

Real-World Robot Control by Deep Active Inference with a Temporally Hierarchical World Model

Kentaro Fujii¹ and Shingo Murata¹, *Member, IEEE*

Abstract—Robots in uncertain real-world environments must perform both goal-directed and exploratory actions. However, most deep learning-based control methods neglect exploration and struggle under uncertainty. To address this, we adopt deep active inference, a framework that accounts for human goal-directed and exploratory actions. Yet, conventional deep active inference approaches face challenges due to limited environmental representation capacity and high computational cost in action selection. We propose a novel deep active inference framework that consists of a world model, an action model, and an abstract world model. The world model encodes environmental dynamics into hidden state representations at slow and fast timescales. The action model compresses action sequences into abstract actions using vector quantization, and the abstract world model predicts future slow states conditioned on the abstract action, enabling low-cost action selection. We evaluate the framework on object-manipulation tasks with a real-world robot. Results show that it achieves high success rates across diverse manipulation tasks and switches between goal-directed and exploratory actions in uncertain settings, while making action selection computationally tractable. These findings highlight the importance of modeling multiple timescale dynamics and abstracting actions and state transitions.

Index Terms—Cognitive Control Architectures, Imitation Learning, Representation Learning, World Model, Active Inference

I. INTRODUCTION

WITH recent advances in deep learning-based robot control methods, there is growing expectation for the realization of robots capable of achieving a wide range of human-like goals [1]–[3]. In real-world environments, the presence or arrangement of objects required for a task is often uncertain, and current robots struggle to cope with such uncertainty [4]. In contrast, humans can not only act toward achieving goals but also explore to resolve environmental uncertainty—e.g., by searching for the location of an object—thereby adapting effectively to uncertain situations [5], [6].

To realize robots capable of both goal-directed and exploratory actions, we focus on deep active inference [7]–[10]—a deep learning-based framework grounded in a computational theory that accounts for various cognitive functions [5], [11], [12]. However, deep active inference faces two key

challenges: (1) its performance heavily depends on the capability of the framework to represent environmental dynamics [13], and (2) the computational cost is prohibitively high [9], making it difficult to apply to real-world robots.

To address these challenges, we propose a deep active inference framework comprising a world model, an action model, and an abstract world model. The world model learns hidden state transitions to represent environmental dynamics from human-collected robot action and observation data [14]–[16]. The action model maps a sequence of actual actions to one of a learned set of abstract actions, each corresponding to a meaningful behavior (e.g., moving an object from a dish to a pan) [17]. The abstract world model learns the relationship between the state representations learned by the world model and the abstract action representations learned by the action model [18]. By leveraging the abstract world model and the abstract action representations, the framework enables efficient active inference.

To evaluate the proposed method, we conducted robot experiments in real-world environments with uncertainty. We investigated whether the framework could reduce computational cost, enable the robot to achieve diverse goals involving the manipulation of multiple objects, and perform exploratory actions to resolve environmental uncertainty.

II. RELATED WORK

A. Learning from Demonstration (LfD) for Robot Control

LfD is a method to train robots by imitating human experts, providing safe, task-relevant data for learning control policies [19]–[24]. A key advancement contributing to recent progress in LfD for robotics is the idea of generating multi-step action sequences, rather than only single-step actions [1]–[3], [17], [25]. However, a major challenge in LfD is the difficulty of generalizing to environments with uncertainty, even when trained on large amounts of expert demonstrations [4]. In this work, we focus on the approach that uses quantized features extracted from action sequences [17], and treat the extracted features as abstract action representations.

B. World Model

A world model captures the dynamics of the environment by modeling the relationship between data (observations), their latent causes (hidden states), and actions. They have recently attracted significant attention in the context of model-based reinforcement learning [14], [15], especially in artificial agents and robotics [26]. However, when robots learn using a world model, their performance is constrained by the model’s capability to represent environmental dynamics [27], [28]. In

Manuscript received: June 17, 2025; Revised October 21, 2025; Accepted November 15, 2025.

This paper was recommended for publication by Tetsuya Ogata upon evaluation of the Associate Editor and Reviewers’ comments. This work was supported by JST PRESTO (JPMJPR22C9), JSPS KAKENHI (JP24K03012), and Mori Manufacturing Research and Technology Foundation.

¹Kentaro Fujii and Shingo Murata are with the School of Integrated Design Engineering, Keio University oakwood.n14.4sp@keio.jp, murata@elec.keio.ac.jp

Digital Object Identifier (DOI): see top of this page.

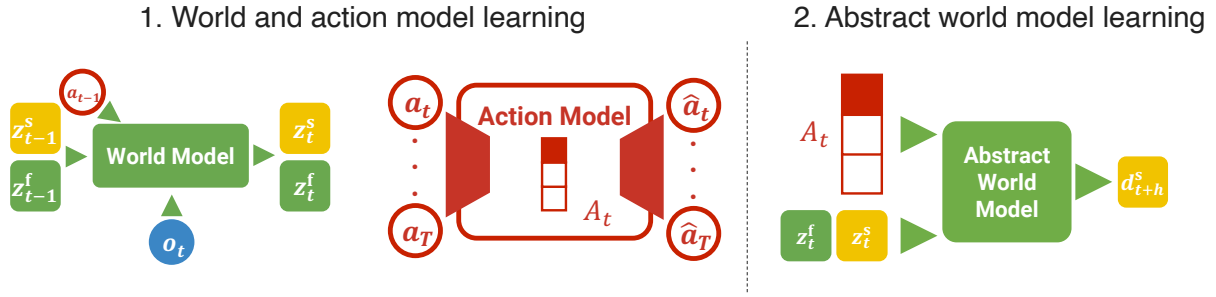


Fig. 1. The overview of the proposed framework. The framework comprises a world model, an action model, and an abstract world model. Here, key variables are visualized: observation o_t and action a_t are processed by the world model to infer hierarchical hidden states z_t^s, z_t^f . The action model compresses action sequences into abstract actions A_t . The abstract world model uses A_t to predict the future slow deterministic state d_{t+h}^s .

particular, learning long-term dependencies in the environment remains a challenge. One solution is to introduce temporal hierarchy into the model structure [27], [29]–[31]. Furthermore, by incorporating abstract action representations that capture slow dynamics, the model can more efficiently predict future observations and states [18]. Temporal hierarchy can be introduced by differentiating state update frequencies [27], [29], [30] or modulating time constants of state transitions [16], [32], [33]. In this work, we adopt the latter to better represent slow dynamics in our world model [31].

III. THE FORMULATION OF ACTIVE INFERENCE

The free-energy principle [5], [6], [11] is a computational principle that accounts for various cognitive functions. According to this principle, human observations o are generated by unobservable hidden states z , which evolve in response to actions a , following a partially observable Markov decision process [5]. The brain is assumed to model this generative process with the world model. Under the free-energy principle, human perception and action aim to minimize the surprise $-\log p(o)$. However, since directly minimizing surprise is intractable, active inference instead minimizes its tractable upper bound, the variational free energy [5], [6].

Perception can be formulated as the minimization of the following variational free energy at time step t [9], [34], [35]:

$$\begin{aligned} \mathcal{F}(t) &= D_{\text{KL}}[q(z_t) \| p(z_t)] - \mathbb{E}_{q(z_t)}[\log p(o_t | z_t)] \\ &\geq -\log p(o_t). \end{aligned} \quad (1)$$

Here, $q(z_t)$ denotes the approximate posterior over the hidden state z_t , $D_{\text{KL}}[q(\cdot) \| p(\cdot)]$ is the Kullback–Leibler (KL) divergence. Note that the first line of (1) is equivalent to the negative evidence lower bound [36], [37].

Action can be formulated as the minimization of expected free energy (EFE), which extends variational free energy to account for future states and observations. Let $\tau > t$ be a future time step. The EFE is defined as follows [35]:

$$\begin{aligned} \mathcal{G}(\tau) &\approx - \underbrace{\mathbb{E}_{q(o_\tau, z_\tau | \pi)}[\log q(z_\tau | o_\tau, \pi) - \log q(z_\tau | \pi)]}_{\text{Epistemic value}} \\ &\quad - \underbrace{\mathbb{E}_{q(o_\tau | \pi)}[\log p(o_\tau | o_{\text{pref}})]}_{\text{Extrinsic value}}. \end{aligned} \quad (2)$$

Here, the expectation is over the observation o_τ because the future observation is not yet available [35], and π indicates

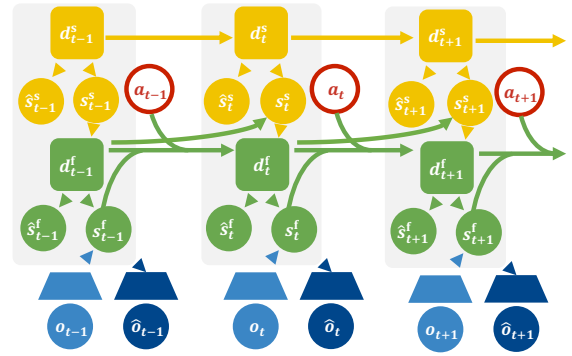


Fig. 2. The world model. It consists of a dynamics model, an encoder, and a decoder. The dynamics model has two different timescales.

the policy (i.e. an action sequence). The variable o_{pref} is referred to as a preference, which encodes the goal, and the distribution $p(o_\tau | o_{\text{pref}})$ is called the prior preference. In (2), the first term referred to as the epistemic value is the mutual information between the state z_τ and the observation o_τ . This term encourages exploratory policies that reduce the uncertainty in the prior belief $q(z_\tau | \pi)$. On the other hand, the second term referred to as the extrinsic value encourages goal-directed policies. Therefore, selecting a policy π that minimizes the EFE can account for both exploratory and goal-directed actions [5], [6], [38].

Conventional active inference requires calculating the EFE over all possible action sequences during task execution, which is intractable for real-world action spaces [6]. Recent works have addressed this by using the EFE as a loss function for training of action generation models [7]–[9], but often ignored exploration capability. In this work, we propose a novel framework focusing on both goal-achievement performance and exploration capability tractably calculating the EFE during task execution.

IV. METHOD

A. Framework

We propose a framework based on deep active inference that enables both goal achievement and exploration. The proposed framework consists of a world model, an action model, and an abstract world model (Fig. 1).

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

1) *World Model*: The world model comprises a dynamics model, an encoder, and a decoder, all of which are trained simultaneously (Fig. 2). As the dynamics model, we utilize a hierarchical model [39], which consists of the slow and fast states as the hidden states $z_t = \{z_t^s, z_t^f\}$ for time step t . Both deterministic d and stochastic s states are defined for each of the slow and fast states $z_t^s = \{d_t^s, s_t^s\}$, $z_t^f = \{d_t^f, s_t^f\}$, respectively. These hidden states are calculated as follows:

Slow dynamics

$$\begin{aligned} \text{Deterministic state:} & \quad d_t^s = f_\theta^s(z_{t-1}^s) \\ \text{Prior:} & \quad \hat{s}_t^s \sim p_\theta^s(s_t^s | d_t^s) \\ \text{Approximate posterior:} & \quad s_t^s \sim q_\theta^s(s_t^s | d_t^s, d_{t-1}^s). \end{aligned}$$

Fast dynamics

$$\begin{aligned} \text{Deterministic State:} & \quad d_t^f = f_\theta^f(s_t^s, z_{t-1}^f, a_{t-1}) \\ \text{Prior:} & \quad \hat{s}_t^f \sim p_\theta^f(s_t^f | d_t^f) \\ \text{Approximate posterior:} & \quad s_t^f \sim q_\theta^f(s_t^f | d_t^f, o_t) \end{aligned} \quad (3)$$

Here, o_t is the observation and a_{t-1} is the action at the previous time step. The approximate posterior of the fast dynamics q_θ^f is conditioned on the observation o_t by receiving its features extracted by the encoder.

The slow and fast deterministic states d_t^s and d_t^f are computed by multiple timescale recurrent neural network parameterized with a time constant [32]. When the time constant is large, the state tends to evolve slowly compared to when the time constant is small. Therefore, by setting the time constant for the slow layer larger than one for the fast layer, the dynamics model represent a temporal hierarchy. The slow and fast stochastic states s_t^s, \hat{s}_t^s and s_t^f, \hat{s}_t^f are represented as one-hot vectors sampled from an approximate posterior or a prior, defined by categorical distributions [40].

The decoder is employed to reconstruct the observation o_t from the hidden state z_t , modeling likelihood $p_\theta(o_t | z_t)$. Simultaneously, a network $p_\theta(d_t^f | z_t^s)$ that predicts the fast deterministic state d_t^f from the slow hidden state z_t^s is also trained. The predicted deterministic state d_t^f is then used to sample the fast stochastic state. By combining both slow and predicted fast hidden states as inputs to the decoder, the dynamics model can represent the observation likelihood $p_\theta(o_t | z_t^s)$ ¹ based on only the slow hidden state z_t^s .

The world model is trained by minimizing the variational free energy $\mathcal{F}(t)$. Here, since the fast deterministic state d_t^f can be regarded as an observation for the slow dynamics, the variational free energies $\mathcal{F}_s(t)$ and $\mathcal{F}_f(t)$ can be computed separately for the slow and fast layers, respectively. Furthermore, we also minimize, as an auxiliary task, the negative log-likelihood of observation o_t given the slow hidden state z_t^s , denoted as $\log p_\theta(o_t | z_t^s)$. In summary, the variational free

energy $\mathcal{F}(t)$ in this work is described as follows:

$$\begin{aligned} \mathcal{F}(t) &= \mathcal{F}_s(t) + \mathcal{F}_f(t) - \log p_\theta(o_t | z_t^s) \\ \mathcal{F}_s(t) &= D_{\text{KL}}[\text{sg}(q_\theta^s(s_t^s | d_t^s, d_{t-1}^s)) \| p_\theta^s(s_t^s | d_t^s)] \\ &\quad - \log p_\theta(d_t^f | z_t^s) \\ \mathcal{F}_f(t) &= D_{\text{KL}}[\text{sg}(q_\theta^f(s_t^f | d_t^f, o_t)) \| p_\theta^f(s_t^f | d_t^f)] \\ &\quad - \log p_\theta(o_t | z_t). \end{aligned} \quad (4)$$

Here, for the KL divergence calculation, we use the KL balancing technique with a weighting factor w [40].

2) *Action Model*: The action model consists of an encoder \mathcal{E}_ϕ and a decoder \mathcal{D}_ϕ composed of multilayer perceptron (MLP), as well as a residual vector quantizer [17], [41], [42] \mathcal{Q}_ϕ with $N_q = 2$ layers. First, the encoder \mathcal{E}_ϕ embeds the action sequence $a_{t:t+h}$ of length h into a low-dimensional feature A_t . Next, the feature A_t is quantized into \hat{A}_t using the residual vector quantizer \mathcal{Q}_ϕ . The residual vector quantizer includes codebooks $\{C_i\}_{i=1}^{N_q}$, each containing K learnable codes $\{c_{i,j}\}_{j=1}^K$. Specifically, the quantized vector at layer i is the code $c_{i,k}$ having the smallest Euclidean distance to the input at layer i . The quantized feature \hat{A}_t is the sum of outputs from each quantization layer $\{\hat{A}_{t,i}\}_{i=1}^{N_q} = \sum_{i=1}^{N_q} c_{i,k}$. Finally, the decoder \mathcal{D}_ϕ reconstructs the quantized feature \hat{A}_t into the action sequence $\hat{a}_{t:t+h}$. In summary, the procedure of the action model is described as follows:

$$\begin{aligned} A_t &= \mathcal{E}_\phi(a_{t:t+h}) \\ \hat{A}_t &= \mathcal{Q}_\phi(A_t) \\ \hat{a}_{t:t+h} &= \mathcal{D}_\phi(\hat{A}_t). \end{aligned} \quad (5)$$

We treat the feature \hat{A}_t , obtained by the action model, as an abstract action representing the action sequence $a_{t:t+h}$.

The encoder \mathcal{E}_ϕ and decoder \mathcal{D}_ϕ of the action model are trained by minimizing the following objective:

$$\begin{aligned} \mathcal{L}_\phi &= \lambda_{\text{MSE}} \|a_{t:t+h} - \hat{a}_{t:t+h}\|_2^2 \\ &\quad + \lambda_{\text{commit}} \sum_{i=1}^{N_q} \left\| (A_t - \sum_i (\hat{A}_{t,i-1})) - \text{sg}(c_{i,k}) \right\|_2^2 \end{aligned} \quad (6)$$

where we assume $\hat{A}_{t,0} = 0$. Moreover, λ_{MSE} and λ_{commit} are coefficients for the reconstruction loss \mathcal{L}_{MSE} and the commitment loss $\mathcal{L}_{\text{commit}}$, respectively. The learning of the codebooks $\{C_i\}_{i=1}^{N_q}$ of the residual vector quantizer \mathcal{Q}_ϕ is performed using exponential moving averages [17], [41].

3) *Abstract World Model*: The abstract world model \mathcal{W}_ψ learns a mapping from the current world model state z_t and an abstract action A_t to the future slow deterministic state d_{t+h}^s . In other words, it provides an abstract representation of state transitions. The model \mathcal{W}_ψ is composed of MLP and takes the abstract action A_t and the current world model state z_t as inputs to predict the slow deterministic state d_{t+h}^s . Here, the input abstract action A_t to \mathcal{W}_ψ can be any of the K^{N_q} combinations of learned codes from the action model, denoted as $\{\hat{A}_n\}_{n=1}^{K^{N_q}}$. Accordingly, for a given current hidden state z_t , the abstract world model \mathcal{W}_ψ predicts K^{N_q} possible future slow deterministic states $\{d_{t+h,n}^s\}_{n=1}^{K^{N_q}}$:

$$\{\hat{d}_{t+h,n}^s\}_{n=1}^{K^{N_q}} = \mathcal{W}_\psi(z_t, \{\hat{A}_n\}_{n=1}^{K^{N_q}}). \quad (7)$$

¹Correctly, this distribution is written as $p_\theta(o_t | z_t^s) = \int p_\theta(o_t | z_t) p_\theta^f(s_t^f | d_t^f) p_\theta(d_t^f | z_t^s) dz_t^f$. We approximate the marginal over the fast states z_t^f with a single Monte Carlo sample.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

The abstract world mode is trained by minimizing the following objective:

$$\mathcal{L}_\psi = \frac{1}{K^{N_q}} \sum_{n=1}^{K^{N_q}} \|\hat{d}_{t+h,n}^s - d_{t+h,n}^s\|_2^2. \quad (8)$$

Here, to obtain the target slow deterministic states $\{d_{t+h,n}^s\}_{n=1}^{K^{N_q}}$, we utilize latent imagination of the world model [15]. To this end, the action sequences $\{\hat{a}_{0:h,n}\}_{n=1}^{K^{N_q}}$ are generated from the code combinations $\{\hat{A}_n\}_{n=1}^{K^{N_q}}$ using the decoder \mathcal{D}_ϕ of the action model. Then, by leveraging the prior distribution over the fast states, the slow deterministic states $\{d_{t+h,n}^s\}_{n=1}^{K^{N_q}}$ at h steps ahead are obtained.

B. Action Selection

To make the EFE $\mathcal{G}(\tau)$ calculation tractable, our framework leverages a learned, finite set of abstract actions $\{\hat{A}_n\}_{n=1}^{K^{N_q}}$, instead of considering all possible (and thus infinite) continuous action sequences.

First, we reformulate (2) in accordance with our world model (for a detailed derivation, see Appendix A):

$$\begin{aligned} \mathcal{G}(\tau) &= -\mathbb{E}_{q_\theta(o_\tau, z_\tau | \pi)} [\log q_\theta(z_\tau | o_\tau, \pi) - \log q_\theta(z_\tau | \pi)] \\ &\quad - \mathbb{E}_{q_\theta(o_\tau | \pi)} [\log p(o_\tau | o_{\text{pref}})] \\ &\approx -\mathbb{E}_{q_\theta(o_\tau, z_\tau | \pi)} [\log q_\theta(s_\tau^f | z_\tau^s, o_\tau) - \log q_\theta(s_\tau^f | z_\tau^s)] \\ &\quad - \mathbb{E}_{q_\theta(o_\tau, z_\tau | \pi)} [\log p(o_\tau | o_{\text{pref}})]. \end{aligned} \quad (9)$$

Here, the joint distribution $q_\theta(o_\tau, z_\tau | \pi)$ can be decomposed as $q_\theta(o_\tau, z_\tau | \pi) = p_\theta(o_\tau | z_\tau) q_\theta(z_\tau^f | z_\tau^s) q_\theta(z_\tau^s | \pi)$ in our proposed framework. Note that, given the distribution $q_\theta(z_\tau^s | \pi)$ over the slow states, all distributions required to compute the EFE $\mathcal{G}(\tau)$ can be obtained using the world model, and thus $\mathcal{G}(\tau)$ becomes computable. Here, we replace the policy π with an abstract action $\hat{A} \in \{\hat{A}_n\}_{n=1}^{K^{N_q}}$, and express the distribution $q_\theta(z_\tau^s | \pi)$ as follows:

$$q_\theta(z_\tau^s | \pi) \approx q_\theta(s_\tau^s | d_\tau^s, \hat{A}) q_\psi(d_\tau^s | \hat{A}). \quad (10)$$

In this way, we can use the abstract world model \mathcal{W}_ψ to predict the slow deterministic state d_τ^s at $\tau = t + h$ from the abstract action \hat{A} . Using the predicted deterministic state d_τ^s , we can obtain the slow prior $q_\theta(z_\tau^s | \pi)$ and compute the EFE. When computing the EFE, the prior preference $p(o_\tau | o_{\text{pref}})$ is assumed to follow a Gaussian distribution $\mathcal{N}(o_{\text{pref}}, \sigma^2)$ with mean o_{pref} and variance σ^2 . Therefore, the EFE can be written as follows:

$$\begin{aligned} \mathcal{G}(\tau) &\approx -\mathbb{E}_{q_\theta(o_\tau, z_\tau | \pi)} [\log q_\theta(s_\tau^f | z_\tau^s, o_\tau) - \log q_\theta(s_\tau^f | z_\tau^s)] \\ &\quad - \mathbb{E}_{q_\theta(o_\tau, z_\tau | \pi)} [-\gamma(o_\tau - o_{\text{pref}})^2], \end{aligned} \quad (11)$$

where $\gamma = 1/2\sigma^2$ is the preference precision, which balances the epistemic and extrinsic values, and the expectations in the EFE are approximated via Monte Carlo sampling [38].

To generate actual robot actions, we first use the abstract world model \mathcal{W}_ψ to predict the slow deterministic states $\{d_{t+h,n}^s\}_{n=1}^{K^{N_q}}$ at h steps into the future for all abstract actions $\{\hat{A}_n\}_{n=1}^{K^{N_q}}$, given the current world model state z_t . Next, we predict slow hidden states $\{z_{t+h,n}^s\}_{n=1}^{K^{N_q}}$ based on the

predicted slow deterministic states by using (10). Then, for each predicted state, we compute the EFE and select the abstract action that yields the minimum EFE. The selected abstract action is then decoded into an action sequence $\hat{a}_{t:t+h}$ by the action model, and the robot executes this sequence.

V. EXPERIMENTS

A. Environment Setup

To investigate whether the proposed framework enables both goal achievement and exploration in real-world environments—where multiple objects can be manipulated and uncertainty arises from their placement—we conducted an experiment using a robot shown in Fig. 4 (left) [43], [44]. The robot had six degrees of freedom, one of which is the gripper. A camera (RealSense Depth Camera D435; Intel) was mounted opposite to the robot to capture a view of both the robot and its environment. From the viewpoint of the camera, a simple dish, a pot, and a pan were placed on the right, center, and left, respectively, and a pot lid was placed closer to the camera than the center pot. Additionally, the environment was configured such that a blue ball, a red ball, or both could be present. Note that, therefore, uncertainty arose when the lid was closed, as the pot might or might not contain a blue or red ball in this environment.

As training data, we collected object manipulation data by demonstrating the predetermined eight patterns of policies (Fig. 4(right)). Each demonstration consists of a sequence of two patterns of policies. For all valid combinations—excluding those in which the policy would result in no movement (e.g., performing action 3 twice in a row)—we collected five demonstrations per combination by teleoperating the robot in a leader–follower manner. There are 36 valid action combinations for environments containing either a blue ball or a red ball, and 72 combinations for environments containing both. Each sequence contains 100 time steps of joint angles and camera images recorded at 5 Hz. Therefore, each pattern of policies had roughly 50 time steps. The original RGB images were captured, resized and clipped to 64×80 . In this experiment, the robot action a_t is defined as the absolute joint angle positions, and the observation o_t is defined as the camera image.

B. Interpretation of the Model Components

In this experiment, we expected the slow hidden states z_t^s to represent the overarching progress of the task, such as where the balls and the lid were placed. In contrast, we expected the fast hidden states z_t^f represents more immediate, transient information. On the other hand, we expected abstract actions A_t to represent a meaningful behavior learned from the demonstration data. In an ideal case, an abstract action corresponds to one of the eight policy patterns in Fig. 4(right), such as moving the ball from the dish to the pan.

C. Experimental Criteria

Capability of abstract world model: We evaluated the capability of the abstract world model. First, we compared

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

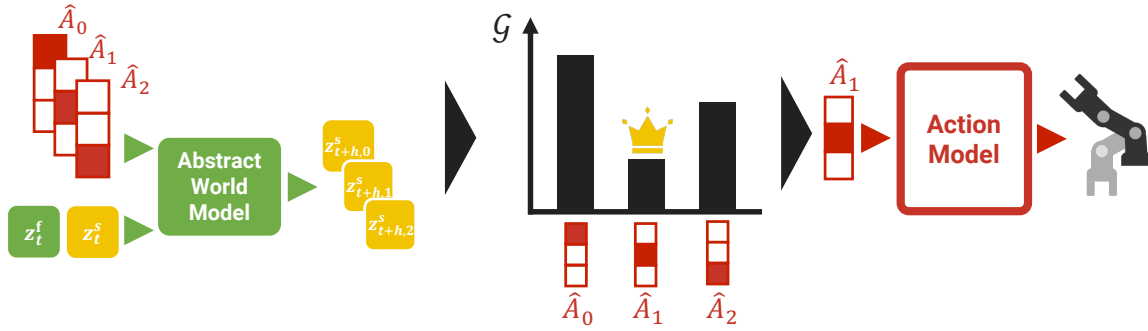


Fig. 3. Action selection based on the minimization of EFE. First, future states are predicted for multiple abstract actions. Then, the EFE is calculated for each of the predicted future states. Finally, the robot execute action sequence reconstructed from the abstract action that yields the lowest EFE.

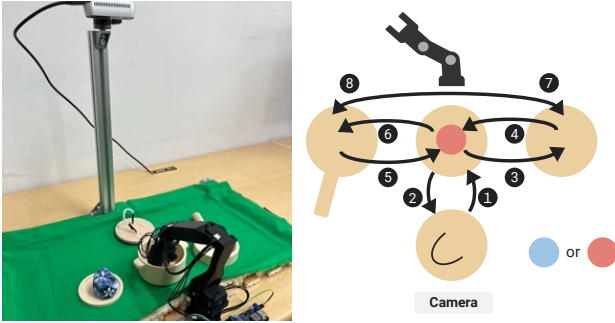


Fig. 4. Experimental environment (left) and policy patterns included in the collected dataset (right). The environment contains either a blue ball, a red ball, or both. The dataset includes demonstrations of eight different policy patterns involving the movement of the lid and the balls.

the computation time of our proposed framework against that of conventional deep active inference approaches [9], [13], [38], which predicts future states with the world model by sequentially inputting the action sequence $\hat{a}_{0:h}$ reconstructed from an abstract action \hat{A} via the action model.

Second, we evaluated whether different predictions can be generated from the same initial state for each abstract action learned by the action model. We also examined whether the observed outcomes resulting from executing actual actions generated from a specific abstract action are consistent with the predictions made by the abstract world model.

Goal achievement performance:

We evaluated the success rate on ball- (140 trials) and lid-manipulation (24 trials) tasks with varying object configurations, such as moving a particular ball or manipulating a lid. A trial was considered successful if the target object was placed in its specified goal position within 50 time steps.

Environment exploration: We evaluated whether the proposed framework can generate not only goal-directed actions but also exploratory actions from an uncertain initial situation. To this end, we set up a scenario in which the blue ball is initially placed in the pan and the lid is closed, creating uncertainty about whether the red ball is present inside the pot. In this scenario, when taking an exploratory action, it was expected that the robot would open the lid to resolve the uncertainty.

D. Baseline and Ablation

In the goal-achievement performance experiment, we compared our proposed framework with a baseline and two ablations described as follows:

- **Goal-conditioned diffusion policy (GC-DP).** As a baseline, we implemented a diffusion policy with a U-Net backbone [1], [45]. In our implementation, this policy predicted a 48-step future actions based on the two most recent observations and a goal observation. To stabilize actions, we apply an exponential moving average of weight 0.7 to the generated actions.
- **Non-hierarchical.** As an ablation study, the world model is replaced by a non-hierarchical dynamics model [40]. In this variant, the hidden state z_t consists of a single-level deterministic state d_t and a stochastic state s_t , where the deterministic state is computed using a gated recurrent unit [46].
- **No abstract world model (AWM).** As an ablation study, the robot does not use the abstract world model for planning. Instead, it calculates the EFE directly over actual action sequences decoded by the action model.

We did not perform an ablation on the action model itself, as our framework relies on it to generate the set of candidate actions (either abstract or actual) for evaluation, making it a core, indispensable component.

VI. RESULTS

A. Capability of abstract world model

Our proposed framework required only 2.37 ms to evaluate all candidate abstract actions, in contrast to 71.8 ms for a sequential evaluation of conventional deep active inference approaches. This demonstrates the higher computational tractability of our proposed framework.

As shown in Fig. 5, different abstract actions lead to distinct predictions. Moreover, for example, by using an action sequence generated from the abstract action represented by $c_{1,2} + c_{2,7}$, the ball was successfully moved from the dish to the pan, consistent with the predicted observation (Fig. 5). These results suggest that the abstract world model has learned the dependency between abstract actions and the resulting state transitions, even without directly referring to actual action sequences. However, the prediction associated with the

A. Predictions by the abstract world model

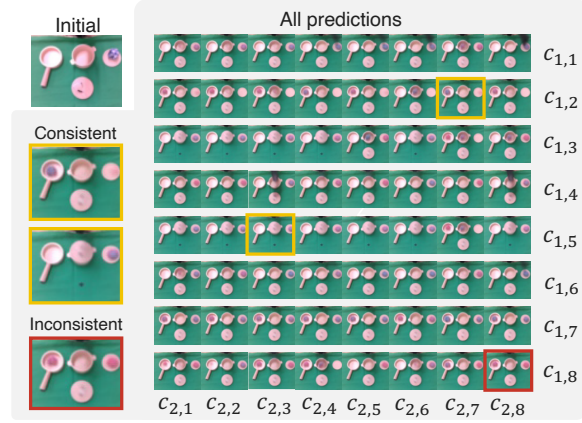
B. Actual robot action ($c_{1,2} + c_{2,7}$)

Fig. 5. Example of predicted observations using the abstract world model and actual robot actions. (A) Predicted observations for each abstract action. Here, each $c_{i,j}$ denotes the j -th code in the i -th layer of the action model. The yellow box highlights an example prediction that is consistent with the initial observation, while the red box indicates an inconsistent prediction. (B) Actual observations corresponding to the action sequence generated from the abstract action \hat{A} represented by $\hat{A} = c_{1,2} + c_{2,7}$ at each time step.

TABLE I
SUCCESS RATE (%).

Manipulation target	Ball		Lid		Total
	Red	Blue	Opening	Closing	
Proposed	61.4	74.3	75.0	100.0	70.7
GC-DP	18.6	25.7	25.0	50.0	24.4
Non-hierarchical	41.4	51.4	75.0	58.3	51.2
No AWM	40.0	21.4	83.3	66.7	37.2

abstract action \hat{A} represented by $\hat{A} = c_{1,8} + c_{2,8}$ in Fig. 5 shows red balls placed on both the dish and the pan, which is inconsistent with the initial condition in which only a blue ball was present. This abstract action corresponded to moving a ball from the center pot to the pan. Since this action was not demonstrated when the pot was empty, the abstract world model may have learned incorrect dependencies for unlearned action–environment combinations.

B. Goal achievement performance

Table I shows the success rates of our proposed framework on goal-directed action generation, evaluated on tasks involving specific ball and lid manipulations. The proposed method outperformed the baseline and the ablations across all goal conditions except the Lid-Opening goal, achieving a total success rate of over 70%. As a qualitative example, Fig. 6 illustrates the EFE calculation for a scenario where the goal is to move a ball from a dish to a pan. The abstract action with the lowest EFE correctly predicts the desired outcome, and executing the actual actions derived from this abstract action led to successful task completion. This overall result

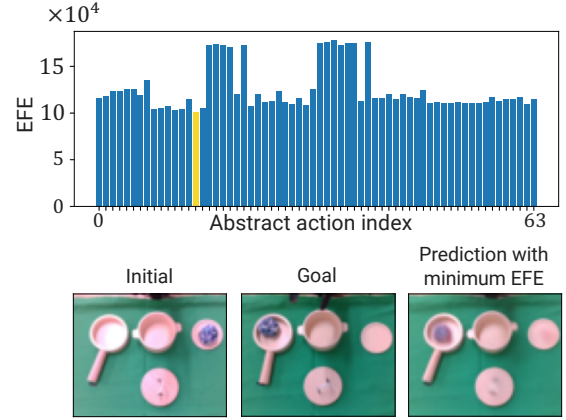


Fig. 6. Example of EFE computed for each abstract action. **Top:** EFE values computed for all 64 abstract actions. The action with the lowest EFE is highlighted as a yellow bar. **Bottom:** From left to right, the images show the initial observation, goal, and predicted observation resulting from the abstract action with the lowest EFE.

TABLE II
EFE VALUES FOR TWO REPRESENTATIVE ABSTRACT ACTIONS
(GOAL-DIRECTED AND EXPLORATORY) IN THE UNCERTAIN SCENARIO.

Preference precision	Goal-directed	Exploratory
$\gamma = 10^2$	4.21×10^4	14.5×10^4
$\gamma = 10^{-4}$	-4.67×10^0	-6.11×10^0

confirms that selecting abstract actions by minimizing the EFE is effective for goal achievement.

The failures in our framework were mainly due to inconsistent world model predictions, which misled the robot into believing an inappropriate action would succeed. For example, the proposed framework selected actions to grasp nothing but place the (non-grasped) target object at the appropriate location. In contrast, the GC-DP, Non-hierarchical, and No AWM all exhibited lower success rates. The GC-DP frequently failed in grasping and placing objects. Both ablations suffered from more prediction inconsistencies than our full model, highlighting the importance of temporal hierarchy and action/state abstraction. The lower performance of the No AWM ablation suggests that action abstraction was a particularly critical component for success.

C. Environment exploration

For simplicity, we computed the EFE for two abstract actions: moving the blue ball from the pan to the dish (goal-directed), and opening the lid (exploratory), as summarized in Table II. When preference precision γ was set to 10^2 , the EFE for the goal-directed action became lower, and thus the robot moved the blue ball from the pan to the dish. In contrast, when preference precision γ was set to 10^{-4} , the EFE for the exploratory action became lower, and thus the robot opened the lid. These results indicate that the proposed framework can assign high epistemic value to exploratory actions that provide new information, and that exploratory actions can be induced by appropriately adjusting the preference precision γ .

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

VII. CONCLUSIONS

In this work, we introduced a deep active-inference framework that combines a temporally-hierarchical world model, an action model utilizing vector quantization, and an abstract world model. By capturing dynamics in a temporal hierarchy and encoding action sequences as abstract actions, the framework makes the action selection based on active inference computationally tractable. Real-world experiments on object-manipulation tasks demonstrated that the proposed framework outperformed the baseline in various goal-directed settings, as well as the ability to switch from goal-directed to exploratory actions in uncertain environments.

Despite these promising results, several challenges remain:

1) The action model used a fixed sequence length, which may not be optimal. 2) The model’s predictive capability decreases for action-environment combinations not present in the dataset. 3) While we validated the capability to take exploratory actions, we did not evaluate their effectiveness in solving tasks and the switching to exploratory behavior still relies on a manually tuned hyperparameter.

Future work will focus on extending the framework to address these limitations. An immediate step is to evaluate our framework in environments that require multi-step action selection and where exploration is necessary to solve the task. Other promising directions include developing a mechanism for adaptive switching between goal-directed and exploratory modes, and extending the action model to represent variable-length action sequences. Ultimately, this work represents a significant step toward the long-term goal of creating more capable robots that can operate effectively in uncertain real-world environments such as household tasks by leveraging both goal-directed and exploratory behaviors.

APPENDIX A EFE DERIVATION

We show the detailed derivation of EFE in our framework:

$$\begin{aligned}
 \mathcal{G}(\tau) &= -\mathbb{E}_{q_{\theta}(o_{\tau}, z_{\tau} | \pi)} [\log q_{\theta}(z_{\tau} | o_{\tau}, \pi) - \log q_{\theta}(z_{\tau} | \pi)] \\
 &\quad - \mathbb{E}_{q_{\theta}(o_{\tau}, z_{\tau} | \pi)} [\log p(o_{\tau} | o_{\text{pref}})] \\
 &= -\mathbb{E}_{q_{\theta}(o_{\tau}, z_{\tau} | \pi)} [\log q_{\theta}(z_{\tau}^f | z_{\tau}^s, o_{\tau}) q_{\theta}(z_{\tau}^s | \pi) \\
 &\quad - \log q_{\theta}(z_{\tau}^f | z_{\tau}^s) q_{\theta}(z_{\tau}^s | \pi)] \\
 &\quad - \mathbb{E}_{q_{\theta}(o_{\tau}, z_{\tau} | \pi)} [\log p(o_{\tau} | o_{\text{pref}})] \\
 &= -\mathbb{E}_{q_{\theta}(o_{\tau}, z_{\tau} | \pi)} [\log q_{\theta}(z_{\tau}^f | z_{\tau}^s, o_{\tau}) - \log q_{\theta}(z_{\tau}^f | z_{\tau}^s)] \\
 &\quad - \mathbb{E}_{q_{\theta}(o_{\tau}, z_{\tau} | \pi)} [\log p(o_{\tau} | o_{\text{pref}})] \\
 &= -\mathbb{E}_{q_{\theta}(o_{\tau}, z_{\tau} | \pi)} [\log q_{\theta}(s_{\tau}^f | d_{\tau}^f, o_{\tau}) q_{\theta}(d_{\tau}^f | z_{\tau}^s) \\
 &\quad - \log q_{\theta}(s_{\tau}^f | d_{\tau}^f) q_{\theta}(d_{\tau}^f | z_{\tau}^s)] \\
 &\quad - \mathbb{E}_{q_{\theta}(o_{\tau}, z_{\tau} | \pi)} [\log p(o_{\tau} | o_{\text{pref}})] \\
 &\approx -\mathbb{E}_{q_{\theta}(o_{\tau}, z_{\tau} | \pi)} [\log q_{\theta}(s_{\tau}^f | z_{\tau}^s, o_{\tau}) - \log q_{\theta}(s_{\tau}^f | z_{\tau}^s)] \\
 &\quad - \mathbb{E}_{q_{\theta}(o_{\tau}, z_{\tau} | \pi)} [\log p(o_{\tau} | o_{\text{pref}})].
 \end{aligned} \tag{12}$$

APPENDIX B ADDITIONAL EXPERIMENTS

To validate the scalability of our framework, we further evaluated our framework on the CALVIN D benchmark [47],

TABLE III
SUCCESS RATE IN CALVIN ENVIRONMET (%).

task	Slider	Drawer	Lightbulb	LED	Total
Proposed	43.8	93.8	0.0	11.8	37.5
GC-DP	1.6	68.3	52.6	16.7	34.8

TABLE IV
HYPERPARAMETERS OF OUR PROPOSED FRAMEWORK

Name	Symbol	Value
World Model		
Training data sequence length	—	75
Slow dynamics		
Deterministic state dimensions	—	32
Stochastic state dimensions × classes	—	4 × 4
Time constant	—	32
Fast dynamics		
Deterministic state dimensions	—	128
Stochastic state dimensions × classes	—	8 × 8
Time constant	—	4
KL balancing	w	0.8
Action Model		
Layers of MLP	—	2
Hidden dimensions of MLP	—	128
Action sequence length	h	50
Codebook size	K	8
Abstract action dimensions	—	32
Learning coefficients	$\lambda_{\text{MSE}}, \lambda_{\text{commit}}$	1.0, 5.0
Abstract World Model		
Layers of MLP	—	2
Hidden dimensions of MLP	—	512

which provides various unstructured human data. Although this environment can serve language goal conditioning, we used only image-based goal conditioning.

For this environment, we compared our proposed framework with the GC-DP. The evaluation was conducted on eight tasks: move_slider_left/right (Slider), open/close_drawer (Drawer), turn_on/off_lightbulb (Lightbulb), and turn_on/off_led (LED). A trial was considered successful if the task was completed within 150 timesteps. Our proposed framework used the same hyperparameters as in our primary experiments, but the GC-DP was trained to predict a 28-step future action sequence from four step observations and re-planned every 16 steps.

As shown in Table III, our proposed method consistently outperformed GC-DP on the Slider and Drawer tasks, as well as on the average success rate across all tasks. These results suggest that our approach, which leverages a temporally hierarchical world model and abstract actions, is robust and effective not only in our primary setup but also in more complex, long-horizon manipulation scenarios.

APPENDIX C HYPERPARAMETERS

We show hyperparameters in our experiments in Table IV.

REFERENCES

- [1] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, “Diffusion policy: Visuomotor policy learning via action diffusion,” *The International Journal of Robotics Research*, p. 02783649241273668, 2023.
- [2] H. Etukuru, N. Naka, Z. Hu, S. Lee, J. Mehu, A. Edsinger, C. Paxton, S. Chintala, L. Pinto, and N. M. M. Shafiqullah, “Robot utility models: General policies for zero-shot deployment in new environments,” *arXiv preprint arXiv:2409.05865*, pp. 1–28, 2024.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

- [3] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, “ π_0 : A vision-language-action flow model for general robot control,” *arXiv preprint arXiv:2410.24164*, pp. 1–17, 2024.
- [4] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence, “Interactive language: Talking to robots in real time,” *IEEE Robotics and Automation Letters (RA-L)*, pp. 1–8, 2023.
- [5] K. Friston, F. Rigoli, D. Ognibene, C. Mathys, T. FitzGerald, and G. Pezzulo, “Active inference and epistemic value,” *Cognitive Neuroscience*, vol. 6, no. 4, pp. 187–214, 2015.
- [6] K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, and G. Pezzulo, “Active inference: A process theory,” *Neural Computation*, vol. 29, pp. 1–49, 2017.
- [7] B. Millidge, “Deep active inference as variational policy gradients,” *Journal of Mathematical Psychology*, vol. 96, p. 102348, 2020.
- [8] Z. Fountas, N. Sajid, P. Mediano, and K. Friston, “Deep active inference agents using monte-carlo methods,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 662–11 675, 2020.
- [9] P. Mazzaglia, T. Verbelen, and B. Dhoedt, “Contrastive active inference,” in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 13 870–13 882.
- [10] K. Fujii, T. Isomura, and S. Murata, “Real-world robot control based on contrastive deep active inference with demonstrations,” *IEEE Access*, vol. 12, pp. 172 343–172 357, 2024.
- [11] K. Friston, “The free-energy principle: a unified brain theory?” *Nature reviews neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.
- [12] P. Schwartenbeck, J. Passacker, T. U. Hauser, T. H. FitzGerald, M. Kronbichler, and K. J. Friston, “Computational mechanisms of curiosity and goal-directed exploration,” *Elife*, vol. 8, p. e41703, 2019.
- [13] N. Sajid, P. Tigas, A. Zakharov, Z. Fountas, and K. Friston, “Exploration and preference satisfaction trade-off in reward-free learning,” *arXiv preprint arXiv:2106.04316*, pp. 1–23, 2021.
- [14] D. Ha and J. Schmidhuber, “Recurrent world models facilitate policy evolution,” in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, 2018, pp. 1–13.
- [15] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, “Learning latent dynamics for planning from pixels,” in *International Conference on Machine Learning*, vol. 97. PMLR, 2019, pp. 2555–2565.
- [16] A. Ahmadi and J. Tani, “A novel predictive-coding-inspired variational rnn model for online prediction and recognition,” *Neural Computation*, vol. 31, no. 11, pp. 2025–2074, 2019.
- [17] S. Lee, Y. Wang, H. Etukuru, H. J. Kim, N. M. M. Shafiqullah, and L. Pinto, “Behavior generation with latent actions,” *arXiv preprint arXiv:2403.03181*, pp. 1–18, 2024.
- [18] C. Gumbsch, N. Sajid, G. Martius, and M. V. Butz, “Learning hierarchical world models with adaptive temporal abstractions from discrete latent dynamics,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [19] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, “Recent advances in robot learning from demonstration,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 297–330, 2020.
- [20] A. Correia and L. A. Alexandre, “A survey of demonstration learning,” *Robotics and Autonomous Systems*, vol. 182, p. 104812, 2024.
- [21] M. Zare, P. M. Kebria, A. Khosravi, and S. Nahavandi, “A survey of imitation learning: Algorithms, recent developments, and challenges,” *IEEE Transactions on Cybernetics*, vol. 54, no. 12, pp. 7173–7186, 2024.
- [22] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson, “Implicit behavioral cloning,” in *Conference on Robot Learning*. PMLR, 2022, pp. 158–168.
- [23] P. Lancaster, N. Hansen, A. Rajeswaran, and V. Kumar, “Modem-v2: Visuo-motor world models for real-world robot manipulation,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 7530–7537.
- [24] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, “Bc-z: Zero-shot task generalization with robotic imitation learning,” in *Conference on Robot Learning*. PMLR, 2022, pp. 991–1002.
- [25] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, “Learning fine-grained bimanual manipulation with low-cost hardware,” in *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023, pp. 1–22.
- [26] T. Taniguchi, S. Murata, M. Suzuki, D. Ognibene, P. Lanillos, E. Ugur, L. Jamone, T. Nakamura, A. Ciria, B. Lara, *et al.*, “World models and predictive coding for cognitive and developmental robotics: frontiers and challenges,” *Advanced Robotics*, vol. 37, no. 13, pp. 780–806, 2023.
- [27] W. Cai, T. Wang, J. Wang, and C. Sun, “Learning a world model with multitimescale memory augmentation,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2022.
- [28] F. Deng, J. Park, and S. Ahn, “Facing off world model backbones: Rnns, transformers, and s4,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 1–27, 2024.
- [29] T. Kim, S. Ahn, and Y. Bengio, “Variational temporal abstraction,” *Advances in Neural Information Processing Systems*, vol. 32, pp. 1–10, 2019.
- [30] V. Saxena, J. Ba, and D. Hafner, “Clockwork variational autoencoders,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 246–29 257, 2021.
- [31] K. Fujii and S. Murata, “Hierarchical latent dynamics model with multiple timescales for learning long-horizon tasks,” in *2023 IEEE International Conference on Development and Learning (ICDL)*, 2023, pp. 479–485.
- [32] Y. Yamashita and J. Tani, “Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment,” *PLoS Computational Biology*, vol. 4, no. 11, p. e1000220, 2008.
- [33] A. Spieler, N. Rahaman, G. Martius, B. Schölkopf, and A. Levina, “The expressive leaky memory neuron: an efficient and expressive phenomenological neuron model can solve long-horizon tasks,” in *The Twelfth International Conference on Learning Representations*, 2024, pp. 1–25.
- [34] P. Mazzaglia, T. Verbelen, O. Çatal, and B. Dhoedt, “The free energy principle for perception and action: A deep learning perspective,” *Entropy*, vol. 24, no. 2, p. 301, 2022.
- [35] R. Smith, K. J. Friston, and C. J. Whyte, “A step-by-step tutorial on active inference and its application to empirical data,” *Journal of Mathematical Psychology*, vol. 107, p. 102632, 2022.
- [36] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” pp. 1–14, 2013.
- [37] D. J. Rezende, S. Mohamed, and D. Wierstra, “Stochastic backpropagation and approximate inference in deep generative models,” in *International Conference on Machine Learning*. PMLR, 2014, pp. 1278–1286.
- [38] K. Igari, K. Fujii, G. W. Haddon-Hill, and S. Murata, “Selection of exploratory or goal-directed behavior by a physical robot implementing deep active inference,” in *5th International Workshop on Active Inference*, 2024, pp. 1–14.
- [39] K. Fujii and S. Murata, “Hierarchical latent dynamics model with multiple timescales for learning long-horizon tasks,” in *2023 IEEE International Conference on Development and Learning (ICDL)*, 2023, pp. 479–485.
- [40] D. Hafner, T. P. Lillicrap, M. Norouzi, and J. Ba, “Mastering atari with discrete world models,” in *International Conference on Learning Representations*, 2021, pp. 1–26.
- [41] A. Van Den Oord, O. Vinyals, *et al.*, “Neural discrete representation learning,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 1–10, 2017.
- [42] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, “Soundstream: An end-to-end neural audio codec,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.
- [43] A. Koch, “Low-cost robot arm,” https://github.com/AlexanderKoch-Koch/low_cost_robot, 2024.
- [44] R. Cadene, S. Alibert, A. Soare, Q. Gallouedec, A. Zouitine, and T. Wolf, “Lerobot: State-of-the-art machine learning for real-world robotics in pytorch,” <https://github.com/huggingface/lerobot>, 2024.
- [45] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.
- [46] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” in *Neural Information Processing Systems 2014 Workshop on Deep Learning*, 2014, pp. 1–9.
- [47] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, “Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks,” *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 3, pp. 7327–7334, 2022.