ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230550

VMBRL3: A Simple Visual Model-Based Reinforcement Learning Framework for Continuous Control

Jian Wang[a], Haitao Wang[a] and Hejun Wu[a;*]

[a]School of Computer Sience and Engineering, Sun Yat-Sen University, Guangdong, China

Abstract. Unsupervised pre-training has demonstrated its potential for accurately constructing world models in visual model-based reinforcement learning (MBRL). However, such MBRL approaches exhibit limited generalizability, thereby limiting their practicality in diverse scenarios. These methods produce models that are restricted to the specific task they were trained on, and are not easily adaptable to other tasks. In this work, we introduce a powerful unsupervised pre-training reinforcement learning (RL) framework called VMBRL3, which improves the generalization ability of visual MBRL. VMBRL3 employs task-agnostic videos to pre-train both the autoencoder and world model without access to actions or rewards information. The fine-tuned world model can then be applied to a range of downstream reinforcement learning tasks, allowing for rapid adaptation to diverse environments and facilitating policy learning. We demonstrate that our framework significantly improves generalization ability in a variety of manipulation and locomotion tasks. Furthermore, VMBRL3 doubles the sample efficiency and overall performance compared to previous visual methods of MBRL.

1 Introduction

Model-based reinforcement learning (MBRL) has become a popular approach for improving sample efficiency in reinforcement learning by utilizing a world model to enable faster learning with fewer interactions with the environment [28, 38]. In real-world applications, using visual images as input to construct task states without requiring expert knowledge can facilitate the development of more robust and adaptable agents capable of learning directly from raw sensory data. However, using visual images as input can increase the need for the agent to interact with the environment. As such, MBRL has emerged as a prevalent method for solving complex visual control tasks, such as robotics control with visual input and video games [27, 46, 23, 9, 20, 8, 7].

The mainstream approach to model construction in MBRL involves unsupervised learning, which has shown promising results for visual control tasks [39, 37]. For instance, Dreamer [13] and DreamerV2 [15] use an autoencoder (AE) to extract features, upon which a world model is built. Moreover, several MBRL methods use unsupervised pre-training and fine-tuning to learn a world model [30, 26]. Specifically, the world model is pre-trained with RL data and then fine-tuned online for the same RL task, leading to improved performance.

Nonetheless, a generalization issue arises in the above-mentioned construction of world models. The learned world model can only be utilized for a specific task and may not generalize well to other tasks. Consequently, this results in inefficient data utilization and significant waste of computational resources. For instance, the representative MBRL algorithm DreamerV2 [15] faces challenges in employing the world model obtained from the source environment to the target environment, as depicted in Figure 1.

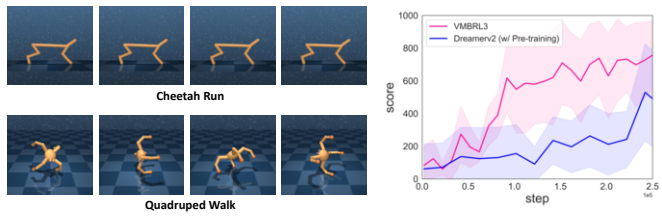


Figure 1: (Left) Some examples of Cheetah Run and Quadruped Walk tasks. (Right) The learning curves of VMBRL3 and DreamerV2, which learned world models on Cheetah Run and fine-tuned on Quadruped Walk, respectively.

A promising solution to overcome the generalization challenge is to adopt a data-driven approach that utilizes all available data to enhance the world model's ability to generalize. Recent breakthroughs in deep learning, exemplified by the impressive development of models such as ChatGPT and GPT-4 [2, 29], have relied heavily on massive amounts of data. While such an approach is well-established in supervised learning, it is less common in RL, where learning typically occurs online. We assume that by leveraging easily accessible task-agnostic data to acquire a good representation and world model, one can augment the generalization ability. Leveraging a good representation, the world model can simulate the environment with greater precision. Then, fine-tuning the resulting model for RL tasks generates concise future state representations, conducive to efficient policy learning.

In this work, we propose a powerful unsupervised pre-training RL framework, Visual Model-based Reinforcement Learning in 3 Stages (VMBRL3), which improves the generalization of visual MBRL. VMBRL3 has the ability to pre-train a world model using task-agnostic data. By subsequently fine-tuning the model during downstream tasks, it is rapidly adaptable to current environmental conditions, resulting in more precise predictions of environmental states and future trajectories for policy learning purposes.

Specifically, **Stage 1** employs unsupervised learning using offline, task-agnostic videos to train an autoencoder that captures fine-grained observation representations. The observation images are projected onto a low-dimensional abstract space, enabling the subse-

* Corresponding Author. wuhejun@mail.sysu.edu.cn

quent learning of a world model based on this space. **Stage 2** continues to utilize the dataset from Stage 1, where action and reward information is omitted. The objective of this stage is to enable the world model to comprehend the dynamics within the low-dimensional abstract space using the unsupervised learning approach. **Stage 3** performs online fine-tuning of the autoencoder and world model in downstream RL tasks. Our primary objective is to attain high performance with minimal online data through enhanced sample efficiency. To this end, we leverage the agent interaction data to fine-tune the autoencoder and world model, while the actors and critics exclusively rely on the updated model for learning.

Our experimental findings validate the successful generalization of VMBRL3 across a variety of tasks. Furthermore, VMBRL3 outperforms existing visual reinforcement learning methods, exhibiting a significant 2.0x improvement in sample efficiency and overall performance. We also provide a comprehensive analysis of our experimental results, highlighting the strengths and weaknesses of our proposed approach and suggesting promising avenues for future research.

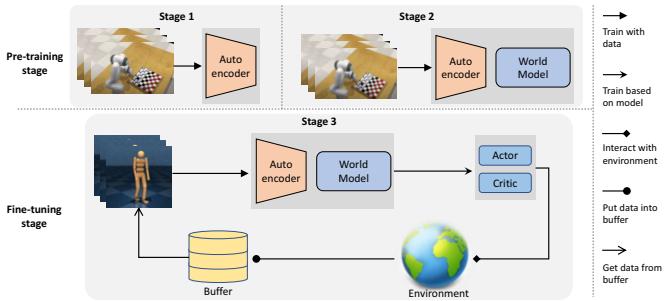


Figure 2: The design of VMBRL3. In stage 1, we leverage task-agnostic videos to pre-train an autoencoder and capture fine-grained observation representations. In stage 2, we use task-agnostic RL transitions without action and reward to construct a task-agnostic world model, using the same dataset as stage 1. In stage 3, we fine-tune the world model using data obtained from agent interactions with downstream RL tasks. Following this, the actor and critic are trained based solely on the updated world model.

2 Related Work

Model-based Reinforcement Learning. In recent years, considerable efforts have been made in the field of deep reinforcement learning to enhance the sampling efficiency of algorithms[45]. Among various research directions, model-based reinforcement learning has gained recognition as a highly promising approach for improving sample efficiency[28, 38]. MBRL can be classified into three main categories based on the method used by the model. The first approach involves planning optimal actions with the model, such as model predictive control (MPC)[4] and Monte Carlo tree search (MCTS)[3, 5, 33, 34]. The second approach uses the model to generate simulated samples for policy learning or value approximation[27, 46, 23, 9, 15, 16]. The third approach involves differentiable dynamic models, which can be used to optimize policy and value networks through differentiation[20, 8, 7]. Our research belongs to the second category, which has demonstrated promising potential in achieving high sample efficiency[23].

Unsupervised Learning. Unsupervised learning is a technique of representation learning that involves using unlabeled data to learn features. It obviates the need for manual labeling by designing pseudo-supervised tasks to improve the generalization and transferability of the learned features. Recent years have witnessed significant advances in unsupervised learning across various domains, including computer vision and natural language processing [19, 11, 6, 10, 18]. In the context of computer vision, the success of representation learning has inspired the application of reinforcement learning to visual tasks [43, 14, 25, 24]. In reinforcement learning, unsupervised representation learning has also been studied to improve the efficiency of sampling [21].

Pre-training in Reinforcement Learning. One of the primary challenges of industrial RL applications is the high computational cost associated with RL training. For instance, replicating the results of AlphaStar [36] may cost millions of dollars [1]. Pre-training can mitigate this challenge by leveraging pre-trained world models or pre-trained representations. For example, IBC [12] is a pre-training method that leverages behavior cloning and model learning to enhance the efficiency and generalization capabilities of RL. Although it is capable of generalizing across various tasks, its performance is constrained by the quality and quantity of available human demonstrations. MVP [40] pre-trains image representations with a large number of offline datasets for subsequent reinforcement learning tasks. MVP uses a multitude of unrelated image pre-training, however, MVP can not get the long sequential information. APV [32] uses action-free videos to pre-train the world model, enabling faster learning in new tasks. Nevertheless, APV neglects state representation learning, resulting in a performance deficit. FIST [17] learns basic skills in the pre-training environment and then adapts to new tasks by adjusting the objective and reward function of the final task. As a result, subsequent tasks and pre-training environments may not be identical, but they should be different variations of the same task distribution. Our work, VMBRL3, leverages task-agnostic time-series data to pre-train both the representation learning module and world model, and then fine-tunes in the downstream RL tasks. Since our representation learning module and world model are exposed to a wider distribution of time-series data while pre-training, it remains effective whatever distribution the policy might induce during the training of the agent.

3 Method

We now present the design of our framework VMBRL3. The overall framework is shown in Figure 2. Both stage 1 and stage 2 pre-training use unsupervised methods. Stage 3 incorporates action and reward information from downstream tasks by fine-tuning the world model.

3.1 Preliminaries

A vision-based control task can be formally defined as a partially observable Markov decision process (POMDP) with a discrete time step denoted by $t \in [1, T]$, represented by a tuple $(\mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. The observation space $\mathcal{O}$ corresponds to the visual image, $\mathcal{A}$ represents the action space, $\mathcal{P}(o_t|o_{<t}, a_{<t})$ refers to the transition dynamics that map the sequence of past observations and actions to the current observation, $\mathcal{R}(r_t|o_{<t}, a_{<t})$ is the reward function that maps the history of observations and actions to the current reward, and $\gamma \in [0, 1)$ is the discount factor. At each time step, the agent performs actions according to the policy $\pi(a_t|o_{\leq t}, a_{<t})$. The objective is to find a policy that maximizes the expected cumulative discounted rewards: $\max_\pi \mathbb{E}\pi[\sum^\infty t = 0[\gamma^t r_t]]$.

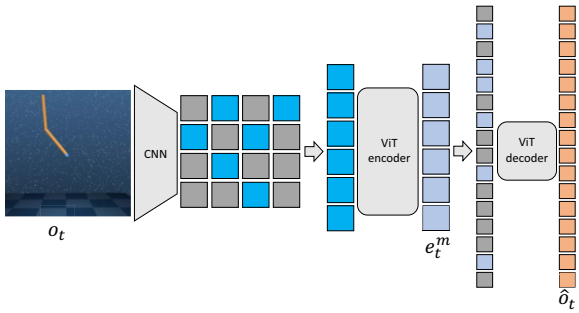


Figure 3: Illustration of masked visual representation. This model can significantly reduce the required training time and efficiently extract feature vectors from visual observations when the mask ratio m is set to a high value. These low-dimensional feature vectors can then be used to learn a world model.

3.2 Stage 1: Masked Visual Representation

In this stage, we use a slightly modified masked autoencoder (MAE) [18] as our backbone network for representation learning. MAE is an unsupervised visual representation technique that can be used to train an autoencoder to reconstruct the original pixels by randomly masking patches of the observed signal. Unlike classical autoencoders, MAE uses an asymmetric design that allows the encoder to operate only on the partial, observed signal (without mask tokens) and a lightweight decoder that reconstructs the full signal from the latent representation and mask tokens. MAE method follows the approach of ViT[11], which involves dividing an observation $o_t \in \mathbb{R}^{H\times W\times C}$ into regular, non-overlapping patches $v_t \in \mathbb{R}^{N\times(K^2C)}$. Here, H, W, and C refer to the height, width, and channel of the observation. K is the patch size and $N = H \times W/K^2$ is the number of patches. A subset of patches is then randomly masked with a ratio of m to reconstruct the input of MAE.

However, prior work has found that using commonly-used pixel patch masking for ViT-based models can make it difficult to capture fine-grained details within patches, such as small objects [31]. To address this limitation, we have incorporated a convolution stem [11, 41] into our approach, which processes the observation through a sequence of convolutional layers followed by a flatten layer, resulting in a feature tensor v_t^c. We then mask v_t^c with a ratio of m ($m = 75\%$ in our experiment) and input it into the ViT encoder to obtain the potential representation e_t^m. By using mask tokens and e_t^m to reconstruct the original pixels of the observation through a ViT decoder, we can learn effective visual representations that capture both high-level and fine-grained features. Note that in the world model learning phase, the masked ratio m is set to 0. As a result, the observation o_t is directly fed into the autoencoder without any masking and gets the latent representation e_t^0. We provide a schematic illustration of our proposed approach in Figure 3.

$$\begin{cases} \text{Convolution stem:} & v_t^c = f_\delta^{CNN}(o_t) \\ \text{ViT encoder:} & e_t^m \sim p_\delta(e_t^m | v_t^c, m) \\ \text{ViT decoder:} & \hat{o}_t \sim p_\delta(\hat{o}_t | e_t^m) \end{cases} \tag{1}$$

The loss function is equation 2.

$$\mathcal{L}(\delta) \doteq E_{p_\delta(e_t^m|v_t^c,m)}[-\ln p_\delta(\hat{o}_t|e_t^m)] \tag{2}$$

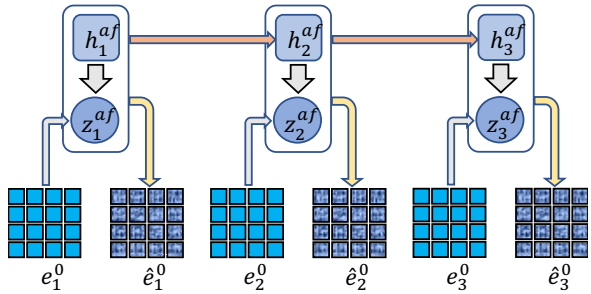


Figure 4: The schematic diagram depicts the task-agnostic latent model, which takes the features generated by the autoencoder with a masked ratio of 0 as input. Based on these features, the model reconstructs the input features and predicts the potential state of the next time step.

3.3 Stage 2: Task-agnostic World Model

To facilitate generalization across diverse tasks, we drew inspiration from APV [32] and performed pre-training of the world model without incorporating action and reward information. This task-agnostic variant of the latent dynamic model[15] comprises four key components: (i) a sequence model that predicts the deterministic state h_t^{tf} at the current moment based on the random state z_{t-1}^{tf} and deterministic state h_{t-1}^{tf} at the previous time step, (ii) a posterior model that predicts the posteriori random state z_t^{tf} from the deterministic state h_t and observation representation e_t^0 generated by masked autoencoder, (iii) a dynamics predictor that predicts the prior random state $\hat{z}_t^{tf}$ from the deterministic state h_t^{tf}, and (iv) a representation predictor that reconstructs the observed representations $\hat{e}_t^0$. The model can be summarized as follow (see equation 3 and Figure 4):

$$\begin{cases} \text{Sequence model:} & h_t^{tf} = f_\phi(h_{t-1}^{tf}, z_{t-1}^{tf}) \\ \text{Posterior model:} & z_t^{tf} \sim q_\phi(z_t^{tf} | h_t^{tf}, e_t^0) \\ \text{dynamics predictor:} & \hat{z}_t^{tf} \sim p_\phi(\hat{z}_t^{tf} | h_t^{tf}) \\ \text{Representation predictor:} & \hat{e}_t^0 \sim p_\phi(\hat{e}_t | h_t^{tf}, z_t^{tf}) \end{cases} \tag{3}$$

This model receives the observation representation from the masked autoencoder as input and maintains a latent state consisting of both deterministic and random components. Specifically, the posterior stochastic state integrates information from the observation, while the priori stochastic state is solely predicted by the deterministic state. During training, our objective is to minimize the difference between the prior and posterior random states. This allows us to efficiently predict the future states in the latent space and avoid the need for a decoder to predict future observation representations during inference. Additionally, we ensure that the representation predictor can reconstruct observation representations from both the deterministic and random states. To achieve our goals, we optimize the model by minimizing the negative variational lower bound, as shown in equation 4, where β_{tf} represents a weight hyperparameter, and T is the length of training sequences in a minibatch.

$$\mathcal{L}(\phi) \doteq E_{q_\phi(z_{1:T}^{tf}|e_{1:T}^0)}\left[\sum_{i=1}^{T}\left(\underbrace{-\ln p_\phi(e_t^0|h_t^{tf}, z_t^{tf})}_{\text{Representation predictor loss}} + \beta_{tf}\underbrace{\text{KL}\left[q_\phi(z_t^{tf}|h_t^{tf}, e_t^0) \parallel p_\phi(\hat{z}_t^{tf}|h_t^{tf})\right]}_{\text{Task-agnostic model loss}}\right)\right] \tag{4}$$

3.4 Stage 3: Fine-tuning in Downstream RL Tasks

To adapt a pre-trained task-agnostic prediction model for various visual control tasks, we fine-tune it into a task-conditional prediction model by incorporating additional information such as actions and rewards during fine-tuning. While a straightforward approach is to initialize a task-conditional model with the pre-trained task-agnostic model and add a reward predictor, this method tends to erase valuable information from the pre-training. To address this, we propose a stacked architecture that combines a task-conditional model with the pre-trained task-agnostic model, as depicted in Figure 5. The task-conditional model is defined by Equation 5.

The task-conditional model displays significant similarities to the task-agnostic model in terms of its dynamics predictor and representation predictor. However, a key difference is that the sequence model in the task-conditional model includes action information as an additional input. Another significant difference is that the posterior model input in the task-conditional model is based on the state of the task-agnostic model (h_t^{tf}, z_t^{tf}) rather than the observation representation e_t^0. Additionally, the task-conditional model integrates a reward predictor to enable reward prediction during data imagining. During the inference phase, the sequence model within the task-conditional model is utilized to forecast future states within the latent space. This provides an effective means of leveraging pre-trained information while simultaneously tailoring the model to specific tasks.

$$\begin{cases} \text{Sequence model:} & h_t = f_\theta(h_{t-1}, z_{t-1}, a_{t-1}) \\ \text{Posterior model:} & z_t \sim q_\theta(z_t | h_t, h_t^{tf}, z_t^{tf}) \\ \text{dynamics predictor:} & \hat{z}_t \sim p_\theta(\hat{z}_t | h_t) \\ \text{Representation predictor:} & \hat{e}_t^0 \sim p_\theta(\hat{e}_t^0 | h_t, z_t) \\ \text{Reward predictor:} & \hat{r}_t \sim p_\theta(\hat{r}_t | h_t, z_t) \end{cases} \tag{5}$$

The loss function during fine-tuning in target tasks is as follows:

$$\begin{aligned} \mathcal{L}(\phi, \theta) \doteq E_{q_\theta(z_{1:T} | e_{1:T}^0, a_{1:T}), q_\phi(z_{1:T}^{tf} | e_{1:T}^0)} \Bigg[\\ \sum_{i=1}^{T} \Bigg(\underbrace{-\ln p_\theta(e_t^0 | h_t, z_t)}_{\text{Representation predictor loss}} \underbrace{-\ln p_\theta(r_t | h_t, z_t)}_{\text{Reward predictor loss}} \\ + \beta_{pre} \underbrace{\mathrm{KL}\left[q_\phi(z_t^{tf} | h_t^{tf}, e_t^0) \parallel p_\phi(\hat{z}_t^{tf} | h_t^{tf}) \right]}_{\text{Task-agnostic model loss}} \\ + \beta \underbrace{\mathrm{KL}\left[q_\theta(z_t | h_t, h_t^{tf}, z_t^{tf}) \parallel p_\theta(\hat{z}_t | h_t) \right]}_{\text{Task-conditional model loss}} \Bigg) \Bigg] \end{aligned} \tag{6}$$

Here, we initialize the representation predictor $p_\theta(\hat{e}_t^0 | h_t, z_t)$ with a pre-trained predictor $p_\phi(\hat{e}_t | h_t^{tf}, z_t^{tf})$. Excpet for the sequence model that employs the GRU network, the remaining components are built using the MLP network. During fine-tuning, we set $\beta_{tf} = 0$ in our experiments and only use the task-conditional model for future imagination. For behavior learning, we predict future states using only the task-conditional model. We adopt the actor-critic learning approach proposed in DreamerV2[15] which involves learning values using imagined rewards from future hypothetical states, and a policy that maximizes the values.

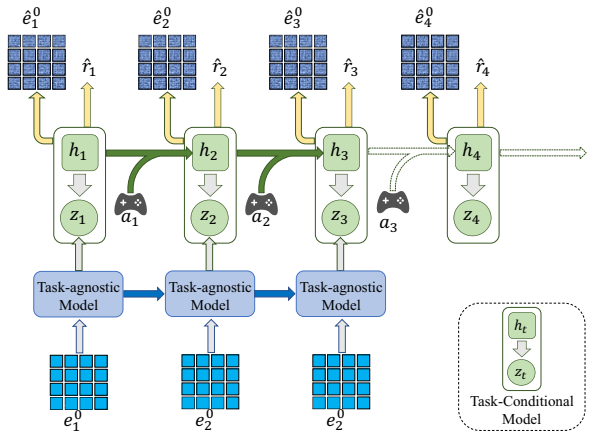


Figure 5: The schematic diagram shows the process of fine-tuning. Specifically, the state of task-agnostic models (h_t^{tf}, z_t^{tf}) and actions serve as inputs, yielding the state of the task-conditional model (h_t, z_t). Subsequently, based on (h_t, z_t), predictions are made with respect to observation representation and reward. Rather than relying on observation, the state of the task-conditional model is employed in subsequent policy learning. Notably, arrows delineated with dotted lines signify model-based forecasts of future states.

4 Experiments

To verify the generalization ability of our VMBRL3, we leverage videos from RLBench[22] for pre-training and evaluated its performance on a diverse range of tasks in various domains, including DeepMind Control suite [35] and Meta-world [44]. As the optimal performance of algorithms varies across different domains and tasks, we compare our approach with state-of-the-art algorithms specific to their respective domains. In addition, we performed ablation experiments to further analyze the individual components of our model.

Implementation Details. We use visual observations of $64 \times 64 \times 3$. For the convolution stem, we use a stack of three convolutional layers with a kernel size of 4 and stride of 2, followed by a linear projection layer. Our approach employs a 4-layer Vision Transformer (ViT) encoder and a 3-layer ViT decoder. In both task-agnostic and task-conditional world models, we built our implementation based on DreamerV2[15]. To accept a sequence of autoencoder representations as inputs, we replace the convolutional neural network (CNN) encoder and decoder with a 2-layer Transformer encoder and decoder.

During the pre-training stage, we utilize 0 vectors to mark all actions, which removes any action information and eliminates the reward prediction module. For the fine-tuning phase, we set the masked ratio to 0 and freeze the parameters of the task-agnostic model. We combine the state of the task-agnostic model with observation representation to use as input for the task-conditional model. During the policy learning phase, we randomly sample a batch of initial states from the replay buffer, and then use the task-conditional model and current policy to predict the next 15 steps. The resulting imagined data is then utilized to train the value and policy networks. We use the same loss function as DreamerV2[15] for both the policy and value networks, and use identical hyperparameters for each benchmark.

Pre-training Dataset. We employ pre-training data consisting of videos from robotic manipulation tasks in RLBench[22]. This dataset comprises 7 $\sim$ 10 demonstrations rendered utilizing five camera views within 99 tasks, resulting in a total of 3,789 videos. We train

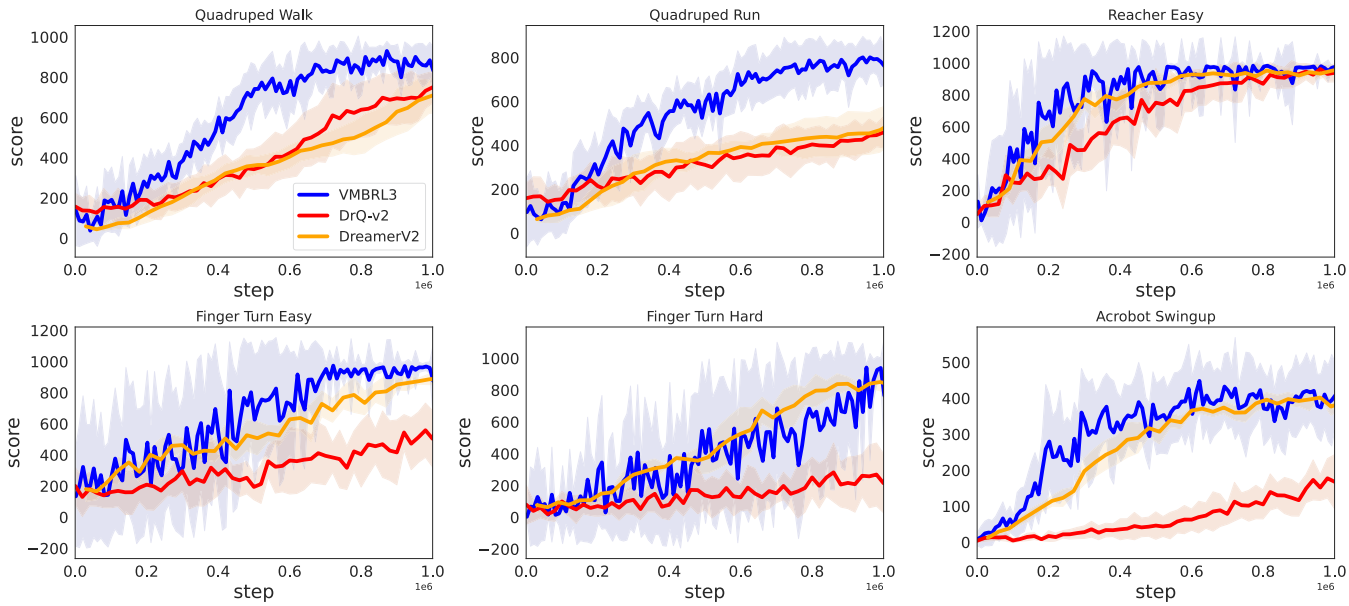


Figure 6: Our experimental results on the medium-level tasks of the DeepMind Control suite demonstrate that VMBRL3 consistently outperforms DreamerV2 and DrQ-v2 in terms of sample efficiency and performance, with only 1e6 environment steps used in our experiments.

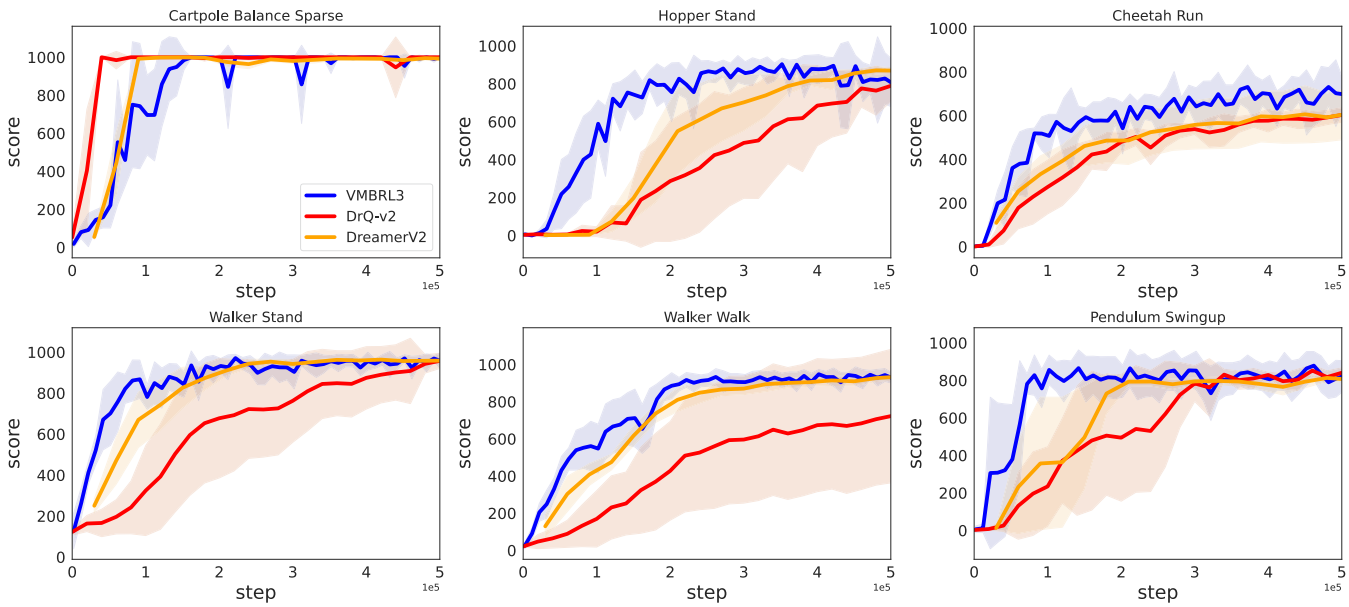


Figure 7: The results of our experiments on the easy tasks of the DeepMind Control suite are presented in the learning curve. We interacted with 5e5 environment steps on these tasks and found that VMBRL3 outperformed the other methods.

the task-agnostic video prediction model by minimizing the objective specified in Equation 4 over 600,000 gradient steps.

4.1 DeepMind Control Suite Experiments

Our evaluation is performed on a diverse range of tasks extracted from the DeepMind Control suite, a commonly adopted benchmark utilized for continuous visual control tasks. Within the task of this benchmark, we define the trajectory length to be 1000 steps, action repetitions to be 2, and observations to consist of $64 \times 64 \times 3$ images. For each task, we employ three random seeds and evaluate the performance of the current policy 5 times every 1000 steps. Dur-

Table 1: Comparison methods are shown. The techniques listed are categorized into model-free RL methods (MFRL) and model-based RL methods (MBRL). Additionally, the table presents the methods adopted for learning the model, as well as the number of parameters utilized. The parameter count for VMBRL3 is 36M.

	Classification	Model building method	Parameter Count
DreamerV2	MBRL	RSSM	22M
DrQ-v2	MFRL	None	7M
APV	MBRL	Action-free/conditional world model	45M
MWM	MBRL	Masked world model	24M

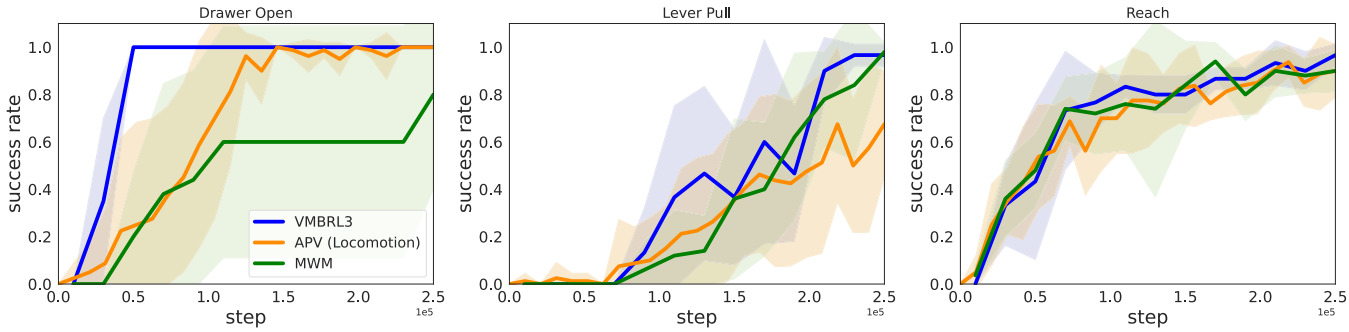


Figure 8: The learning curve obtained from our experiments on Meta-world demonstrates that VMBRL3 exhibits superior sample efficiency compared to APV and MWM. The solid line represents the mean success rate, while the shaded regions indicate the confidence intervals, obtained from three independent runs. At every 1×10^4 time step, we evaluated the current policy 10 times and computed the success rate ($\frac{\text{the number of successes}}{10}$).

ing our evaluation, we compare our proposed approach against several established baseline models, including DreamerV2[15], a state-of-the-art model-based reinforcement learning algorithm specifically designed for visual control tasks. Furthermore, we also compare our approach against DrQ-v2[42], a state-of-the-art model-free method that has been optimized for application within DeepMind Control.

Experimental Results. We adopt the same approach as DrQ-v2 to classify the DMC tasks into three levels of difficulty, and evaluate our proposed method against DreamerV2 and DrQ-v2 separately on the easy and medium difficulty levels. The results for the easy tasks are presented in Figure 7, and those for the medium tasks are shown in Figure 6.

Our experimental results demonstrate that VMBRL3 achieves comparable performance with fewer training steps compared to state-of-the-art model-free and model-based approaches. In certain tasks, it achieves near double the performance of DrQ-v2 and Dreamerv2 under the same number of environment steps. These experimental results provide evidence that VMBRL3 effectively generalizes the prior knowledge acquired from pretraining to downstream tasks, thus enhancing the learning speed in those tasks. Specifically, VMBRL3 leverages pretraining on task-agnostic data to acquire prior knowledge about environmental latent state changes. With the presence of prior knowledge, VMBRL3 facilitates the rapid learning of transition probabilities in the potential state space of downstream tasks, thereby promoting more effective learning of policies and value functions.

4.2 *Meta-world Experiments*

We conducted additional experiments to evaluate the performance of VMBRL3 on a variety of vision-based robotic manipulation tasks derived from the Meta-world environment. Each manipulation task had an episode length of 500 steps with an action repeat of 2, and we executed three random seeds for each task. The performance of the current policy was evaluated ten times for every 10,000 steps. For this benchmark evaluation, we compared our proposed method against APV[32] and MWM[31], which are currently the best performing methods on Meta-world tasks.

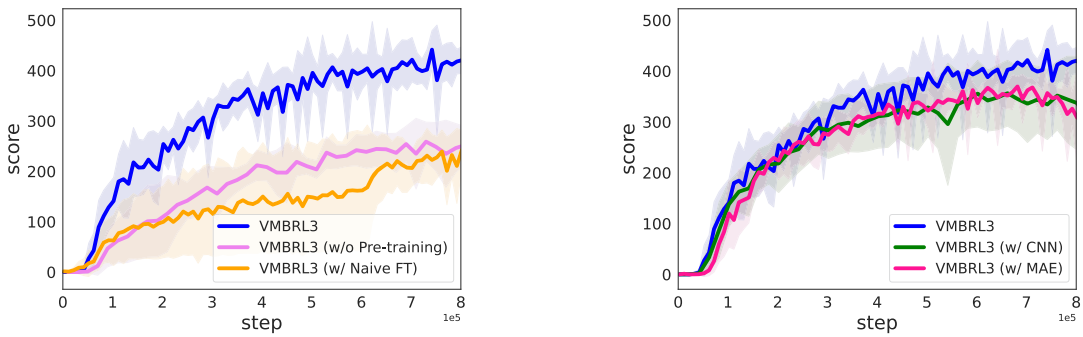
Experimental Results. In Figure 8, success rates are presented for the three algorithms instead of scores. Our experimental results demonstrate that our proposed method achieves superior sampling efficiency and more competitive performance compared to the baseline algorithms. For instance, in the Drawer Open task, VMBRL3 achieves a 100% success rate at the 30,000th step, while the other two baselines require more steps to achieve similar performance, with greater variance and less stability. Similarly, in the Lever Pull task, VMBRL3 achieves almost a 100% success rate while APV only achieves approximately 60%. These task results further validate the generalization capability of our proposed framework across diverse tasks. Additionally, these experimental findings demonstrate that the acquired prior knowledge through pretraining encapsulates general patterns of state transitions within the potential state space. This knowledge can be effectively applied to various tasks through simple fine-tuning, resulting in significant performance improvements and enhanced sample efficiency.

4.3 *Ablation Experiments*

To assess the effectiveness of the techniques proposed in the VMBRL3, we conduct several ablation experiments. In the continuous control task of DMC, we compare VMBRL3 with a naive fine-tuning scheme (VMBRL3 (w/ Naive FT)) that initializes the task-conditional model using the pre-trained parameters of the task-agnostic model. Furthermore, we compare VMBRL3 with a scheme that directly trains the task-conditional model without pre-training in downstream tasks (VMBRL3 (w/o Pre-training)). The results illustrated in Figure 9a demonstrate that the unsophisticated fine-tuning approach yields similar results to the curve when no pre-training is employed. This phenomenon exemplifies the rapid degradation of pre-trained knowledge through this method. On the other hand, VMBRL3 effectively leverages the acquired prior knowledge from pre-training and successfully applies it to downstream tasks.

We also conduct a performance comparison between using a CNN (VMBRL3 (w/ CNN)) and the original MAE method (VMBRL3 (w/ MAE)) to process images. As depicted in Figure 9b, the proposed MVBRL3 outperforms both CNN and the original MAE method. Additionally, due to the relatively simplistic nature of the DMC task image, the CNN network can attain performance comparable to the original MAE method. In contrast, VMBRL3 demonstrates enhanced performance by effectively capturing the intrinsic details of visual images.

In addition, we conducted a comparison of the impact of various pre-training datasets on the VMBRL3 performance. For this purpose, we collected 1800 videos from 7 tasks in DMC as a pre-training dataset. The task-agnostic video prediction model was trained by minimizing the objective function presented in Equation 4 over 600,000 gradient steps. Figure 10 presents the experimental results. Pre-training using the same task set can improve sample efficiency



(a) Abaltion of pre-training

(b) Abaltion of representation

Figure 9: The ablation analysis of VMBRL3. (a) Comparing the performance of VMBRL3 without pre-training and with naive pre-training. (b) Comparing the performance of VMBRL3 with CNN and with original MAE.

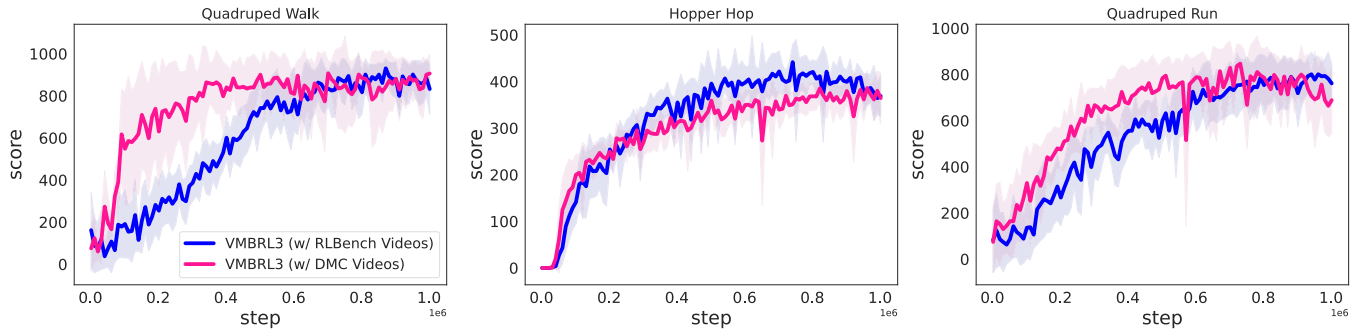


Figure 10: The learning curves were obtained for the DMC task when pre-trained using data from both RLBench and DMC. It was observed that pre-training on the same task led to modest improvements in performance.

to some extent, but the improvement is not significant. When dealing with tasks with high sampling costs or sampling difficulties, pre-training with VMBRL3 on unrelated task datasets can achieve similar performance.

5 Conclusion

In summary, VMBRL3 effectively utilizes unsupervised pre-training to learn the world model and exhibits strong generalization capabilities across diverse tasks. Moreover, VMBRL3 outperforms prior methods in terms of sampling efficiency and task performance on a range of vision-based control tasks. The compact feature extraction and cross-task pre-training of our approach enable faster and more accurate model learning, enhance sample efficiency, and reduce the number of required environment interactions. Our experimental results indicate that our method represents a promising direction for MBRL in vision-based tasks, and we anticipate that this work will inspire further research in this area.

Moving forward, we aim to explore methods for enabling models to acquire task-specific information in natural contexts. Our current observation representation framework uses pixel-level reconstruction as the optimization target, which treats all image regions equally. In natural environments, it is essential to disregard task-irrelevant information and instead prioritize task-related information, even at the expense of reconstruction accuracy.

Acknowledgements

We would like to express our heartfelt gratitude to Professor H. Wu for his exceptional guidance and continuous support throughout this research. We are also immensely grateful to Mr. Wang for his valuable insights and constructive suggestions. Additionally, we acknowledge the support provided by the National Natural Science Foundation of China (Grant No. 62272497 to H. Wu) and the Science and Technology Program of Guangzhou, China (No. 202002020045). We sincerely appreciate the contributions of all individuals and organizations who have played a significant role in the successful completion of this conference paper.

References

[1] Alekh Agarwal, Yuda Song, Wen Sun, Kaiwen Wang, Mengdi Wang, and Xuezhou Zhang, 'Provable benefits of representational transfer in reinforcement learning', *arXiv preprint arXiv:2205.14571*, (2022).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., 'Language models are few-shot learners', *Advances in neural information processing systems*, (2020).

[3] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton, 'A survey of monte carlo tree search methods', *IEEE Transactions on Computational Intelligence and AI in games*, (2012).

[4] Eduardo F Camacho and Carlos Bordons Alba, *Model predictive control*, Springer science & business media, 2013.

[5] Guillaume Chaslot, Sander Bakkes, Istvan Szita, and Pieter Spronck, 'Monte-carlo tree search: A new framework for game ai', in *Proceed-*

ings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, (2008).

[6] Xinlei Chen, Saining Xie, and Kaiming He, 'An empirical study of training self-supervised vision transformers', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021).

[7] Jonas Degrave, Michiel Hermans, Joni Dambre, et al., 'A differentiable physics engine for deep learning in robotics', *Frontiers in neurorobotics*, (2019).

[8] Marc Deisenroth and Carl E Rasmussen, 'Pilco: A model-based and data-efficient approach to policy search', in *Proceedings of the 28th International Conference on machine learning (ICML-11)*, (2011).

[9] Fei Deng, Ingook Jang, and Sungjin Ahn, 'Dreamerpro: Reconstruction-free model-based reinforcement learning with prototypical representations', in *International Conference on Machine Learning*, (2022).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (2019).

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., 'An image is worth 16x16 words: Transformers for image recognition at scale', in *9th International Conference on Learning Representations*. OpenReview.net, (2021).

[12] Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson, 'Implicit behavioral cloning', in *Conference on Robot Learning*, pp. 158–168. PMLR, (2022).

[13] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi, 'Dream to control: Learning behaviors by latent imagination', (2020).

[14] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson, 'Learning latent dynamics for planning from pixels', in *International conference on machine learning*, (2019).

[15] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba, 'Mastering atari with discrete world models', in *9th International Conference on Learning Representations*, (2021).

[16] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap, 'Mastering diverse domains through world models', *arXiv preprint arXiv:2301.04104*, (2023).

[17] Kourosh Hakhamaneshi, Ruihan Zhao, Albert Zhan, Pieter Abbeel, and Michael Laskin, 'Hierarchical few-shot imitation with skill transition models', *arXiv preprint arXiv:2107.08981*, (2021).

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, 'Masked autoencoders are scalable vision learners', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022).

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, 'Momentum contrast for unsupervised visual representation learning', in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (2020).

[20] Nicolas Heess, Gregory Wayne, David Silver, Timothy Lillicrap, Tom Erez, and Yuval Tassa, 'Learning continuous control policies by stochastic value gradients', (2015).

[21] Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu, 'Reinforcement learning with unsupervised auxiliary tasks', *arXiv preprint arXiv:1611.05397*, (2016).

[22] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison, 'Rlbench: The robot learning benchmark & learning environment', *IEEE Robotics and Automation Letters*, (2020).

[23] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine, 'When to trust your model: Model-based policy optimization', (2019).

[24] Michael Laskin, Aravind Srinivas, and Pieter Abbeel, 'Curl: Contrastive unsupervised representations for reinforcement learning', in *International Conference on Machine Learning*, (2020).

[25] Alex X Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine, 'Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model', (2020).

[26] Hao Liu and Pieter Abbeel, 'Unsupervised active pre-training for reinforcement learning', (2021).

[27] Yuping Luo, Huazhe Xu, Yuanzhi Li, Yuandong Tian, Trevor Darrell, and Tengyu Ma, 'Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees', in *7th International Conference on Learning Representations*, (2019).

[28] Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al., 'Model-based reinforcement learning: A survey', *Foundations and Trends® in Machine Learning*, (2023).

[29] OpenAI, 'GPT-4 technical report', *CoRR*, (2023).

[30] Max Schwarzer, Nitarshan Rajkumar, Michael Noukhovitch, Ankesh Anand, Laurent Charlin, R Devon Hjelm, Philip Bachman, and Aaron C Courville, 'Pretraining representations for data-efficient reinforcement learning', *Advances in Neural Information Processing Systems*, (2021).

[31] Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel, 'Masked world models for visual control', (2023).

[32] Younggyo Seo, Kimin Lee, Stephen L James, and Pieter Abbeel, 'Reinforcement learning with action-free pre-training from videos', in *International Conference on Machine Learning*. PMLR, (2022).

[33] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al., 'Mastering the game of go with deep neural networks and tree search', *nature*, (2016).

[34] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al., 'Mastering the game of go without human knowledge', *nature*, (2017).

[35] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al., 'Deepmind control suite', *arXiv preprint arXiv:1801.00690*, (2018).

[36] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al., 'Grandmaster level in starcraft ii using multi-agent reinforcement learning', *Nature*, (2019).

[37] Niklas Wahlström, Thomas B Schön, and Marc Peter Deisenroth, 'From pixels to torques: Policy learning with deep dynamical models', *arXiv preprint arXiv:1502.02251*, (2015).

[38] Tingwu Wang, Xuchan Bao, Ignasi Clavera, Jerrick Hoang, Yeming Wen, Eric Langlois, Shunshi Zhang, Guodong Zhang, Pieter Abbeel, and Jimmy Ba, 'Benchmarking model-based reinforcement learning', *arXiv preprint arXiv:1907.02057*, (2019).

[39] Manuel Watter, Jost Springenberg, Joschka Boedecker, and Martin Riedmiller, 'Embed to control: A locally linear latent dynamics model for control from raw images', *Advances in neural information processing systems*, (2015).

[40] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik, 'Masked visual pre-training for motor control', *arXiv preprint arXiv:2203.06173*, (2022).

[41] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick, 'Early convolutions help transformers see better', (2021).

[42] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto, 'Mastering visual continuous control: Improved data-augmented reinforcement learning', in *The Tenth International Conference on Learning Representations*, (2022).

[43] Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus, 'Improving sample efficiency in model-free reinforcement learning from images', in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2021).

[44] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine, 'Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning', in *Conference on robot learning*, (2020).

[45] Yang Yu, 'Towards sample efficient reinforcement learning.', in *IJCAI*, (2018).

[46] Qi Zhou, HouQiang Li, and Jie Wang, 'Deep model-based reinforcement learning via estimated uncertainty and conservative policy optimization', in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2020).