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　Ziwoo 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

深層強化学習は多くの分野で成功を収めている一方で，大量の環境相互作用を必要とするサンプル効率の低さや，長期的依存を伴うタスクにおける学習の不安定性が依然として課題である [HRL survey, HRL benefit]。この課題に対し，環境ダイナミクスを潜在空間上に学習し，内部モデル上で将来状態を予測するワールドモデルが提案されてきた [WorldModels, PlaNet]。特に Dreamer 系列 [DreamerV1, DreamerV2, DreamerV3] は，潜在状態空間上での想像学習により高いサンプル効率を実現している。さらに，長期的な依存関係を扱うために，時間抽象化を導入する階層的強化学習（HRL）が研究されてきた [HRL survey, HRL benefit]。近年では，世界モデルと階層構造を統合した手法が提案されており，Director [Director] や Hieros [Hieros] は，多時間スケールでの計画と想像を同時に実現する枠組みとして報告されている。これらの手法は，ベースラインの世界モデルを上回る性能を示している点で注目されている。しかしながら，性能向上が報告されている一方で，階層的世界モデルが内部でどのような抽象表現を獲得しているのか，あるいは真に階層的な想像や長期計画が実現されているのかについての実証的検証は十分ではない。特に，世界モデル自体を階層化した構造がどのように機能しているのかは明らかになっていない。そこで本研究では，階層的世界モデルの代表的手法である Hieros [Hieros] に着目し，その性能評価および内部状態の可視化を通じて，階層的想像学習の実態を検証する。これにより，階層的世界モデルの現状を整理し，構造的課題を明らかにすることを目的とする。

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。また，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。

連絡先: 三好理輝，rm2278@cam.ac.uk

3. 関連研究

3.1 世界モデル

世界モデルとは，環境の遷移ダイナミクスを学習し，内部モデル上で将来状態を予測することで方策学習を効率化する枠組みである [WorldModels, PlaNet]。強化学習において環境はマルコフ決定過程（MDP）[SuttonBarto] として定式化される：

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma) \tag{1}$$

ここで，$\mathcal{S}$ は状態空間，$\mathcal{A}$ は行動空間，$P(s_{t+1} \mid s_t, a_t)$ は状態遷移確率，$r(s_t, a_t)$ は報酬関数，$\gamma \in [0, 1)$ は将来報酬に対する割引率を表す。世界モデルは遷移確率 P を近似することで内部モデル上で軌道生成（imagination）を可能にする [WorldModels, DreamerV1]。この内部シミュレーションに基づく学習はサンプル効率を大きく向上させることが知られている [DreamerV1]。

3.1.1 世界モデルの分類

世界モデルは大きく以下の 3 種類に分類できる：

- 状態予測モデル（State Transition Model）
- 観測予測モデル（Observation Prediction Model）
- 状態空間モデル（State-Space Model）

状態予測モデルは現在状態から次状態を直接予測する枠組みであり，観測予測モデルは高次元観測空間を直接予測するが，学習安定性や表現効率の面で課題がある [WorldModels]。これに対し状態空間モデルは観測を低次元の潜在空間に写像し，その潜在空間上でダイナミクスを学習する手法である [PlaNet]。近年の世界モデル研究の主流はこの状態空間モデルである。

3.1.2 Dreamer

Dreamer 系列 [DreamerV1, DreamerV2, DreamerV3] は潜在状態空間において環境ダイナミクスを学習する代表的な状態空間モデルである。Recurrent State-Space Model (RSSM)

を用いて観測から潜在表現を構築し，その潜在空間上で将来状態および報酬を予測する。最大の特徴は，実環境ではなく学習済み世界モデル内部で将来軌道を生成し，imagination に基づいて Actor-Critic を最適化する点にある。DreamerV2 では離散潜在表現が導入され [DreamerV2]，DreamerV3 ではスケール適応性および安定性が向上している [DreamerV3]。

3.1.3 TD-MPC2

TD-MPC2 [TD-MPC2] は潜在世界モデルとモデル予測制御（MPC）を統合した連続制御向け手法であり，TD-MPC [TD-MPC] の拡張として提案された。Dreamer が imagination に基づく方策学習を主軸とするのに対し，TD-MPC2 は学習された潜在モデルを用いてオンラインで行動系列を計画する点に特徴がある。モデル学習と時間差分学習を統合することで，連続制御タスクにおいて高い精度と安定性を示している [TD-MPC2]。

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。

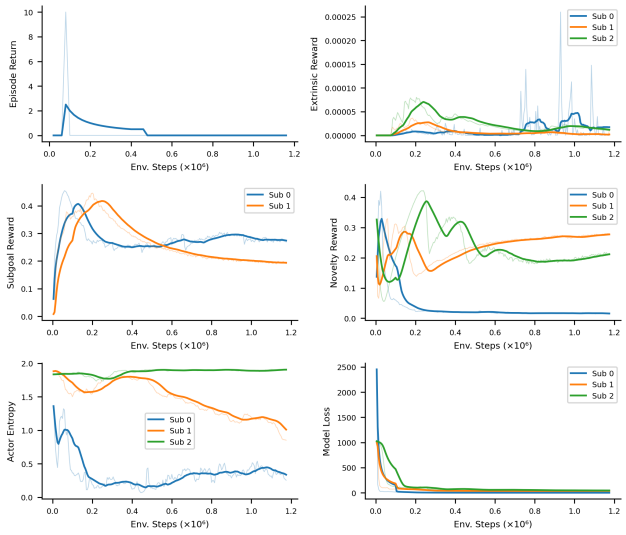


図 1: Visual Pinpad 環境における Hieros の学習および内部報酬曲線。

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）。

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。

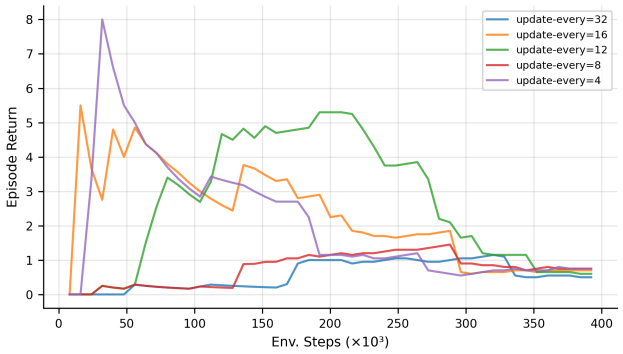


図 2: 異なる subactor-update-every パラメータでの学習曲線の比較。

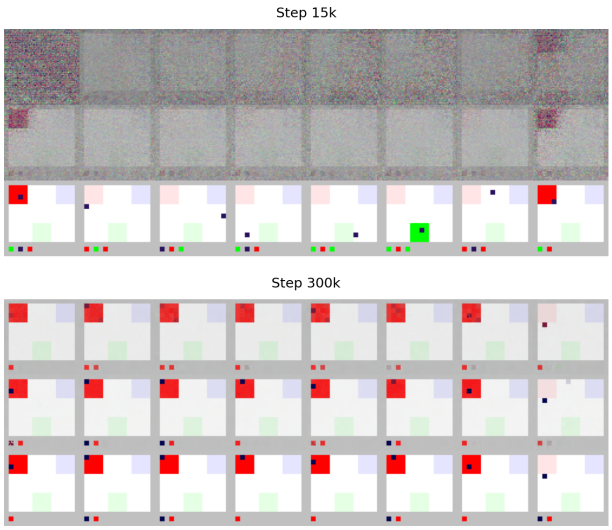


図 3: 学習過程におけるサブゴール可視化の時間的変化。典型的でわかりやすい例として 15k ステップと 300k ステップを抽出した。

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, subgoal 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。

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。図 3 と同様、フルシーケンスを学

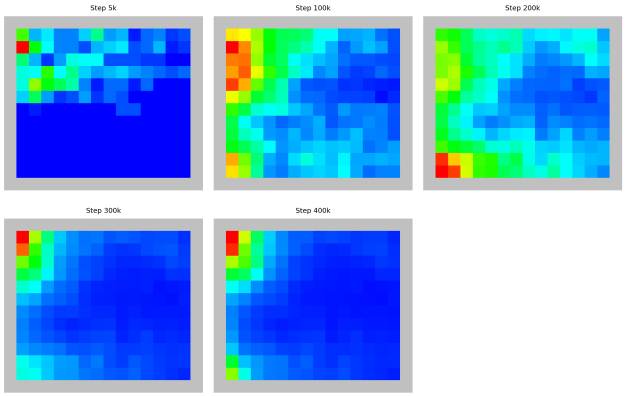


図 4: 探索位置のヒートマップの時間的変化。

習したモデルはなく、赤のみのサブゴールを提案するか、直感的には役に立たないサブゴールを学習するかであった。

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ったが、顕著に学習が進むことはなかった。サブゴールの可視化については、図 3 と同様、フルシーケンスを学習したモデルはなく、赤のみのサブゴールを提案するか、直感的には役に立たないサブゴールを学習するかであった。

4.5 Pinpad の報酬設計の変更

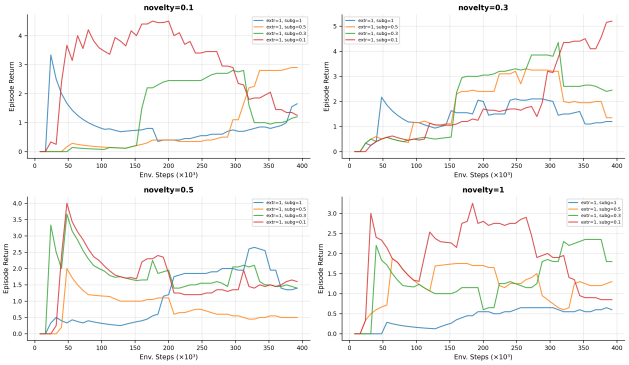
pinpad-easy-three タスクにおいて、報酬設計を変更した実験結果。以下の 7 種類の報酬設計を比較した：

- **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が報酬（$2^{\text{tail-match}}$）
- **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み報酬 (元の pinpad)
- **progressive_steep**: 急勾配な指数関数的報酬（$3^{\text{tail-match}}$）
- **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）

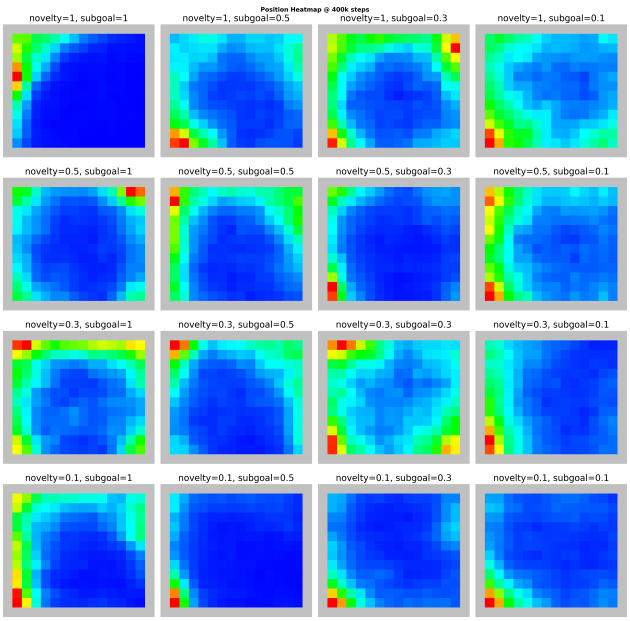
図 3 と同様、フルシーケンスを学習したモデルはなく、赤のみのサブゴールを提案するか、直感的には役に立たないサブゴールを学習するかであった。なお、ここには掲載しないが、探索位置のヒートマップからも方策が特定の領域に収束していることが確認された。

4.6 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



(a) 異なる報酬割り当て係数の学習曲線



(b) 異なる報酬割り当て係数の探索位置のヒートマップ

図 5: 報酬割り当て係数の変更による学習への影響。400k ステップ時の探索の偏りを示している。

元の論文と同じで (batch size, batch length) が (16, 64)、他は (16, 16) で検証を行った。

図 9 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、400k ステップ時点での高解像度方策画像から抽出した連続フレームを確認すると単純な行動パターンが繰り返されていることが確認される。特に Freeway では、スコアは高いものの連続したフレームを確認すると前進動作のみが繰り返されており、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。

1

2

3

4

5

6

7

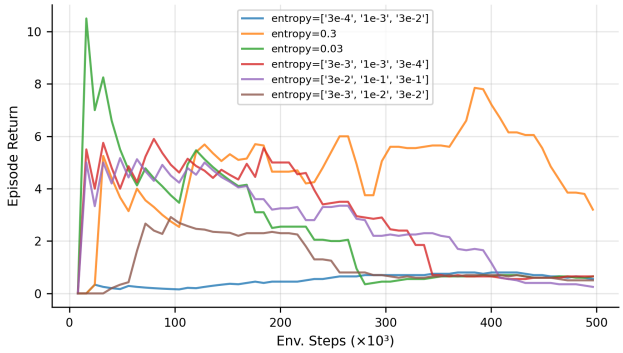


図 6: 異なる actor_entropy パラメータでの学習曲線の比較。

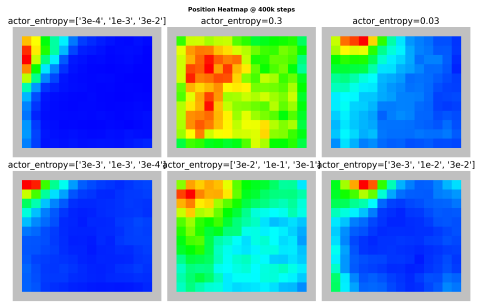


図 7: エントロピー調整実験における探索位置のヒートマップの時間的変化。

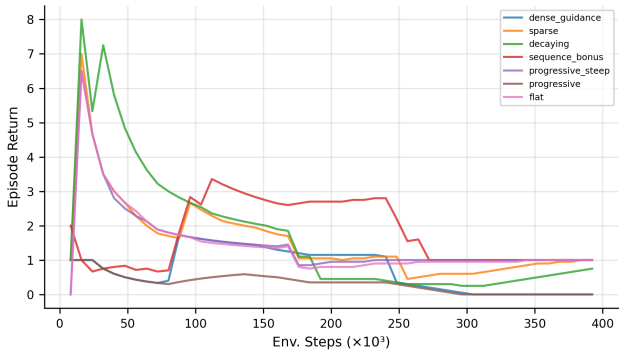


図 8: Pinpad-easy_three タスクにおける異なる報酬設計での学習曲線の比較。

8
9
10

4.7 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ったため、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 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と考える。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の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。

5. まとめ

1
2
3
4
5
6
7
8
9
10

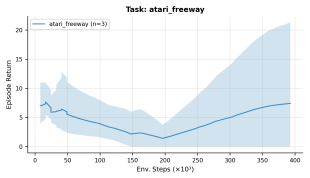
6. 展望

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。

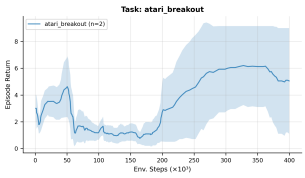
1
2
3
4
5

参考文献

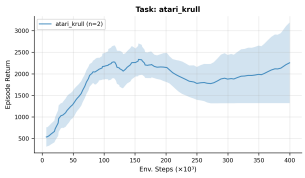
[Director] Hafner, D. et al.: Deep Hierarchical Planning from Pixels, NeurIPS (2022).



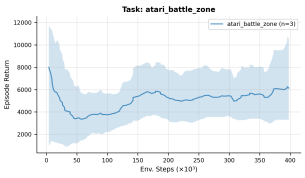
(a) Freeway: Score



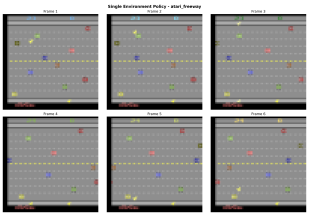
(b) Breakout: Score



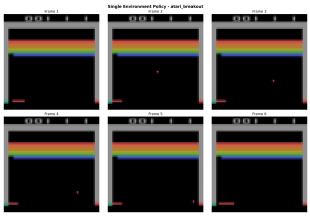
(c) Krull: Score



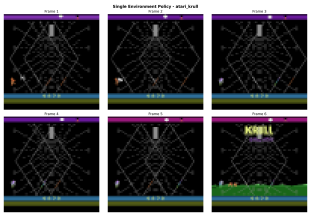
(d) Battle Zone: Score



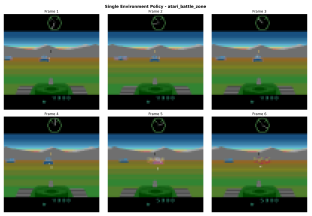
(e) Freeway: Policy



(f) Breakout: Policy



(g) Krull: Policy



(h) Battle Zone: Policy

図 9: 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 400k ステップ時点での方策画像（5 フレーム間隔で抽出）を示す。Freeway では単調な前進動作のみが繰り返されていることが確認できる。

[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] Gumbsch, 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).

[WorldModels] Ha, D. and Schmidhuber, J.: World Models, NeurIPS (2018).

[PlaNet] Hafner, D. et al.: Learning Latent Dynamics for Planning from Pixels, ICML (2019).

[SuttonBarto] Sutton, R. S. and Barto, A. G.: Reinforcement Learning: An Introduction, MIT Press (2018).

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。

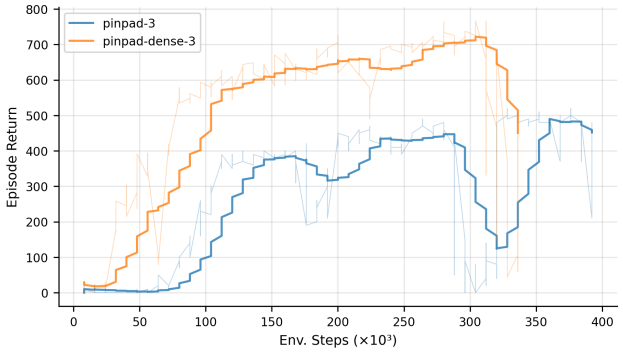


図 10: Director による Pinpad-3 および Pinpad-dense-3 の学習曲線。

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。

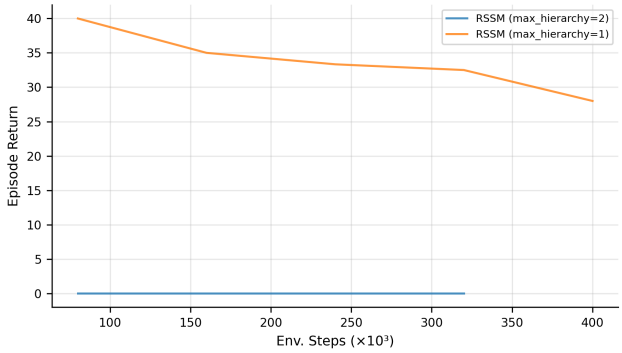


図 11: RSSM を用いた Hieros の Pinpad での学習曲線。

学習が進まなかったため、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ることから、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。