Focus-Then-Decide: Segmentation-Assisted Reinforcement Learning

Chao Chen[1,2*], **Jiacheng Xu**[1,2*], **Weijian Liao**[1,2], **Hao Ding**[1,2], **Zongzhang Zhang**[1,2†], **Yang Yu**[1,2], **Rui Zhao**[3]

[1] National Key Laboratory for Novel Software Technology, Nanjing University, China
[2] School of Artificial Intelligence, Nanjing University, China
[3] Tencent Robotics X, Shenzhen, China
{chenc, xujc, liaowj, dingh}@lamda.nju.edu.cn, {zzzhang, yuy}@nju.edu.cn, rui.zhao.ml@gmail.com

Abstract

Visual Reinforcement Learning (RL) is a promising approach to achieve human-like intelligence. However, it currently faces challenges in learning efficiently within noisy environments. In contrast, humans can quickly identify task-relevant objects in distraction-filled surroundings by applying previously acquired common knowledge. Recently, foundational models in natural language processing and computer vision have achieved remarkable successes, and the common knowledge within these models can significantly benefit downstream task training. Inspired by these achievements, we aim to incorporate common knowledge from foundational models into visual RL. We propose a novel **F**ocus-**T**hen-**D**ecide (FTD) framework, allowing the agent to make decisions based solely on task-relevant objects. To achieve this, we introduce an attention mechanism to select task-relevant objects from the object set returned by a foundational segmentation model, and only use the task-relevant objects for the subsequent training of the decision module. Additionally, we specifically employed two generic self-supervised objectives to facilitate the rapid learning of this attention mechanism. Experimental results on challenging tasks based on DeepMind Control Suite and Franka Emika Robotics demonstrate that our method can quickly and accurately pinpoint objects of interest in noisy environments. Consequently, it achieves a significant performance improvement over current state-of-the-art algorithms.

Project Page: https://www.lamda.nju.edu.cn/chenc/FTD.html

Code: https://github.com/LAMDA-RL/FTD

Introduction

Human-like intelligence, holding the prospect of liberating humans from repetitive labor, has long been one of the goals pursued by the machine learning community. Reinforcement Learning (RL), as a general decision-making framework, has the potential to enable machines to generate human-like decision-making behaviors (Sutton and Barto 2018; Wu and Zhang 2023). Research has shown that visual input constitutes the majority of human information intake (Goldstein and Cacciamani 2021); therefore, visual RL methods targeting such input have garnered widespread attention and are potential methods for achieving human-like intelligence. Living up to expectations, visual RL algorithms have consistently made remarkable progress in areas such as gaming (Hessel et al. 2018), robotic control (Andrychowicz et al. 2020), and autonomous driving (Chen, Li, and Tomizuka 2022).

However, there is still a substantial gap if we want to apply visual RL algorithms to practical scenarios. One reason is that current visual RL algorithms struggle to train an effective policy in complex and noisy real-world environments. Consider the most influential visual RL benchmarks, such as Atari (Bellemare et al. 2012) and DeepMind Control Suite (Tassa et al. 2018). Agents trained in these clean, structured scenarios merely need to leverage all the information they receive to make decisions. However, when trained under complex, unstructured real-world conditions, agents may be overwhelmed by the massive influx of information, lacking the ability to identify which parts are useful, thus leading to learning failure (Stone et al. 2021; Xu et al. 2023b). Discerning which concepts are task-relevant and worth focusing on, and which are task-irrelevant and should be ignored, allowing learning to proceed undisturbed in any scenario, is a human internalized ability. This ability is also essential for advanced visual intelligent agents.

Recent works have sought to address the challenge of learning in complex scenarios filled with disturbances (Zhang et al. 2021; Fu et al. 2021; Wang et al. 2022b; Xu et al. 2023a). The main approach of previous methods involves utilizing auxiliary representation objectives to extract task-relevant information from the original observations while disregarding noise. These loss functions are commonly optimized over the unstructured pixel space, necessitating sophisticated design to ensure proper functioning. We identify two potential problems. The first pertains to the proposal of representation learning targets, which may contain erroneous inductive biases or approximation errors due to their complexity, leading to biased information extraction. The second problem concerns the optimization in pixel space, which is overly fine-grained and does not align with human perception mechanisms. Using prior knowledge, humans perceive and make decisions at a more coarser object level. Considering these issues, we aim to propose a visual perception and extraction method that is both simply designed and more human-like.

Similar to the human ability to rapidly leverage existing knowledge to learn new tasks, foundation models pre-trained

*These authors contributed equally.
†Zongzhang Zhang is the corresponding author.

on broad datasets have demonstrated exceptional abilities in knowledge transfer to diverse downstream tasks (Bommasani et al. 2022), achieving great success in both natural language processing (OpenAI 2023) and computer vision (Kirillov et al. 2023) domains. Inspired by this, our paper introduces a new paradigm for visual RL based on foundational segmentation models. The goal is to efficiently train RL algorithms in noisy perceptual environments, thus narrowing the gap between current visual RL algorithms and the demands of training in real-world environments.

We adopt the perspective that the state-space of a Markov Decision Process (MDP) can be represented in terms of objects (Yi et al. 2022), and propose a Focus-Then-Decide (FTD) framework. The goal is for the agent to automatically distinguish between task-relevant and task-irrelevant objects, focusing on the former for decision-making. Initially, the plug-and-play style foundational segmentation model is naturally employed to effectively partition the complete perceptual state into object-level fragments. Then, we introduce a novel attention mechanism for calculating the task relevance of each object. Finally, the decision-making agent is trained on the visual representation that excludes task-irrelevant objects. Additionally, we specifically select two simple self-supervised losses to enhance the focus on task-relevant objects.

We summarize our contributions as follows:

1. We introduce a novel perspective for visual RL problems, viewing observations as compositions of task-relevant and task-irrelevant objects. We then naturally incorporate a foundational segmentation model into visual RL algorithms, enabling the partitioning of observations into distinct sets of objects.
2. We propose a novel Focus-Then-Decide framework for visual RL, utilizing an attention selector to assess the task relevance of each object returned by the foundational segmentation model, subsequently integrating task-relevant perceptions into the training of decision models.
3. In complex visual noisy scenarios based on DeepMind Control Suite and Franka Emika Robotics, the experimental results demonstrate that our method can quickly and effectively identify the task-relevant parts in a noisy environment, thus robustly learning in these conditions, and achieving significant performance improvements compared to state-of-the-art algorithms.

Background

In this section, we briefly describe the background of reinforcement learning and foundational segmentation model.

Reinforcement Learning

Traditional Reinforcement Learning (RL) considers the task in the form of a Markov Decision Process (MDP) composed of a 4-tuple $(\mathcal{S}, \mathcal{A}, R, P)$, where $\mathcal{S}$ is the state space, s_t denotes state at timestep t; $\mathcal{A}$ is the action space, a_t denotes the action at timestep t; $R(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, which maps state-action pair to a real number; and $P(s,a,s') : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the transition function. The aim of RL is to train an agent with policy $\pi(a|s)$ that maximizes the expected discounted cumulative reward (a.k.a. return) $\mathbb{E}_\pi[\sum_t \gamma^t r_t]$, where $\gamma \in (0,1]$ is the discount factor. In visual RL tasks, the agent cannot access the compact low-dimensional state, instead, it receives an image-based observations $o_t = \mathcal{O}(s_t)$, which are usually high-dimensional. Here, $\mathcal{O}$ is the mapping function from state to observation. Corresponding, policy should be written as $\pi(a \mid o)$.

In this paper, we use Soft Actor-Critic (SAC) (Haarnoja et al. 2018) as the basic RL algorithm. SAC is a widely used off-policy actor-critic algorithm that optimizes a stochastic policy. The objective of SAC is to maximize the weighted sum of cumulative reward and policy entropy, $\mathbb{E}_{s_t,a_t\sim\pi}\left[\sum_t r_t + \alpha\mathcal{H}\left(\pi\left(\cdot \mid o_t\right)\right)\right]$, where $\mathcal{H}\left(\pi\left(\cdot \mid o_t\right)\right)$ denotes the entropy of policy, α is a learnable factor.

Foundational Segmentation Model

Image segmentation (Szeliski 2022) has long been an essential field in computer vision, with many subfields such as semantic segmentation (Shotton et al. 2006), instance segmentation (Lin et al. 2014), etc. In this work, we primarily focus on the instance segmentation task, where the model takes an image as input and outputs all objects that can be segmented from the image. Many works have contributed to progress in this field, including Swin (Liu et al. 2021), ViTDet (Li et al. 2022), etc.

A recent breakthrough in image segmentation is Segment Anything Model (SAM) (Kirillov et al. 2023). It utilizes a straightforward network architecture and trains the model on a vast dataset containing over one billion masks. The trained model exhibits remarkable capability in zero-shot transfer, allowing the foundational segmentation model to be directly applied to unseen images and adapted for downstream tasks. Later, FastSAM (Zhao et al. 2023) and MobileSAM (Zhang et al. 2023) attempt to distill the large SAM model into smaller ones, sacrificing a small amount of performance for faster inference speed. In our work, MobileSAM is used as the foundational segmentation model.

Our Method

We propose a novel Focus-Then-Decide (FTD) framework that allows the agent to make decisions based on task-relevant objects alone, thereby enabling effective training in noisy environments. In this section, we first present a novel way of defining the problem in noisy visual environments by perceiving high-dimensional observation as the rendering of an object collection. This makes it convenient to naturally integrate a fundamental segmentation model into the visual RL process. Then, we provide a brief overview of the overall workflow of the FTD framework. Next, we introduce the design philosophy and implementation method of the attention selector module, which is used to obtain perception inputs that contain only task-relevant information. Finally, we introduce two self-supervised losses that substantially enhance the learning speed of the attention selector, and explain how to combine them with the RL objective in the decision stage.

Problem Formulation

We adopt the perspective that the state-space of MDP can be represented in terms of objects (Yi et al. 2022), further

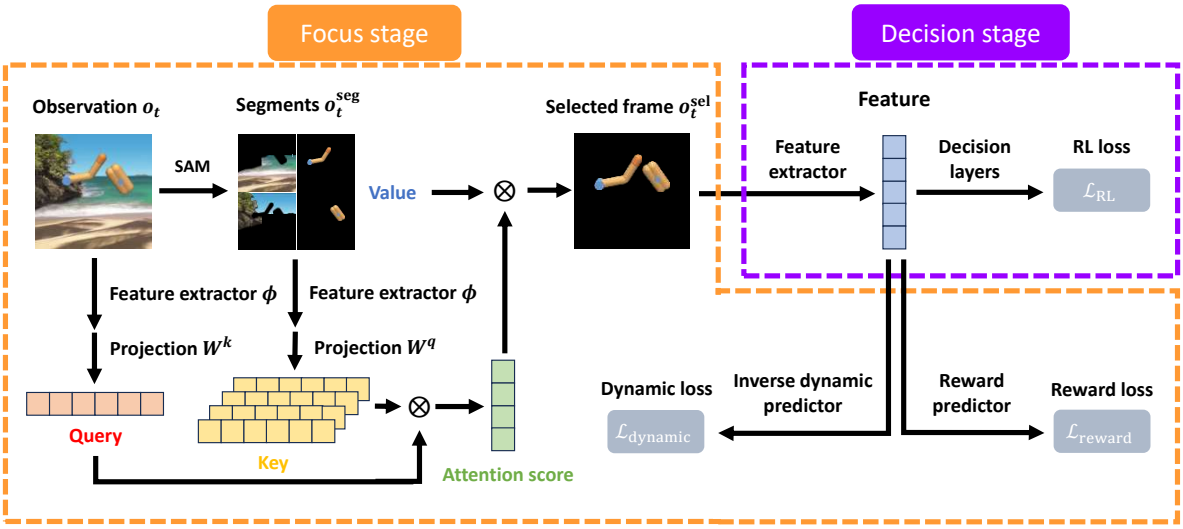

Figure 1: Illustration of the proposed Focus-Then-Decide (FTD) framework. The focus stage of FTD is enclosed in an orange box. A noisy observation o_t is first passed to the foundational segmentation model to obtain o_t^{seg}. Query and Key are calculated based on the original observation and segments respectively, with the segments also being directly used as Value. The selected frame o_t^{sel} is then determined by the attention selector. The decision stage is highlighted in a purple box. The entire network is trained by minimizing a combination of RL loss, Dynamic loss, and Reward loss.

distinguishing objects into task-relevant objects and task-irrelevant objects, introducing a novel perspective for visual RL problems. We view any state $s \in \mathcal{S}$ as a collection of the states of all objects within that state, $s = \cup_{i=1}^{k} s^{\text{obj}_i}$, where k is the number of objects in the state, s^{obj_i} is the state of the i-th object, and the observation corresponding to state s is given by $o = \mathcal{O}(s)$, satisfying $o = \cup_{i=1}^{k} \mathcal{O}(s^{\text{obj}_i})$. In noisy and complex environments, observations may contain a large number of task-irrelevant objects. We define the set of task-relevant objects as s^{obj^+}, and task-irrelevant objects as s^{obj^-}, satisfying $s = s^{\text{obj}^+} \cup s^{\text{obj}^-}$, $s^{\text{obj}^+} \cap s^{\text{obj}^-} = \emptyset$, $\pi^*(s) = \pi^*(s^{\text{obj}^+})$, and the removal of any $s^{\text{obj}_i} \in s^{\text{obj}^+}$ leads to $\pi^*(s) \neq \pi^*(s^{\text{obj}^+} \setminus \{s^{\text{obj}_i}\})$, where π^* represents the optimal policy. Existing robust visual representations aim to obtain task-relevant observations $o^+ = \mathcal{O}(s^{\text{obj}^+})$ from any observation o, but high-dimensional representation learning is a challenging issue, especially in perception environments filled with disturbances.

The zero-shot transfer capability of foundational segmentation models has been extensively validated and employed in numerous downstream tasks. Given a foundational segmentation model and an observation o, we assume that the segmentation model can obtain a set of all object observations $\{o^{\text{obj}_i} | i = 1, 2, \ldots, k\}$, where each o^{obj_i} corresponds to an s^{obj_i}. By employing this foundational segmentation model, the original problem of mapping o to o^+ is transformed from a high-dimensional image space mapping challenge into a search problem in the object-level combinatorial space, thereby greatly reducing the solution space.

For simplicity, we will use s^i and o^i to represent s^{obj_i} and o^{obj_i} in the following text.

Method Overview

Figure 1 shows the Focus-Then-Decide (FTD) framework. The focus stage is bounded with an orange-colored box and the decision stage is in a purple-colored box.

At each timestep t, the agent receives a distracted input frame. Initially, a foundational segmentation model processes the input to yield a batch of segments. Subsequently, an attention selector evaluates each segment, assigning an attention score to generate a selectively focused frame. Traditional RL algorithms are then applied to this selected frame. The network is updated by integrating the losses derived from both the RL process and two self-supervised objectives. In the following sections, we will delve into the details of the attention selector and the self-supervised objectives.

Attention Selector as Focus Stage

For the segmentation results containing both task-relevant and task-irrelevant parts, it is a very natural idea to filter out task-irrelevant ones, and we refer to the part with such function as "selector". More specifically, the selector is responsible for filtering and recombining the segments to form a new undisturbed frame.

As a selector, the most important thing is to integrate global information, rather than just process each segment separately. Take the finger-spin environment in Figure 1 as an example, to rotate the spiner with the finger, the selector should simultaneously pay attention to the finger and the spiner, missing either of these two will lead to deficient policy. Another thing that needs to be noticed is that the selector should be compatible with different numbers of segments, since the number of segments varies depending on the complexity of observation.

To realize a selector that can integrate global information

and is tolerant to the change in the number of segments, a natural way is to use the attention mechanism. We propose a novel attention mechanism to handle segmentation results, and refer to this selector as the attention selector.

At each timestep t, observation $o_t \in \mathbb{R}^{C\times H\times W}$ (C, H, W are respectively channels, height, and width) is passed to segmentation models and get k segments $\left\{o_t^i | i = 1, 2, \ldots, k\right\}$. Here we concatenate the segments along the channel level and denote the concatenated result as $o_t^{\text{seg}} \in \mathbb{R}^{kC\times H\times W}$. Then o_t and o_t^{seg} are passed through feature extractor ϕ and get their embeddings $\phi(o_t) \in \mathbb{R}^{1\times D}$ and $\phi(o_t^{\text{seg}}) \in \mathbb{R}^{k\times D}$, where D denotes the dimension of embedding.

After getting the embeddings of o_t and o_t^{seg}, n linear projections will be applied to them respectively, where n denotes the number of attention heads. And we denote the projection as W_i^q and W_i^k, $i \in \{1, 2, \cdots, n\}$. Thus we can get the "query" vector Q (red vector in Figure 1) and the "key" vector K (yellow vector in Figure 1):

$$Q_i = W_i^q(\phi(o_t)) \in \mathbb{R}^{1\times D}, i \in \{1, 2, \cdots, n\}, \quad (1)$$

$$K_i = W_i^k(\phi(o_t^{\text{seg}})) \in \mathbb{R}^{k\times D}, i \in \{1, 2, \cdots, n\}. \quad (2)$$

Like the traditional implementation of attention, Q and K are multiplied and then go through the Softmax function. The attention score A (green vector in Figure 1) is obtained by averaging the Softmax result of all attention heads:

$$A = \frac{1}{n}\sum\nolimits_{i=1}^{n} \text{Softmax}(Q_i K_i^T) \in \mathbb{R}^{1\times k}. \quad (3)$$

As for the "value" vector V, instead of using the embedding of segments, we directly use the segments themselves (segments marked in blue in Figure 1). We make such changes because the i-th value of A exactly means the attention paid to the i-th segment. Therefore, multiplying the attention value to its corresponding segment is equivalent to changing the saliency of the segment. The higher the attention value, the more pronounced the segment will be, and consequently, the more important the segment becomes.

Finally, we can directly get the selected frame o_t^{sel} by summing up the multiplied result of each segment:

$$o_t^{\text{sel}} = \sum\nolimits_{i=1}^{k} A_i o_t^i, \quad (4)$$

where A_i denotes the i-th value of A.

An additional benefit is that this attention mechanism can lead to better interpretability. In previous works, interpretability can only be acquired as a pixel-level mask, where the highlighted pixels represent the area of interest (Bertoin et al. 2022; Wu, Khetarpal, and Precup 2021; Mott et al. 2019). However, such mask is often not clear and needs humans to further judge the exact object the agent is focusing on. On the contrary, the selected frame produced by the attention selector can accurately locate the object of interest.

Self-supervised Objectives of the Attention Selector

We find that solely depending on RL loss to update the attention selector is sample-inefficient, some self-supervised objectives help accelerate the learning. Compared with previous works that introduce complex self-supervised objectives (Zhang et al. 2021; Wang et al. 2022b), FTD approaches the RL problem from an objective perspective, allowing simple self-supervised objectives to perform well.

Reward prediction As a main component of MDP, reward is used in many works to assist agent learning (Fu et al. 2021; Tomar et al. 2021). In our method, reward is also a useful signal. If the selector can correctly identify task-relevant parts and recombine them into frames without distraction, then o_t^{sel} can more easily predict reward r_t. As an auxiliary task of the selector, we introduce a reward predictor $\hat{\mathcal{R}}$, which needs to predict r_t given the selected frame o_t^{sel} and a_t. The loss function of the reward predictor can be written as:

$$\mathcal{L}_{\text{reward}} = \left(r_t - \hat{\mathcal{R}}\left(o_t^{\text{sel}}, a_t\right)\right)^2. \quad (5)$$

Inverse dynamic prediction Transition-related prediction is used in many works (Wang et al. 2022b; Tomar et al. 2021). Considering that forward dynamic prediction is hard to learn and may cause representation collapse, we use inverse dynamic prediction (Pathak et al. 2017) to learn transition information. Given two adjacent selected frames o_t^{sel} and o_{t+1}^{sel}, inverse dynamic predictor $\hat{\mathcal{P}}$ is designed to predict the action a_t. Its loss function can be written as:

$$\mathcal{L}_{\text{dynamic}} = \left(a_t - \hat{\mathcal{P}}\left(o_t^{\text{sel}}, o_{t+1}^{\text{sel}}\right)\right)^2. \quad (6)$$

Integration of Learning Objectives in Decision Stage

The selected frame o_t^{sel} given by the attention selector is then passed to the decision stage, which is designed as a sequential combination of feature extractor and decision layers.

As shown in Figure 1, we let the reward predictor $\hat{\mathcal{R}}$, inverse dynamic predictor $\hat{\mathcal{P}}$, and decision layers share the same attention selector and feature extractor. By simultaneously optimizing the above two unsupervised objectives and the RL objective, the selector can not only acquire task information brought by the RL objective but also learn the straightforward information of reward and transition to accelerate the learning.

Since the gradient is conductive throughout the entire network, the whole method can be trained in an end-to-end way by minimizing the total loss function $\mathcal{L}_{\text{total}}$:

$$\mathcal{L}_{\text{total}} = \eta_1\mathcal{L}_{\text{RL}} + \eta_2\mathcal{L}_{\text{reward}} + \eta_3\mathcal{L}_{\text{dynamic}}, \quad (7)$$

where all η_i are hyper-parameters and set to 1 by default.

Experiments

In this section, we begin by conducting experiments on eight tasks within the widely-recognized DeepMind Control Suite benchmark (Tassa et al. 2018). These experiments aim to demonstrate the superior performance of our method in noisy environments and to showcase the accuracy and visualization effects of the attention selector in identifying task-relevant objects. Subsequently, we carry out additional experiments in the more complex Franka Emika Robotics simulation environment (Yuan et al. 2023), highlighting the potential for its application in real-world training scenarios. Finally, we undertake a comprehensive ablation study to validate the effectiveness of each individual module. For details in the experiment, please refer to the appendix[1].

[1]https://www.lamda.nju.edu.cn/chenc/AAAI24-Appendix.pdf

Task setting Our task setting aims at simulating the real condition that agent is trained in a natural scene with varies task-irrelevant distractions, and we realize it by playing video as the background of observation. Previous works use grayscale image as the backgrounds (Fu et al. 2021; Wang et al. 2022b) or repeat a single video (Zhang et al. 2021), which may differ from real situation. In order to make the environment more realistic, a larger video dataset is played as background in RGB mode (Hansen, Su, and Wang 2021). Specifically, 80 color video clips will loop on the background during training, and 20 clips for testing. The played videos include a variety of types, ranging from indoor to outdoor.

Baseline methods We compare FTD with several baselines. DrQ-v2 (Yarats et al. 2021) is a widely used data augmentation method, which shows some effect in previous work. For visual distraction tasks, it is a mainstream idea to use the behavioral metric (Le Lan, Bellemare, and Castro 2021; Zhang et al. 2021) based on reward information to shape the representation space. According to different behavioral metrics defined, there are three methods worth considering: DBC (Zhang et al. 2021), MICo (Castro et al. 2021), and Q^2-learning (Liao, Zhang, and Yu 2023). Denoised-MDP (Wang et al. 2022b) is a representative model-based method specifically proposed for tasks involving noisy visual observation. According to the policy optimization mechanism, there are two kinds of Denoised-MDP implementations in the experiments: one backpropagating via dynamics (Denoised (D)) and the other using SAC on a learned MDP (Denoised (S)).

DeepMind Control

We choose DeepMind Control (DMC) Suite, a widely used benchmark, as our first environment. It is a comprehensive set of physical simulation environments that contains many control tasks, ranging from single object to multiple objects.

Tasks We select eight tasks from six DMC environments, respectively pendulum-swingup (ps), cartpole-swingup (cs), finger-spin (fs), hopper-stand (hs), hopper-hop (hh), cheetah-run (cr), walker-walk (ww), and walker-run (wr). The upper part of Figure 2 are examples of the distracted input.

Experiment results Table 1 shows the performance of FTD and baselines, and FTD ranks the first in five of eight tasks. It can be seen that when facing tasks with complex objects (walker, cheetah), the advantage of FTD is more obvious, while in tasks of simple-shaped objectives and sparse rewards, FTD may perform inferior to MICo and Q^2-learning.

As a representative of data augmentation methods, DrQ-v2 hardly acquires any improvement during training. This indicates that when facing environments distracted by task-irrelevant objects, the priors brought by the data augmentation method (e.g., random clip implies that the margin of observation is not important for decision making) are not sufficient to assist the agent in completing the task (Tomar et al. 2021). Consistent with the experiments conducted in previous work (Wang et al. 2022b), DBC fails in all eight environments. This indicates that the latent dynamic prediction and bi-simulation metric learning of DBC are challenging in noisy environments. By improving metric learning, MICo and Q^2-learning show effectiveness in tasks with simple-shaped objects like pendulum-swingup and cartpole-swingup but fail in more complex tasks like cheetah-run and walker-walk. This implies that learning such metrics in complex environments remains difficult.

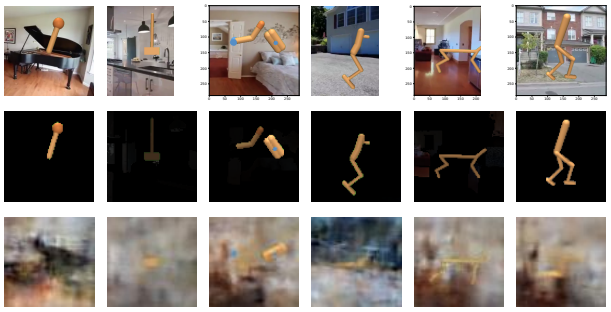

Figure 2: (Upper) Distracted observations of six DMC environments. (Middle) Selected frames of the attention selector. (Lower) Reconstruction frames of Denoised (D).

The two implementations of Denoised-MDP roughly perform the same, they both show some effect in the walker-walk and cheetah-run environments, but fail in other environments. Considering that Denoised-MDP is a model-based method, which may have advantage in sample-efficiency, it still performs worse than the model-free method FTD. Both Denoised-MDP and FTD require the agent to distinguish between task-relevant and task-irrelevant objects. Denoised-MDP operates in latent space, while FTD directly conducts on segmented observations, leading to differences in performance, especially when multiple task-relevant objects exist.

Further discussion of attention selector To visually demonstrate the performance of the attention selector, we plot the selection result o_t^{sel} in Figure 2. When facing distracted observations, the attention selector can correctly choose segments, thus producing clear images with only task-relevant segments highlighted. For comparison, We also plot the reconstruction frames of Denoised (D). It can be observed that these frames are full of noise and often miss important parts of the object (e.g., the finger in the finger environment and the pole in the cartpole environment). The objects reconstructed are also not accurate (e.g., the direction of the rotational body in the finger environment and the motion of legs in the cheetah environment), demonstrating that Denoised-MDP has not learned to separate task-relevant parts effectively.

To better illustrate the performance of the attention selector, we calculated the success rate of SAM, the success rate of selector, and the overall success rate across different tasks in Table 2. SAM's success rate is defined as the percentage of frames in which the task-relevant objects are within the segments produced by SAM. The selector's success rate is calculated based on the ratio of frames in which the task-relevant objects are given the highest value to the frames in which the task-relevant objects are correctly segmented. For example, in the walker environment, the attention selector is judged as correct if, and only if, the attention values of the walker is the highest. The overall success rate is calculated as

Task	DrQ-v2	DBC	MICo	Q^2-learning	Denoised (D)	Denoised (S)	FTD (Ours)
ps	284.7 ± 387.8	145.4 ± 180.5	176.3 ± 214.2	**546.4 ± 366.8**	0.0 ± 0.0	0.1 ± 0.2	498.8 ± 282.8
cs	137.9 ± 103.9	200.0 ± 85.5	**289.0 ± 47.5**	283.8 ± 39.4	180.6 ± 28.0	169.2 ± 36.8	207.3 ± 26.1
fs	1.1 ± 1.4	2.3 ± 2.9	494.9 ± 219.9	276.0 ± 212.2	19.2 ± 9.5	1.5 ± 1.2	**591.5± 146.0**
hs	2.7 ± 3.2	7.6 ± 7.6	203.4 ± 53.9	**249.7 ± 51.5**	39.8 ± 14.6	33.2 ± 13.8	112.0 ± 41.6
hh	0.0 ± 0.0	1.1 ± 1.4	0.7 ± 1.0	43.9 ± 15.8	4.3 ± 2.0	3.1 ± 3.5	**65.0 ± 26.6**
cr	1.3 ± 1.3	13.4 ± 4.0	16.5 ± 12.9	60.6 ± 29.0	40.4 ± 28.0	60.5 ± 33.0	**228.9± 43.1**
ww	30.6 ± 14.8	31.4 ± 8.3	142.5 ± 45.8	126.8 ± 65.4	109.5 ± 72.0	145.9 ± 24.9	**395.7± 48.9**
wr	24.4 ± 9.6	22.6 ± 4.4	81.8 ± 46.2	48.3 ± 25.0	42.9 ± 25.3	25.3 ± 15.2	**185.8 ± 12.3**

Table 1: Performance comparison of FTD and baselines on DeepMind Control (mean ± std). ps=pendulum_swingup, cs=cartpole_swingup, fs=finger_spin, hs=hopper_stand, hh=hopper_hop, cr=cheetah_run, ww=walker_walk, wr=walker_run.

Task	SAM	Selector	Overall
ps	0.98 ± 0.00	0.90 ± 0.01	0.88 ± 0.01
cs	0.50 ± 0.00	0.88 ± 0.00	0.44 ± 0.00
fs	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
hs	0.96 ± 0.00	0.95 ± 0.00	0.90 ± 0.00
hh	0.96 ± 0.00	0.95 ± 0.00	0.92 ± 0.00
cr	0.96 ± 0.00	0.98 ± 0.00	0.94 ± 0.00
ww	0.99 ± 0.00	0.97 ± 0.00	0.95 ± 0.00
wr	0.98 ± 0.00	0.97 ± 0.00	0.96 ± 0.00

Table 2: Success rates for SAM, Selector, and Overall.

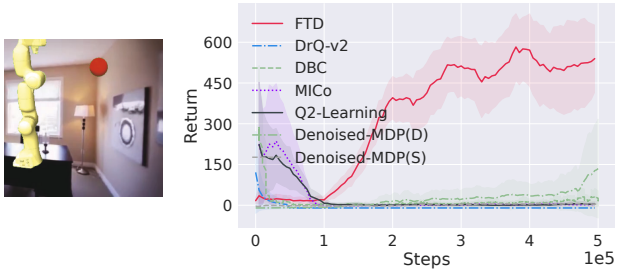

Figure 3: (Left) Franka-reach environment. (Right) Performance comparison of FTD and baselines on Franka-reach.

the product of SAM's success rate and the selector's success rate. The results are averaged over 20 episodes of interaction with the environment, covering all 20 video clips used for testing. The success rate of the attention selector exceeds 90% in seven out of eight tasks, demonstrating FTD's ability in capturing task-relevant objects. There exists a positive correlation between SAM's and selector's success rate, with a Pearson correlation coefficient of 0.72. This indicates that the performance of the foundation model used is critical to the learning of the attention selector. When we combine Table 1 and Table 2, we can also observe a positive correlation between the performance of FTD and the overall success rate. We find that the overall success rates are high in most tasks, leading to the best performance of FTD. However, for tasks such as pendulum-swingup, cartpole-swingup, and hopper-stand, their overall success rates are lower due to the low SAM's and/or selector's success rates, which is also directly reflected in the final performance of FTD.

Franka Emika Robotics

To evaluate FTD on more practical and realistic environments, we choose Franka Emika Robotics (Yuan et al. 2023) as our second experiment environment, and the task is called franka-reach. As shown in the left part of Figure 3, franka-reach is a task of manipulating a robotic arm to reach a certain area marked with a red ball. Compared with DMC tasks, franka-reach is a three-dimensional task, which places higher demands for the attention selector, requiring it to correctly classify task-relevant objects from different angles. In addition, franka-reach has significantly more joints than that of DMC tasks, and as such foundational segmentation model are more likely to segment the object of interest into multiple parts, which will cause greater difficulties to the selector.

The right part of Figure 3 shows the results of FTD and the baselines in this environment. We can see that all six baselines can hardly gain any reward. While in contrast, FTD successfully acquires the ability to catch the ball. This is consistent with the our previous claim that the advantage of FTD is more obvious in tasks with complex objects, and we believe that the performance of FTD on franka-reach further demonstrates its ability to learn in near-real situations.

Ablation Study

To identify the contribution of the key components of FTD, we conduct ablation study on FTD and its three variants:

- **FTD w/o SEL** is FTD without attention selector, and segments will be directly used as the input of SAC. Self-supervised objectives are retained.
- **FTD w/o SSO** is FTD without self-supervised objectives, but attention selector is retained.
- **FTD w/o SEL & SSO** is FTD without attention selector and self-supervised objectives. It is equivalent to substitute the original input frame of SAC with segmentation results stacked on channel dimension.

Experiment is conducted on the finger-spin environment. Figure 4 is the performance of FTD and its ablation vari-

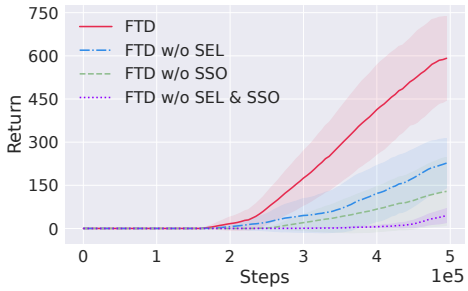

Figure 4: Ablation result on the finger-spin environment.

ants. From it we can see that FTD significantly outperforms other methods. The large performance gap between FTD and FTD w/o SEL shows that the proposed attention selector has obvious advantages than traditional CNN in tackling segmentation results. From the comparison between FTD and FTD w/o SSO, we can observe that simply introducing two common self-supervised losses can greatly aid the learning of attention selector, which shows the generality of FTD. Lastly, FTD w/o SEL & SSO can hardly acquire any reward, which means simply substituting input with segmentation is not enough, and our method better realizes the potential of foundational segmentation model in RL.

Related Work

Visual RL with distracted input Although visual RL has achieved great success in many simulation environments, when applied to environments with distracted visual input (e.g., play video on background), traditional visual RL algorithms face huge performance decline (Stone et al. 2021). Many works try to close this gap. DBC (Zhang et al. 2021) is the first method trying to solve this task. By fitting the bi-simulation metric where the distance between two states is only dependent on the probabilistic sequence of rewards, it is promising to learn a representation that is robust to distractions. However, DBC relies on precise latent dynamic prediction which can be inherently hard to learn. And the learning of the metric itself is not robust to the online policy learning process (Kemertas and Aumentado-Armstrong 2021; Liao, Zhang, and Yu 2023), which can result in less informative representation space. MICo resolves the first dilemma via sampling next states instead of predicting them and Q^2-learning decouples the process of metric learning and policy learning to make the metric robust and provides informative supervised signal to representation learning. Another way to handle distracted input is to explicitly distinguish between task-relevant parts and task-irrelevant parts. TIA (Fu et al. 2021) reformulates the MDP to explicitly separates states into reward-relevant parts and reward-irrelevant parts. Similar to the RSSM used in Dreamer (Hafner et al. 2020), TIA learns a world model based on the above two parts. Denoised MDP (Wang et al. 2022b) further extends the decomposition of the state into four parts based on both controllability and correlation with rewards. Similarly, Iso-Dream (Pan et al. 2022) and Iso-Dream++ (Pan et al. 2023) decouples state into controllable, uncontrollable, and time-invariant part. However, the reconstruction process in these methods are hard and time-consuming.

Some data augmentation methods that are proposed to improve generalization performance and data efficiency, like RAD (Yarats, Kostrikov, and Fergus 2021) and DrQ-v2 (Yarats et al. 2021), also demonstrate certain effectiveness when facing distracted input (Tomar et al. 2021), which may give credit to the priors implied in augmentations. As for more complex visual inputs, their abilities are limited.

Foundation models for visual RL Foundational models pretrained on diverse data at scale have demonstrated exceptional abilities in knowledge transfer to various downstream tasks (Bommasani et al. 2022). Researchers posit that RL can leverage common knowledge from foundational models to solve tasks faster and generalize better (Yang et al. 2023). Numerous ongoing efforts aim to replicate this success for visual RL. Currently, two primary approaches stand out: one involves directly using the foundational model as a pretrained feature extractor, such as VRL3 (Wang et al. 2022a), which leverages pretrained representations to increase the sample efficiency in the offline-to-online fine-tuning process, and PIE-G (Yuan et al. 2022), which employs fixed pretrained representations to exhibit good generalization performance. The other approach is to rely on multimodal foundation models to assist in planning or reward definition, like PaLM-E (Driess et al. 2023), which utilizes multimodal models for high-level planning, hoping that language-based descriptions of actions will generalize better than low-level motor controls, and MineDojo (Fan et al. 2022), which defines reward information by aligning image and language modalities in the agent's observation and goal space. However, none of them have specifically considered the problem of noise interference in the environment, which is an unavoidable situation when training in real-world conditions.

Conclusion and Future Work

In this work, we propose a novel Focus-Then-Decide (FTD) framework to tackle the challenge of learning in noisy environments by integrating a foundational segmentation model into visual RL. A novel attention selector is employed to focus on task-relevant objects, which are then utilized for the training of the decision-making module. Experimental results on the DeepMind Control Suite and Franka Emika Robotics indicate that our method can efficiently identify task-relevant objects, thereby achieving strong performance. We hope our work will inspire the community to further investigate the use of foundational models to address complex visual RL tasks.

For future work, we will explore more self-supervised objectives, especially those targeted at the attention selector, to further enhance the FTD's performance. Concerning the permutation invariance and the variable quantity of segments, we will try to process them using some network structures specifically designed for this kind of set-input. Additionally, since the performance of FTD is contingent on the segmentation accuracy of the foundational segmentation model, and considering that segmentation speed and accuracy may not be compatible, it is also warranted to research more suitable foundational segmentation models for RL downstream tasks.

Acknowledgements

We thank the reviewers for their insightful and valuable comments. We thank Manjie Xu, Pengyuan Wang, Fuxiang Zhang, Zehua Xia, and Jiafei Lyu for their helpful discussions and support. This work is supported by the National Key R&D Program of China (2022ZD0114804), the National Science Foundation of China (62276126, 62250069), and the Tencent AI Lab (RBFR2023011).

References

Andrychowicz, M.; Baker, B.; Chociej, M.; Józefowicz, R.; McGrew, B.; Pachocki, J.; Petron, A.; Plappert, M.; Powell, G.; Ray, A.; Schneider, J.; Sidor, S.; Tobin, J.; Welinder, P.; Weng, L.; and Zaremba, W. 2020. Learning Dexterous In-hand Manipulation. *International Journal of Robotics Research*, 39: 3–20.

Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2012. The Arcade Learning Environment: An Evaluation Platform for General Agents. *arXiv preprint arXiv:1207.4708*.

Bertoin, D.; Zouitine, A.; Zouitine, M.; and Rachelson, E. 2022. Look Where You Look! Saliency-guided Q-networks for Visual RL Tasks. *arXiv preprint arXiv:2209.09203*.

Bommasani, R.; Hudson, D. A.; Adeli, E.; and et al. 2022. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258v3*.

Castro, P. S.; Kastner, T.; Panangaden, P.; and Rowland, M. 2021. MICo: Improved Representations via Sampling-based State Similarity for Markov Decision Processes. In *Advances in Neural Information Processing Systems*, 30113–30126.

Chen, J.; Li, S. E.; and Tomizuka, M. 2022. Interpretable End-to-end Urban Autonomous Driving With Latent Deep Reinforcement Learning. *IEEE Transactions on Intelligent Transportation Systems*, 23: 5068–5078.

Driess, D.; Xia, F.; Sajjadi, M. S. M.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; Huang, W.; Chebotar, Y.; Sermanet, P.; Duckworth, D.; Levine, S.; Vanhoucke, V.; Hausman, K.; Toussaint, M.; Greff, K.; Zeng, A.; Mordatch, I.; and Florence, P. 2023. PaLM-E: An Embodied Multimodal Language Model. *arXiv preprint arXiv:2303.03378*.

Fan, L.; Wang, G.; Jiang, Y.; Mandlekar, A.; Yang, Y.; Zhu, H.; Tang, A.; Huang, D.; Zhu, Y.; and Anandkumar, A. 2022. MineDojo: Building Open-ended Embodied Agents with Internet-scale Knowledge. In *Advances in Neural Information Processing Systems*, 18343–18362.

Fu, X.; Yang, G.; Agrawal, P.; and Jaakkola, T. S. 2021. Learning Task Informed Abstractions. In *International Conference on Machine Learning*, 3480–3491.

Goldstein, E. B.; and Cacciamani, L. 2021. *Sensation and perception*. Cengage.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft Actor-critic: Off-policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *International Conference on Machine Learning*, 1856–1865.

Hafner, D.; Lillicrap, T. P.; Ba, J.; and Norouzi, M. 2020. Dream to Control: Learning Behaviors by Latent Imagination. In *International Conference on Learning Representations*.

Hansen, N.; Su, H.; and Wang, X. 2021. Stabilizing Deep Q-learning with ConvNets and Vision Transformers under Data Augmentation. *arXiv preprint arXiv:2107.00644*.

Hessel, M.; Modayil, J.; van Hasselt, H.; Schaul, T.; Ostrovski, G.; Dabney, W.; Horgan, D.; Piot, B.; Azar, M. G.; and Silver, D. 2018. Rainbow: Combining Improvements in Deep Reinforcement Learning. In *AAAI Conference on Artificial Intelligence*, 3215–3222.

Kemertas, M.; and Aumentado-Armstrong, T. 2021. Towards Robust Bisimulation Metric Learning. In *Advances in Neural Information Processing Systems*, 4764–4777.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.; Dollár, P.; and Girshick, R. B. 2023. Segment Anything. *arXiv preprint arXiv:2304.02643*.

Le Lan, C.; Bellemare, M. G.; and Castro, P. S. 2021. Metrics and Continuity in Reinforcement Learning. In *AAAI Conference on Artificial Intelligence*, 8261–8269.

Li, Y.; Mao, H.; Girshick, R. B.; and He, K. 2022. Exploring Plain Vision Transformer Backbones for Object Detection. In *European Conference on Computer Vision*, 280–296.

Liao, W.; Zhang, Z.; and Yu, Y. 2023. Policy-independent Behavioral Metric-based Representation for Deep Reinforcement Learning. In *AAAI Conference on Artificial Intelligence*, 8746–8754.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, 740–755.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *International Conference on Computer Vision*, 9992–10002.

Mott, A.; Zoran, D.; Chrzanowski, M.; Wierstra, D.; and Rezende, D. J. 2019. Towards Interpretable Reinforcement Learning Using Attention Augmented Agents. In *Advances in Neural Information Processing Systems*, 12329–12338.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774v3*.

Pan, M.; Zhu, X.; Wang, Y.; and Yang, X. 2022. Iso-dream: Isolating and leveraging noncontrollable visual dynamics in world models. In *Advances in Neural Information Processing Systems*, 23178–23191.

Pan, M.; Zhu, X.; Zheng, Y.; Wang, Y.; and Yang, X. 2023. Model-Based Reinforcement Learning with Isolated Imaginations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–15.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-driven Exploration by Self-supervised Prediction. In *International Conference on Machine Learning*, 2778–2787.

Shotton, J.; Winn, J. M.; Rother, C.; and Criminisi, A. 2006. *TextonBoost*: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation. In *European Conference on Computer Vision*, 1–15.

Stone, A.; Ramirez, O.; Konolige, K.; and Jonschkowski, R. 2021. The Distracting Control Suite - A Challenging Benchmark for Reinforcement Learning from Pixels. *arXiv preprint arXiv:2101.02722*.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction (Second Edition)*. MIT Press.

Szeliski, R. 2022. *Computer Vision: Algorithms and Applications (Second Edition)*. Springer.

Tassa, Y.; Doron, Y.; Muldal, A.; Erez, T.; Li, Y.; de Las Casas, D.; Budden, D.; Abdolmaleki, A.; Merel, J.; Lefrancq, A.; Lillicrap, T. P.; and Riedmiller, M. A. 2018. DeepMind Control Suite. *arXiv preprint arXiv:1801.00690*.

Tomar, M.; Mishra, U. A.; Zhang, A.; and Taylor, M. E. 2021. Learning Representations for Pixel-based Control: What Matters and Why? *arXiv preprint arXiv:2111.07775*.

Wang, C.; Luo, X.; Ross, K. W.; and Li, D. 2022a. VRL3: A Data-driven Framework for Visual Deep Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 32974–32988.

Wang, T.; Du, S. S.; Torralba, A.; Isola, P.; Zhang, A.; and Tian, Y. 2022b. Denoised MDPs: Learning World Models Better Than the World Itself. In *International Conference on Machine Learning*, 22591–22612.

Wu, C.; and Zhang, Z. 2023. Surfing Information: The Challenge of Intelligent Decision-Making. *Intelligent Computing*, 2: Article 0041.

Wu, H.; Khetarpal, K.; and Precup, D. 2021. Self-supervised Attention-aware Reinforcement Learning. In *AAAI Conference on Artificial Intelligence*, 10311–10319.

Xu, M.; Jiang, G.; Liang, W.; Zhang, C.; and Zhu, Y. 2023a. Active Reasoning in An Open-world Environment. In *Advances in Neural Information Processing Systems*.

Xu, M.; Jiang, G.; Liang, W.; Zhang, C.; and Zhu, Y. 2023b. Interactive Visual Reasoning under Uncertainty. In *Advances in Neural Information Processing Systems*.

Yang, S.; Nachum, O.; Du, Y.; Wei, J.; Abbeel, P.; and Schuurmans, D. 2023. Foundation Models for Decision Making: Problems, Methods, and Opportunities. *arXiv preprint arXiv:2303.04129*.

Yarats, D.; Fergus, R.; Lazaric, A.; and Pinto, L. 2021. Mastering visual continuous control: Improved data-augmented reinforcement learning. *arXiv preprint arXiv:2107.09645*.

Yarats, D.; Kostrikov, I.; and Fergus, R. 2021. Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels. In *International Conference on Learning Representations*.

Yi, Q.; Zhang, R.; Peng, S.; Guo, J.; Hu, X.; Du, Z.; Zhang, X.; Guo, Q.; and Chen, Y. 2022. Object-category Aware Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 36453–36465.

Yuan, Z.; Xue, Z.; Yuan, B.; Wang, X.; Wu, Y.; Gao, Y.; and Xu, H. 2022. Pre-trained Image Encoder for Generalizable Visual Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 13022–13037.

Yuan, Z.; Yang, S.; Hua, P.; Chang, C.; Hu, K.; Wang, X.; and Xu, H. 2023. RL-ViGen: A Reinforcement Learning Benchmark for Visual Generalization. *arXiv preprint arXiv:2307.10224*.

Zhang, A.; McAllister, R. T.; Calandra, R.; Gal, Y.; and Levine, S. 2021. Learning Invariant Representations for Reinforcement Learning without Reconstruction. In *International Conference on Learning Representations*.

Zhang, C.; Han, D.; Qiao, Y.; Kim, J. U.; Bae, S.; Lee, S.; and Hong, C. S. 2023. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. *arXiv preprint arXiv:2306.14289*.

Zhao, X.; Ding, W.; An, Y.; Du, Y.; Yu, T.; Li, M.; Tang, M.; and Wang, J. 2023. Fast Segment Anything. *arXiv preprint arXiv:2306.12156*.