# MOSER: Learning Sensory Policy for Task-specific Viewpoint via View-conditional World Model

**Shenghua Wan**, **Hai-Hang Sun**, **Le Gan** and **De-Chuan Zhan***
School of Artifcial Intelligence, Nanjing University, China
National Key Laboratory for Novel Software Technology, Nanjing University, China
{wansh,sunhh}@lamda.nju.edu.cn, {ganle,zhandc}@nju.edu.cn

## Abstract

Reinforcement learning from visual observations is a challenging problem with many real-world applications. Existing algorithms mostly rely on a single observation from a well-designed fixed camera that requires human knowledge. Recent studies learn from different viewpoints with multiple fixed cameras, but this incurs high computation and storage costs and may not guarantee the coverage of the optimal viewpoint. To alleviate these limitations, we propose a straightforward View-conditional Partially Observable Markov Decision Processes (VPOMDPs) assumption and develop a new method, the **MO**del-based **SE**nsor controlle**R** (**MOSER**). MOSER jointly learns a view-conditional world model (VWM) to simulate the environment, a sensory policy to control the camera, and a motor policy to complete tasks. We design intrinsic rewards from the VWM without additional modules to guide the sensory policy to adjust the camera parameters. Experiments on locomotion and manipulation tasks demonstrate that MOSER autonomously discovers task-specific viewpoints and significantly outperforms most baseline methods.

## 1 Introduction

Deep reinforcement learning (RL) has achieved remarkable successes in challenging domains, from game-playing [Mnih *et al.*, 2015; Schrittwieser *et al.*, 2020] to drug discovery [Pereira *et al.*, 2021]. Many RL works now focus on learning skills directly from visual inputs like images and videos, which are common in real-world settings such as robotics manipulations [Zhu *et al.*, 2020; Zhan *et al.*, 2021] and autonomous driving [Hu *et al.*, 2022]. Typically, autonomous agents acquire visual observations through static, fixed cameras to train policies [Levine *et al.*, 2018]. In contrast, humans dynamically adjust their eye movements to find optimal viewpoints that provide the most helpful information to facilitate task completion [Dodge, 1903].

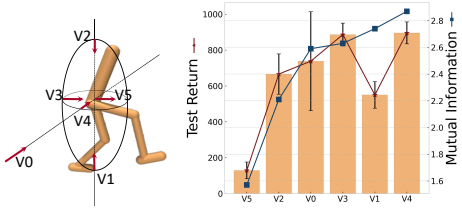



Figure 1: Left: Illustration of six different camera viewpoints of the Walker2d task. Right: The relationship between the test return and the mutual information $I(O;S)$ for different viewpoints. We collect 20 trajectories with the final policy trained on the Walker2d task from each viewpoint and record the mean and standard deviation of the test returns and the mutual information for each. Policies trained on different viewpoints have different test performances. The Walker2d task favors viewpoints 3 and 4, which capture more details of the agent's legs compared to other views.

In most reinforcement learning environments, camera parameters are predefined by the designers without much concern about their explicit impact on subsequent learning. Different views provide varying amounts of task-relevant information, thus optimal viewpoint design requires effort and depends on expert knowledge. To provide a straightforward intuition of this, we conduct a proof-of-concept experiment. We quantify this effect by collecting image trajectories from well-trained policies under six viewpoints in the Mujoco Walker2d environment [Todorov *et al.*, 2012]. We measure the mutual information (MI) between image observations and ground truth states using the MINE[1] method [Belghazi *et al.*, 2018]. As shown in Figure 1, overall, the policies trained from viewpoints with high mutual information have relatively high test performance, indicating that viewpoint is crucial for the final RL performance. On the other hand, many studies explore the multi-view RL paradigm [Li *et al.*, 2019; Kinose *et al.*, 2022] to improve performance with more cameras, but at the cost of increased computation and storage. This raises an intriguing question: can we design an algo-

*Corresponding author.

[1]It is worth noting that the MINE method is only used in the proof-of-concept experiment, and not used to estimate the MI in main experiments. More details are in Appendix D.1.

rithm that ***automatically determines the optimal viewpoint*** for RL tasks, thus improving the final performance ***without expert knowledge or additional cameras?***

This problem poses two challenges: 1) we need to train a stable *motor policy* that adapts to the real-time changes of the observation viewpoint; 2) we need to provide suitable signals that guide the *sensory policy* to find a good viewpoint. We propose a new method, the **MO**del-based **SE**nsor controlle**R** (**MOSER**), that simultaneously learns a view-conditional world model to simulate the environment, a sensory policy to control the camera, and a motor policy to complete the task. The view-conditional world model overcomes the two challenges by: 1) incorporating viewpoint information in the observation decoding process to exclude the view-related information from the latent states; 2) providing three types of intrinsic rewards from its dynamics model, reward model, and image decoder to direct the sensory policy toward the optimal viewpoint.

We evaluate MOSER on eight tasks in two environments: DeepMind Control Suite [Tassa *et al.*, 2018] and Jaco Arm [Laskin *et al.*, 2021]. MOSER outperforms the baseline methods on almost all tasks and approaches the performance of model-based multi-view RL methods on most tasks. We conduct ablations to analyze the effect of different design choices and interpret the adapted camera parameters in each task. These experiments demonstrate that MOSER can find optimal viewpoints during RL training to improve performance across diverse tasks.

Our contributions can be summarized as follows:

1. To the best of our knowledge, we are the first to introduce the sensory policy into model-based RL approaches for autonomously changing camera parameters.
2. We propose an effective **MO**del-based **SE**nsor controlle**R** (**MOSER**) method that simultaneously learns the view-conditional world model, sensory policy, and motor policy to find the task-specific viewpoint.
3. Our approach performs excellently on several locomotion and manipulation tasks, demonstrating the significance of actively changing viewpoints in visual RL tasks.

## 2 Related Work

**Paradigm of Visual Reinforcement Learning.** In many real-world problems, agents can only observe high-dimensional inputs, such as images or videos, rather than the underlying states. Previous works have defined this problem from various perspectives; for example, POMDP [Hausknecht and Stone, 2015] models it as a partially observable problem. HCMDP [Ma *et al.*, 2022] considers high-dimensional input as a combination of mappings of task-relevant state and task-irrelevant contexts. Block MDP [Zhang *et al.*, 2020] assumes that observation generation is the concatenation of noise and state variables. However, these definitions neglect an important factor - how images are captured. Different perspectives can lead to different visual representations of the same object, directly affecting the task-relevant information in observations. We propose a View-conditional Partially Observable Markov Decision Processes (VPOMDPs) assumption incorporating viewpoint factors into the observation generation process.

**Multi-view Reinforcement Learning.** Many studies exploit observations from multiple cameras to improve RL task performance. MVRL [Li *et al.*, 2019] proposed a framework and model-free/based solutions for multi-view reinforcement learning. Multi-view Dreaming [Kinose *et al.*, 2022] uses contrastive learning to embed multi-view observations in a shared latent space. MV-MWM [Seo *et al.*, 2023] trains a masked multi-view autoencoder and world model. The model-free LookCloser [Jangir *et al.*, 2022] integrates multi-view observations via cross-view attention. The Fuse2Control [Hwang *et al.*, 2023] learns the underlying state space model from multi-view observation sequences with an information-theoretic method, showing the effectiveness in aggregating task information across many views. While these approaches outperform single-view methods, they increase computation and storage costs and may not guarantee the coverage of the best viewpoint. In contrast, our method receives observations from a movable camera from one viewpoint at a time instead of multiple cameras.

**Active Vision in Reinforcement Learning.** Actively seeking optimal viewpoints is essential for agents performing visual RL tasks. Previous work has enabled control of hand and eye movements to avoid occlusions and complete manipulation tasks [Cheng *et al.*, 2018; Grimes *et al.*, 2023]. SUGARL [Shang and Ryoo, 2023] proposes active perception of acquired images by selectively focusing on certain image regions. The recent work SAM-RL [Lv *et al.*, 2023] uses a differentiable physics simulator and rendering and learns a sensing-aware Q function for sensory and motor action selection. Our method fundamentally differs from previous work: (1) We use a sensory policy decoupled from the motor policy and change camera parameters guided by intrinsic signals. (2) We efficiently train the motor policy inside a View-conditional World Model, simulating the environment under viewpoint changes.

## 3 Preliminary

**View-conditional Partially Observable Markov Decision Processes.** A standard RL problem can be defined as a Markov decision process (MDP), which is a tuple of $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, p, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the transition function, $p$ is the initial state distribution, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function, and $\gamma$ is the discount factor. The goal of RL is to learn an optimal policy $\pi^\star(a|s)$ that maximizes the expected cumulative discounted return $R_\mathcal{M}(\pi) = \mathbb{E}_{s_0 \sim p(\cdot), a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1})\right]$.

To solve RL problems with incomplete observations such as high-dimensional images, prior works [Hafner *et al.*, 2019; Hafner *et al.*, 2020; Yarats *et al.*, 2021b; Bharadhwaj *et al.*, 2022] adopt partially observable Markov decision processes (POMDPs) assumption which additionally introduces an observation space $\mathcal{O}$ and generates observations by an emission function $\phi : \mathcal{S} \to \mathcal{O}$. We

modify the emission function as $\phi(o|s,c)$ and reformulate a new assumption named view-conditional partially observable Markov decision processes (VPOMDPs), where observations are generated based on the ground-truth state $s$ and the sensor parameter $c$.

**Model-based Reinforcement Learning.** Model-based reinforcement learning (MBRL) learns a model of the environment's transition dynamics $p(s_{t+1}|s_t, a_t)$ and reward function $r(s_t, a_t)$ from experience and leverages this model to plan actions by searching over possible future states [Janner *et al.*, 2019; Wang *et al.*, 2019]. Model-based methods enable sample-efficient learning as the agent can learn from the simulated environment instead of the real one [Moerland *et al.*, 2023]. This is especially beneficial when the input is high-dimensional images [Hafner *et al.*, 2019; Hafner *et al.*, 2020; Hafner *et al.*, 2021], as MBRL can obtain a low-dimensional surrogate environment to reduce the storage cost and improve the learning efficiency. Our method adopts a model-based approach to learn the dynamics of the environment and obtain an abstract state representation based on historical information, stabilizing policy training under varying viewpoints.

**Soft Actor Critic.** Soft Actor-Critic (SAC) [Haarnoja *et al.*, 2018; Haarnoja *et al.*, 2019] is an off-policy actor-critic deep RL algorithm based on the maximum entropy framework that optimizes a stochastic policy to maximize expected reward while also maximizing entropy. The agent learns a policy network $\pi_\theta(a_t|s_t)$ and two Q-networks $Q_\psi(s_t, a_t)$ to optimize the expected return $J(\pi) = \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[r(s_t, a_t) + \alpha\mathbb{H}(\pi(\cdot|s_t))]$, where $\rho_\pi$ is the state-action marginal distribution, $r(s_t, a_t)$ is the reward, and $\mathbb{H}$ represents entropy [Haarnoja *et al.*, 2018]. The policy and Q-functions are updated by backpropagating the gradients of the expected return via the reparameterization trick, while the temperature parameter $\alpha$ is automatically adjusted to ensure sufficient exploration [Haarnoja *et al.*, 2019]. The two Q-networks mitigate positive bias and improve training stability. We choose SAC, a typical off-policy algorithm, to update the sensory policy since it can efficiently utilize the collected sensory data. Any off-policy algorithm can instantiate our sensory policy.

# 4 Methods

In this section, we first introduce the model components and training objective of the View-conditional World Model (VWM) (Section 4.1). We then detail the designed intrinsic rewards for sensory policy training from three aspects, active mutual information maximization, next-frame prediction, and future reward maximization (Section 4.2). Finally, we use the actor-critic algorithm to train the motor policy by imagining in the learned world model (Section 4.3). We present an overview of our training process in Figure 2a, the detailed algorithm in Appendix B, the network architectures in Appendix D.3, and the hyper-parameters in Appendix F.

## 4.1 Training View-conditional World Model

To enable the agent to train stably under varying viewpoints, we propose the View-conditional World Model (VWM), a variant of the world model [Ha and Schmidhuber, 2018; Hafner *et al.*, 2019; Hafner *et al.*, 2020]. We model the image decoder of the VWM as the emission function under the VPOMDPs assumption to separate the viewpoint information from the latent state. The image decoder $p_\phi\left(\hat{o}_t|s_t, a^s_{t-1}\right)$ reconstructs the original image $o^c_t$ from the current latent state $s_t$ and the previous sensory action $a^s_{t-1}$. The other components of VWM are as follows:

$$\begin{aligned}
&\text{Encoder:} && s_t \sim q_\psi\left(s_t|s_{t-1}, a^m_{t-1}, o^c_t\right)\\
&\text{Dynamics model:} && \hat{s}_t \sim p_\theta\left(\hat{s}_t|s_{t-1}, a^m_{t-1}\right)\\
&\text{Reward Decoder:} && \hat{r}_t \sim p_\alpha\left(\hat{r}_t|s_t\right)
\end{aligned}$$

The encoder infers latent state $s_t$ from the previous latent state $s_{t-1}$, previous motor action $a^m_{t-1}$, and current observation $o^c_t$. The dynamics model predicts the latent state $s_t$ without observation $o^c_t$, enabling the model to forecast future states in the latent space. The reward decoder estimates rewards for possible future states. We jointly optimize the VWM parameters by minimizing the negative variational lower bound [Hafner *et al.*, 2020] as below:

$$\begin{aligned}
\mathcal{L}_\mathcal{M} = \ & \beta\mathrm{KL}[q_\psi(s_t|s_{t-1}, a^m_{t-1}, o^c_t)||p_\theta(\hat{s}_t|s_{t-1}, a^m_{t-1})]\\
& - \ln p_\phi(o^c_t|s_t, a^s_{t-1}) - \ln p_\alpha(r_t|s_t),
\end{aligned} \tag{1}$$

where $\beta$ is a hyperparameter to weigh the KL-divergence term. The detailed derivation of the loss is in Appendix C.

## 4.2 Learning Sensory Policy

The desiderata of the optimal view captured by the camera for real-world RL tasks are threefold. First, the view should be informative, containing rich proprioceptive information about the ground truth state. Second, the view between adjacent frames should exhibit temporal coherence, as consistent and continuous visual inputs can stabilize the world model training and policy learning. Finally, the view should inform high potential future rewards for the policy. In this section, we formulate these desiderata into three types of intrinsic rewards that encourage the sensory policy to take proper actions.

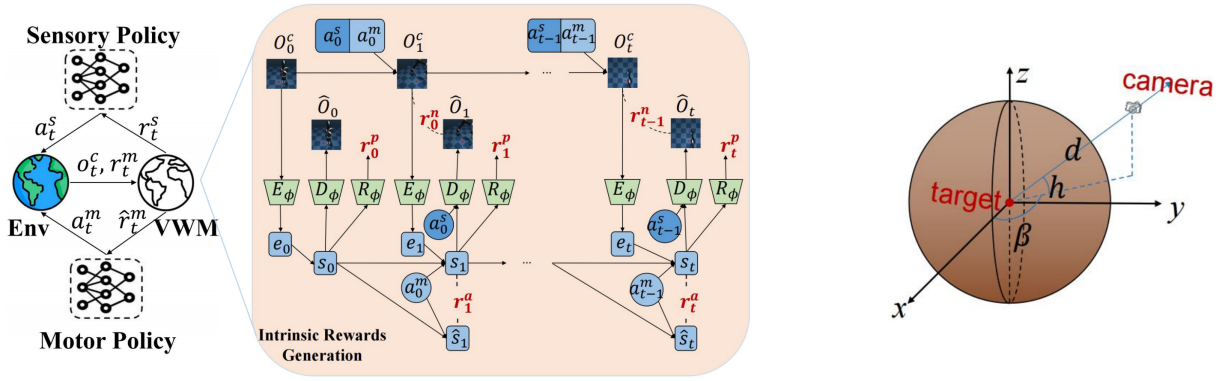
**Active Mutual Information Maximization (AMIM).** Based on the result of the proof-of-concept experiment in Figure 1, viewpoints containing rich information about the ground-truth state benefit policy training. We employ the sensory policy to select actions that adjust camera parameters to maximize the mutual information between the image observations $o^c$ and latent states $s$, conditioned on the previous states and actions.

$$a^{s\star}_t = \underset{a^s_t}{\operatorname{argmax}}\, \mathbb{I}(o^c_{t+1}; s_{t+1}|s_t, a^m_t, a^s_t) \tag{2}$$

Directly optimizing this mutual information is challenging. We can decompose it into two terms:

$$\begin{aligned}
& \mathbb{I}(o^c_{t+1}; s_{t+1}|s_t, a^m_t, a^s_t)\\
= \ & \mathbb{H}(s_{t+1}|s_t, a^m_t, a^s_t) - \mathbb{H}(s_{t+1}|s_t, a^m_t, a^s_t, o^c_{t+1})\\
= \ & \underbrace{\mathbb{H}(s_{t+1}|s_t, a^m_t)}_{(i)diversity} - \underbrace{\mathbb{H}(s_{t+1}|s_t, a^m_t, o^c_{t+1})}_{(ii)uncertainty}
\end{aligned} \tag{3}$$

Equation (3) tells us that maximizing the original mutual information is equivalent to maximizing the above information




(a) The framework of MOSER.

(b) Specification of camera parameters.

Figure 2: (a) The main framework of the MOSER method: 1) Left: The motor policy interacts with the environment, generates trajectories for training, and optimizes itself in the latent state space. The sensory policy selects the optimal viewpoint through intrinsic reward guidance, which requires view awareness in the RL process. 2) Right: We construct the world model following DreamerV2 [Hafner *et al.*, 2021] and add the viewpoint condition to the image decoder to form the VWM. The VWM generates three types of intrinsic rewards $r_t^a$, $r_t^n$, and $r_t^p$ for sensory policy updating, without additional modules. (b) For a specific target, the camera pose is parameterized as $(d, \beta, h)$, corresponding to the radius $d$, azimuth angle $\beta$, and elevation angle $h$.

gain, which maximizes the diversity in latent state space and minimizes the uncertainty after obtaining the observation under a selected viewpoint. We omit the sensory action $a_t^s$ in the above two entropy terms for two reasons: (i) in the VWM's latent state space, predicting the next state $s_{t+1}$ does not require sensory actions; (ii) $a_t^s$ only affects the generation of the image observation $o_{t+1}^c$ and can be absorbed into it. Therefore, to maximize the mutual information, we only need to minimize the second term in Equation (3), which depends on the sensory policy's actions. The sensory policy should select the action that produces the most informative observation $o_{t+1}^c$, reducing the uncertainty of the posterior distribution as much as possible. To encourage this behavior, we define the AMIM reward as:

$$r_t^a = -\mathbb{H}(s_{t+1}|s_t, a_t^m, o_{t+1}^c), \tag{4}$$

where $o_{t+1}^c$ is generated by the underlying emission function $\phi(o_{t+1}^c|s_{t+1}, a_t^s)$ of the environment, and $a_t^s$ is determined by the sensory policy. We can effectively estimate the posterior entropy in Equation (4) by the encoder $q_\psi(s_{t+1}|s_t, a_t^m, o_{t+1}^c)$ of VWM, which outputs the mean and standard deviation of a Gaussian distribution.

**Next-Frame Prediction (NFP).** The key challenge of exploring viewpoints is that timely changing camera parameters can cause inconsistent observations for model and policy learning. To alleviate abrupt changes in adjacent observations, the sensory policy should output predictable actions that can help predict the next frame from the latent state. Based on this, we design the NFP reward as the log-likelihood of the next frame prediction by the VWM's image decoder, as follows:

$$r_t^n = \log p_\phi(o_{t+1}^c|s_{t+1}, a_t^s) \tag{5}$$

**Future Reward Maximization (FRM).** The sensory policy should take actions that generate image observations that imply high rewards. To achieve this, we use the VWM's reward decoder $R_\phi$ to predict the potential reward associated with the selected observation, without introducing any additional modules. We formulate the FRM reward as a prediction of the potential task reward on the image observation $o_{t+1}^c$, which is controlled by the sensory action $a_t^s$ and the latent state $s_{t+1}$:

$$r_t^p = R_\phi(o_{t+1}^c), \text{where } o_{t+1}^c = \phi(s_{t+1}, a_t^s) \tag{6}$$

**Sensory Policy Training.** We update the sensory policy through the SAC algorithm [Haarnoja *et al.*, 2018] based on the weighted sum of three intrinsic rewards ($r^a, r^n$, and $r^p$). In principle, the sensory policy can be optimized with any off-policy RL algorithm. The critic of the sensory policy fits the value function based on the bootstrapped intrinsic rewards and guides the actor to select actions based on the current camera parameters $c_t$. The components of the sensory policy are shown below:

$$\text{Intrinsic Reward:} \quad r_t^s = \alpha_1 r_t^a + \alpha_2 r_t^n + \alpha_3 r_t^p$$

$$\text{Sensory Actor:} \quad a_t^s \sim \pi_s(a_t^s|c_t)$$

$$\text{Sensory Critic:} \quad v_s(c_t, a_t^s) \approx \mathbb{E}_{\pi_s(\cdot|c_t)}\Big[\sum_{k=t}^{T} \gamma^{k-t} \log r_k^s\Big]$$

where $\alpha_1, \alpha_2, \alpha_3$ are the weights of different intrinsic rewards. The detailed training algorithm of the sensory policy is in Algorithm 2 of Appendix B, and the values of $\alpha_1, \alpha_2, \alpha_3$ are in Table 3 of Appendix F.

### 4.3 Learning Motor Policy

We adopt the behavior learning mechanism of Dreamer [Hafner *et al.*, 2020] to optimize the motor policy within the VWM. We only train the motor policy on the imagined trajectories inside the VWM to improve sample efficiency. Specifically, the critic of the motor policy estimates the value function using the decoded reward and bootstraps, and the actor

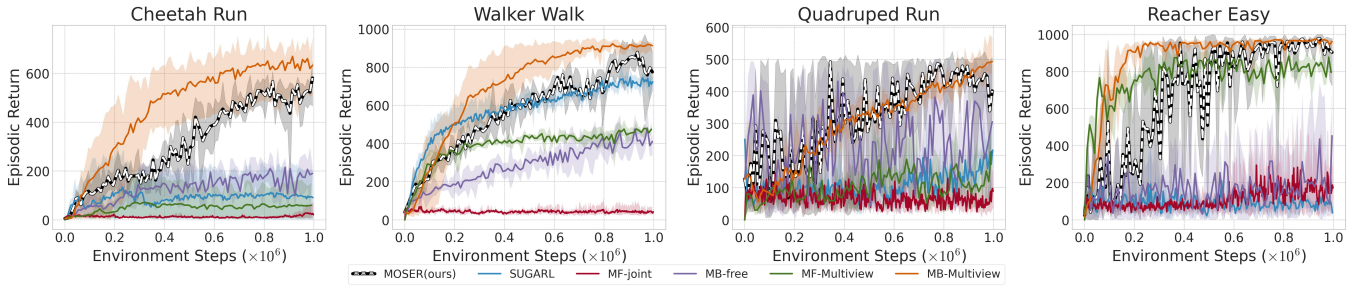



Figure 3: Performance results of our method MOSER and the five baseline methods over four seeds in four visual control tasks of DMC. The solid curves show the average episodic returns and the shaded region indicates the performance range under different runs. MOSER consistently beats the compared methods except for MB-multiview in almost all environments. It is worth noting that multi-view methods receive twice as much data as single-view methods, as they obtain two image observations of distinct views per step.

samples actions based on the imagined latent state, maximizing the expected return.

$$\text{Motor Actor:} \qquad a_t^m \sim \pi_m(a_t^m|s_t)$$

$$\text{Motor Critic:} \quad v_m(s_t) \approx \mathbb{E}_{\pi_m(\cdot|s_t)}\Big[\sum_{k=t}^{T} \gamma^{k-t} \log r_k^m\Big]$$

## 5 Experiments

We conduct several experiments to answer the following research questions (RQs):

1. Effectiveness (RQ1): How effective is MOSER in continuous control tasks?
2. Ablation (RQ2): How do the critical designs of MOSER contribute to the final performance?
3. Specialization (RQ3): Can MOSER identify specific viewpoints for different environments and tasks?
4. Optimality (RQ4): Can the best camera parameters found by MOSER improve the performance of the current algorithms?

### 5.1 Experiments Setup

**Locomotion.** We select four visual control tasks from DeepMind Control Suite [Tassa *et al.*, 2018]: *cheetah run*, *walker walk*, *reacher easy*, and *quadruped run*. These tasks pose challenges like sparse rewards, contact dynamics, and 3D scenes. The shape of image observation is $64 \times 64 \times 3$ for all tasks.

**Manipulation.** We test the method's ability to control cameras and adapt to different goals while performing simple manipulation tasks on the Jaco Arm, a 6-DOF robotic arm with a three-finger gripper. The Jaco Arm has been used in previous unsupervised RL works [Laskin *et al.*, 2021]. This environment is challenging due to the sparse reward, as shown in prior work [Laskin *et al.*, 2021; Yarats *et al.*, 2021a]. We consider four tasks with different target positions: *top left*, *top right*, *bottom left*, and *bottom right*. The shape of observations is the same as that in the Locomotion tasks.

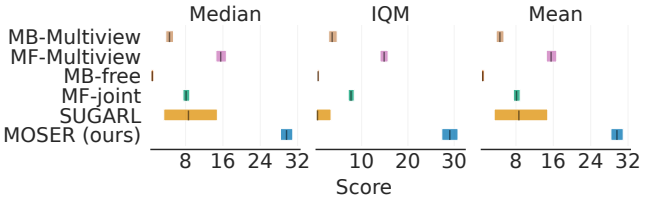



Figure 4: Aggregate metrics on Jaco Arm tasks of MOSER and the baselines with 95% confidence intervals. Higher mean, median and IQM scores are better. MOSER consistently outperforms the compared methods in Jaco Arm tasks.

**Settings of camera parameters.** As depicted in Figure 2b, each sensory state is represented by the tuple $(d, \beta, h)$ in a spherical coordinate system, where $d$ is the distance from the target to the camera, $\beta$ is the azimuth angle, and $h$ is the elevation angle. For both locomotion and manipulation tasks, the sensory state space is continuous, initialized at $(3, 90, -45)$ and bounded within the ranges of $[0, 10]$ for $d$, $[0, 180]$ for $\beta$, and $[-90, 90]$ for $h$. The sensory action space is also continuous, spanning $[-1, 1]$ in all three dimensions. The examples of camera parameters are shown in Appendix A.

**Baselines.** We compare our method with several baselines on both DMControl Suite and Jaco Arm:

- **SUGARL**: A visual reinforcement learning method that trains a sensory policy to actively select the optimal region from the original image observation as the input for the motor policy [Shang and Ryoo, 2023].
- **MF-joint**: A model-free RL method that jointly trains sensory and motor policies based on the environmental reward. We use the DrQ algorithm [Yarats *et al.*, 2021b] as the backbone, which is a state-of-the-art model-free visual RL method.
- **MB-free**: A model-based RL method that learns a sensory policy with the environmental reward and trains the motor policy inside the model. We employ the DreamerV2 algorithm [Hafner *et al.*, 2021], an MBRL method that plans through a learned world model.
- **MF-multiview**: A model-free RL method with multi-view observation. We choose the DrQ algorithm [Yarats

*et al.*, 2021b] as the backbone and concatenate the first- and third-person image observations at each step as inputs.

- **MB-multiview**: A model-based RL method with multi-view observation. We select the DreamerV2 algorithm [Hafner *et al.*, 2021] as the backbone and process the inputs the same way as MF-multiview.

## 5.2 Effectiveness (RQ1)

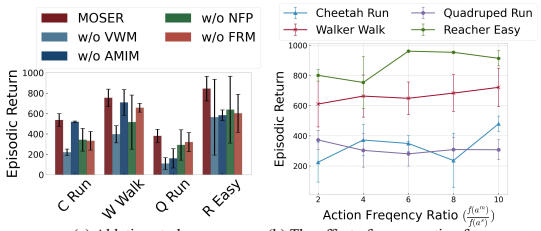
**Locomotion.** As shown in Figure 3, MOSER enhances performance in tasks with single fixed-camera observations, outstrips the model-free multi-view approach, and rivals the model-based multi-view method in nearly all tasks. The MF-joint method struggles with task execution and viewpoint optimization due to fluctuating observations from changing perspectives. Interestingly, without explicit sensory rewards, the MB-free method outperforms the MF-joint method, possibly because its long-term predictions enable accurate proprioceptive state estimation and facilitate motor control under viewpoint changes. The SUGARL method, although it seeks ideal image regions, is hindered by static camera settings. MOSER exceeds SUGARL across all tested environments by actively selecting the best viewpoints. It also surpasses the MF-multiview in four environments, suggesting that a single optimal viewpoint can be as informative as multiple ones. While MOSER nears the performance of the MB-multiview method in three tasks, it lags in *Cheetah Run*. Notably, the MB-multiview benefits from utilizing twice the amount of data compared to MOSER, as it receives two observations from distinct views at each step. This substantial increase in data volume significantly contributes to its performance advantage. Our objective is not to supplant the multiview approach with MOSER universally but to enhance the efficacy of existing visual RL methods by a learnable viewpoint.

**Manipulation.** In Figure 4, the MOSER method outperforms all other methods with fixed camera observation inputs, including single-view and multi-view, on Jaco Arm. This is because the manipulation task requires the agent to control the arm to reach the goal, which demands dynamic viewpoint changes to capture the relative position between the arm and the goal and contribute to policy adaptation. Multi-view methods perform significantly better than single-view methods on the Jaco Arm task since multiple viewpoint inputs can provide richer task-relevant information. Overall, MOSER achieves superior results by actively selecting suitable viewpoints to optimize the raw input for the motor policy.

## 5.3 Ablation (RQ2)

To evaluate the contribution of different components of MOSER, we conduct ablation studies on the viewpoint condition, the designed intrinsic rewards, and the frequency of executing sensory actions.

**Impact of viewpoint condition.** To examine the influence of viewpoint conditioning, we modify the image decoder in VWM to reconstruct the image observation from the latent state only. In Figure 5, MOSER with WM shows a significant performance drop on multiple tasks compared to the original




(a) Ablation study. (b) The effect of sensor action frequency.

Figure 5: (a) The ablation study of the viewpoint condition on the world model and three types of intrinsic rewards. Removing any one will result in decreased performance. (b) Investigation of the effect of sensory action frequency on policy performance. The horizontal axis shows the frequency of sensory actions, with $k = 2, 4, 6, 8, 10$ representing sensory actions executed once every $2, 4, 6, 8, 10$ motor actions, respectively. Larger values indicate less frequent sensory actions. Policies trained with different sensory action frequencies show different performances. We run all experiments on DMC tasks with error bars indicating performance standard deviation across four seeds.

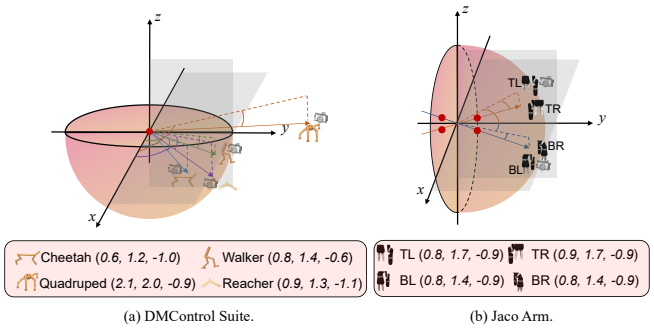
MOSER, especially on the *Quadruped Run* task. This indicates that incorporating the viewpoint condition in the world model of MOSER is essential for stable training and achieving high performance.

**Effects of intrinsic rewards.** To investigate the individual effects of different intrinsic rewards - AMIM, NFP, and FRM - we conduct ablation studies on each separately within DMC environments. In Figure 5a, removing the AMIM reward leads to a significant performance decrease in the DMC tasks, indicating that a good viewpoint needs to maximize the amount of information about the proprioceptive state. Eliminating the NFP reward results in a high variance of performance, revealing that predicting the following observation from the previous state and current sensory action helps smooth and continuous viewpoint changes. Removing FRM results in a remarkable performance drop across environments, confirming good viewpoints should enable high rewards. Each component is crucial to learning good viewpoints across diverse control tasks. We present the change of these intrinsic rewards during training in Appendix E.4.

**Impact of sensory action frequency.** Different sensory action frequencies can affect training stability and final performance. We investigate the effects of different sensory action frequencies on experimental performance. The results in Figure 5b show that agents prefer high sensory action frequency to adapt to the environment in complex environments such as *Quadruped Run*. In contrast, in 2D locomotion environments like *Cheetah Run* and *Walker Walk*, changing camera parameters every 10 steps leads to optimal performance, suggesting that low sensory action frequency is more suitable for these environments.

## 5.4 Specialization (RQ3)

We present the camera parameters selected by MOSER for four DMC tasks in Figure 6a, illustrating the preference for different viewpoints in each environment. Cameras are positioned closer to the agent in tasks like *Cheetah Run*, *Walker*




(a) DMControl Suite. (b) Jaco Arm.

Figure 6: The optimal camera parameters selected by MOSER for tasks of (a) DMControl Suite and (b) Jaco Arm. For each task, we report the average parameters $(d, \beta, h)$ with $\beta$ and $h$ expressed in radians[2] for clarity.

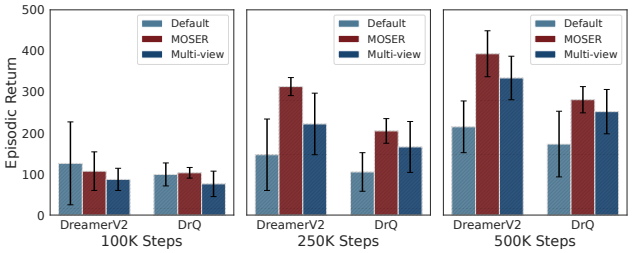



Figure 7: Performance comparison of DreamerV2 and DrQ algorithms on the *Quadruped Run* task with three camera parameter settings: default, MOSER-selected, and multi-view. The episodic returns are recorded over 100K, 250K, and 500K steps, with error bars representing the standard deviation across four runs. The MOSER-selected camera setting enhances both methods' performance at the intermediate and final stages.

*Walk*, and *Reacher Easy* compared to *Quadruped Run*, to accommodate the need for a broader view in 3D locomotion tasks. Variations in azimuth angles across tasks highlight the unique optimal viewpoints required for each. Elevation angles, consistently negative, suggest a universal preference for a bottom-up perspective to observe leg movements. These findings underscore MOSER's capability to tailor viewpoints to specific tasks.

Figure 6b details the camera parameters selected by MOSER for four Jaco Arm tasks, revealing shared and distinct viewpoint preferences within the same environment under varied objectives (the red balls). All four tasks display similarities in camera distances and elevation angles while differing in azimuth angles, notably demonstrating symmetry between the *Top* and *Bottom* tasks (with the y-axis as the axis of symmetry). Such variations in azimuth are attributable to the target's horizontal positioning, influencing the optimal viewpoint. These results confirm MOSER's proficiency in discerning both common and subtle task characteristics to determine ideal camera perspectives.

### 5.5 Optimality (RQ4)

We validate the optimality of the searched camera parameters in *Quadruped Run*, a challenging 3D locomotion task. Specifically, we compare three environmental settings: (1) camera parameters chosen by MOSER, (2) original default camera parameters, and (3) two distinct cameras with default views. We run DreamerV2 and DrQ algorithm on these three environments with all other settings identical and record the episodic return for 100K, 250K, and 500K steps. In Figure 7, all methods exhibit performance gains to varying degrees in the environment with camera parameters selected by MOSER. With pre-training viewpoints by MOSER, the performances of both algorithms surpass that of human-defined multiviews. These results reveal that learning from an optimal viewpoint can improve the performance of current visual RL algorithms, thus corroborating our initial motivation.

[2]Angles in degrees can be converted to radians by multiplying by $\frac{\pi}{180}$.

## 6 Conclusion

In visual RL research, learning from a fixed single-view observation requires expert knowledge to determine the viewpoint, and learning from a multi-view input will incur additional computing and storage costs. To tackle this problem, we introduce the view-conditional partially observable Markov decision processes assumption to incorporate the view-related parameters into the emission function and develop the view-conditional world model based on this assumption. We propose the MOdel-based SEnsor controlleR (MOSER) method, which learns the sensory policy to control the camera parameters and optimizes the motor policy to complete the RL task simultaneously. We evaluate the performance of MOSER on four locomotion tasks and four manipulation tasks. Moreover, we conduct ablation studies to analyze the contribution of viewpoint conditioning and different intrinsic rewards. We explore the suitable sensory action frequency in diverse tasks. MOSER can select task-specific viewpoints and improve the performance of the visual RL methods with a single fixed camera.

One limitation of MOSER lies in its performance on specific tasks compared to the MB-multiview method. The latter benefits from a richer data regime with two distinct images as input at each step. Our objective is not to replace the multi-view method with MOSER but to offer a possible improvement of autonomously selecting optimal viewpoints when using model-based multi-view methods. We will explore this further in future work. Furthermore, our current experiments involve task-centric observations, such as those in DMControl environments. Robotics manipulation and some navigation tasks typically require more flexibility from the camera. We plan to explore other scenes [Yang *et al.*, 2023] beyond the Jaco Arm task to demonstrate MOSER's capabilities. Additionally, real-world scenarios often present distracting inputs impacting sensory and motor policy optimization. One straightforward approach is to separately model task-relevant and irrelevant dynamics [Fu *et al.*, 2021; Wan *et al.*, 2023] within the view-conditional world model, without modifying other aspects of MOSER. Further investigation of this is a promising direction for future work.

## Acknowledgments

We thank Han-Jia Ye, Shaowei Zhang, Minghao Shao, Yucen Wang, Ziyuan Chen, and Olivia Yan for their valuable discussions. This work was partly supported by the National Science and Technology Major Project under Grant No. 2022ZD0114805 and partly by the Postgraduate Research & Practice Innovation Program of Jiangsu Province.

## References

[Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 530–539, Stockholmsmässan, Stockholm, Sweden, 2018. PMLR.

[Bharadhwaj *et al.*, 2022] Homanga Bharadhwaj, Mohammad Babaeizadeh, Dumitru Erhan, and Sergey Levine. Information prioritization through empowerment in visual model-based RL. In *International Conference on Learning Representations*, 2022.

[Cheng *et al.*, 2018] Ricson Cheng, Arpit Agarwal, and Katerina Fragkiadaki. Reinforcement learning of active vision for manipulating objects under occlusions. In *2nd Annual Conference on Robot Learning, CoRL 2018*, volume 87 of *Proceedings of Machine Learning Research*, pages 422–431, Zürich, Switzerland, 2018. PMLR.

[Dodge, 1903] Raymond Dodge. Five types of eye movement in the horizontal meridian plane of the field of regard. *American Journal of Physiology-Legacy Content*, 8:307–329, 1903.

[Fu *et al.*, 2021] Xiang Fu, Ge Yang, Pulkit Agrawal, and Tommi S. Jaakkola. Learning task informed abstractions. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 3480–3491, Virtual Event, 2021. PMLR.

[Grimes *et al.*, 2023] Matthew Koichi Grimes, Joseph Modayil, Piotr W. Mirowski, Dushyant Rao, and Raia Hadsell. Learning to look by self-prediction. *Trans. Mach. Learn. Res.*, 2023, 2023.

[Ha and Schmidhuber, 2018] David Ha and Jürgen Schmidhuber. World models, 2018.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1856–1865, Stockholmsmässan, Stockholm, Sweden, 2018. PMLR.

[Haarnoja *et al.*, 2019] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications, 2019.

[Hafner *et al.*, 2019] Danijar Hafner, Timothy P. Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 2555–2565, Long Beach, California, USA, 2019. PMLR.

[Hafner *et al.*, 2020] Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia, 2020. OpenReview.net.

[Hafner *et al.*, 2021] Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *9th International Conference on Learning Representations, ICLR 2021*, Virtual Event, Austria, 2021. OpenReview.net.

[Hausknecht and Stone, 2015] Matthew J. Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps. In *2015 AAAI Fall Symposia*, pages 29–37, Arlington, Virginia, USA, 2015. AAAI Press.

[Hu *et al.*, 2022] Anthony Hu, Gianluca Corrado, Nicolas Griffiths, Zachary Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, and Jamie Shotton. Model-based imitation learning for urban driving. In *Advances in Neural Information Processing Systems*, volume 35, pages 20703–20716, 2022.

[Hwang *et al.*, 2023] HyeongJoo Hwang, Seokin Seo, Youngsoo Jang, Sungyoon Kim, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Information-theoretic state space model for multi-view reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[Jangir *et al.*, 2022] Rishabh Jangir, Nicklas Hansen, Sambaran Ghosal, Mohit Jain, and Xiaolong Wang. Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. *IEEE Robotics Autom. Lett.*, 7(2):3046–3053, 2022.

[Janner *et al.*, 2019] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12498–12509, 2019.

[Kinose *et al.*, 2022] Akira Kinose, Masashi Okada, Ryo Okumura, and Tadahiro Taniguchi. Multi-view dreaming: Multi-view world model with contrastive learning, 2022.

[Laskin *et al.*, 2021] Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, and Pieter Abbeel. URLB: unsupervised rein-

forcement learning benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, virtual, 2021.

[Levine *et al.*, 2018] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *Int. J. Robotics Res.*, 37(4-5):421–436, 2018.

[Li *et al.*, 2019] Minne Li, Lisheng Wu, Jun Wang, and Haitham Bou-Ammar. Multi-view reinforcement learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 1418–1429, Vancouver, BC, Canada, 2019.

[Lv *et al.*, 2023] Jun Lv, Yunhai Feng, Cheng Zhang, Shuang Zhao, Lin Shao, and Cewu Lu. SAM-RL: sensing-aware model-based reinforcement learning via differentiable physics-based simulation and rendering. In *Robotics: Science and Systems XIX*, Daegu, Republic of Korea, 2023.

[Ma *et al.*, 2022] Guozheng Ma, Zhen Wang, Zhecheng Yuan, Xueqian Wang, Bo Yuan, and Dacheng Tao. A comprehensive survey of data augmentation in visual reinforcement learning, 2022.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature.*, 518(7540):529–533, 2015.

[Moerland *et al.*, 2023] Thomas M. Moerland, Joost Broekens, Aske Plaat, and Catholijn M. Jonker. Model-based reinforcement learning: A survey. *Found. Trends Mach. Learn.*, 16(1):1–118, 2023.

[Pereira *et al.*, 2021] Tiago Pereira, Maryam Abbasi, José Luís Oliveira, Bernardete Ribeiro, and Joel Arrais. Optimizing blood-brain barrier permeation through deep reinforcement learning for de novo drug design. *Bioinform.*, 37(Supplement):84–92, 2021.

[Schrittwieser *et al.*, 2020] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy P. Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature.*, 588(7839):604–609, 2020.

[Seo *et al.*, 2023] Younggyo Seo, Junsu Kim, Stephen James, Kimin Lee, Jinwoo Shin, and Pieter Abbeel. Multi-view masked world models for visual robotic manipulation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 30613–30632. PMLR, 2023.

[Shang and Ryoo, 2023] Jinghuan Shang and Michael S. Ryoo. Active reinforcement learning under limited visual observability, 2023.

[Tassa *et al.*, 2018] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite, 2018.

[Todorov *et al.*, 2012] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012*, pages 5026–5033, Vilamoura, Algarve, Portugal, 2012. IEEE.

[Wan *et al.*, 2023] Shenghua Wan, Yucen Wang, Minghao Shao, Ruying Chen, and De-Chuan Zhan. Semail: eliminating distractors in visual imitation via separated models. In *International Conference on Machine Learning*, pages 35426–35443. PMLR, 2023.

[Wang *et al.*, 2019] 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.

[Yang *et al.*, 2023] Yang Yang, Yurui Huang, Weili Guo, Baohua Xu, and Dingyin Xia. Towards global video scene segmentation with context-aware transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 3206–3213, 2023.

[Yarats *et al.*, 2021a] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 11920–11931, Virtual Event, 2021.

[Yarats *et al.*, 2021b] Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *9th International Conference on Learning Representations, ICLR 2021*, Virtual Event, Austria, 2021. OpenReview.net.

[Zhan *et al.*, 2021] Albert Zhan, Ruihan Zhao, Lerrel Pinto, Pieter Abbeel, and Michael Laskin. A framework for efficient robotic manipulation. In *Deep RL Workshop NeurIPS 2021*, 2021.

[Zhang *et al.*, 2020] Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. Invariant causal prediction for block mdps. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pages 11214–11224, Virtual Event, 2020. PMLR.

[Zhu *et al.*, 2020] Henry Zhu, Justin Yu, Abhishek Gupta, Dhruv Shah, Kristian Hartikainen, Avi Singh, Vikash Kumar, and Sergey Levine. The ingredients of real world robotic reinforcement learning. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia, 2020. OpenReview.net.