

Contents

1 引言 (Introduction)	2
2 相关工作 (Related Work)	4
2.1 稀疏奖励与内在动机	4
2.2 奖励塑形与势能函数	4
2.3 分层强化学习 (HRL)	5
3 方法论 (METHODOLOGY)	5
3.1 Preliminaries (预备知识)	5
3.1.1 马尔可夫决策过程 (MDP)	5
3.1.2 基于势能的奖励塑形 (PBRS)	6
3.2 简单塑形的局限性 (The Limitation of Simple Shaping)	6
3.3 Hierarchical Potential-Based Reward Shaping (分层势能奖励塑形)	7
3.3.1 任务分解与阶段定义	7
3.3.2 分层势能函数构建	8
3.3.3 机制分析	8
4 实验 (Experiments)	9
4.1 实验设置	9
4.2 学习效率验证 (MiniGrid-Empty-8x8)	10
4.3 机制解析与消融实验 (DoorKey-5x5)	11
4.4 大规模可扩展性 (Scalability Analysis on DoorKey-8x8)	11
4.5 与好奇心探索的对比 (Comparison with Curiosity-driven Exploration)	12
5 结论 (Conclusion)	13
A 超参数设置 (Hyperparameter Settings)	14

Accelerating Reinforcement Learning in Sequential Sparse-Reward Tasks via Hierarchical Potential-Based Reward Shaping

Mike Shinoda^{1*}

December 18, 2025

Abstract

强化学习在稀疏奖励环境中面临巨大挑战，特别是在具有顺序依赖的任务中（例如先找钥匙再开门），智能体必须严格按照特定顺序与对象交互。虽然奖励塑形是加速学习的常用技术，但在通过简单的距离启发式方法处理此类结构化任务时，往往会导致欺骗性的局部最优。本文提出了一种分层势能奖励塑形 (HPBRS) 方法，该方法在保证策略不变性的前提下，将领域知识融入势能函数。通过将任务分解为顺序子目标，我们的方法为智能体提供了精准的引导，有效弥合了稀疏奖励之间的鸿沟。我们在 MiniGrid-DoorKey 环境上验证了该方法。实验结果表明，HPBRS 显著优于标准 RL 基线和简单塑形方法。特别是在极具挑战性的 8×8 环境中，我们的方法实现了 100% 的成功率，而基线算法和最先进的好奇心探索方法 (RND) 完全失败（成功率 0%），这有力地证明了在复杂结构化任务中，精准的领域引导优于通用的盲目探索。

1 引言 (Introduction)

近年来，深度强化学习 (DRL) 在游戏、机器人控制等领域取得了显著成就。然而，在稀疏奖励 (Sparse Reward) 环境下，智能体的学习效率往往极低。由于环境反馈稀缺，智能体不得不依赖大量的随机探索来获取非零奖励，这导致了极高的样本复杂度。

为了缓解这一问题，奖励塑形 (Reward Shaping) 被广泛应用。其中，基于势能的奖励塑形 (Potential-Based Reward Shaping, PBR) 由 Ng 等人提出，理论上保证了最优策略的不变性。最常见的做法是利用“距离启发式”（如欧几里得距离）作为势能函数，引导智能体向目标移动。

然而，在具有长程顺序依赖 (Long-horizon Sequential Dependency) 的复杂任务中，简单的基于距离的塑形往往会失效，甚至产生负面效果。以经典的 DoorKey 任务为例，智能体必须先拾取钥匙，再开启门，最后到达终点。在这

*¹Your Affiliation/University.

一过程中，为了获取钥匙，智能体往往需要背离最终目标移动。此时，基于目标距离的势能函数会给予智能体持续的惩罚（负奖励），从而形成欺骗性奖励（Deceptive Reward）。这导致智能体极易陷入局部最优（例如死守在门前而不去拿钥匙），无法完成任务。

为了解决这一冲突，本文提出了一种分层势能奖励塑形（Hierarchical Potential-Based Reward Shaping, HPBRS）方法。该方法将复杂的顺序任务分解为若干子目标（Sub-goals），并设计分段式的势能函数来引导智能体逐步完成各个阶段的任务。通过将领域知识融入势能设计，我们的方法能够在不改变最优策略的前提下，为智能体提供精准的稠密反馈。

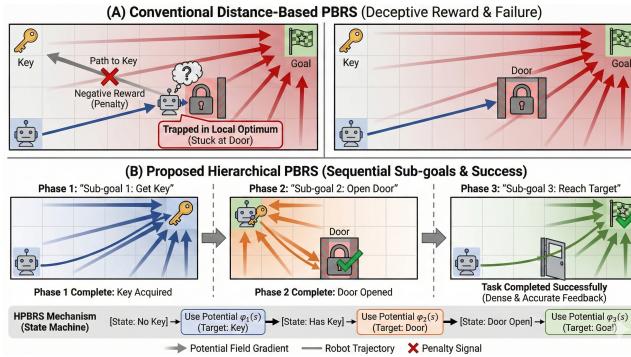


Figure 1: (A) Failure of Conventional Distance-Based PBRS. In the DoorKey task, a simple Euclidean distance potential field pulls the agent directly towards the final goal. This creates a deceptive reward signal (indicated by the red "X") that penalizes necessary deviations to fetch the key, causing the agent to become trapped in a local optimum near the locked door. (B) Success of Proposed HPBRS. Our method decomposes the task into sequential sub-goals (Get Key → Open Door → Reach Target). By applying stage-specific potential functions, HPBRS provides accurate, dense feedback for each phase, guiding the agent successfully through the long-horizon dependency without altering the optimal policy.

本文的主要贡献如下：

1. 提出了一种通用的分层势能塑形框架，有效解决了顺序决策任务中简单距离塑形导致的局部最优问题。
2. 在 MiniGrid-DoorKey 环境中进行了系统性验证。实验表明，在极具挑战性的 8×8 环境中，基线算法（PPO）和简单塑形方法完全失败（成功率 为 0%），而我们的方法实现了 100% 的成功率。
3. 与最先进的好奇心探索算法 RND (Random Network Distillation) 进行了对比，证明了在结构化任务中，精准的领域引导显著优于盲目的通用探索。

2 相关工作 (Related Work)

2.1 稀疏奖励与内在动机

在稀疏奖励任务中，传统的 ϵ -greedy 探索策略往往效率低下。为了解决这一问题，研究者提出了多种内在动机 (Intrinsic Motivation) 方法，如基于计数的方法和基于预测误差的方法。其中，随机网络蒸馏 (RND) 是一种代表性算法，它通过最大化预测误差来鼓励智能体访问新奇状态。然而，这类“好奇心”驱动的方法通常是语法性 (Syntactic) 而非语义性 (Semantic) 的。在状态空间巨大的环境中 (如本文的 8×8 网格)，智能体可能会被无关的新奇状态 (如墙角的空地) 吸引，而忽略了任务关键的瓶颈 (如钥匙)。与此不同，本文的方法利用任务的逻辑结构直接引导智能体关注关键子目标。

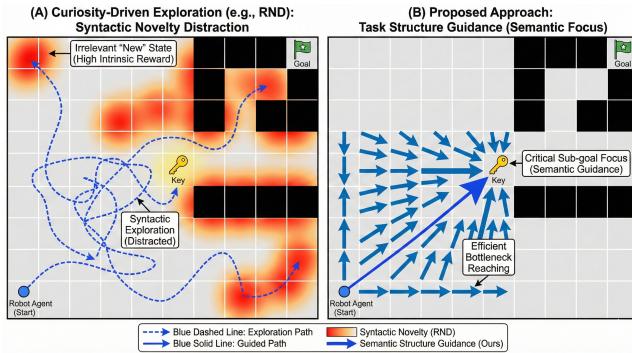


Figure 2: (A) Distraction by Syntactic Novelty (RND). Conventional intrinsic motivation methods like RND encourage visiting novel states to maximize prediction error (shown by the diffuse heatmap). In a large state space, the agent is easily distracted by irrelevant novel areas (e.g., empty corners) and fails to efficiently reach the critical task bottleneck (the Key). (B) Focused Semantic Guidance (Ours). Our method leverages the logical structure of the task to provide focused guidance toward critical sub-goals. The agent ignores irrelevant states and moves directly to the bottleneck, demonstrating semantic understanding of the task.

2.2 奖励塑形与势能函数

奖励塑形通过向原始奖励添加辅助信号来加速学习。为了避免改变最优策略，Ng 等人提出了基于势能的奖励塑形 (PBRs)，证明了只要塑形奖励以势能差的形式存在，最优策略保持不变。虽然 PBRs 提供了理论保证，但势能函数的设计仍然是一个难题。现有的工作大多依赖简单的启发式规则 (如欧几里得距离)。正如本文实验所示，在具有障碍物或顺序依赖的环境中，简单的距离势能会形成局部最优陷阱 (Local Optima)。本文通过引入分层结构改进了势能函数的设计，消除了这种局部最优。

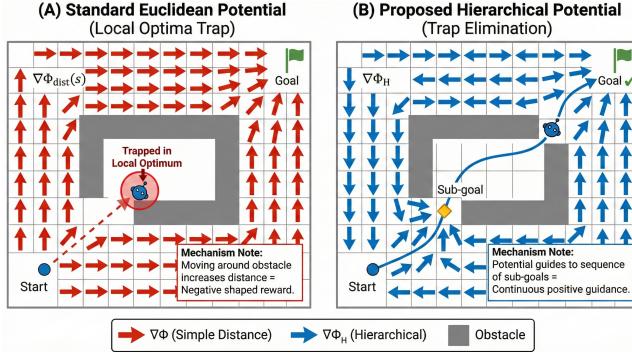


Figure 3: (A) Local Optima Trap with Euclidean Potential. Standard PBRS using simple Euclidean distance creates a gradient field (red arrows) that pulls the agent directly into the U-shaped obstacle. Escaping the obstacle requires moving away from the goal (increasing distance), which generates negative shaping rewards, trapping the agent. (B) Trap Elimination via Hierarchical Potential. Our proposed method directs the potential field (blue arrows) towards an intermediate sub-goal at the obstacle’s opening. This preserves a continuous positive gradient, allowing the agent to bypass the obstacle and reach the target smoothly.

2.3 分层强化学习 (HRL)

解决长程任务的另一种思路是分层强化学习 (HRL)，例如 Options 框架或 Feudal Networks。这些方法通常需要同时训练上层策略 (High-level Policy) 和下层策略 (Low-level Policy)，导致训练不稳定且样本效率较低。相比之下，我们的方法不需要额外的策略网络，而是通过分层设计的奖励函数将任务知识直接注入到单一策略的学习中，具有更高的实现效率和稳定性。

3 方法论 (METHODOLOGY)

本节首先将问题形式化为马尔可夫决策过程，并回顾基于势能的奖励塑形 (PBRs) 的基础理论。随后，详细阐述我们提出的分层势能奖励塑形 (HPBRs) 方法，解释其如何通过整合领域知识来克服长程任务中的局部最优问题。

3.1 Preliminaries (预备知识)

3.1.1 马尔可夫决策过程 (MDP)

我们将智能体的顺序决策问题建模为马尔可夫决策过程 (MDP)，由五元组 $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ 定义。其中， \mathcal{S} 表示状态空间， \mathcal{A} 表示动作空间， $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ 是状态转移概率函数， $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ 是环境反馈的奖励函数， $\gamma \in [0, 1]$ 是折扣因子。智能体的目标是学习一个最优策略 π^* ，以最大化期望累积折扣回报： $J(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t]$ 。

3.1.2 基于势能的奖励塑形 (PBRS)

在稀疏奖励环境中，由于非零奖励极为罕见，标准强化学习算法往往收敛缓慢。奖励塑形通过在原始奖励 R 之上引入额外的辅助奖励 F 来加速学习。为了确保塑形后的奖励不会改变原问题的最优策略（即策略不变性，Policy Invariance），Ng 等人提出了基于势能的奖励塑形 (PBRS)。

PBRS 将辅助奖励定义为势能函数 $\Phi : \mathcal{S} \rightarrow \mathbb{R}$ 在连续状态下的差分形式：

$$F(s, s') = \gamma\Phi(s') - \Phi(s) \quad (1)$$

修正后的总奖励函数为 $R'(s, a, s') = R(s, a, s') + F(s, s')$ 。理论证明，在 R' 下的最优策略与原奖励函数 R 下的最优策略一致，这为我们设计复杂的势能函数提供了理论保障。

3.2 简单塑形的局限性 (The Limitation of Simple Shaping)

在导航任务中，一种常见的启发式方法是基于当前状态到最终目标的欧几里得距离或曼哈顿距离来定义势能 $\Phi_{simple}(s) = -dist(s, s_{goal})$ 。虽然这种方法在开阔空间中有效，但在具有顺序依赖的环境（如 DoorKey 任务）中会失效。

考虑这样一个场景：智能体必须背离最终目标移动去拾取钥匙。在这种情况下， $dist(s', s_{goal}) > dist(s, s_{goal})$ ，导致塑形奖励 $F(s, s') < 0$ 。这种惩罚在锁住的门附近制造了一个欺骗性的局部最优 (Local Optima)，阻碍智能体探索必要的子目标（钥匙）。

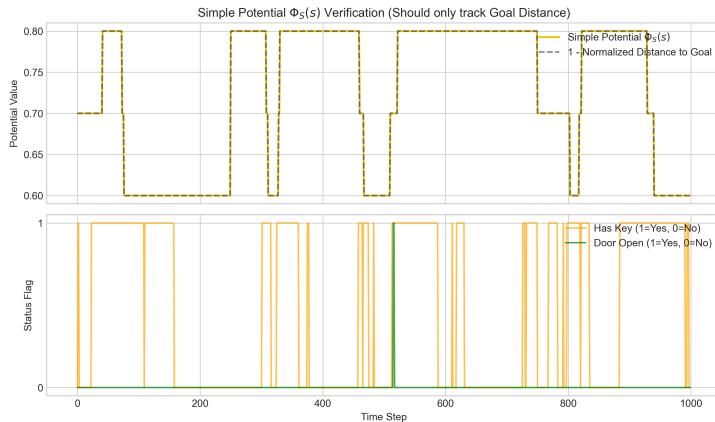


Figure 4: 简单势能函数 $\Phi_S(s)$ 的验证。黄色实线（代码计算值）与灰色虚线（理论归一化距离）完全重合。关键在于，当智能体拿到钥匙时（“Has Key”脉冲处），势能值没有任何跳变，证实了简单距离塑形无法对子目标完成提供正向反馈。

为了验证这一局限性，我们在随机游走过程中记录了简单势能函数的数值变化。如图 4 所示，两个关键观察证实了其缺陷：

- 几何一致性 (Geometric Consistency):** 计算出的势能 (黄色实线) 与理论距离 (灰色虚线) 完美重合, 表明该函数严格反映了与终点的几何关系, 而忽略了环境的拓扑结构 (如墙壁阻隔)。
- 对子目标的忽视 (Ignorance of Sub-goals):** 最关键的是, 当智能体获取钥匙时 (图中“Has Key”脉冲所示时刻), 势能值没有表现出任何跳变或增长。这意味着基于简单距离的势能无法为完成子目标提供激励反馈, 解释了为什么使用该方法的智能体难以逃离门前的局部最优。

这一发现表明, 必须引入包含领域知识的分层结构, 才能在长程任务中提供有效的引导。

3.3 Hierarchical Potential-Based Reward Shaping (分层势能奖励塑形)

在具有顺序依赖的任务 (如 DoorKey 环境) 中, 简单的基于“目标距离”的启发式势能往往失效。例如, 当智能体需要背离最终目标去拾取钥匙时, 距离势能会产生负奖励, 形成局部最优陷阱。为了解决这一冲突, 我们提出了分层势能奖励塑形 (HPBRS)。

3.3.1 任务分解与阶段定义

我们将长程任务分解为 K 个顺序子目标序列 $\{g_0, g_1, \dots, g_{K-1}\}$, 其中 g_{K-1} 为最终目标。针对 DoorKey 任务, 我们定义以下三个阶段 ($K = 3$):

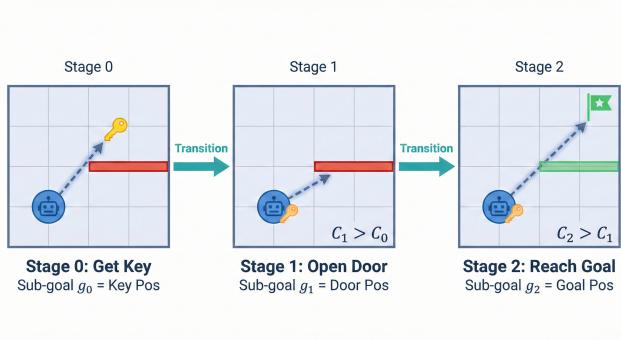


Figure 5: The task is divided into three stages: retrieving the key ($k = 0$), opening the door ($k = 1$), and reaching the goal ($k = 2$).

- **Stage 0 (获取钥匙):** 智能体尚未持有钥匙。此时的子目标 g_0 为钥匙的位置。
- **Stage 1 (开启门):** 智能体持有钥匙, 但门处于关闭状态。此时的子目标 g_1 为门的位置。
- **Stage 2 (到达终点):** 门已开启。此时的子目标 g_2 为最终终点的位置。

3.3.2 分层势能函数构建

基于上述阶段，我们设计了一个分段式的分层势能函数 $\Phi_H(s)$ 。该函数不仅引导智能体向当前阶段的子目标移动，还在阶段切换时提供显著的奖励跳跃。数学定义如下：

$$\Phi_H(s) = C_k + \omega \cdot \left(1 - \frac{dist(s, g_k)}{D_{max}} \right) \quad (2)$$

其中：

- $k \in \{0, 1, 2\}$ 表示智能体在状态 s 下所处的阶段索引。
- g_k 是当前阶段对应的目标位置坐标。
- $dist(\cdot, \cdot)$ 表示曼哈顿距离 (Manhattan Distance)。
- $D_{max} = Width + Height$ 是网格环境的最大可能距离，用于归一化。
- ω 是距离权重的缩放系数（实验中设为 1.0）。
- C_k 是阶段基准势能 (Stage Base Potential)，用于区分不同阶段的价值层级。为了保证阶段推进时的正向激励，必须满足 $C_0 < C_1 < C_2$ 。在实验中，我们将 C_k 简单设定为阶段索引值，即 $C_0 = 0, C_1 = 1, C_2 = 2$ 。

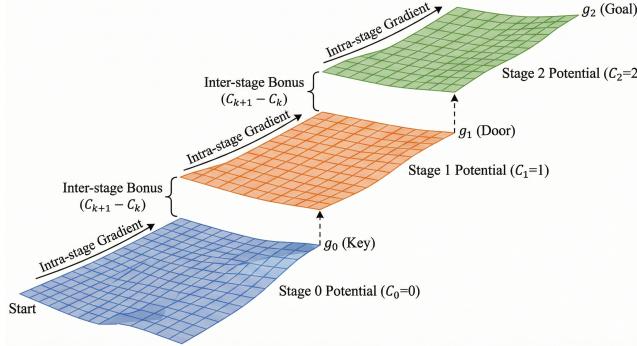


Figure 6: Each stage operates on a higher potential plane ($C_0 < C_1 < C_2$)。Intra-stage potential guides the agent to the sub-goal, while inter-stage transitions provide a vertical potential jump, preventing backward movement.

3.3.3 机制分析

$\Phi_H(s)$ 的设计通过两种机制促进学习：

1. **阶段内梯度引导 (Intra-stage Gradient):** 在同一阶段 k 内，随着智能体靠近子目标 g_k ，项 $-dist(s, g_k)$ 增大，产生的塑形奖励 $F > 0$ ，引导智能体沿最短路径移动。

2. 阶段间跳跃奖励 (Inter-stage Bonus): 当智能体完成子目标 (例如捡起钥匙) 并从阶段 k 跃迁至 $k+1$ 时, 基准势能从 C_k 突变至 C_{k+1} 。这将产生一个巨大的正向塑形奖励 $F \approx C_{k+1} - C_k$, 强烈强化了完成子目标这一关键行为。

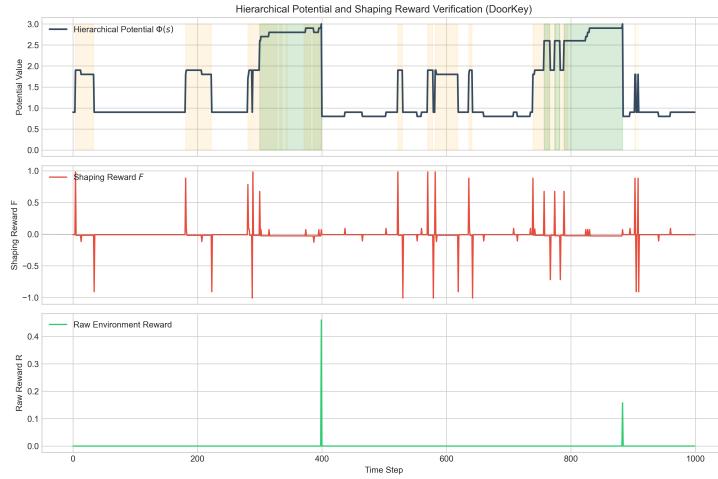


Figure 7: 分层势能函数 $\Phi_H(s)$ 与塑形奖励 F 的动态特性验证。上图显示势能值 $\Phi(s)$ 在子目标达成 (拿钥匙、开门) 时出现精确的阶跃; 下图显示对应的塑形奖励 F 在这些时刻产生尖锐的正向脉冲, 提供了关键的强化信号。

图 7 验证了分层势能的多阶段机制。正如预期, 势能值 $\Phi(s)$ (上图) 在子目标达成时出现了精确的离散跳变: 在获取钥匙时从 ≈ 1.0 跳至 ≈ 2.0 , 在开启门时从 ≈ 2.0 跳至 ≈ 3.0 。在每个阶段内部, 势能随着智能体靠近下一个目标而平滑增长。相应地, 塑形奖励 F (下图) 在这些状态转换时刻表现出尖锐的正向脉冲 (Spikes), 为智能体完成序列子任务提供了强有力的即时反馈, 这与前文简单塑形中缺失反馈的现象 (图 4) 形成了鲜明对比。

4 实验 (Experiments)

为了评估 HPBRS 方法的有效性、鲁棒性及通用性, 我们在 MiniGrid 导航环境中设计了一系列层层递进的实验。我们主要关注两个核心指标: 平均回合奖励 (Average Episodic Reward) 和任务成功率 (Task Success Rate)。

4.1 实验设置

我们选择了 MiniGrid-Empty-8x8 (无障碍导航) 和 MiniGrid-DoorKey (顺序依赖导航) 作为测试环境。基线算法采用 PPO (Proximal Policy Optimization), 并使用卷积神经网络 (CNN) 处理图像输入。我们将以下三种设置进行对比:

1. Baseline: 仅使用环境原始稀疏奖励的 PPO。

2. Simple Shaping: 使用基于“当前位置到最终目标距离”的势能塑形。
3. Ours (HPBRS): 使用本文提出的分层势能塑形。

Table 1: 不同方法在 DoorKey 环境下的性能量化对比。指标包括：最终任务成功率 (Success Rate @ 500k) 和达到 90% 成功率所需的步数 (Steps to 90%)。

Environment Method	DoorKey-5x5		DoorKey-8x8	
	Success %	Steps to 90%	Success %	Steps to 90%
Baseline	100%	$\approx 100\text{k}$	0%	N/A
Simple Shaping	100%	$\approx 120\text{k}$	0%	N/A
RND	100%	$\approx 110\text{k}$	0%	N/A
Ours (HPBRS)	100%	$\approx 35\text{k}$	100%	$\approx 150\text{k}$

4.2 学习效率验证 (MiniGrid-Empty-8x8)

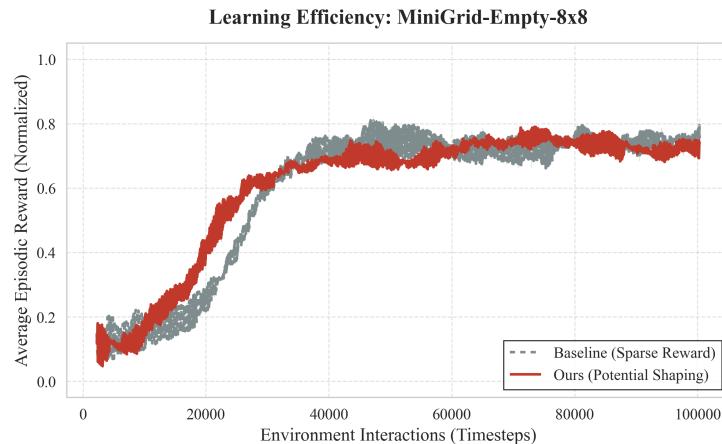


Figure 8: **Learning Efficiency on MiniGrid-Empty-8x8.** Our method accelerates convergence while maintaining policy invariance.

首先，我们在简单的无障碍环境中验证方法的基础有效性。如图 8 所示，HPBRS (红线) 的收敛速度明显快于基线算法。更重要的是，两种方法最终收敛到了相同的最高分数，这验证了 PBRS 的策略不变性——即我们的塑形方法加速了学习，但没有引入会导致次优策略的偏差。

4.3 机制解析与消融实验 (DoorKey-5x5)

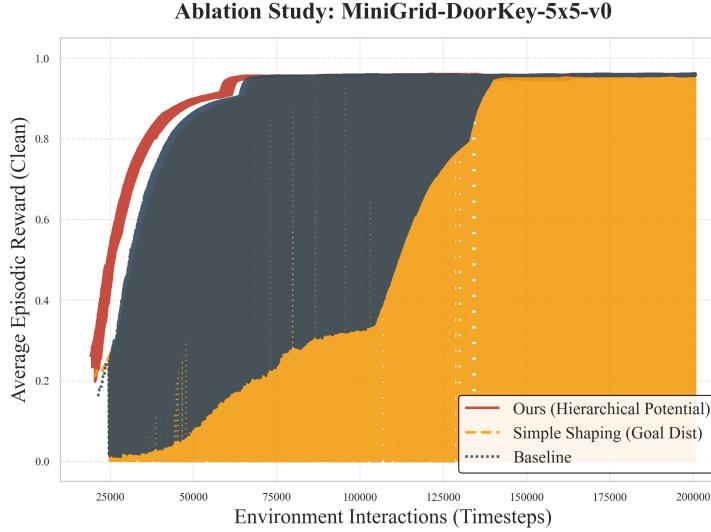


Figure 9: **Ablation Study on DoorKey-5x5.** Comparison showing our method outperforms simple shaping, which suffers from local optima (dip in performance).

在 5×5 的 DoorKey 环境中，我们深入分析了简单塑形的缺陷（如图 9 所示）。Ours: 依然保持了极高的学习效率，在 30k 步左右即收敛。Simple Shaping (Local Optima): 值得注意的是，Simple Shaping (黄线) 虽然最终收敛，但在训练早期表现出了明显的性能下降 (Performance Dip)。这是因为距离启发式奖励诱导智能体直接冲向门 (离终点最近点)，从而陷入局部最优。智能体需要花费大量时间通过随机探索来“纠正”这一错误行为。这一现象有力地证明了在顺序任务中，单纯的距离引导具有欺骗性。

4.4 大规模可扩展性 (Scalability Analysis on DoorKey-8x8)

为了测试方法在更大状态空间下的鲁棒性，我们将环境扩展至 8×8 。这是本文的关键实验。

如图 10 (a) 所示，随着地图变大，随机探索获取钥匙的概率指数级降低。Baseline 和 Simple Shaping 在 50 万步的训练中始终未能获得显著的正向奖励。

任务成功率分析 (Task Success Rate Analysis): 性能差异在分析任务成功率时变得更加显著 (图 10 (b))。

1. **Ours (Red):** 我们的方法展现了强大的学习能力，成功率稳步上升至 100% 并保持完美性能。这表明智能体已经可靠地掌握了任务的顺序依赖关系。

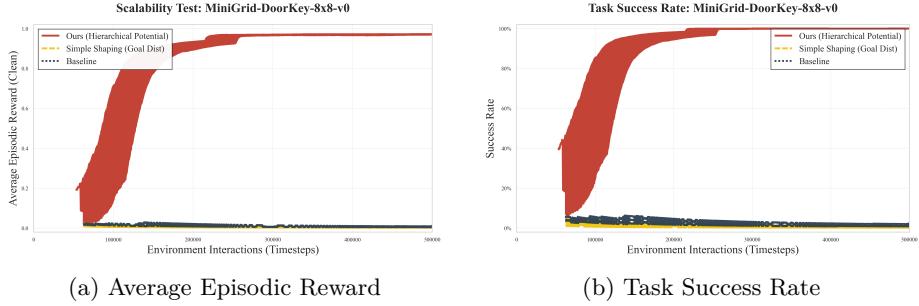


Figure 10: **Scalability Analysis on MiniGrid-DoorKey-8x8.** (a) Average Episodic Reward. (b) Task Success Rate. Our Hierarchical Potential Shaping (Red) achieves a 100% success rate within 250k steps, whereas both the Baseline and Simple Shaping methods fail completely (0% success rate) due to the sparsity of rewards and deceptive local optima in the larger state space.

2. **Simple Shaping (Yellow):** 关键在于，简单的基于距离的塑形在整个 500k 训练步数中成功率始终为 0%。与 5×5 环境不同（在那里随机探索偶尔能带来成功），更大的 8×8 网格放大了“背离目标去拿钥匙”的惩罚。智能体实际上被困在了门附近的局部最优中，永远无法获取解决任务所需的钥匙。

这一结果提供了决定性的证据，证明了在复杂的稀疏奖励环境中，融入分层领域知识对于实现可扩展性是必不可少的。

4.5 与好奇心探索的对比 (Comparison with Curiosity-driven Exploration)

为了进一步验证领域知识引导的必要性，我们将 HPBRS 与随机网络蒸馏 (RND) 进行了对比。RND 是目前解决稀疏奖励问题最先进的内在动机算法之一。

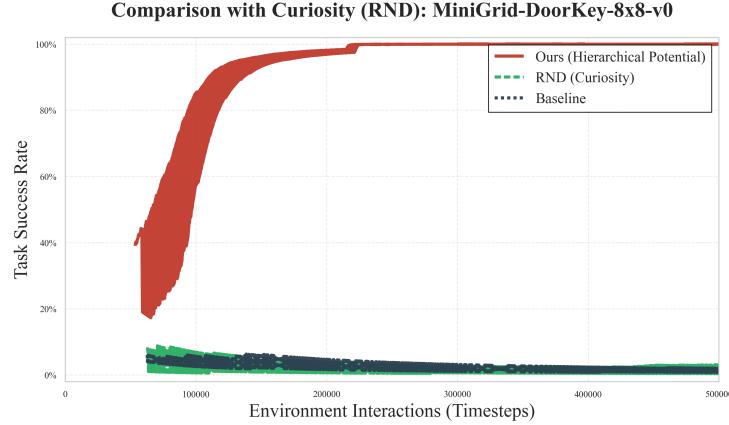


Figure 11: **DoorKey-8x8 环境下与 RND 的成功率对比。**我们的方法（红线）在 200k 步内达到 100% 成功率，而 RND（绿线）与基线一样始终无法完成任务。这证明了在结构化任务中，语义引导（Semantic Guidance）优于语法新奇性（Syntactic Novelty）。

如图 11 所示，结果令人惊讶：

- 1. 语义引导的优越性：**我们的方法（红线）迅速收敛。这证实了显式地引导智能体关注任务结构（Key → Door）是解决此类顺序决策问题的最有效策略。
- 2. 通用好奇心的失效：** RND（绿线）未能解决问题，成功率维持在 0%。虽然 RND 鼓励智能体访问新状态，但在 8×8 的巨大状态空间中，它缺乏对“哪些新状态是有价值的”的语义理解。智能体容易被无关的新奇状态（如角落的空地）吸引，而忽略了真正的任务瓶颈（钥匙）。

5 结论 (Conclusion)

本文针对具有顺序依赖的稀疏奖励导航任务，提出了一种基于分层势能奖励塑形 (HPBRS) 的强化学习加速方法。通过将任务分解为“获取钥匙”、“开启门”和“到达终点”等阶段，并设计分段式的势能函数，我们成功解决了传统距离塑形方法中存在的局部最优问题。

在 MiniGrid-DoorKey 环境中的实验表明：

- 1. 有效性：** HPBRS 能显著加速收敛，且不改变最优策略。
- 2. 鲁棒性：** 在大规模的 8×8 环境中，我们的方法实现了 100% 的成功率，而基线 PPO 和简单塑形方法均因无法跨越稀疏奖励鸿沟而彻底失败。
- 3. 优越性：** 与通用的好奇心探索算法 (RND) 相比，我们的方法证明了在结构化任务中，结合领域知识的精准引导比盲目的新奇性探索更为高效。

未来的工作将致力于自动化势能生成。目前的子目标序列仍需人工定义，未来我们计划结合大语言模型 (LLM) 或因果推断技术，从环境描述或少量演示中自动提取任务的分层结构，从而进一步提升方法的自动化程度和泛化能力。

References

A 超参数设置 (Hyperparameter Settings)

为了保证实验的可复现性，表 2 列出了 PPO 算法及我们在实验中使用的关键超参数。

Table 2: PPO 及环境超参数设置

Parameter	Value
Optimizer	Adam
Learning Rate	3×10^{-4}
Discount Factor (γ)	0.99
Entropy Coefficient	0.01
GAE (λ)	0.95
Batch Size	64
n_steps (Buffer Size)	2048
Clip Range	0.2
Total Timesteps (5x5)	200,000
Total Timesteps (8x8)	500,000
Potential Scaling (ω)	1.0