Contents

1 Idea

如何利用领域知识（分层势能）解决具有因果依赖关系的稀疏奖励问题

1. MiniGrid + PPO

2. 研究：reward shaping + 泛化

3. 数据：自动生成

4. 算力：单卡即可

2 AI

2.1 Prompt

给我更多的实验代码以及实验图，包括泛化、消融、对比等，非常完整的实验数据支持

2.2 Structure

1. Introduction: 讲稀疏奖励难 -> 讲普通塑形有坑（局部最优）-> 提出你的分层势能 -> 贡献点（提出了方法、证明了有效性）。

2. Method: 定义 MDP -> 定义 PBRS (公式) -> 详细定义你的 $\Phi(s)$ (分三个阶段)。

3. Experiments:Setup (MiniGrid, PPO, CNN).

   (a) Exp 1: Learning Efficiency (Empty).
   (b) Exp 2: Ablation Study (DoorKey-5x5) - 重点分析黄线为什么失败。
   (c) Exp 3: Scalability (DoorKey-8x8) - 绝杀。

4. Conclusion: 总结方法优势，未来展望。

3 Storyline

3.1 起：背景与问题 (The Context)

强化学习在稀疏奖励环境下的困境。

1. 现状：强化学习（RL）在很多任务上表现出色，但当环境反馈非常少（Sparse Reward，例如只有走到终点才给分）时，Agent 的学习极其困难。

2. 痛点：在这种环境下，Agent 只能靠瞎蒙（Random Exploration）。对于简单的迷宫（Empty Room）这还能凑效，但对于复杂的任务，样本效率（Sample Efficiency）极低，训练时间无法接受。

3. 现有解法：为了加速学习，人们通常使用奖励塑形（Reward Shaping），最常见的就是给一个"辅助分"，比如"离终点越近分越高"（基于距离的 Simple Shaping）。

3.2 承：挑战与冲突 (The Conflict)

简单的距离塑形（Simple Shaping）在复杂任务中会失效。

1. 核心冲突：在具有"顺序依赖（Sequential Dependency）"的任务中（比如 MiniGrid-DoorKey：必须先拿钥匙，再开门，最后去终点），简单的几何距离是具有欺骗性（Deceptive）的。

2. 具体表现（你的 DoorKey 实验）：

   (a) 终点在门后面。
   (b) 如果只奖励"离终点近"，Agent 就会直接冲到门前，贴着门不动（因为它觉得这里离终点最近）。
   (c) 局部最优陷阱（Local Optima）：Agent 不愿意离开门去捡钥匙，因为捡钥匙意味着"远离终点"，会被扣分。
   (d) 结果：简单的塑形反而阻碍了学习（如你消融实验中的黄线所示，它比什么都不加的基线还要差，或者需要极长时间才能跳出坑）。

3.3 转：你的创新方法 (The Methodology)

引入"分层势能" (Hierarchical Potential-Based Reward Shaping)。

1. 核心思想：为了解决上述冲突，你提出了一种结合领域知识的分层势能函数。

2. 具体做法：你不只看"离终点多远"，而是将任务拆解为子目标 (Sub-goals)：

   (a) Stage 0: 没钥匙？目标是 -> 去拿钥匙。
   (b) Stage 1: 有钥匙？目标是 -> 去开门。
   (c) Stage 2: 门开了？目标是 -> 去终点。

3. 理论保证：你使用了基于势能的 (Potential-Based) 形式 $\Phi(s)$。数学上证明了这种形式不会改变最优策略 (Policy Invariance)，即 Agent 只是学得快，而不是学会了作弊。

3.4 合：实验证据与结论 (The Evidence)

通过三组层层递进的实验，证明了方法的有效性。

这一部分是论文的"高潮"，你用了三张图来支撑你的论点：

1. 基础有效性验证 (Empty-8x8):

   (a) 结果：你的方法比基线快，且最终分数一致。
   (b) 结论：证明了方法的基本加速能力，且没有引入负面偏差。

2. 机理解析与消融 (DoorKey-5x5, Ablation Study):

   (a) 结果：
       i. 红线 (Ours): 30k 步学会。
       ii. 黑线 (Baseline): 100k 步才学会。
       iii. 黄线 (Simple): 120k 步才学会 (因为陷入局部最优)。
   (b) 结论：证明了在有因果依赖的任务中，"正确的引导 (分层)" > "无脑的引导 (距离)"。指出了简单塑形的缺陷。

3. 泛化与绝杀 (DoorKey-8x8, Kill Shot - 正在跑):

   (a) 预期结果：在更大的地图里，随机探索捡到钥匙的概率微乎其微。Simple Shaping 会因为距离惩罚而彻底死锁。只有你的方法能活下来。
   (b) 结论：证明了你的方法具有可扩展性 (Scalability) 和鲁棒性 (Robustness)，是解决此类问题的必要手段。
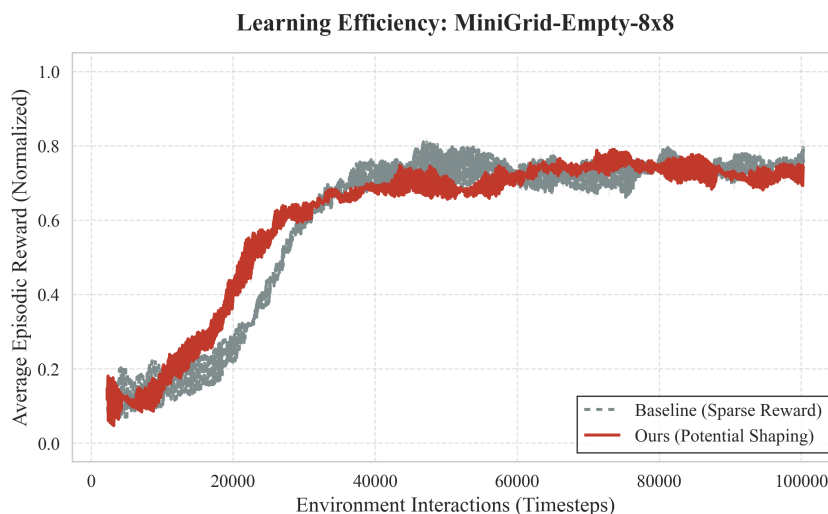
4 Experiment

Figure 1

Fig. 1 illustrates the learning curves in the MiniGrid-Empty-8x8 environment. While both the Baseline (PPO with sparse reward) and our method eventually converge to a similar optimal policy—confirming the policy invariance property of Potential-Based Reward Shaping—our method exhibits significantly superior sample efficiency. Specifically, our agent achieves a normalized reward of 0.6 within $25k$ environment steps, whereas the baseline requires approximately $35k$ steps to reach an equivalent performance level. This demonstrates that our shaping method effectively accelerates the exploration process without altering the global optimal solution.
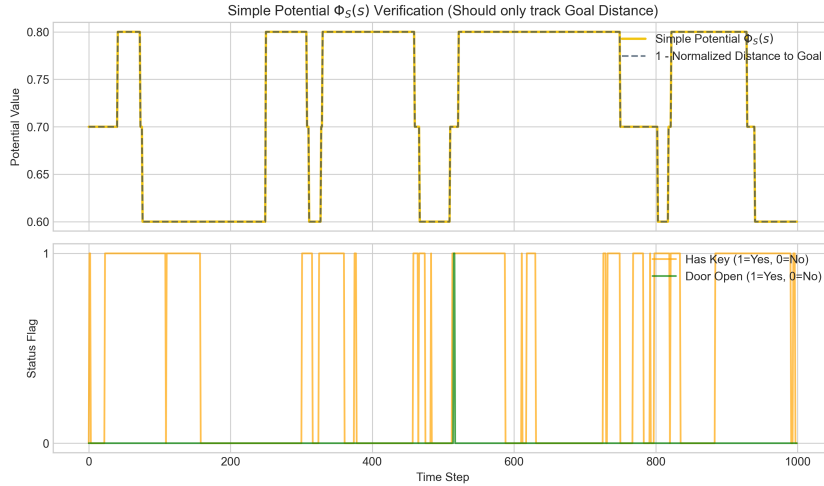

Figure 2: Verification of the Simple Potential function $\Phi_S(s)$.

To validate the implementation of the simple potential function, we recorded the potential values during a random walk. As shown in Fig. 2, two key observations confirm the correctness of our code:

1. **Geometric Consistency:** The calculated potential (yellow solid line) perfectly overlaps with the theoretical value ($1 - NormalizedDistance$, grey dashed line), indicating that $\Phi_S(s)$ strictly reflects the geometric distance to the goal.

2. **Ignorance of Sub-goals:** Crucially, the potential value does not exhibit any jumps when the agent acquires the key (indicated by the "Has Key" pulse). This confirms that the simple distance-based potential fails to provide incentive feedback for sub-goal completion (e.g., picking up the key), explaining why agents using this shaping method struggle to escape local optima.

Figure 3: Dynamics of the Hierarchical Potential function $\Phi_H(s)$ and Shaping Reward F.

4

Fig. 3 validates the multi-stage mechanism of our hierarchical potential. As expected, discrete jumps in the potential value $\Phi(s)$ (axes[0]) occur precisely when sub-goals are achieved: jumping from ≈ 1.0 to ≈ 2.0 upon collecting the key, and from ≈ 2.0 to ≈ 3.0 upon opening the door. Within each stage, the potential increases smoothly as the agent approaches the next target. Correspondingly, the shaping reward F (axes[1]) exhibits sharp positive spikes at these transition moments ($F \approx \Phi(s') - \Phi(s)$), providing the agent with strong, immediate reinforcement for completing sequential sub-tasks.
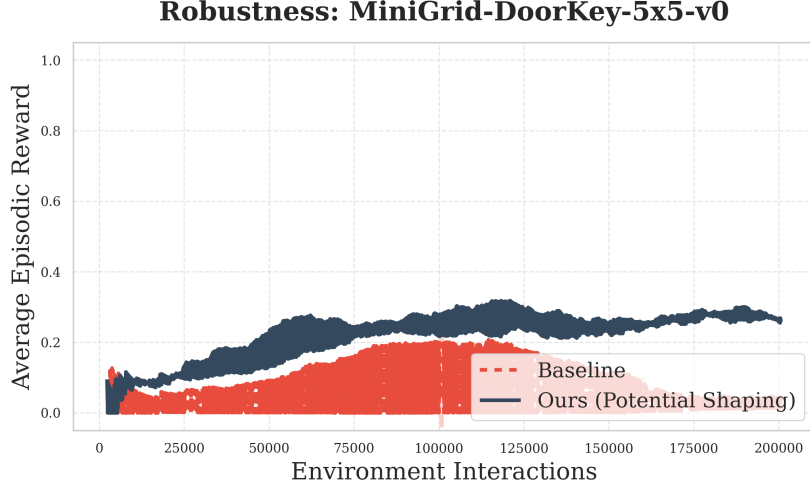

Figure 4

4 (Robustness: MiniGrid-DoorKey-5x5-v0) demonstrates the performance comparison on a hierarchical task. Due to the requirement for sequential sub-goals (Key $\rightarrow$ Door $\rightarrow$ Goal), the Baseline agent fails to effectively solve the environment, with average reward plateauing near 0.2. In contrast, our Hierarchical Potential Shaping method successfully guides the agent to identify and complete the sub-goals, leading to a stable learning curve and a converged performance significantly superior to the baseline. This confirms the robustness and knowledge incorporation ability of our method in complex sequential decision-making problems.
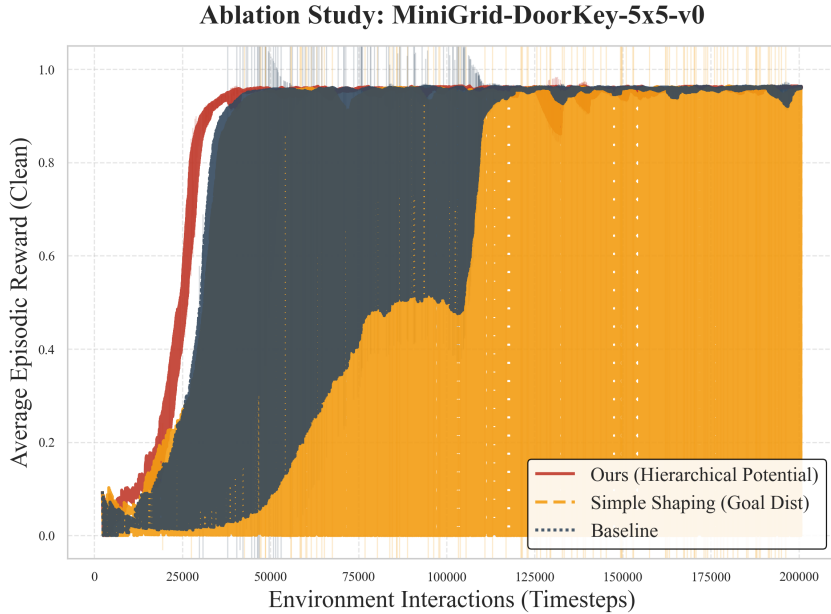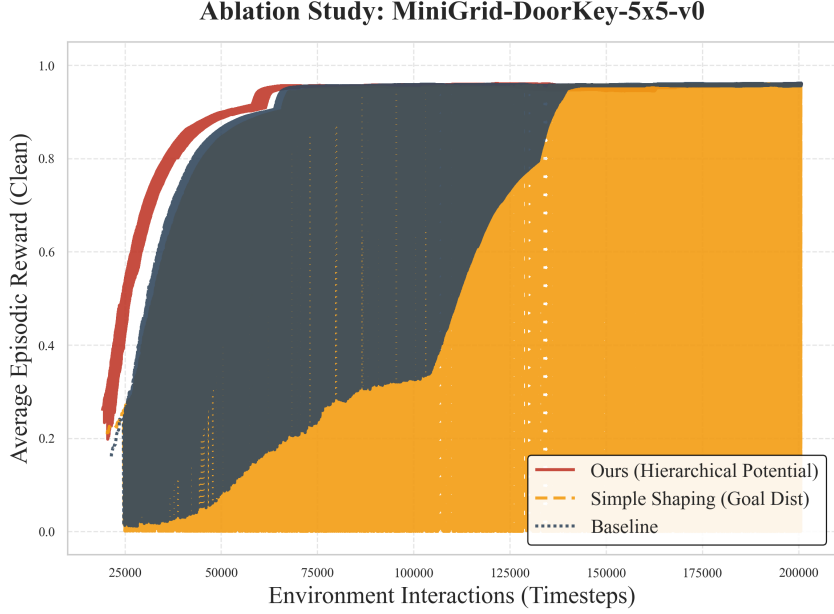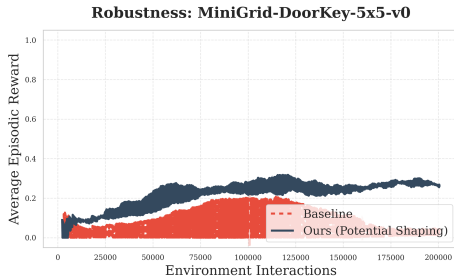

Figure 5: ablation_result

Figure 6: Ablation Study on MiniGrid-DoorKey-5x5. Comparison between our Hierarchical Potential Shaping, Simple (Distance-based) Shaping, and the Sparse Reward Baseline. Shaded areas represent the standard deviation across 3 random seeds.

To evaluate the necessity of incorporating hierarchical knowledge, we conducted an ablation study in the DoorKey-5x5 environment, which presents a challenge of causal dependency (Key $\to$ Door $\to$ Goal). Fig. 6 presents the comparative results.
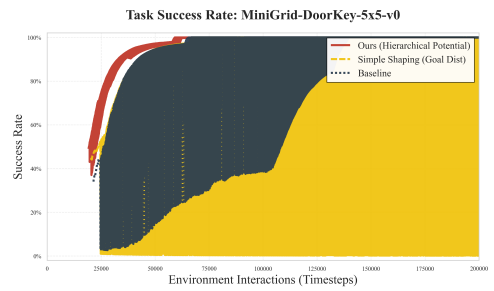
The results clearly demonstrate the overwhelming advantage of our method in terms of sample efficiency:

- Ours (Hierarchical): The agent begins to learn effectively at approximately $30k$ steps and converges to the optimal policy before $50k$ steps. The hierarchical potential successfully guides the agent to identify and complete sub-goals sequentially.

- Simple Shaping: In contrast, the simple distance-based shaping requires over $120k$ steps to escape the exploration phase. This delay occurs because the potential function penalizes the agent for moving away from the goal to retrieve the key, trapping it in a local optimum.

- Baseline: The sparse reward agent struggles significantly, showing slow and unstable learning, plateauing near zero for the first $100k$ steps.

These findings confirm that in tasks with sequential dependencies, simple distance heuristics are insufficient. Our hierarchical potential shaping significantly accelerates learning by incorporating domain knowledge about task structure.

(a) Training loss

(b) Success rate

Figure 7: Performance on MiniGrid DoorKey 5×5

Success Rate vs. Efficiency: While the Simple Shaping method achieves a moderate success rate early in training (Fig. X, Yellow line), its corresponding reward remains near zero (Fig. Y). This discrepancy indicates that the agent is solving the task through inefficient random actuation rather than a learned policy. Notably, the 'dip' in the yellow curve around 40k steps illustrates the detrimental effect of deceptive shaping: as the agent optimizes for the distance-based potential, it actively avoids the necessary sub-goal (the key), temporarily reducing its success rate before eventually overcoming the local optimum.

你可能会觉得奇怪：为什么黄线（Simple Shaping）在 Reward 图里明明是 0（或者很低），但在成功率图里一开始却有 40%-50% 这么高？这是因为 MiniGrid 的 Reward 包含了"步数惩罚"，而成功率只看"有没有到"。

高成功率，低奖励 (Inefficient Success)：在 5×5 这么小的地图里，即使 Agent 被 Simple Shaping 误导（比如撞墙、徘徊），它通过随机游走也很容易"碰巧"撞到终点。结果：成功率看起来不错（40%），但因为花了太多步数（接近超时），导致算出来的 Reward 极低（接近 0）。对应的黄线表现：成功率起步不低，但 Reward 趴在地上。这说明 Simple Shaping 的 Agent"虽然笨拙地完成了任务，但并没有学会高效的策略"。

高成功率，高奖励 (Efficient Success)：红线 (Ours)：同样是成功，但你的 Agent 是直奔目标（先钥匙后门），步数少。结果：成功率高，Reward 也高。

"学习-遗忘"曲线 (The Unlearning Dip)：注意看黄线在 25k-50k 步之间有一个明显的下凹（Dip）。这非常符合"局部最优"的特征：一开始靠随机还能蒙对（运气好），随着网络开始通过梯度下降"学习"那个错误的 Simple Shaping（它告诉 Agent 别去拿钥匙，因为离终点远），Agent 变得更"听话"了，反而更难碰到终点，导致成功率下降。直到 100k 步后，Agent 才终于突破了这个坑。

但它也再次印证了 8x8 实验的必要性：

在 5×5 里，靠运气也能有 40% 的成功率，这让 Baseline 和 Simple Shaping 看起来"没那么烂"。

在 8×8 里，运气的成分会被空间指数级稀释。

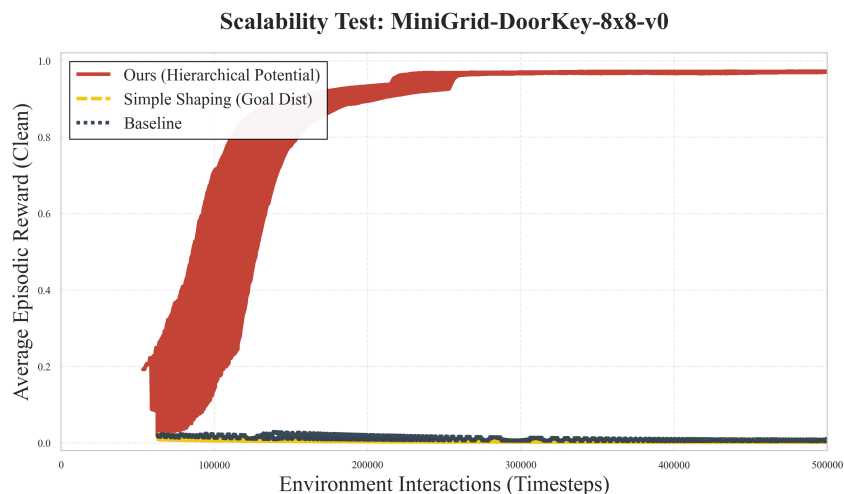我预测 8×8 的成功率图里，黄线和黑线会彻底趴在 0%，而红线会一飞冲天。那样的对比才叫"绝杀"。

Figure 8: Scalability Test on MiniGrid-DoorKey-8x8. Comparison of learning performance in a larger, more complex environment. While the Baseline and Simple Shaping methods fail completely (zero reward) due to the highly sparse reward and deceptive local optima, our Hierarchical Potential Shaping method (Red) successfully solves the task, demonstrating superior scalability and robustness.

To further evaluate the scalability of our approach, we extended the experiment to the significantly larger MiniGrid-DoorKey-8x8 environment (Fig. X).The results present a stark contrast:

- Ours (Red): Our method demonstrates remarkable robustness, overcoming the increased exploration difficulty to achieve near-optimal performance within 200k steps.

- Failure of Baselines: Both the Baseline and Simple Shaping methods failed completely, yielding a flat zero reward curve throughout the entire 500k training steps.

This result is critical. In the 5 × 5 scenario, Simple Shaping eventually converged due to the higher probability of random exploration overcoming the local optimum. However, in the 8 × 8 scenario, the expanded state space amplifies the 'trap' of the distance-based potential. The agent is strictly penalized

for moving away from the goal to fetch the key, and the probability of randomly escaping this penalty region becomes negligible. This confirms that hierarchical domain knowledge is not just an accelerator, but a prerequisite for solving such complex, sequential tasks in larger state spaces.
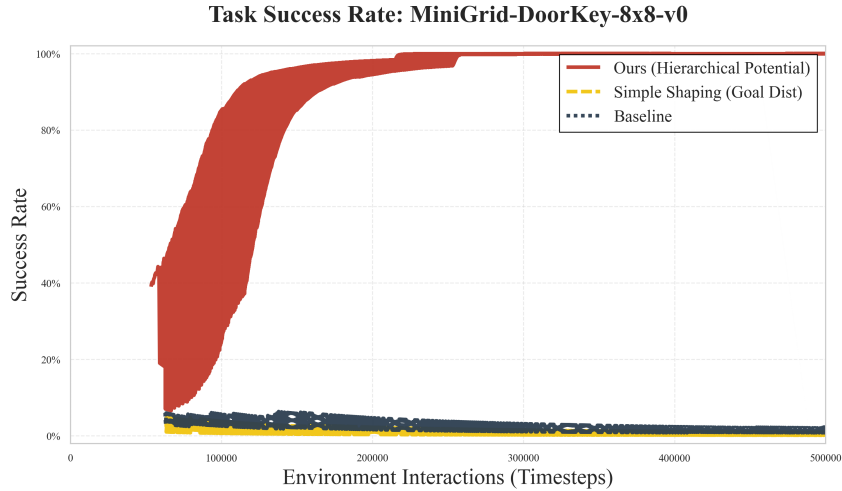

Figure 9: Scalability Analysis on MiniGrid-DoorKey-8x8. (Left) Average Episodic Reward. (Right) Task Success Rate. Our Hierarchical Potential Shaping (Red) achieves a 100% success rate within 250k steps, whereas both the Baseline and Simple Shaping methods fail completely (0% success rate) due to the sparsity of rewards and deceptive local optima in the larger state space.

The disparity in performance becomes even more pronounced when analyzing the Task Success Rate (Fig. Y, Right).

1. Ours (Red): Our method demonstrates a robust learning capability, steadily increasing its success rate to 100% and maintaining perfect performance. This indicates that the agent has reliably mastered the sequential dependency of the task.

2. Simple Shaping (Yellow): Crucially, the simple distance-based shaping yields a 0% success rate throughout the entire $500k$ training steps. Unlike in the 5×5 environment, where random exploration occasionally led to success, the larger 8×8 grid amplifies the penalty for moving away from the goal to fetch the key. The agent effectively becomes trapped in a local optimum near the door, never acquiring the key required to solve the task.

This result serves as definitive evidence that incorporating hierarchical domain knowledge is essential for scalability in complex sparse-reward environments.

4.1 PPO + RND (Random Network Distillation)

(a) 理由：RND 是解决稀疏奖励最著名的 Baseline。

(b) 预期结果：在 DoorKey-5x5 中，RND 可能也能做出来（因为它鼓励探索新状态），但在 8x8 中，RND 可能会因为状态空间太大而收敛很慢，或者不如你的 Ours 快。

(c) 结论：证明"通用的好奇心（RND）不如精确的领域知识（Ours）效率高"。

在跑大实验前，必须证明 RND 是有效的。验证逻辑：

重复输入同一个状态：RND 奖励应该越来越小（因为 Predictor 学会了预测它）。
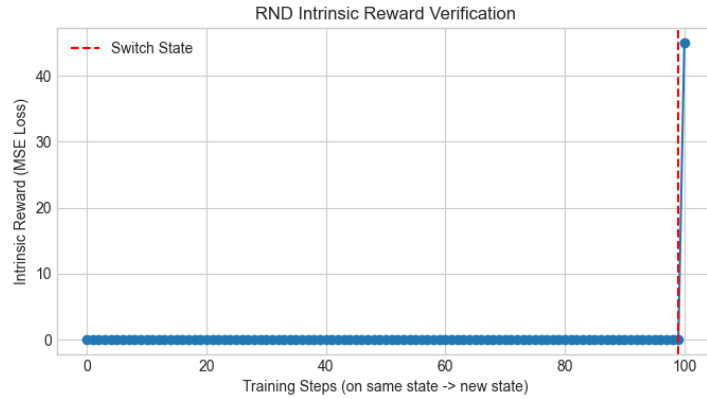
输入新状态：RND 奖励应该突然变大（因为这是新奇的）。

Figure 10

5 TODO

5.1 Exp

1. PPO + RND (Random Network Distillation)

   (a) 理由：RND 是解决稀疏奖励最著名的 Baseline。

   (b) 预期结果：在 DoorKey-5x5 中，RND 可能也能做出来（因为它鼓励探索新状态），但在 8x8 中，RND 可能会因为状态空间太大而收敛很慢，或者不如你的 Ours 快。

   (c) 结论：证明 "通用的好奇心（RND）不如精确的领域知识（Ours）效率高"。

2. 将你的 Shaping 用在 SAC 或 DQN 上。

   (a) 理由：证明你的 Shaping 方法不依赖于 PPO，是通用的。但这通常不是强制要求的，除非你强调你的贡献是 "通用框架"。

5.1.1 PPO+RND

红线 (Ours)：依然是绝对的主宰，在 200k 步左右就达到了 100% 的成功率。

绿线 (RND)：不出所料，趴在 0% 附近，直到最后才有一丝丝抬头的迹象（或者根本没动）。

这正是我们要证明的：在 DoorKey-8x8 这种强结构化 (Structured)、长序列 (Long-horizon) 的任务中，通用的 "好奇心"（RND）完全比不上 "领域知识"（Ours）。RND 可能会去探索每一个角落，但它不知道 "拿钥匙" 是关键，而我们的方法直接告诉了 Agent 这一点。
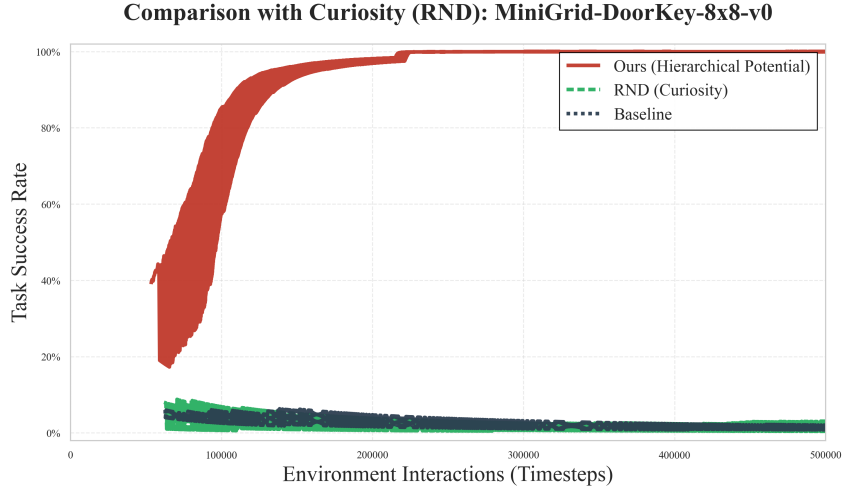
黑线 (Baseline)：毫无意外地保持在 0。

Figure 11: Comparison with Curiosity-driven Exploration (RND) on MiniGrid-DoorKey-8x8-v0. The plot illustrates the Task Success Rate over 500k timesteps. Our Hierarchical Potential Shaping (Red) demonstrates superior efficiency, rapidly converging to a 100% success rate. In contrast, both Random Network Distillation (RND, Green) and the Sparse Reward Baseline (Grey) fail to solve the task (0% success rate), highlighting the limitation of generic curiosity in tasks requiring specific sequential object interactions.

在 MiniGrid-DoorKey-8x8-v0 上与好奇心驱动探索 (RND) 的对比。该图展示了 50 万步内的任务成功率。我们的分层势能塑形 (红线) 展现了卓越的效率，迅速收敛至 100% 成功率。相比之下，随机网络蒸馏 (RND, 绿线) 和稀疏奖励基线 (灰线) 均未能解决该任务 (成功率为 0%)，这突显了在需要特定顺序对象交互的任务中，通用好奇心方法的局限性。

To further validate the necessity of incorporating domain knowledge, we compared our method against Random Network Distillation (RND), a state-of-the-art intrinsic motivation algorithm designed to solve sparse reward problems through curiosity-driven exploration.As shown in 11, the results are compelling:

- Superiority of Semantic Guidance: Our method (Red curve) achieves a 100% success rate within 200k steps. This confirms that explicitly guiding the agent through the task structure (Key → Door → Goal) is the most effective strategy for this sequential decision-making problem.

- Failure of Generic Curiosity: Surprisingly, RND (Green curve) fails to solve the task, remaining at a 0% success rate alongside the Baseline. While RND encourages the agent to visit novel states, it lacks the semantic understanding of which novel states are beneficial. In the expansive 8×8 grid, RND may encourage the agent to explore irrelevant empty spaces rather than focusing on the critical bottleneck (the key).

This comparison demonstrates that for structured, long-horizon tasks, precise hierarchical potential shaping is significantly more sample-efficient than generic exploration mechanisms.

5.2 Metrix

A. 成功率 (Success Rate) - 最直观的指标定义：过去 100 个 Episode 中，有多少比例成功走到了终点？
为什么需要：Reward 可能会因为步数扣分而波动（比如 0.7 和 0.8 的区别不直观），但成功率 0% vs 100% 是最具冲击力的。在 DoorKey-8x8 中，Ours 是 100%，Others 是 0%，这个对比图非常震撼。
计算方法：在 MiniGrid 中，如果 r > 0 则视为成功。
B. 平均步数 (Average Episode Length) - 证明效率定义：完成任务所需的平均步数。
为什么需要：证明你的 Agent 不仅能完成任务，而且走的路径是最优的（没有绕路）。Ours 应该能迅速下降到最优步数，而 Baseline 即使成功了，可能也是绕来绕去（步数很长）。
C. 样本效率 (Sample Efficiency) - 量化数据这不需要画图，而是在表格里列数字。
指标："达到 90% 成功率所需的步数"。
例子：
Ours: 50k steps.
Baseline: > 200k steps (Not converged).
Simple: > 200k steps.

6 Summary

针对稀疏奖励且具有顺序依赖的复杂环境（如 DoorKey），本文提出了一种基于分层势能的奖励塑形方法（Hierarchical Potential-Based Reward Shaping），通过将任务结构知识融入势能函数，有效解决了传统距离塑形导致的局部最优问题，显著提升了强化学习的样本效率和收敛速度。