社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Toward Human Cognition-inspired High-Level Decision Making For Hierarchical Reinforcement Learning Agents

Rousslan Fernand Julien DOSSA† and Takashi MATSUBARA††

† Graduate School of System Informatics, Kobe University,
1-1 Rokko-dai-cho, Nada-ku, Kobe-shi, Hyogo, 657-8501 Japan
†† Graduate School of Engineering Science, Osaka University,
1-3 Machikaneyama-cho, Toyonaka-shi, Osaka, 560-8531 Japan

Abstract Hierarchical reinforcement learning (HRL) methods aim to leverage the concept of temporal abstraction to efficiently solve long-horizon, sequential decision-making problems with sparse and delayed rewards. However, the decision-making process of the agent in most HRL methods is often based directly on low-level observations, while also using fixed temporal abstraction. We propose the hierarchical world model (HWM), which can capture more flexible high-level, temporally abstract dynamics, as well as low-level dynamics of the system. We posit such model is a natural extension to the HRL framework toward a decision-making process closer to that of humans.

Key words Reinforcement learning, hierarchical reinforcement learning, world models, temporal abstraction, hierarchically organized behavior

1. Introduction

Deep reinforcement learning (DRL) has proven to be a powerful set of automation methods, able to solve a gamut of tasks varying in complexity [1]〜[4], [23]. Still, conventional DRL methods can be very sample inefficient when applied to long-horizon, sequential decision-making tasks, which usually overlap with sparse and delayed rewards problems, further increasing their complexity.

The hierarchical reinforcement learning (HRL) framework aims to improve the efficiency of conventional (flat) RL by introducing temporal abstraction in the decision-making process of an agent [7], [8]. Namely, an HRL agent is structured as a hierarchy of policies, where each level acts at a coarser time scale than the level below. In practice, this allows HRL agents to explore the state space more efficiently by leveraging low-level policies (also referred to as *options* or *skills*), allowing for a more efficient sequential decision making [8] in long-horizon tasks. Concurrently, temporal abstraction allows for a better credit assignment through time, which is especially helpful in long-horizon, sparse and delayed reward tasks [8], [9]. While existing HRL methods have empirically demonstrated a considerable improvement in efficiency over conventional DRL methods [9]〜[13], most HRL methods rely on *fixed length temporal abstraction*. Moreover, the decision-making occurring at higher levels in the agent's hierarchy is still often based on the *observations at the lowest level*.

On the other hand, a growing body of studies in human behavior, cognitive science, neuroscience, and computational biology suggests that human behavior is hierarchically organized [14]. Namely, not only is the representation of knowledge structured into different levels of abstraction but so is the planning process itself, as well as the execution of the plans [16], [17]. It would thus be desirable to have HRL agents endowed with the ability to explicitly plan and act at a different level of abstraction [15], [18].

In this work, we combine world modeling methods [21]〜[24] with the framework of variational temporal abstraction [28] to propose the *hierarchical world model* (HWM). The proposed model is geared toward capturing more flexible, temporally abstract dynamics, while also modeling the low-level dynamics of the system, as per standard model-based RL. Owing to its hierarchical structure, the proposed model inherently provides (1) a *temporally abstract state representation* summarizing an arbitrary number of lower-level states, and (2) an adaptive temporal abstraction mechanism to divide long-horizon, sequential decision tasks into smaller tasks of variable lengths.

2. Background

2.1 Reinforcement Learning and Hierarchical Reinforcement Learning

We base ourselves on the formalism of a finite time horizon partially observable Markov decision processes (POMDP), which we denote by the tuple $(\mathbb{S}, \mathbb{A}, P_{s,a}, R, \mathbb{O}, P_o, H)$, where $\mathbb{S}$ is the state set, $\mathbb{A}$ is the action set, $P_{s,a} : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \to [0, 1]$ is the state transition

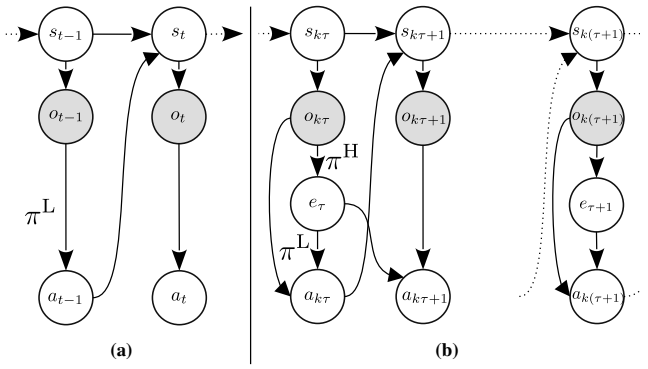

Figure 1 Graphical models incorporating the system dynamics and the decision-making of the agent of (a) flat RL and (b) HRL agent, respectively.

probability, $R : \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ the reward function, $\mathbb{O}$ the set of partial observations, $P_o : \mathbb{S} \to \mathbb{O}$ the emission probability distribution, and H the horizon (maximum episode length). The decision process of an RL agent can be represented by a stochastic policy $\pi : \mathbb{S} \times \mathbb{A}$, which defines the probability of action a being selected given a state s. From a probabilistic perspective [5], [6], the POMDP modeling the dynamics of the system and the decision making of the agent can be formally expressed as the following generative process:

$$p(O, S, A) = \prod_{t=1}^{H} p(o_t|s_t)\, p(s_t|s_{t-1}, a_{t-1})\, \pi(a_t|o_t) \qquad (1)$$

In the hierarchical RL (HRL) framework, the policy of the agent decomposed into hierarchically structured component policies. The lowest level in the hierarchy operates at the finest time scale, the same as a flat RL policy. On the other hand, the higher levels in the hierarchy operate at a coarser time scale, thus resulting in temporal abstraction. Consider the simplest case of a two-level hierarchical agent composed of π^{H} and π^{L} respectively representing the high and low-level policies. π^{H} selects a *skill* on which π^{L} will be conditioned. In practice, π^{H} is usually conditioned on the low-level observations, and its time scale is often set to a fixed length we denote as k hereafter [9]∼[12]. Denoting the *skill* space by $\mathbb{E}$, and augmenting the generative process of the POMDP in Eq. 1, we obtain:

$$p(O, S, A, E) = \prod_{\tau=0}^{H/k} \pi^{\mathrm{H}}(e_\tau|o_{k\tau}) \prod_{t=k\tau}^{k(\tau+1)} p(o_t|s_t) \; p(s_t|s_{t-1}, a_{t-1})\, \pi^{\mathrm{L}}(s_t|o_t, e_\tau). \qquad (2)$$

The graphical models for Eq. 1 and Eq. 2 are illustrated in Fig. 1 (a) and Fig. 1 (b), respectively.

2. 2 Hierarchically organized behavior

A growing body of studies [14]∼[17] at the intersection of neuroscience, cognitive science, psychology, and computational biology suggests that the human decision-making process is hierarchically organized. For example, when faced with a task such as *making a trip abroad*, we divide it in a sequence of sub-tasks such as *booking the flight*, *packing the luggage*, *driving to the airport*, *boarding the plane*, and so on. Each sub-task can be further divided into sub-sub-tasks, down to the finest granularity of actions such as bodily movements. This concept is referred to as *temporal abstraction* and allows us to efficiently learn, plan, and act in a wide gamut of activities. It is a fundamental principle underlying the HRL framework introduced in 2. 1, where each of the aforementioned sub-tasks would be realized by learning and executing the appropriate *skill*.

While existing HRL methods [8]∼[13] do structure the decision-making process of the agent hierarchically, the *skill* selection at higher levels in the hierarchy is more often than not based on the observations at the finest level of the hierarchy. In the case of example task *making a trip abroad*, this is akin to having the high-level policy decide to *drive to the airport* while using a very exhaustive representation of the current state of the agent, instead of the more abstract state *luggage packed and ready to go to the airport*.

Following this line of thought, [16], [17] suggest that our hierarchical decision-making process is intertwined with a corresponding *temporally abstract state representation*. Intuitively, such abstracted state representation would align with intermediate sub-goals, milestones, or *bottleneck states* that contribute to solving the overall task, thereby greatly simplifying planning, exploration, and execution.

The ability to create long-term plans and strategies without necessarily interacting with the world happens to also be tied to such abstract state representation. This is made possible by leveraging our internal model of the world, which is theorized to also be hierarchical. For example, we can both imagine the detailed process of *folding shirt* (low-level), as well as the more abstract process of *putting folded shirts into the suitcase* (high-level). Botvinick et al. empirically demonstrated the benefit of having a *temporally abstract model* [15]. They proposed a model-based HRL variant of the *options framework* [8], where the standard HRL agent is augmented with a temporally abstract model that allows the agent to directly plan at a coarser time scale. This resulted in a great improvement in performance, but also sample efficiency. However, the *skills*, and the temporally abstract model were manually engineered, thus limiting the generality of the proposed approach.

As suggested by Barto et al. [18], it would be desirable to endow HRL agents with *an additional mechanism that allows autonomous extraction of temporally abstract state, and the corresponding temporal dynamics model*. Moving toward a human cognition-inspired decision-making process, we turn ourselves to recent methods for dynamics learning that fall under the umbrella of *world modeling*, to complement standard HRL agents with a general, end-to-end mechanism for discovery and learning of temporally abstract state representations and dynamics.

2. 3 World models

The broad range of methods that have marked the recent resurgence of model-based RL (MbRL) are referred to as *world models* [21]∼[24]. Such methods have not only demonstrated either competitive or superior performance to the leading model-free RL

methods but also improved the sample efficiency of the agents.

One of the key factors behind the success of world modeling methods is the introduction of unsupervised learning objectives [19] to learn more compact representations of the internal state belief s_t in Fig. 1. This allowed to separate the decision-making component (policy) from the raw, and usually noisy pixel-based observations, greatly simplifying the learning process. Additionally, world modelling techniques leverage Recurrent Neural Networks (RNN) [20] to approximate the state transition dynamics $p(s_t|s_{t-1}, a_{t-1})$ from Eq. 1. Such approximation can then be used to collect samples in an *imaginary* environment in the place of the real environment, leading to a drastic improvement in sample efficiency.

Through the series of *Dreamer* agents [23], [24], Hafner et al. take this concept one step further by seamlessly incorporating reward prediction and an actor-critic [2], [3] policy into the model. Leveraging the differentiable dynamics of the latter allows *Dreamer* agents to directly improve their policy in an end-to-end manner over simulated trajectories.

However, most world model methods only estimate the internal state belief and transition dynamics at the finest time scale. Inspired from the theorized hierarchically structured model of humans presented in 2. 2, we seek to augment conventional modeling methods with a hierarchical structured dynamics model.

2. 4 Variational temporal abstraction

The variational temporal abstraction (VTA) [28] framework was introduced as a discovery method for temporally hierarchical structure and representation in sequential data. Formally, VTA assumes the existence of a sequence of observations $O = \{o_1, o_2, ...o_H\}$ of length H that can be decomposed into N non-overlapping subsequences $O = (O_1, O_2, ..., O_N)$, such that each sub-sequence $O_i = \{o^i_{1:l_i}\}$ has length l_i, and $\sum_{i=1}^{N} l_i = H$. Each observation o_t is generated from the corresponding *low-level state* w_t, such that each observation sub-sequence O_i is associated with a low-level state sub-sequence W_i. Finally, each low-level state sub-sequence W_i is assumed to have been generated from a *temporally abstract state* z_i.

To efficiently separate sub-sequences W_i corresponding to different z_i, the VTA framework leverages a binary random variable M referred to as *binary boundary indicator* instead of modeling both the number of sub-sequences N and their lengths L. At an arbitrary time-step t, the binary indicator m_t specifies whether or not a new sub-sequence starts at time step $t + 1$. The generative process for an observed sequence O is formally defined as follows:

$$p(O, W, Z, M) = \prod_{t=1}^{H} p(o_t|w_t)\, p(w_t|z_t, w_{t-1}, m_{t-1}) \\ p(z_t|z_{t-1}, m_{t-1})\, p(m_t|w_t). \tag{3}$$

The *temporally abstract* state z and the low-level state abstraction w are approximated using the proposed hierarchical recurrent state-space model, which is trained using sequential variational inference [19].

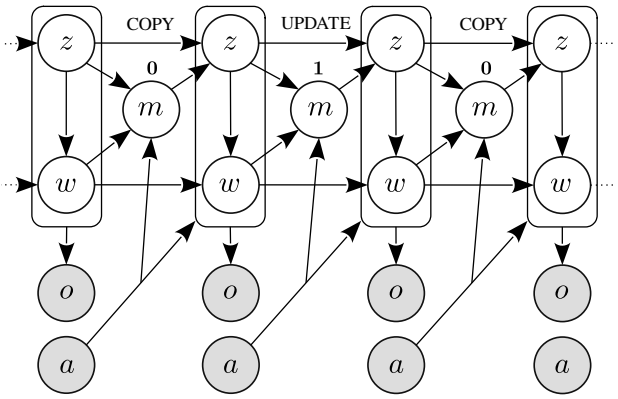

Figure 2 Dynamic Bayesian network representing the generative process that is approximated by the proposed HWM.

While it does provide an adaptable mechanism to learn temporally abstracted dynamics and the corresponding abstract states, it cannot be directly incorporated into a hierarchically structured decision-making process. Namely, it does not model the influence of the actions of neither low nor high-level actions originating as causal components of the observed sequences.

3. Proposed method

3. 1 Theoretical model

In this work, we propose the *hierarchical world model* (HWM), which combines conventional world modeling techniques and VTA to capture flexible, temporally abstract dynamics, and the corresponding state representations structured hierarchically.

First, we adopt world model structure of the Dreamer agents [23], [24] introduced in Section 2. 3. For a given task formalized as a POMDP, such a world model approximates the internal belief state s_t as a single random variable using sequential variation inference. Following the VTA [28] framework, we assume that the internal belief state s_t itself can be decomposed into the hierarchically structured components w_t, and z_t, with the boundary indicator m_t determining when the temporally abstract transition takes place.

Since we are mainly interested in learning the dynamics of the environment, we assume that the sequence of observations is generated under a fixed policy π, which influence is hereafter represented by the random variable A. Based on Eq. 1, and Eq. 3 we obtain the following generative process the proposed model is based upon:

$$p(O, W, Z, M|A) = \prod_{t=1}^{H} p(o_t|w_t, z_t)\, p(w_t|z_t, w_{t-1}, a_{t-1}) \\ p(z_t|z_{t-1}, m_{t-1}, w_{t-1}, a_{t-1})\, p(m_t|w_t, z_t, a_t). \tag{4}$$

The corresponding graphical model is documented as Fig. 2.

3. 2 Learning and inference

The generative process proposed in Eq. 4 is modeled using a

hierarchical recurrent state-space Model [28]. More specifically, $p_\theta(o_t|w_t, z_t)$ is parameterized as corresponds to the decoder of a variational auto-encoder [19]. The temporally abstract transition is modeled by $p_\theta(z_t|z_{t-1}, m_{t-1}, \kappa_{t-1})$ using deep neural networks. First, κ_t encodes all the low-level states and actions of the current segment corresponding to z_t using a gated recurrent unit (GRU).

To improve the modeling of long-term dynamics, z_t is decomposed into a deterministic component c_t and and stochastic component v_t [22]∼[24], [28]. The deterministic transitions for c_t are modeled using the following rule:

$$c_t = \begin{cases} c_{t-1} & \text{if } m_{t-1} = 0 \text{ (COPY)} \\ f_{\text{z-rnn}}(z_{t-1}, \kappa_{t-1}, \ c_{t-1}) & \text{otherwise (UPDATE)} \end{cases}$$

where $f_{\text{z-rnn}}$ is a GRU neural network. The stochastic component v_t is implemented as a Normal distribution: $v_t \sim \mathcal{N}(\mu_v(c_t), \sigma_v(c_t))$, where μ_v and σ_v are parameterized by their respective densely connected feed-forward neural networks. The temporally abstract state z_t is thus obtained by concatenating c_t and v_t and feeding it through a single dense layer.

Similarly, the low-level state w_t is also decomposed into deterministic and a stochastic components, respectively denoted by h_t and y_t. Unlike in VTA [28], h_t is seamlessly updated using the rule:

$$h_t = f_{w\text{-rnn}}\big(w_{t-1}, a_{t-1}, z_t, \ \ h_{t-1}\big),$$

where $f_{\text{w-rnn}}$ is the GRU neural network associated with the low-level state transitions. The stochastic component y_t is implemented as a normal distribution $y_t \sim \mathcal{N}(\mu_y(h_t), \sigma_y(h_t))$, where μ_y and σ_y are parameterized by their respective densely connected feed-forward neural networks. The concatenation of h_t and y_t is then used to represent the low-level state.

Finally, the prior boundary detector $p_\theta(m_t|w_t, z_t, a_t)$ is parameterized using a densely connected neural network with a final sigmoid activation function.

The hierarchy of state representations and the corresponding dynamics is inferred [19] using the parameterized variational distribution $q_\phi(Z, W, M|O, A)$. The latter is decomposed as follows:

$$q_\phi(Z, W, M|O, A) = q_\phi(M|O) \prod_{t=1}^{H} q_\phi(w_t|z_t, M, O)\, q_\phi(z_t|M, O),$$

with $q_\phi(M|O) = \prod_t \text{Bernoulli}(m_t|\sigma(\varphi(O)))$, where σ is the sigmoid function, and φ is temporal convolution operation over a sequence of pairs of observation and actions. Both $q_\phi(w_t|z_t, M, O)$ and $q_\phi(z_t|M, O)$ are approximated using the mean field approximation-based method proposed by Kim et al. [28].

The parameter vectors θ and ϕ are learned by maximizing the variational lower bound (VLB) derived from Eq. 4 as follows:

$$\log p(O|A) \geq \mathop{\mathbb{E}}_{q_\phi}\Big[\log p_\theta(O|Z, W)\Big] - \text{KL}\Big[q_\phi(Z, W, M|O, A) \,||\, p_\theta(Z, W, M|A)\Big]. \tag{5}$$

For a given sequence of observation-action pairs $\{(o_t, a_t)\}_{t=1}^{H}$, the VLB derived in Eq. 5 is appromixated as:

$$J(\theta, \phi) = \sum_{t=1}^{H} \log p_\theta(o_t|w_t, z_t) - \text{KL}\big[q_\phi(z_t) \,||\, p_\theta(z_t)\big] - \text{KL}\big[q_\phi(w_t) \,||\, p_\theta(w_t)\big] - \text{KL}\big[q_\phi(m_t) \,||\, p_\theta(m_t)\big] \tag{6}$$

The first term in Eq. 6 corresponds to the reconstruction objective across the observed sequence. The last three terms correspond to the Kullback-Leibler divergence between the generative and variational distributions used to approximate temporally abstract state z_t dynamics, the low-level state w_t's dynamics, and the sequence segmentation based on the boundary indicator m_t, respectively.

4. Experimental setting

The experiments are grounded in one of the representative task of the HRL domain referred to as *Four Rooms* [8], illustrated in Fig. 3 (a). We built on top of the publicly available implementation referred to as the *MiniGrid-FourRooms-v*0 environment, provided by Chevalier-Boisvert et al. [30]. By default, this environment provides RBG images as observations to the RL agent.

In *Four Rooms*, the agent (red triangle in Room 3 of Fig. 3) has to reach the exit (green tile in *Room* 1 of Fig. 3) within a maximal episode length of $H = 100$ steps. The layout of the maze, which determines the position of the *doors* (*Rooms* 4,5,6*, and* 7 in Fig. 3) is fixed across all episodes. Both the starting position of the agent and the goal are set randomly at the beginning of each episode. The action space is simplified to three actions: *turn left*, *turn right*, and *move forward*.

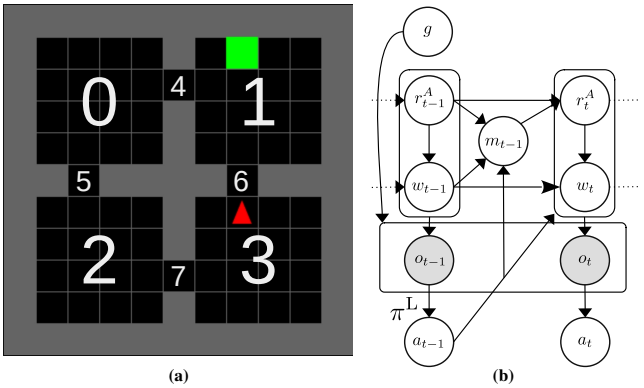

Figure 3 (a) Screenshot of the *MiniGrid-FourRooms-v*0 task. We assign a number from 0 to 7 to each room for later reference. (b) Dynamic Bayesion Network illustrating the internal dynamics of *MiniGrid-FourRooms-v*0 task and the decision-making process of a flat RL agent.

For our purpose, we consider a *decomposed state representation*

of s_t, illustrated in Fig. 1 (b). Namely, we can efficiently describe the state of the whole maze using 3 random variables. First, let G encode the room of the goal, as well as the relative x and y coordinates of the goal in said room. Next, the agent position in the maze can be decomposed into the room of the agent, denoted as the random variable R^A, and the variable W that encodes the relative x and y coordinates of the agent in R^A, as well as the direction it is facing. This is derived from the observation that the information encoded by G is fixed across an episode, while the room of the agent R^A changes less frequently when compared to the x and y coordinates, or the direction of the agent encoded in W. Notice that this compact representation can be matched with the two-level hierarchy of the HWM proposed in Fig. 2, by collapsing both G and R^A into the variable Z, because they both change at a slower pace than W. This is motivated by the experimental results of Saxena et al. [29], stipulating that in a hierarchically structured recurrent state-space model, slowly changing information in the observations are encoded into higher levels of the latent variable hierarchy, while fast-changing components are encoded at the lower levels.

To demonstrate how the proposed HWM captures an adaptable temporally abstracted state dynamics, we first train an RL agent instantiated as Proximal Policy Optimization (PPO) algorithm [2] using the reference implementation provided in the *Stable Baselines* 3 RL algorithm library [31]. The pre-trained PPO agent is used to generate a dataset of $25,000$ observation-action pairs (o_t, a_t), corresponding to $1,736$ distinct episode trajectories. This dataset is then used to train an instance of the proposed model to maximize the objective function derived in Eq. 6.

5. Results

Recall that the purpose of the HWM is to provide (1) a *temporally abstract state representation* summarizing an arbitrary number of lower-level states while at the same time filling the role of a world model [21], [23], [24]. Additionally, the model should also provide (2) an adaptive temporal abstraction mechanism to divide trajectories into coherent sub-sequences [28].

To evaluate the proposed method, a trained instance of the proposed HWM following the setting described in Section 4. is fed trajectories of observation-action pairs (o_t, a_t). For each trajectory, the first temporally abstract state z_0 is infered using the learned posterior $q_\phi(z_t)$. For $t > 0$, the temporal transition is modeled using the learned $p_\theta(z_t)$. The lower-level state w_t at each step is inferred using the posterior $q_\phi(w_t)$, identically to Dreamer agents [23], [24]. The boundary indicator m_t is estimated using the learned prior distribution $p_\theta(m_t)$. Finally, each observation o_t is reconstructed using the learned decoder $p_\theta(o_t|w_t, z_t)$.

From line (a) and (e) in Fig. 4, we observe that the HWM manages to accurately reconstructs the provided observations, while modeling the high and low-level state transitions, thus satisfying the requirement (1). Requirement (2) is also satisfied, as the prior boundary

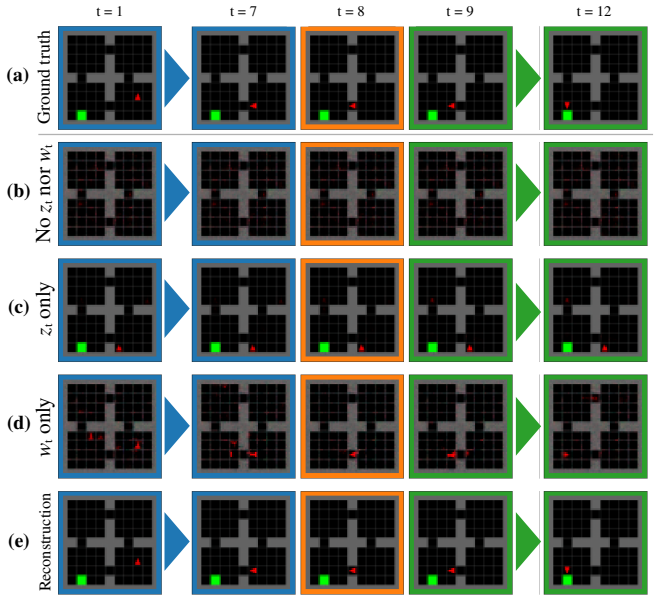

Figure 4 Illustrative example of the reconstruction and segmentation of an arbitrary trajectory performed by the proposed HWM. Each color represents a segment that corresponds to a temporally abstract state s_t. The segmentation is performed using the learned prior boundary distribution $m_t \sim p_\theta(m_t)$. Intermediate steps of long segments are ommited

indicator $p_\theta(m_t)$ can accurately predict the change in the room of the agent. In this specific case, the proposed trajectory segmentation also coincides with the intuited abstracted dynamics we derived for the *FourRooms* task, as illustrated in Fig. 3 (b).

Through the lines (b), (c), and (d) in Fig. 4, we aim to illustrate what part of the reconstructed observation each of the latent variables z_t and w_t is responsible for. From (b), we observe that only the layout of the maze is reconstructed when we pass zeros as input to the decoder $p_\theta(o_t|w_t, z_t)$. When conditioning the decoder on the high-level state z_t only, the goal tile, as well as a blurry depiction of the agent appear in the reconstructions, illustrated by the line (c). On the other hand, when conditioning the decoder on w_t only, not only is the goal absent from the frame but the position of the agent in the maze becomes ambiguous, as illustrated in line (d). These results suggest that the information about the goal and the room of the agent tend to be encoded at the higher level by z_t, while other, more *granular*, information is encoded at the lower level by w_t.

One caveat would be that the role of encoding information about the agent seems to be shared by the combination of w_t and z_t. Namely, on line (c) at time step $t = 8$, the blurry agent is depicted in the room corresponding to the previous segment of $t \in [1, 7]$. Concurrently, on line (d) at time steps $t \in \{7, 8, 9\}$, the agent is localized in the correct room, albeit with a less defined depiction. This could be solved by further refining the HWM's structure and using methods such as [26], [27] to disentangle the information captured

at different levels of the hierarchy.

6. Discussion

In this work, we leverage the recent progress in world modeling methods and the framework of variation temporal abstraction to derive the hierarchical world model (HWM). The proposed model captures both temporally abstract and granular dynamics, as well as the corresponding hierarchically structured state representations. By design, the HWM thus maintains the ability of world model methods [21]〜[24] to provide a compact state representation in the form of z_t and w_t for standard MbRL agents. Moreover, we postulate that the temporally abstract state z_t and the corresponding temporally abstract dynamics is a simplified analog to the mechanism that allows humans to conceptualize plans at a higher level of abstraction, as previously motivated in Section 2. 2.

Furthermore, the proposed model also provides an adaptive temporal abstraction mechanism to divide long-horizon, sequential decision tasks into smaller tasks of variable lengths. We posit that such an adaptive mechanism will allow the relaxation of the higher-level decision-making process of the HRL agent, thus building toward a more human cognition-based, hierarchically structured decision-making process.

Ongoing and future endeavors will consist in developing this framework by incorporating a hierarchically structured decision-making process into the proposed HWM, which we expect to improve the sample efficiency of HRL and MbRL while expanding the range of tasks that can be reliably solved by RL-based agents.

Acknowledgement

This work was partially supported by JST-Mirai Program (JPMJMI20B8), Japan. Rousslan F.J. Dossa thanks the Japanese Ministry of Education, Culture, Sports, Science, and Technology (MEXT) for its scholarship grant from the year 2017 to 2023.

References

[1] Mnih et al., "Playing Atari with Deep Reinforcement Learning", *ArXiv*, vol. abs/1312.5602, 2013.

[2] Schulman et al., "Proximal Policy Optimization Algorithms", *ArXiv*, vol. abs/1707.06347, 2017.

[3] Haarnoja et al., "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor", *International Conference on Machine Learning*, 2018.

[4] Dabney et al., "Implicit Quantile Networks for Distributional Reinforcement Learning", *International Conference on Machine Learning*, 2018.

[5] Levine et al., "Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review", *ArXiv*, vol. abs/1805.00909, 2018.

[6] Sun et al., "Tutorial and Survey on Probabilistic Graphical Model and Variational Inference in Deep Reinforcement Learning", *ArXiv*, vol. abs/1908.09381, 2019.

[7] Dayan et al., "Feudal Reinforcement Learning", *Neural Information Processing Systems*, 1992.

[8] Sutton et al., "Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning", *Artificial Intelligence*, 1999.

[9] Vezhnevets et al., "FeUdal Networks for Hierarchical Reinforcement Learning", *ArXiv*, vol. arXiv:1703.01161, 2017.

[10] Kulkarni et al., "Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation", *Arxiv*, vol. abs/1604.060572v2, 2016.

[11] Florenza et al., "Stochastic Neural Networks for Hierarchical Reinforcement Learning", *International Conference on Learning Representations*, 2017.

[12] Haarnoja et al., "Latent Space Policies for Hierarchical Reinforcement Learning", *International Conference on Learning Representations*, 2018.

[13] Li et al., "Sub-policy Adaptation for Hierarchical Reinforcement Learning", *International Conference on Learning Representations*, 2020.

[14] Botvinick et al., "Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective", *Cognition* vol. 113, 2009.

[15] Botvinick et al., "Model-based hierarchical reinforcement learning and human action control", *Philosophical Transactions of The Royal Society, Biological Science*, 2014.

[16] Tomov et al., "Discovery of hierarchical representations for efficient planning", *PLOS, Computational Biology*, vol. 16, 2009.

[17] Xia et al., "Temporal and state abstractions for efficient learning, transfer and composition in humans", *American Psychological Association*, 2021

[18] Barto et al., "Recent Advances in Hierarchical Reinforcement Learning", *Discrete Event Dynamic Systems*, 2003.

[19] Kingma et al., "Auto-Encoding Variational Bayes", *International Conference on Learning Representations*, 2014.

[20] Hochreiter et al., "Long Short-Term Memory", *Neural Computation*, 1997.

[21] Ha et al., "Recurrent World Models Facilitate Policy Evolution", *Neural Information Processing Systems*, 2018.

[22] Hafner et al., "Learning Latent Dynamics for Planning from Pixels", *International Conference on Machine Learning*, 2019.

[23] Hafner et al., "Dream to Control: Learning Behaviors by Latent Imagination", *International Conference on Learning Representations*, 2020.

[24] Hafner et al., "Mastering Atari with Discrete World Models", *International Conference on Learning Representations*, 2021.

[25] Higgins et al., "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework", *International Conference on Learning Representations*, 2017.

[26] Higgins et al., "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework", *International Conference on Learning Representations*, 2017.

[27] Zhao et al., "InfoVAE: Balancing Learning and Inference in Variational Autoencoders", *Association for the Advancement of Artifical Intelligence*, 2018.

[28] Kim et al., "Variational Temporal Abstraction", *Neural Information Processing Systems*, 2019.

[29] Saxena et al., "Clockwork Variational Autoencoders", *Neural Information Processing Systems*, 2021.

[30] Chevalier-Boisvert et al., "Gym-Minigrid", *Github Repository*, 2018. `https://github.com/maximecb/gym-minigrid`

[31] Raffin et al., "Stable Baselines3", *Github Repository*, 2019. `https://github.com/DLR-RM/stable-baselines3`