

階層的世界モデルの現状と課題：Hieros の限界と将来の展望

Current Status and Challenges in Hierarchical World Models: Identifying Limitations of Hieros and Future Prospects

三好 理輝 ^{*1} 劉 智優 ^{*2} 山田 達也 ^{*3}

Riki Miyoshi Jiu You Yamada Tatsuya

*¹ケンブリッジ大学

University of Cambridge

*²電気通信大学

University of Electro-Communications

*³大阪大学

The University of Osaka

ここに
アブストラクトを
5行程度
書き
ます

1. はじめに

2. 研究背景・目的

階層的強化学習 (HRL) は、探索において有益であると知られており [HRL benefit]、効率的な報酬割り当て、転移学習、解釈性などといった点で優位性があると考えられている [HRL survey]。

また、世界モデルは高いサンプル効率を実現し、エージェントが少ない環境との相互作用で学習することを可能にする。

一方で、階層性と世界モデルを組み合わせた研究は少ない。この 2 つの要素を組み合わせた数少ないモデルとして、Director[Director] や Hieros[Hieros] があり、ベースラインとして使用されている世界モデルを上回る性能を示している。一方で、解釈性やモデルをスケールさせる方向での研究はされておらず、階層的な世界モデルの理論および評価実験は知られていない。

そこで本研究では、階層的世界モデルの現状を理解し課題を特定するために、Hieros の性能評価および内部状態の理解を試みた。

3. 関連研究

3.1 世界モデル

世界モデルは、「説明・数式」。世界モデルは大きく、1. 状態空間モデル、2. 状態予測モデル、3. 観測予測モデル に分けることができる。代表的な世界モデルとして、Dreamer および TD-MPC2 が挙げられる。Dreamer は、状態空間モデルの一つであり、「説明」。

3.2 階層的強化学習 (HRL)

3.3 階層的世界モデル (HWM)

階層的世界モデルも、世界モデルと同様、状態空間モデルである Dreamer の方策を階層化したものが Director である。マネージャーとワーカーの 2 つの方策が存在し、マネージャーはより長い時間軸で目標とする状態を圧縮された形でワーカーに渡し、ワーカーは通常の強化学習と同様の短い時間軸で行動を出力する。ワーカーの損失関数にマネージャーの提示した目標がどの程度達成されたかを組み込むことで、ワーカーとマネー

連絡先: 三好理輝 , rm2278@cam.ac.uk

ジャーの 2 つの方策がそれぞれ役割分担しながら同時に学習できるようにしている。

Director の拡張として Hieros がある。Hieros では、

1. 方策だけでなく世界モデルも階層化している。
2. 2 階層だけでなくそれより多い階層でも実験を行っている。
3. 計算効率を上げるために、状態空間モデルとして RSSM ではなく S5WM を用いている。
4. リプレイバッファからのサンプリングを工夫し、偏りを減らすことと時間計算量の削減を実現している。

という工夫が加えられたモデルが提案されており、Atari において DreamerV3 を超える性能を示した。本稿では方策だけでなく世界モデルも階層化しているという点で Hieros に着目し、その限界と可能性を調べる。

THICK も、Director をベースとして提案された手法である。「説明」。

状態予測モデルである TD-MPC2 を階層的にしたものが Puppeteer である。これは、「説明」

4. 実験・考察

本稿では、階層的世界モデルの Hieros[Hieros] モデルの評価を行った。特に、論文中では検証されていなかった、長期タスクの評価に適している Visual Pinpad 環境での検証を行った。また、Atari でのモデルの内部状態の確認を行った。

4.1 手法

Hieros のモデル検証では、既存のレポジトリ [Hieros Repo] を変更して実験を行った。また、Director ではこのレポジトリ [Director Repo] を用いて実験を行った。RTX-5070Ti または RTX-6000 一つを用いて学習をした。

Pinpad における Hieros の学習状況を調査した。まずは、Hieros のベースラインでの検証を行った。

図からわかるように、報酬が 0 のまま学習が進まないことが確認された。そこで、以下では pinpad の報酬設計を元の実装から、末尾の一一致度が上がるとスコアが与えられるように変更した（例えば、赤 → 赤緑に変化した場合、+1 の報酬、赤緑 → 赤緑緑に変化した場合、-2 の報酬が与えられる）。

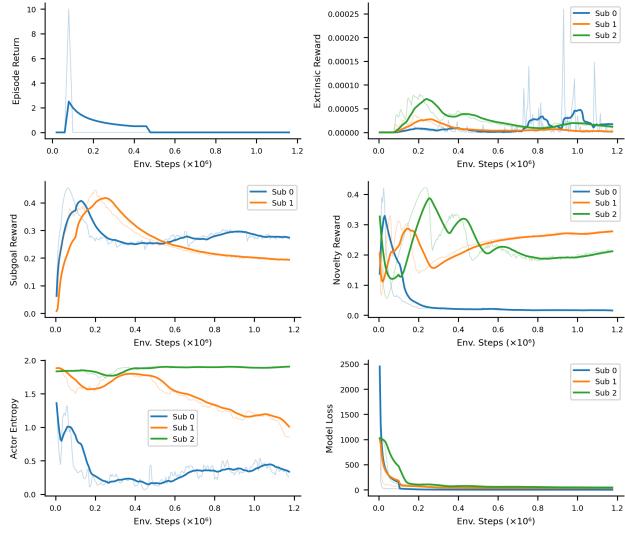


図 1: Visual Pinpad 環境における Hieros の学習および内部報酬曲線。

より詳しく Hieros の性能について調査するために、Hieros の各レイヤが提示している状態の可視化を行った。元の論文では過去の経験を参考にしたノイズを加えての可視化を行っていたが、モデルの予測のみに焦点を当てるために、ノイズなしでの可視化を行った。update-every は 8 で実験を行った。

4.2 時間的抽象化の調整

update-every は、それぞれの階層が、下の階層と比較してどの程度時間的抽象化を行っているかを表している。例えば、update-every=4 では、4 ステップごとに上の階層が下の階層に与えるサブゴールを更新しているという意味である。

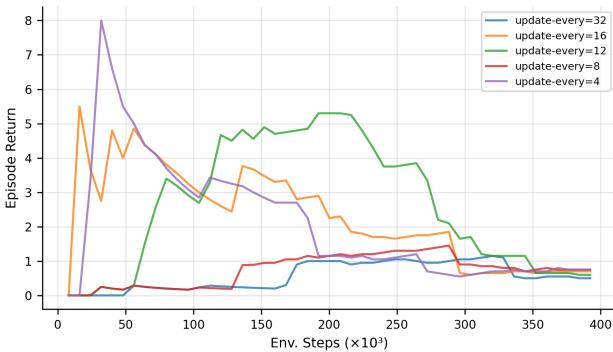


図 2: 異なる subactor-update-every パラメータでの学習曲線の比較。

図からわかるように、多少の差異がある一方で、最終的に報酬が 1 になる程度に方策が収束している。また、提示されているサブゴールも、赤までしか学習されておらず、求めている赤緑青のフルシーケンスは学習されなかった。

4.3 報酬割り当て係数の変更

Hieros では、各レイヤごとの方策は、external reward, sub-goal reward と intrinsic reward の 3 つを用いて学習を行っている。これらの比率を変更することによって、学習がどう変化するかを確認した。

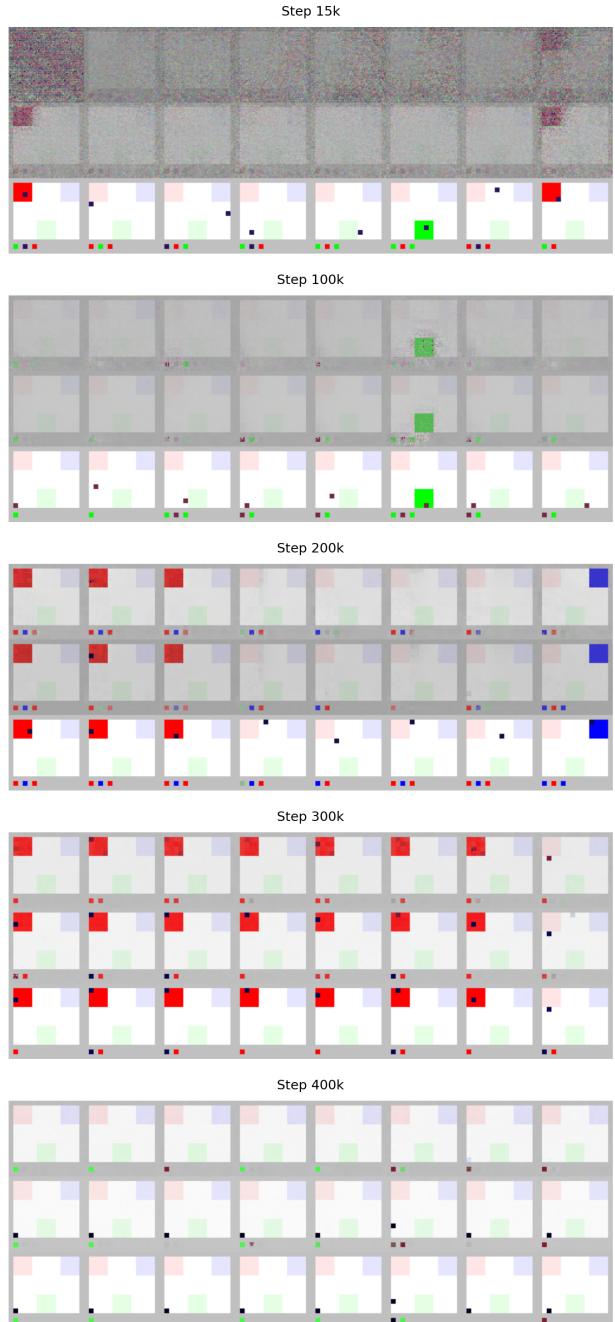


図 3: 学習過程におけるサブゴール可視化の時間的变化。

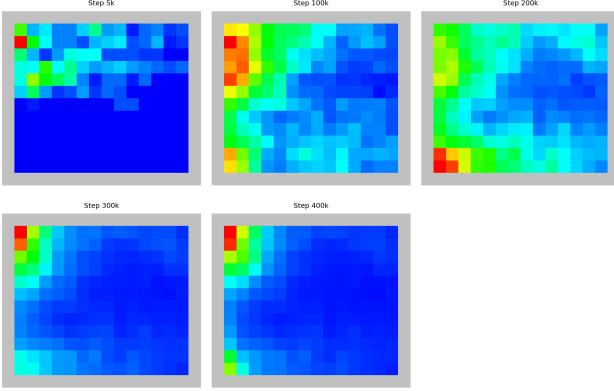


図 4: 探索位置のヒートマップの時間的变化。

限定的な向上が確認でき、報酬割り当て係数を変更することの有効性が示唆される。

4.4 方策エントロピーの変更

方策エントロピーを変更したときの結果を確認した。

エントロピーが大きくなるとより収束しにくくなり報酬も高くなつたが、顕著に学習が進むことはなかった。

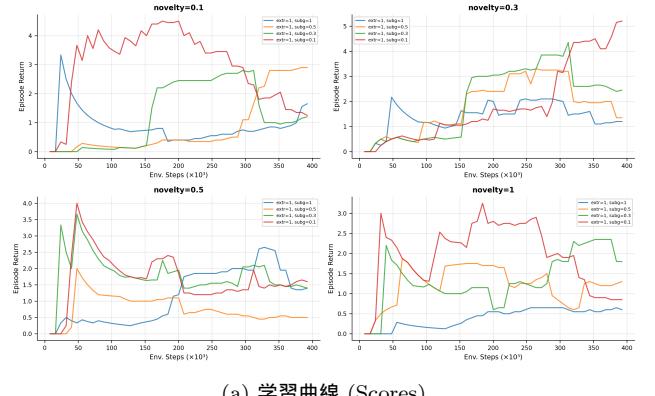
4.5 Pinpad の報酬設計の変更

pinpad-easy-three タスクにおいて、報酬設計を変更した実験結果。以下の 7 種類の報酬設計を比較した：

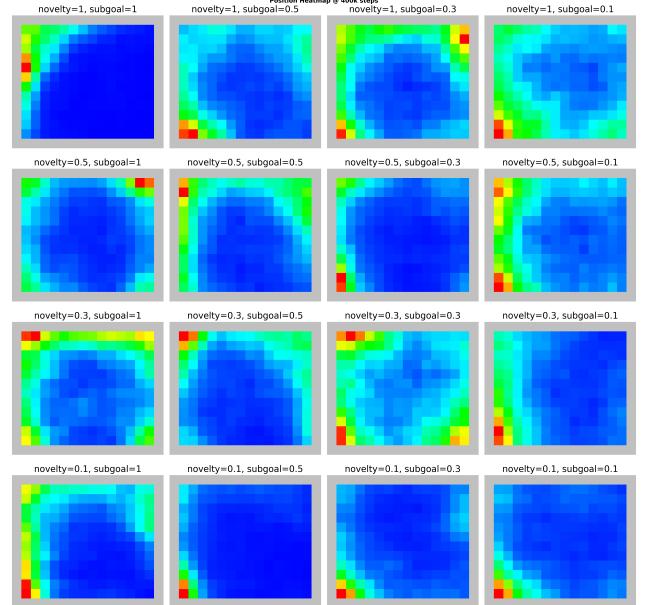
- flat (デフォルト)：タイルを踏んだときの末尾の一致度の増減が報酬
- progressive: タイルを踏んだときの末尾の一致度の増減を指数関数的にしたもののが報酬 (2^{tail_match})
- sequence_bonus: flat の報酬に加え、末尾一致度が増加ごとにボーナス報酬。
- decaying: 時間経過とともに報酬を減衰
- sparse: 完全なシーケンス一致時ののみ報酬 (元の pinpad)
- progressive_stEEP: 急勾配な指数関数的報酬 (3^{tail_match})
- dense_guidance: flat の報酬に加えステップごとの細かい報酬 (正しい方向に進んだ場合+0.1、間違った方向に進んだ場合-0.1)

4.6 考察

実験当初は pinpad で学習が進まない理由が報酬が疎であるからだと考えていたが、Pinpad の報酬設計を密にした時にも学習が進まなかつたため、そうでは無いと考えた。さらに Atari のモデルの可視化もする中で、Atari でも有意義な方策が学習できていないと思われ、Hieros のモデルそのものに、不安定性あるいは改良の余地があるのではないかと考える。一方で、有意義だと思えない学習をしているにも関わらず、なぜ Hieros が Atari で性能が高く出ていたのかの理由の解明も今後の課題である。Atari の最適方策が学習しやすく、[HRL benefit] で提示されているような階層化による探索の改善が関連しているとも考えられるが、検証が必要である。検証結果からは Hieros



(a) 学習曲線 (Scores)



(b) 探索位置のヒートマップ (Heatmap)

図 5: 報酬割り当て係数の変更による学習への影響。

の上手く学習ができておらず汎化性能が低く、より頑健性の高いモデルの理論的、実験的模索が必要であると結論づける。

ベースラインコードにエラーが含まれている可能性は否めない一方で、世界モデルを階層的にすることによって学習が不安定になっている可能性も高い。特に世界モデルは目的関数や誤差蓄積に起因して、また階層的モデルは階層的同時学習による非定常性によってハイパーパラメータに繊細であることが知られており、それらの組み合わせによって学習が不安定になってしまっている可能性もある。この理論的、実証実験的解析は今後の課題である。

4.7 Atari における定性解析

本節では、Atari ベンチマークにおける Hieros の内部挙動を解析する。そのため、学習過程における方策の時間変化を可視化した。なお計算資源の都合上、freeway のみは Hieros の元の論文と同じで (batch size, batch length) が (16, 64)、他は (16, 16) で検証を行つた。

図 14 に示すように、Hieros の Atari 環境における学習は、スコアの観点では論文通り高い性能を示しているが、方策の可視化を行うと単純な行動パターンに収束していることが確認される。特に Freeway(a)(b) では、スコアは高いものの方策

画像を確認すると「前進(Up)」動作のみが学習されており、複雑な方策は学習できていない。このことから、Hieros が報告する高スコアは必ずしも「賢い階層的計画」によるものではなく、環境の特性上、単純な方策でも高報酬が得られる可能性が示唆される。

5.まとめ

6.展望

本稿では Hieros の 3 つの報酬係数を手動で設定し比較を行ったが、HarmonyDream [HarmonyDream] のように、報酬を自動でバランスすると、より多様なタスクにおいて学習ができるようになると予想する。

参考文献

[Director] Hafner, D. et al.: Deep Hierarchical Planning from Pixels, NeurIPS (2022).

[Hieros] Mattes, et al.: Hieros - Hierarchical Imagination on Structured State Space Sequence World Models (2023).

[DreamerV1] Hafner, D. et al.: Dream to Control: Learning Behaviors by Latent Imagination, ICLR (2020).

[DreamerV2] Hafner, D. et al.: Mastering Atari with Discrete World Models, ICLR (2021).

[DreamerV3] Hafner, D. et al.: Mastering Diverse Domains through World Models, Nature (2025).

[TD-MPC] Hansen, N. et al.: Temporal Difference Learning for Model Predictive Control, ICML (2022).

[TD-MPC2] Hansen, N. et al.: TD-MPC2: Scalable, Robust World Models for Continuous Control, ICLR (2024).

[Puppeteer] Hansen, N. et al.: Hierarchical World Models as Visual Whole-Body Humanoid Controllers, ICLR (2025).

[HRL benefit] Nachum, O. et al.: Why Does Hierarchy (Sometimes) Work So Well in Reinforcement Learning?, NeurIPS (2019).

[THICK] Gumsch, C. et al.: Learning Hierarchical World Models with Adaptive Temporal Abstractions from Discrete Latent Dynamics, EWRL (2023).

[Hieros Repo] <https://github.com/Snagnar/Hieros>

[Director Repo] <https://github.com/danijar/director>

[HarmonyDream] Ma, H. et al.: HarmonyDream: Task Harmonization Inside World Models, ICML (2024).

[HRL survey] Pateria, S. et al.: Hierarchical Reinforcement Learning: A Comprehensive Survey, ACM Computing Surveys (2021).

7. Appendix

7.1 Director 実験結果

Director を用いた Pinpad-3 および Pinpad-easy-3 の結果。学習が進んでいることが確認される。pinpad-3 の結果は [Director] で報告されている結果に近いものとなっている。

また、比較のために、作成した pinpad-dense での学習も行い、高い性能が出ることを確認した。これにより、報酬設計には問題がないことが確認された。なお、GPU のメモリの都合上途中までしか学習ができなかった。

7.2 RSSM

Hieros は S5WM を用いているが、ベースラインには RSSM も用意されている。以下は、RSSM を用いて Hieros を pinpad で結果。

学習が進まなかっただため、max-hierarchy=1 のときは 320k で学習を打ち切った。階層数を少ない時により学習が進んでいくことから、階層を加えることがモデルの頑健性を下げていることが推測できる。

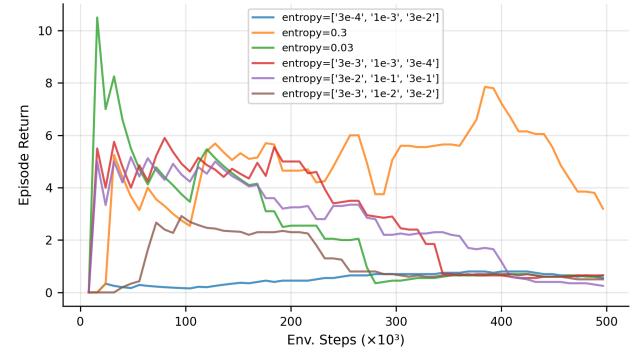
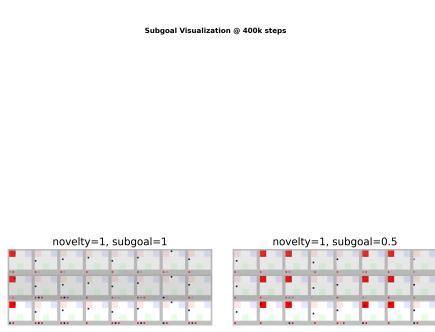
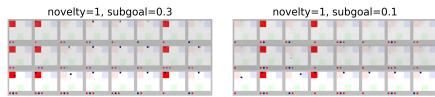


図 7: 異なる actor_entropy パラメータでの学習曲線の比較。



Subgoal Visualization @ 400k steps

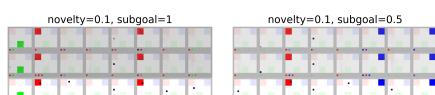
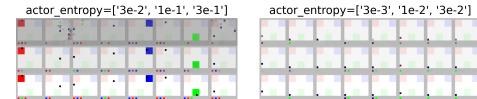
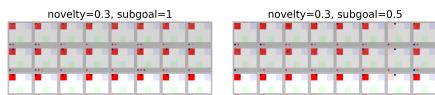
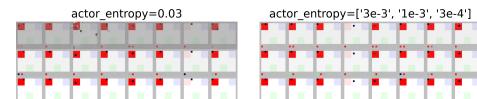
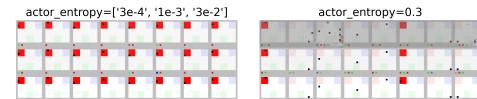
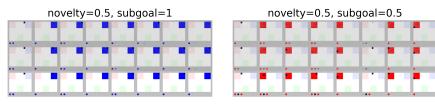


図 6: サブゴールの時間的变化。特定の報酬設定においてサブゴールが固定化される傾向が見られる。

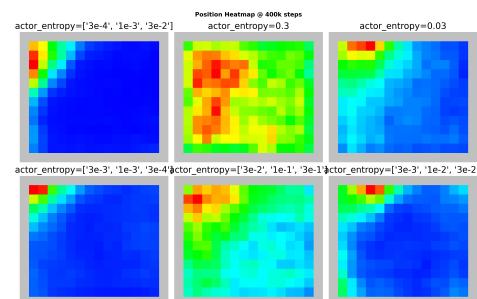


図 9: エントロピー調整実験における探索位置のヒートマップの時間的变化。

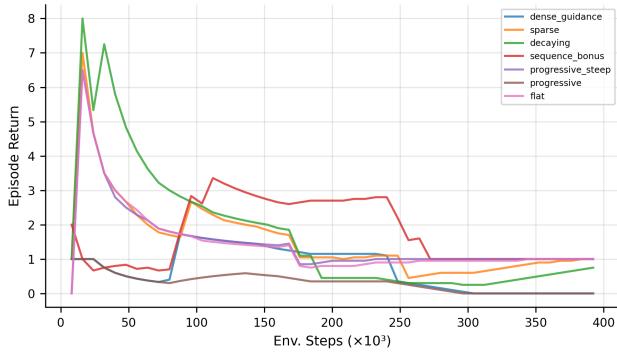


図 10: Pinpad-easy_three タスクにおける異なる報酬設計での学習曲線の比較。

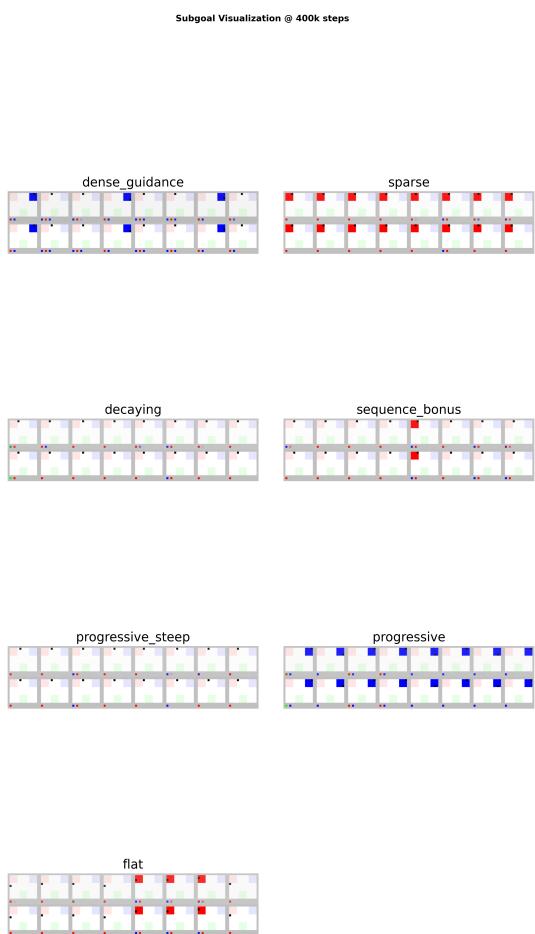


図 11: 報酬設計実験におけるサブゴール可視化の時間的変化。

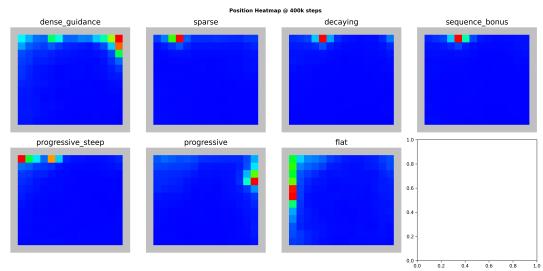


図 12: 報酬設計実験における探索位置のヒートマップの時間的変化。

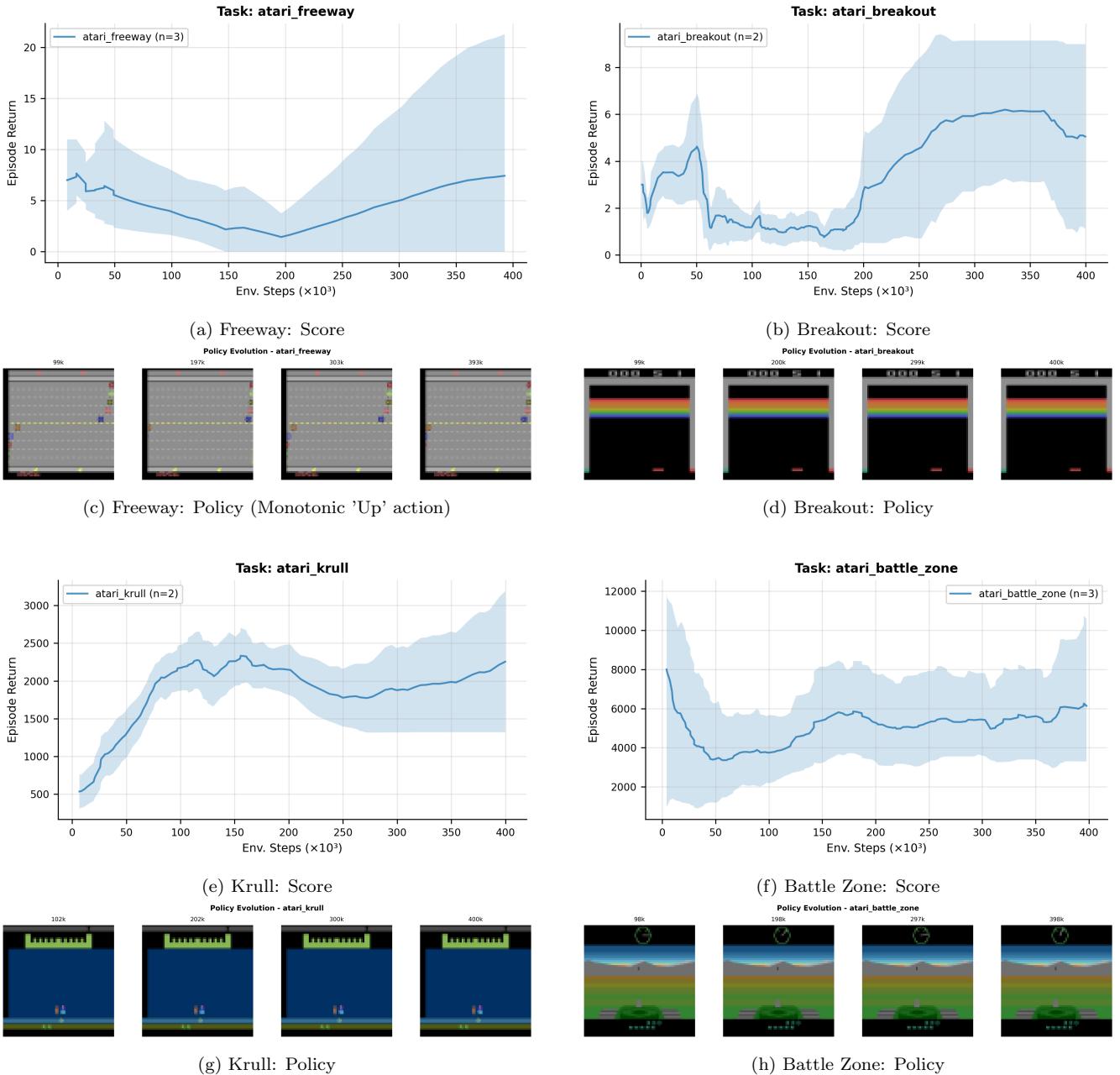


図 13: Atari 100k ベンチマークにおける学習結果。上段が学習曲線、下段が方策の可視化を示す。Freeway などでは早期に方策が固定化されている様子が、大きな画像で確認できる。

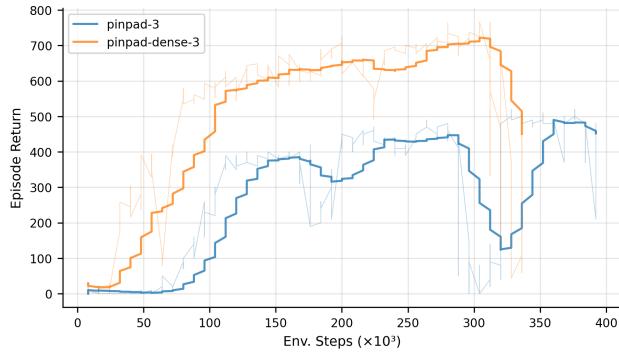


図 14: Director による Pinpad-3 および Pinpad-dense-3 の学習曲線。

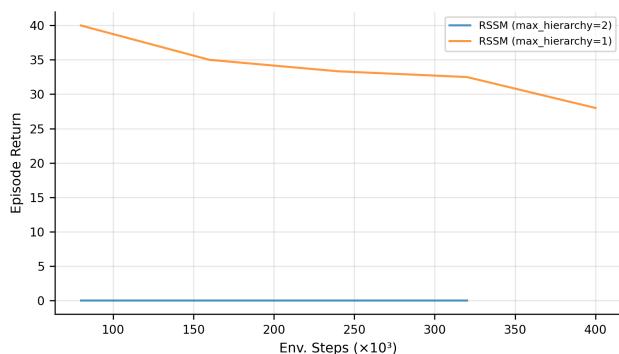


図 15: RSSM を用いた Hieros の Pinpad での学習曲線。