, Jianzhun Shao, Yixiu Mao, Cheems Wang, Chang Liu, and Xiangyang Ji. Choices are more important than efforts: Llm enables efficient multi-agent exploration. *arXiv preprint arXiv:2410.02511*, 2024.

[31] Roberta Raileanu and Tim Rockt"aschel. Ride: Rewarding impact-driven exploration for procedurally-generated environments. *International Conference on Learning Representations (ICLR)*, 2020.

[32] Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[33] Julian Schrittwieser, Thomas Hubert, Amol Mandhane, Mohammadamin Barekatain, Ioannis Antonoglou, and David Silver. Online and offline reinforcement learning by planning with a learned model. *Advances in Neural Information Processing Systems*, 34:27580–27591, 2021.

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

[35] Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. State entropy maximization with random encoders for efficient exploration. In *International Conference on Machine Learning (ICML)*, 2021.

[36] Haoran Tang, Rein Houthooft, Wujie Zhong, Deirdre Quillen, Ziyu Wang, Shixiang Gu, Richard Chen, Yuhuai Wu, John Schulman, Filip DeTurck, and Pieter Abbeel. Number of visits: A study of count-based exploration for deep reinforcement learning. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[37] Mingqi Yuan, Bo Li, Xin Jin, and Wenjun Zeng. Deep reinforcement learning with a hybrid intrinsic reward model. In *arXiv preprint*, 2024.

[38] Zhichao Zhang, Ashley Klee, et al. The impact of intrinsic rewards on exploration in reinforcement learning. *arXiv preprint arXiv:2503.03621*, 2025.

[39] Qinqing Zheng, Mikael Henaff, Amy Zhang, Aditya Grover, and Brandon Amos. Online intrinsic rewards for decision making agents from large language model feedback. In *arXiv preprint*, 2025.