階層的世界モデルにおける現状と課題：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 *3
Riki Miyoshi　Jiu You　Third Author's Name

*1ケンブリッジ大学　*2所属
University of Cambridge　Affiliation #2 in English

階層的強化学習（HRL）は、探索効率、報酬割り当て、および解釈性の向上において有効であり、世界モデルと統合することで高いサンプル効率が期待される。しかし、既存の階層的世界モデルである Hieros や Director は、固定的な時間抽象化や有意義なスキルの獲得といった点で課題を抱えている。本研究では、Visual Pinpad 環境における Hieros の評価を通じて、疎な報酬環境における実証的な限界を特定した。また、これらの知見に基づき、動的な時間抽象化やピラミッド型構造など階層間の動的な相互作用の改善に向けた将来的な発展可能性を論じる。

1. はじめに

階層的強化学習（HRL）は、効率的な探索やスキルの抽象化を可能にする [1]。世界モデル（World Model）を用いることで、環境との相互作用を抑えた学習が可能となるが、長期依存タスクにおける時間方向の階層化は依然として大きな課題である。本プロジェクトでは、最新の HWM である Hieros [2] を中心に、サーベイ軸（既存 HWM の体系的整理）と実証軸（Hieros の実験評価）の二つの観点から、その実用性と限界を調査し、現在の HWM が抱える本質的な課題を明らかにすることを目的とする。

2. 研究背景・目的

世界モデルに HRL を統合する試みは限定的であり、階層化された世界モデルを採用している既知の例は Hieros のみである [2]。しかし、Hieros には以下の課題が存在する：

- **固定的な時間抽象化**: タスクごとに最適な更新間隔が異なるにもかかわらず固定値を用いているため、効率性と解釈性が制限される。
- **疎な報酬への弱さ**: 報酬が疎な環境（Visual Pinpad 等）での検証が不十分である。

各階層を独立した RL 問題と見なすと [4]、階層間の「動的な相互作用」の改善こそが HWM 進展の鍵であると考える。

3. 関連研究

階層的世界モデル（HWM）は、時間抽象化の扱いに応じて大きく三つの流派に分類できる。

- 行動階層型：世界モデル自体はフラットだが、ポリシーを Manager と Worker に分け、潜在空間でのサブゴール指定を通じて階層性を実現する手法（Director など）である。限界として、上位層が物理的な詳細に引きずられ、真に抽象的な戦略を学習しにくい点が指摘されている。
- 固定時間階層型：上位層ほど更新頻度を低く（k ステップごと）固定する構造（Clockwork VAE など）であり、タスクごとに最適な時間スケールが異なるにもかかわらず固定値を用いるため、柔軟性に欠ける。
- 事象駆動型：予測誤差や離散的なコンテキスト変化（イベント境界）を検出し、動的に階層を切り替える構造（VPR, THICK など）であるが、境界検出の失敗が上位層の学習を崩壊させる「非定常性」のリスクを伴う。

これらに加え、世界モデル自体と Actor-Critic の両方を多層化する Hieros [2] や、最適化自体を階層化する Nested Learning [6] なども、長期記憶保持と抽象表現の観点から重要な関連研究である。

3.1 Hieros の技術的特徴

本研究の主対象である Hieros は、長期依存を扱うための Structured State Space Sequence World Model（S5WM）を採用し、従来の RSSM を S5 ベースの世界モデルに置き換えることで、長期の依存関係を効率的かつ並列的に学習する。また、世界モデルと Actor-Critic の双方を多層化し、上位レイヤが下位レイヤの初期状態やコンテキストを条件付ける階層構造を備える。さらに、リプレイバッファからのサンプリングを時間軸で均衡化する time-balanced sampling を導入することで、学習の安定化を図っている [2]。

4. 評価手法と取り組んだ内容

HWM の現状把握を目的として、以下の検証を実施した。

- **ベースライン評価**: Director および Hieros を用いた Visual Pinpad、Battle Zone 環境での性能評価 [1, 2, 7]。
- **構造的要因の調査**: 階層数やパラメータサイズ削減、および更新頻度 k の変更が学習収束に与える影響の調査 [8]。

特に本研究では、Hieros をベースラインとし、以下の検証を実施した：

- **環境**: Visual Pinpad 3-6 [5, 6] および Battle Zone [**?**]。
- **検証項目**: サブゴールの解釈性、レイヤーごとのパラメータサイズの影響、および階層追加のタイミングが学習に与える効果。

連絡先: 氏名，所属，住所，rm2278@cam.ac.uk

4.1 環境

5. 実験・考察

5.1 学習の収束性と可視化

Visual Pinpad 環境において、Hieros は期待されたスコアを獲得できず、報酬が 0 付近で停滞した（図 1）。

図 1: Visual Pinpad における学習曲線の比較。Hieros は疎な報酬下で局所解に陥り収束しなかった [8, 9]。

探索ヒートマップ（図 2）では、エージェントが開始地点付近の局所に固執し、広域な探索が行われていないことが確認された。

図 2: 探索範囲のヒートマップ。エージェントは開始地点付近に留まっている [8, 10, 11]。

5.2 Hieros の限界特定

解析の結果、以下の要因が特定された：

1. **Actor-Critic の不全**: 外部報酬が疎な場合、内部のサブゴール報酬のみに依存し、探索が迷走する [12, 13]。
2. **サブゴールの崩壊**: 階層間のコサイン類似度を用いた報酬計算が、多様な状態生成を阻害する「Subgoal collapse」を引き起こした。
3. **学習の不安定性**: S5WM 導入による実装の複雑化と、勾配爆発（Exploding gradient）を確認した [7, 12]。

5.3 Hieros の限界分析

実験を通じて以下の限界を特定した。

1. **Actor-Critic の未収束**: 疎な報酬環境下で外部報酬を得られない場合、内部報酬（Subgoal reward）のみに依存し、探索が迷走する [8]。
2. **サブゴールの崩壊**: 階層間のコサイン類似度を用いた報酬計算が、多様な状態生成を阻害する「Subgoal collapse」を引き起こしている可能性が高い。
3. **学習の不安定性**: S5WM 導入による実装の複雑化と、それに伴う勾配爆発（Exploding gradient）の発生を確認した。

6. 将来的な展望

特定された課題に基づき、以下の発展案を提案する。

- **ピラミッド型 HWM**: 低レイヤを大きく、高レイヤを小さくすることで抽象化を強制し、計算資源を最適化する [8]。
- **動的な時間抽象化**: TempoRL [9] を参考に、サブゴール更新間隔 k を動的に変更する「Dynamic k」の導入。
- **Adapter (LoRA) の活用**: TD-MPC2 [10] 等に LoRA [11] を適用し、共通概念をコアに保持しつつタスク固有表現を分離することで、破滅的忘却を防止する。

7. おわりに

本研究は Hieros の限界を実証的に特定し、HWM の安定化には単純な多層化ではなく、動的な相互作用と報酬設計の改善が不可欠であることを示した。

参考文献

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

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

[3] Hansen, N. et al.: TD-MPC2: Scalable Model-Based Reinforcement Learning, arXiv (2023).

[4] ICLR 2024: Learning Hierarchical World Models with Adaptive Temporal Abstractions from Discrete Latent Dynamics (2023).

[5] Vaidyanath, et al.: Clockwork Variational Autoencoders, NeurIPS (2021).

[6] ICLR 2022: Variational Predictive Routing with Nested Subjective Timescales (2021).

[7] NeurIPS 2025: Nested Learning: The Illusion of Deep Learning Architectures (2025).

[8] Hu, E. J. et al.: LoRA: Low-Rank Adaptation of Large Language Models, ICLR (2022).

[9] TempoRL: Learning When to Act, arXiv:2106.05262 (2021).

[10] Rao, R. and Ballard, D.: Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-classical Receptive-field Effects, Nature Neuroscience (1999).

[11] W&B Logs: Project Hieros-hieros, run: womkk4d8, wmk3jlws.

[12] Internal Memo: Hieros Implementation and Pyramidal HWM Experiments (2024-2025).

[13] Why Does Hierarchy (Sometimes) Work So Well in Reinforcement Learning?, arXiv preprint (2019).

[14] On Efficiency in Hierarchical Reinforcement Learning, arXiv preprint (2020).

[15] Hierarchical Reinforcement Learning: A Comprehensive Survey, preprint (2021).

[16] Learning Options via Compression, NeurIPS (2022).

[17] Mastering Diverse Domains through World Models, preprint (2023).

[18] Minigrid & Miniworld: Modular & Customizable Reinforcement Learning Environments for Goal-Oriented Tasks, preprint (2023).

[19] Hieros thesis, thesis (2024).

[20] Hierarchical World Models as Visual Whole-Body Humanoid Controller, preprint (2025).

[21] Training Agents Inside of Scalable World Models, preprint (2025).

[22] Reverse-Engineering Memory in DreamerV3: From Sparse Representations to Functional Circuits, preprint (2025).

[23] Learning Massively Multitask World Models for Continuous Control, preprint (2025).

[24] Temporal Structure of Natural Language Processing in the Human Brain Corresponds to Layered Hierarchy of Large Language Models, preprint (2025).