Generalizing Consistency Policy to Visual RL with Prioritized Proximal Experience Regularization

Haoran Li
Institute of Automation,
Chinese Academy of Sciences
University of Chinese Academy of Sciences
`lihaoran2015@ia.ac.cn`

Zhennan Jiang
Institute of Automation,
Chinese Academy of Sciences
University of Chinese Academy of Sciences
`jiangzhennan2024@ia.ac.cn`

Yuhui Chen
Institute of Automation,
Chinese Academy of Sciences
University of Chinese Academy of Sciences
`chenyuhui2022@ia.ac.cn`

Dongbin Zhao[*]
Institute of Automation,
Chinese Academy of Sciences
University of Chinese Academy of Sciences
`dongbin.zhao@ia.ac.cn`

Abstract

With high-dimensional state spaces, visual reinforcement learning (RL) faces significant challenges in exploitation and exploration, resulting in low sample efficiency and training stability. As a time-efficient diffusion model, although consistency models have been validated in online state-based RL, it is still an open question whether it can be extended to visual RL. In this paper, we investigate the impact of non-stationary distribution and the actor-critic framework on consistency policy in online RL, and find that consistency policy was unstable during the training, especially in visual RL with the high-dimensional state space. To this end, we suggest sample-based entropy regularization to stabilize the policy training, and propose a consistency policy with prioritized proximal experience regularization (CP3ER) to improve sample efficiency. CP3ER achieves new state-of-the-art (SOTA) performance in 21 tasks across DeepMind control suite and Meta-world. To the best of our knowledge, CP3ER is the first method to apply diffusion/consistency models to visual RL and demonstrates the potential of consistency models in visual RL. Our project page is hosted at `https://jzndd.github.io/CP3ER-Page/`.

1 Introduction

RL has achieved remarkable results in many fields, such as video games [1], Go [2], Chess [3, 4] and robotics [5–8]. Since it is hard to parameterize the complex policy distribution over high-dimensional state and continuous action spaces, the performance and stability of visual RL are still unsatisfactory. As the most common policy distribution, Gaussian distribution is easy to sample, but its unimodal nature limits the expressiveness to represent complex behaviors [9]. While complex distributions have the rich, expressive power to improve the exploration ability [10], the difficulty of sampling makes it hard to apply to online RL directly. Parameterizing the complex policy distribution to balance ease of sampling and expressiveness is a bottleneck to improving the efficiency of visual RL.

As an emerging generation model, the diffusion model [11] stands out in fields such as image generation [12–14] and video generation [15, 16] with its ability to model complex distributions and ease of sampling characteristics. These properties have also been explored for learning a complex

---

[*]Corresponding author.

policy [17]. For example, diffusion models are used to imitat e the diverse expert policies [18, 19] or trajectories [20–23] in datasets. In addition, due to their excellent expressive and data generation abilities, diffusion models are often employed to address policy constraints [24–26] and data scarcity [21, 27, 28] in offline RL. Most of these applications are limited to offline learning due to the demand for pre-collected datasets to train diffusion models.

Applying diffusion models in online RL will face different problems than offline RL. Firstly, unlike pre-collected data in offline RL, the data distribution in online RL is non-stationary [29], and it is currently unclear whether this change will impact training diffusion models. Secondly, since the optimal policy distribution is unknown, samples from this distribution are inaccessible, resulting in the ill-posed traditional score matching problem [30]. In addition, the time-inefficiency of diffusion models [31] will become more prominent with a large number of online interactions, leading to unacceptable time costs for online learning. As an efficient diffusion model, the consistency model [32] directly establishes a mapping from noise to denoised data, which is employed for online RL and achieves time efficiency and better performance [33, 34]. These methods simply replace the Gaussian model in the actor-critic framework with the consistency model and train consistency policy with the Q-loss. Although they achieve competitive performance in state-based RL tasks, this training method is incompatible with traditional score matching for diffusion models. Therefore, the question is whether this training framework is suitable for consistency model-based policy training, especially for visual RL tasks with high-dimensional state spaces.

In this paper, we investigate the impact of non-stationary dataset and the actor-critic framework on consistency policy. By analyzing the dormant ratio [29] of the policy network, we find that the non-stationary of training data is not the main factor affecting the instability of consistency policy, while the Q-loss in the actor-critic framework leads to a sharp increase in the dormant ratio of the policy network, resulting in the loss of complex expression ability, which is particularly significant in visual RL tasks. To address the above issues, we suggest sample-based entropy regularization to stabilize the policy training and propose the prioritized proximal experience regularization, which uses weighted sampling to construct an appropriate proxy policy for policy regularization and achieves sample-efficiency consistency policy. Overall, our contributions are as follows:

- We investigate the impact of non-stationary distribution and actor-critic framework on consistency policy in online RL, and find that the Q-loss of the actor-critic framework can impair the expressive ability of the consistency model, leading to unstable policy training. This phenomenon is particularly significant in visual RL tasks.

- We suggest sample-based entropy regularization and propose a consistency policy with prioritized proximal experience regularization (CP3ER) which significantly enhances the stability of policy training with the Q-loss under the actor-critic framework.

- Our proposed method performs new SOTA in 21 visual control tasks, including DeepMind control suite and Meta-world tasks. To our knowledge, our proposed CP3ER is the first method to apply diffusion/consistency models to visual RL.

2 Related Work

2.1 Diffusion Model in Reinforcement Learning

Due to its high-quality sample generation ability and training stability, diffusion models [11] have been widely applied in fields such as image generation, video generation, and natural language processing and have also been promoted in RL. Since the diffusion model can represent complex distribution in datasets, it is commonly used in offline RL to model behavior policies [18, 25, 35] or expected policies [34, 36–38] to meet the requirements of diversity policy constraints and achieve a balance between constraint and exploitation. The diffusion model can also model trajectory distribution [20, 39, 40], achieving specified trajectory generation under different guidance. In addition, diffusion models are also employed to generate data to augment limited training data [27, 28].

Although diffusion models have been widely applied in offline learning, using diffusion models in online learning remains a challenging problem. [41] proposes the concept of action gradient, which uses a value function to estimate the gradient of actions and updates the actions in the replay buffer. The diffusion model-based policy is trained based on the updated actions. [42] employs a diffusion model as the world model to generate complete rollouts at once instead of auto-regressive generation.

[30] introduces the Q-score matching (QSM), which iteratively matches the parameterized score of a policy with the action gradient of its Q-function. Considering the low inference efficiency and long training time of diffusion models in RL training, [33] and [34] use consistency models instead of diffusion models and implement policy training under the actor-critic framework, achieving excellent performance in continuous control tasks.

2.2 Visual Reinforcement Learning

Compared to state-based RL, visual RL is faced with high-dimensional state space and continuous action space and is sensitive to training parameters and random seeds, which leads to unstable training and sample inefficiency. Image data augmentation [43–45] is a common technique to alleviate the above problems. In addition, auxilliary losses are usually combined to improve the efficiency of state representation learning from the image, such as contrastive learning loss [46], state representation loss [47, 48], action and state representation loss [49], and self-supervised loss [50]. Recent works have focused on enhancing the stability of visual RL from a micro perspective of neural networks. For example, [51] proposes the visual dead trial phenomenon and introduces an adaptive regularization method for convolutional features. [52] proposes the concept of dormant neuron phenomenon to explain the behavior of the policy network during RL training. [53] controls the dormant ratio of the policy network during training so that it achieves the SOTA performance on multiple tasks.

3 Preliminary

3.1 Reinforcement Learning

Online RL solves sequential decision problems, typically modeled through Markov Decision Processes (MDP). MDP is represented by 6 tuples $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \rho_0, \gamma)$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{R}$ and $\mathcal{T}$ represent the reward function and state transition function of the environment, respectively. ρ_0 is the initial distribution of the state, and γ is the discount factor. In visual RL, it is difficult for agents to directly obtain the state s_t from the image $o_t \in \mathcal{O}$, where $\mathcal{O}$ is observation space. Therefore, an image encoder $f(\cdot)$ is usually required, and the state is estimated from the image through this encoder. The goal of the agent is to learn an optimal policy π^* and the corresponding encoder f^* to maximize the expected cumulative reward $\mathbb{E}_{\pi(f(\cdot))}[\sum_{t=0}^{\infty} \gamma^t r_t]$ under that policy.

3.2 Consistency Policy

The consistency model [32] is a new diffusion model proposed to address the time inefficiency caused by hundreds of reverse diffusion steps in diffusion models. It replaces the iterative denoising process with learned score functions in traditional diffusion models by constructing a mapping between noise and denoised samples, and directly maps any point on the probability flow ordinary differential equation (ODE) trajectory to the original data in the reverse diffusion process. Thus, it only requires a small number of steps or even one step to achieve the generation from noise to denoised data. Consistency policy [33, 34] is a new policy representation under the actor-critic framework, which replaces traditional Gaussian models with the consistency model and updates the policy by maximizing the state-action value. Consistency policy is defined as

$$\pi_\theta(a_t|s_t) \triangleq c_{skip}(\tau_k)a_t^{\tau_k} + c_{out}(\tau_k)F_\theta(a_t^{\tau_k}, \tau_k|s_t) \tag{1}$$

where $\{\tau_k|k \in [N]\}$ a sub-sequence of time points on the time period $[\epsilon, K]$ with $\tau_1 = \epsilon$ and $\tau_N = K$. $a_t^{\tau_k}$ is the noised action and $a_t^{\tau_k} = a_t + \tau_k z$ where $z \sim \mathcal{N}(0, I)$ is Gaussian noise. F_θ is a trainable network that takes the state s_t as a condition and outputs an action of the same dimension as the input a_t^k. $c_{skip}(\cdot)$ and $c_{out}(\cdot)$ are differentiable functions such that $c_{skip}(\epsilon) = 1$ and $c_{out}(\epsilon) = 0$ to ensure consistency policy is differentiable at $\tau_k = \epsilon$. ϵ is a real number close to 0. To train this policy, [33] directly applies the above policy to the actor-critic framework and updates the policy using the following the Q-loss, which is named Consistency-AC.

$$\mathcal{L}_a(\theta) = -\mathbb{E}_{s_t \sim \mathcal{B}, a_t \sim \pi_\theta}[Q_\phi(s_t, a_t)] \tag{2}$$

where $\mathcal{B}$ is the replay buffer and Q_ϕ is the state-action value function. Compared to diffusion-based policies, consistency policy have significant advantages in inference speed and performance in online learning tasks [33].

3.3 Dormant Ratio of Neural Networks

The expressive ability is crucial for training the policy with RL. [29] proposes the concept of dormant ratio β_r, which quantifies the expression ability of a neural network by calculating the proportion of dormant neurons in the neural networks.

$$\beta_r = \frac{\sum_l H_\tau^l}{\sum_l N^l} \tag{3}$$

where N^l represents the number of neurons in the l-th layer. H_τ^l is the number of neurons in the l-th layer whose score s_i^l is less than τ. The score of each neuron is calculated as follows:

$$s_i^l = \frac{\mathbb{E}_{x\in\mathcal{D}}|h_i^l(x)|}{\frac{1}{N^l}\sum_{k\in l}\mathbb{E}_{x\in\mathcal{D}}|h_k^l(x)|} \tag{4}$$

Here $h_i^l(\cdot)$ is the activation function of the i-th neuron in the l-th layer. $\mathcal{D}$ is the distribution of the input x. In the following sections of this paper, we use the dormant ratio to evaluate the expression ability of consistency policy during the training.

As introduced in [29], the dormant ratio of a neural network indicates the proportion of inactive neurons and is typically used to measure the activity of the network. A higher dormant ratio implies fewer active neurons in the network, implying the network's capacity and expressiveness are damaged. In RL, the episode return is closely related to the dormant ratio of the policy network. A higher dormant ratio results in more lazy action outputs, inactive agent behavior, and lower episode returns; conversely, when policy performance is good, the policy network is usually more active, and the dormant ratio is typically lower. This phenomenon has been reported in [29, 53–55].

4 Is Consistency-AC Applicable to Visual RL?

Does the non-stationary distribution in online RL affect the training of consistency models? Unlike offline RL, online RL does not have pre-collected datasets. The data distribution for training the policy is constantly changing with policy improvement. So, whether this non-stationarity distribution affects the training of consistency models is a question that needs to be explored. In order to investigate the impact of non-stationarity of data for consistency model training, we record the dormant ratio of the policy network during consistency model training under two different settings: online training and offline training. We employ two tasks (MuJoCo Halfcheetah and MuJoCo Walker2d) and conduct 4 random seeds for each setting. In order to eliminate the impact of Q-loss, we follow the behavior clone setting and train the consistency model with consistency loss [32] using data from offline datasets or online replay buffers. The distribution of the data in the replay buffer varies with policy improvement. The results are shown in Figure 1. Although there is a difference in the dormant ratios between online and offline learning settings in the Halfcheetah task, the overall trend is the same. We speculate that this difference is caused by the diversity of the samples included in the dataset. For the Walker2d task, the dormant ratios are nearly the same under two different settings. Therefore, we can infer that the non-stationary distribution of online RL does not significantly affect the consistency model training.
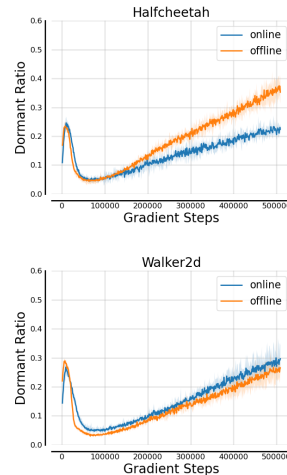

Figure 1: The dormant ratios of the policy under the online and offline training.

Is the actor-critic framework suitable for training consistency policy? The actor-critic framework is a highly effective policy training framework for online RL, in which the policy network achieves policy improvement by maximizing the value function. Some works [33, 34] directly apply consistency models to this framework. Although they achieve good results in RL tasks with low dimensional state spaces, whether the actor-critic training framework is compatible with consistency model training remains a question that needs further investigation. To evaluate the impact of the actor-critic framework on the training of consistency models, we compare the dormant ratios of policy

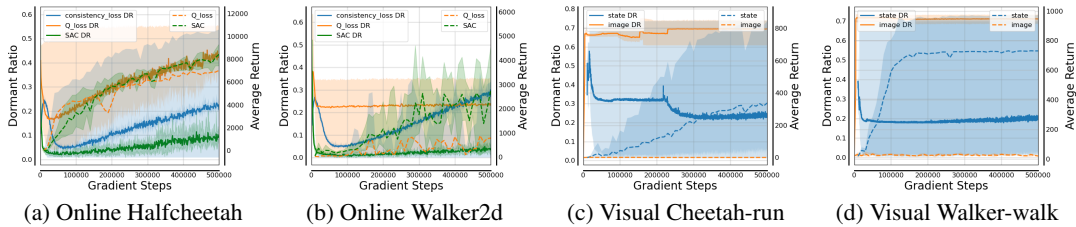| (a) Online Halfcheetah | (b) Online Walker2d | (c) Visual Cheetah-run | (d) Visual Walker-walk |

Figure 2: The dormant ratios of the policy networks with different losses and observations.

networks under the consistency loss and Q-loss settings under the actor-critic framework. The results are shown in (a) and (b) of Figure 2, the solid line shows the dormant ratio of the policy while the dashed line shows the performance of the policy. When training the policy with the consistency loss, the dormant ratios of the network show a trend of first decreasing and then increasing. This means that the policy learns the distribution from the data and then overfits the distribution. When training the policy with the Q-loss, the dormant ratio of the policy network will rapidly increase and maintain a high value, which means that the policy network will quickly fall into local optima, making the policy no longer change. In addition, we can also see that when using the Q-loss, the variance of the dormant ratios is relatively large under different random seeds. When the dormant ratio is low, the policy network can iterate properly to learn good policy. Therefore, we can determine that the Q-loss under the actor-critic framework will destabilize the consistency policy training.

Will high-dimensional state space exacerbate the degradation phenomenon of consistency policy? Compared to RL with low dimensional state space, training stability in visual RL is still a challenge. In order to investigate whether the degradation phenomenon of consistency policy will become more significant under visual RL tasks, we compare the dormant ratios of the policy networks with the state as input and image as input on 2 tasks (Walker-walk and Cheetah-run in DeepMind control suite) under the setting of online learning. During the training process, only the Q-loss was used. To maintain consistency in the settings, we only count the dormant ratio of the multilayer perceptron (MLP) of the policy newtork in the image-based settings. The results are shown in (c) and (d) of Figure 2. Similar to using the state as input, in visual RL with the image as input, most of the neurons in the MLP of the policy network go dormant. Unlike the high variance of the former, the dormant ratios of consistency policy network in visual RL maintain a low variance and a high value. This indicates that there have been almost no successful trials under different random seeds. Therefore, we can infer that visual RL will exacerbate the instability of consistency policy training caused by the Q-loss under the actor-critic framework.

5 Consistency Policy with Prioritized Proximal Experience Regularization

Consistency Policy with Entropy Regularization. To solve the problem of consistency policy quickly falling into local optima under the influence of the Q-loss, we introduce policy regularization to stabilize policy improvement. Here, we employ entropy regularization to constrain policy behavior. The objective of RL is:

$$J(\theta) = \mathbb{E}_{s_t \sim \mathcal{B}, a_t \sim \pi_\theta} \big[\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t)\big] - \eta \mathbb{E}_{s_t \sim \mathcal{B}, a_t \sim \pi_\beta}[\log \pi_\theta(a_t|s_t)] \tag{5}$$

where π_β is the proxy distribution required for policy regularization. Entropy regularization is a commonly method for stabilizing policy training in RL. When π_β is a uniform distribution, the above objective is equal to maximum entropy RL, which maximizes the entropy of the policy while optimizing the return. The prerequisite for this method is to obtain the closed form of the policy distribution to calculate its entropy. However, for diffusion models or consistency models, obtaining the closed form of the policy distribution is very difficult. Thanks to the development of generative models, we can use score matching instead of solving analytic entropy in entropy regularization RL, thus achieving sample-based policy regularization. Therefore, the training loss of consistency policy with the entropy regularization is:

$$\mathcal{L}_a^r(\theta) = -\mathbb{E}_{s_t \sim \mathcal{B}, a_t \sim \pi_\theta}[Q_\phi(s_t, a_t)] + \eta \mathcal{L}_c(\theta) \tag{6}$$
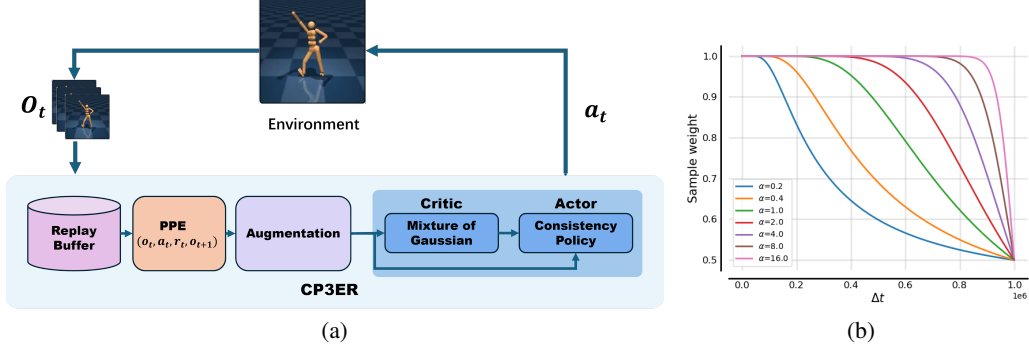
Figure 3: (a) The framework of CP3ER, where PPE is the abbreviation of prioritized proximal experience. (b) The sampling weights β with different α.

where $\mathcal{L}_c$ is consistency loss defined by following:

$$\mathcal{L}_c(\theta) = \mathbb{E}_{k\sim\mathcal{U}(1,N-1),s_t\sim\mathcal{B},a_t\sim\pi_\beta,z\sim\mathcal{N}(0,I)}[\lambda(\tau_k)d(\pi_\theta(s_t,a_t^{\tau_{k+1}},\tau_{k+1}),\pi_{\bar{\theta}}(s_t,a_t^{\tau_k},\tau_k)) \quad (7)$$

Here $\lambda(\cdot)$ is a step-dependent weight function, $d(\cdot,\cdot)$ is a distance metric. Since there is no need to obtain the closed form of the proxy distribution, only the data under that distribution needs to be obtained, making the selection of proxy distribution flexible. The remaining question is how to construct a suitable proxy distribution π_β.

Prioritized Proximal Experience Regularization. When the proxy distribution is uniform, this method approximates the maximum entropy consistency policy (MaxEnt CP). It should be noted that the difference between the proxy distribution and the optimal policy distribution can lead to difficulty in optimizing the above objectives. When the proxy distribution is far from the optimal policy or the proxy distribution is complex, the above optimization objectives require more samples to converge to better results. To better balance training stability and sample efficiency, we propose the prioritized proximal experience regularization (PPER). Specifically, when sampling data from the replay buffer, we design sampling weight β for each data instead of sampling the data uniformly.

$$\beta = \frac{1}{1 + \exp\left(2\alpha - \alpha\frac{2|\mathbf{B}|}{\Delta t}\right)} \quad (8)$$

where α is the hyperparameter and Δt is the interval between the sample generation step and the current step. $|\mathbf{B}|$ is the capacity of the replay buffer. The curves with different α are shown in Figure 3 (b). In the above settings, data closer to the current step will be sampled with a higher probability, while data farther away will have a lower probability. We refer to the above sampling method as a prioritized proximal experience (PPE).

For policy evaluation, we suggest a distributional value function instead of a deterministic value function to ensure the stability and accuracy of the value estimation. Precisely, we follow [56] and use a mixture of Gaussian (MoG) to model the distribution of the state-action value. When MoG is employed for value distribution, the following loss is used to update Q-function.

$$\mathcal{L}_q(\phi) = -\mathbb{E}_{(s_t,s_{t+1})\sim\mathcal{B},a_t\sim\pi_\theta}[\frac{1}{M}\sum_{i=1}^{M}\log Z_\phi^{(s_t,a_t)}(r_t(s_t,a_t)+\gamma z_i')], \text{where} \begin{cases} z_i' \sim Z_{\bar{\phi}}^{(s_{t+1},a_{t+1})} \\ a_{t+1} \sim \pi_\theta(s_{t+1}) \end{cases} \quad (9)$$

where $Z_\phi^{(s_t,a_t)}$ is the estimated value distribution. According to the equation (9), we need to sample M target Q-values z_i' and update the value distribution. Different from [56], we sample only one next action a_{t+1} instead of multiple actions to reduce the time cost and find that this simplification can achieve good experiment results. Considering the simplicity and efficiency of DrQ-v2 [57], our proposed consistency policy with prioritized proximal experience regularization (CP3ER) is built based on DrQ-v2. The framework is shown as Figure 3 (a). We sample the data from the replay buffer, and augment the image with the techniques in DrQ-v2. Thanks to the natural randomness of the action from consistency policy, our method no longer requires additional exploration strategies.

6

In addition, we only used a single Q-network to estimate the mean and variance of the mixture of Gaussian instead of double Q-network. We consider prioritized proximal experience regularization when updating consistency policy and used equation (9) when training the Q-network, which differs from DrQ-v2. The complete algorithm is included in the appendix B.1.

6 Experiments

In this section, we evaluate the proposed method from the following aspects: 1) Does CP3ER have performance advantages compared to current SOTA methods? 2) Can policy regularization improve the behavior of the policy? 3) What is the impact of different modules on the performance?

6.1 Visual Continuous Control Tasks

Environment Setup. We evaluate the methods on 21 visual control tasks from DeepMind control suite [58] and Meta-world [59]. We split these tasks into three domains, including 8 medium-level tasks in the DeepMind control suite, 7 hard-level tasks in the DeepMind control suite, and 6 tasks in the Meta-world. The details of each domain are included in the appendix C.

Baselines. We compare current advanced model-free visual RL methods, including DrQ-v2 [57], ALIX [60] and TACO [61]. The more detailed results are shown in the appendix C.

6.1.1 Does CP3ER have performance advantages compared to current SOTA methods?

Medium-level tasks in DeepMind control suite. We evaluate CP3ER on 8 medium-level tasks [57] in DeepMind control suite. The results are shown in Figure 4. From the left part of Figure 4, it can be seen that compared to the current SOTA methods, our proposed CP3ER has achieved better sample efficiency. It should be noted that TACO uses auxiliary losses of action and state representation during training to improve sample efficiency. Moreover, our proposed CP3ER uses no additional losses or exploration strategies. On the right part of Figure 4, we compare the mean, Interquartile Mean (IQM), median, and optimal gap of these methods. CP3ER has significant advantages in all metrics and has more minor variance. This means that CP3ER has better training stability.
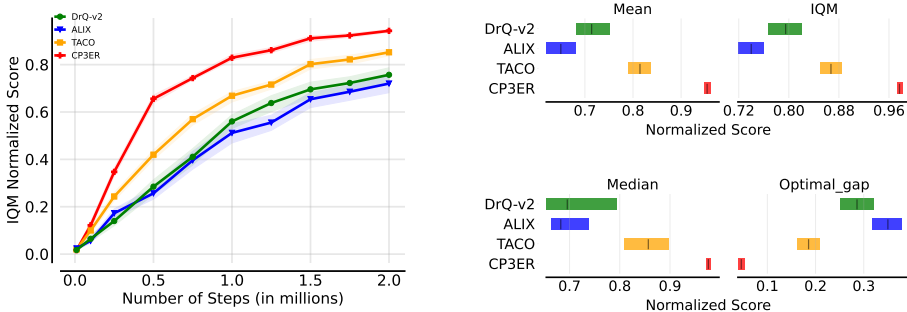

Figure 4: Results on medium-level tasks in DeepMind control suite with 5 random seeds.

Hard-level tasks in DeepMind control suite. We also evaluate CP3ER on 7 challenging tasks [53, 57] in the DeepMind control suite. It should be noted that all tasks here only train 5M frames, rather than the commonly used 30M frames in other work[53, 57]. This means it is a very hard challenge. From the results on the left part of Figure 5, it can be seen that most methods have yet to learn effective policy within 5M frames. Our proposed CP3ER surpasses the performance of all methods without relying on any additional loss or exploration strategies. Moreover, it has significant advantages on all metrics including mean, IQM, median, and optimal gap.

Meta-world. We also evaluated the methods on 6 complex tasks in the Meta-world. The results are shown in Figure 6. We record the success rates of the tasks, and all results are based on the success rates. Compared to other methods, CP3ER can quickly learn effective manipulation policy, and the success rate can reach nearly 100% in all tasks within 2M steps. By comparing the mean, IQM, median, and optimal gap of the success rates, CP3ER has a significant performance advantage with minimal variance.
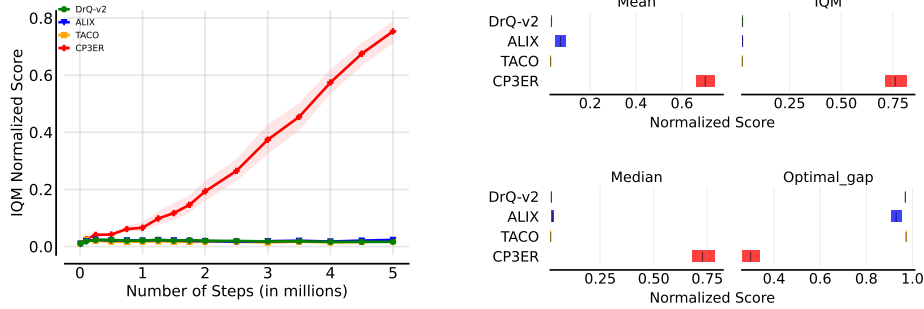
Figure 5: Results on hard-level tasks in DeepMind control suite with 5 random seeds.
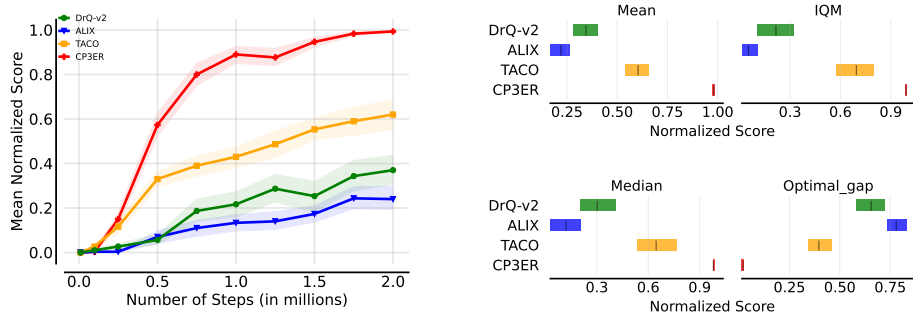

Figure 6: Results on Meta-world tasks with 5 random seeds.

6.2 Ablation Study

6.2.1 Can policy regularization improve the behavior of the policy during training?

Action distribution analysis with toy example. In order to further explore the impact of policy regularization on the training, we borrow the 1D continuous bandit problem [9] to analyze the policy behavior. The green curve in Figure 7 (a) shows the reward function. Within a narrow range of actions, the agent receives higher rewards, while within a broader range, the agent can only receive suboptimal rewards. Therefore, the policy needs strong exploration ability to achieve the highest return. We compare Gaussian policy with entropy regularization (MaxEnt GP), Consistency-AC[33] and consistency policy with entropy regularization (MaxEnt CP), and record the action distribution during the training. As shown in Figure 7, Consistency-AC quickly converges to the local optimal value with the Q-loss. Policy regularization ensures the diversity of action distribution during consistency policy training, preventing the policy from falling into local optima too early. Moreover, consistency policy has robust exploration compared to the Gaussian policy and achieves higher returns.
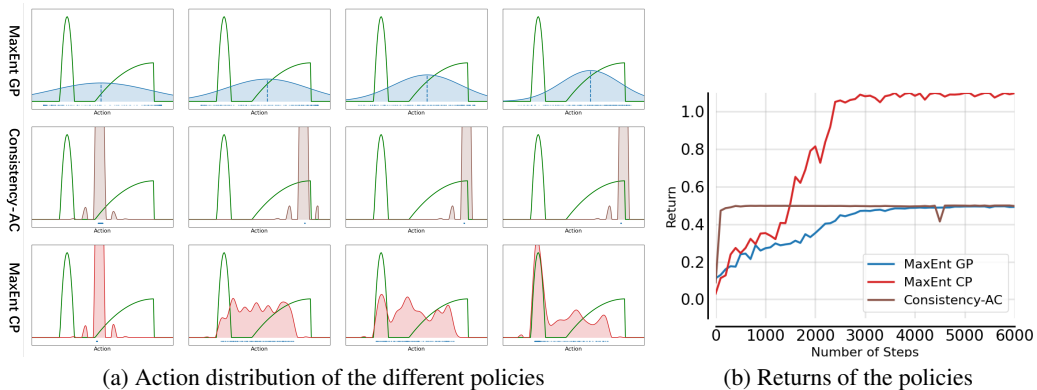

(a) Action distribution of the different policies

(b) Returns of the policies

Figure 7: Results on the toy example. Left part is action distributions during training, while right is returns of different policies.

8

Dormant ratio analysis. We have shown that the Q-loss can rapidly increase the dormant ratio of the consistency policy network, leading to a loss of policy diversity. In order to analyze whether entropy regularization can alleviate the phenomenon, we record the dormant ratios of the policy networks during the training in 3 tasks. The results are shown in Figure 8. Compared to the rapid increase in the dormant rate in Consistency-AC, CP3ER has a lower dormant ratio, which means that entropy regularization can effectively reduce the dormant ratios of consistency policy.
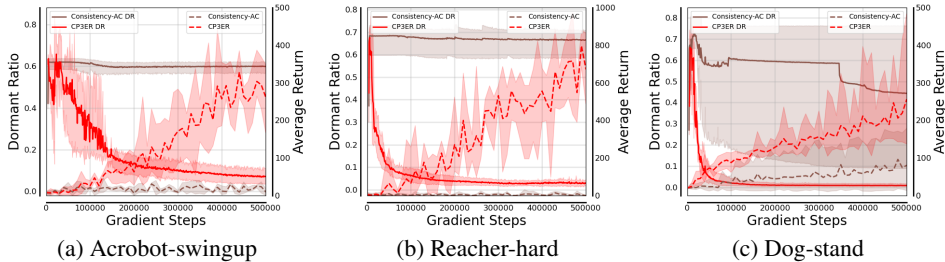

(a) Acrobot-swingup   (b) Reacher-hard   (c) Dog-stand

Figure 8: Dormant ratios of the policy networks on different tasks with 5 random seeds.

6.2.2   What is the impact of different modules on the performance?

We conduct ablation studies in 2 tasks to evaluate the contribution of each module in the proposed method. In addition, to analyze the impact of proxy distribution for policy regularization on performance, we also compare several candidates, including uniform distribution or behavioral policy in the replay buffer. The results are shown in Figure 9. It is noticeable that policy regularization is crucial for consistency policy. Without policy regularization, consistency policy (CP3ER w/o PPER) makes it difficult to learn meaningful behavior in the tasks. The proxy distribution also has an impact on the performance. Using uniform distribution to regularize policies can make the policy (CP3ER w. MaxEnt) improvement difficult, resulting in low sample efficiency. Compared to using behavior distribution in the replay buffer (CP3ER w. URB), the policy (CP3ER) obtained through prioritized proximal experience sampling has a closer distribution to the current policy, making policy optimization easier and resulting in higher sample efficiency. In addition, we find that the performance of CP3ER is significantly better than the baseline (DrQ-v2), indicating that the feasible usage of consistency policy can help solve visual RL tasks.
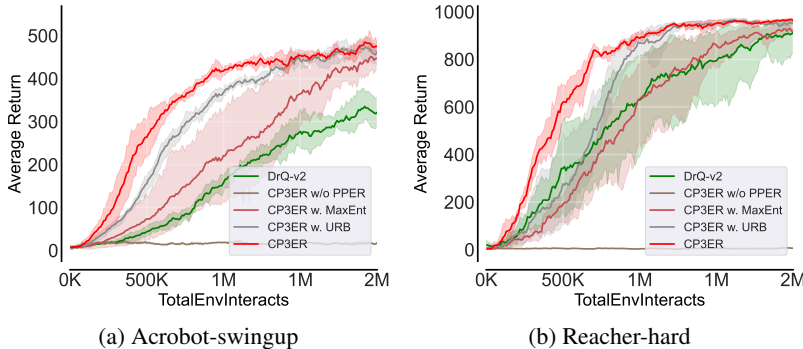

(a) Acrobot-swingup   (b) Reacher-hard

Figure 9: Results of ablation study on 2 visual control tasks with 4 random seeds.

7   Conclusion

In this paper, we analyze the problems faced by extending consistency policy to visual RL under the actor-critic framework and discover the phenomenon of the collapse of consistency policy during training under the actor-critic framework by analyzing the dormant ratio of the neural networks. To address this issue, we propose a consistency policy with prioritized proximal experience regularization (CP3ER) that effectively alleviates the training collapse problem of consistency policy. The method is evaluated on 21 visual control tasks and shows significantly better sample efficiency and performance

than the current SOTA methods. It is worth mentioning that, to the best of our knowledge, our proposed CP3ER is the first method to apply diffusion/consistency models to visual RL tasks.

Our experimental results show that the consistency policy benefits from its expressive ability and ease of sampling, effectively balancing exploration and exploitation in RL with high-dimensional state space and continuous action space. It achieves significant performance advantages without any auxiliary loss and additional exploration strategies. We believe that consistency policy will play an essential role in visual RL. There are still some issues worth exploring in future work. Firstly, auxiliary losses for representation in current visual RL have the potential to improve the performance of consistency policy. Secondly, the diversity of behavior in consistency policy is crucial for RL exploration. This paper only discusses the stability of policy training and does not analyze the diversity of behavior during training, which will help improve the exploration performance of policies. In addition, consistency policy under the on-policy framework is also a direction worth exploring.

References

[1] 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. *Nat.*, 518(7540):529–533, 2015.

[2] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nat.*, 529(7587):484–489, 2016.

[3] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. doi: 10.1126/science.aar6404.

[4] 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. *Nat.*, 588(7839):604–609, 2020.

[5] Haoran Li, Qichao Zhang, and Dongbin Zhao. Deep reinforcement learning-based automatic exploration for navigation in unknown environment. *IEEE Transactions on Neural Networks and Learning Systems*, 31(6):2064–2076, 2020. doi: 10.1109/TNNLS.2019.2927869.

[6] Chuanyu Yang, Kai Yuan, Qiuguo Zhu, Wanming Yu, and Zhibin Li. Multi-expert learning of adaptive legged locomotion. *Sci. Robotics*, 5(49):2174, 2020.

[7] Peter R. Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J. Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, Leilani Gilpin, Piyush Khandelwal, Varun Kompella, HaoChih Lin, Patrick MacAlpine, Declan Oller, Takuma Seno, Craig Sherstan, Michael D. Thomure, Houmehr Aghabozorgi, Leon Barrett, Rory Douglas, Dion Whitehead, Peter Dürr, Peter Stone, Michael Spranger, and Hiroaki Kitano. Outracing champion gran turismo drivers with deep reinforcement learning. *Nat.*, 602(7896): 223–228, 2022.

[8] Guangzheng Hu, Haoran Li, Shasha Liu, Yuanheng Zhu, and Dongbin Zhao. Neuronsmae: A novel multi-agent reinforcement learning environment for cooperative and competitive multi-robot tasks. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2023. doi: 10.1109/IJCNN54540.2023.10191291.

[9] Zhiao Huang, Litian Liang, Zhan Ling, Xuanlin Li, Chuang Gan, and Hao Su. Reparameterized policy learning for multimodal trajectory optimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 13957–13975. PMLR, 2023.

[10] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1352–1361. PMLR, 2017.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022.

[13] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022.

[14] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[15] Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[16] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023.

[17] Zhengbang Zhu, Hanye Zhao, Haoran He, Yichao Zhong, Shenyu Zhang, Yong Yu, and Weinan Zhang. Diffusion models for reinforcement learning: A survey. *CoRR*, abs/2311.01223, 2023.

[18] Tim Pearce, Tabish Rashid, Anssi Kanervisto, David Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, and Sam Devlin. Imitating human behaviour with diffusion models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[19] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu, editors, *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023.

[20] Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9902–9915. PMLR, 2022.

[21] Zhixuan Liang, Yao Mu, Mingyu Ding, Fei Ni, Masayoshi Tomizuka, and Ping Luo. Adaptdiffuser: Diffusion models as adaptive self-evolving planners. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 20725–20745. PMLR, 2023.

[22] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *CoRR*, abs/2310.06114, 2023.

[23] João Carvalho, An T. Le, Mark Baierl, Dorothea Koert, and Jan Peters. Motion planning diffusion: Learning and planning of robot motions with diffusion models. In *IROS*, pages 1916–1923, 2023.

[24] Wonjoon Goo and Scott Niekum. Know your boundaries: The necessity of explicit behavioral cloning in offline RL. *CoRR*, abs/2206.00695, 2022.

[25] Huayu Chen, Cheng Lu, Chengyang Ying, Hang Su, and Jun Zhu. Offline reinforcement learning via high-fidelity generative behavior modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[26] Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. IDQL: implicit q-learning as an actor-critic method with diffusion policies. *CoRR*, abs/2304.10573, 2023.

[27] Cong Lu, Philip J. Ball, Yee Whye Teh, and Jack Parker-Holder. Synthetic experience replay. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[28] Haoran He, Chenjia Bai, Kang Xu, Zhuoran Yang, Weinan Zhang, Dong Wang, Bin Zhao, and Xuelong Li. Diffusion model is an effective planner and data synthesizer for multi-task reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[29] Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 32145–32168. PMLR, 2023.

[30] Michael Psenka, Alejandro Escontrela, Pieter Abbeel, and Yi Ma. Learning a diffusion model policy from rewards via q-score matching. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 41163–41182. PMLR, 21–27 Jul 2024.

[31] Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[32] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32211–32252. PMLR, 23–29 Jul 2023.

[33] Zihan Ding and Chi Jin. Consistency models as a rich and efficient policy class for reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[34] Yuhui Chen, Haoran Li, and Dongbin Zhao. Boosting continuous control with consistency policy. In Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum, editors, *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, pages 335–344. ACM, 2024.

[35] Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. IDQL: implicit q-learning as an actor-critic method with diffusion policies. *CoRR*, abs/2304.10573, 2023.

[36] Bingyi Kang, Xiao Ma, Chao Du, Tianyu Pang, and Shuicheng Yan. Efficient diffusion policies for offline reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[37] Cheng Lu, Huayu Chen, Jianfei Chen, Hang Su, Chongxuan Li, and Jun Zhu. Contrastive energy prediction for exact energy-guided diffusion sampling in offline reinforcement learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22825–22855. PMLR, 2023.

[38] Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[39] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, Tommi S. Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision making? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[40] Haoran Li, Yaocheng Zhang, Haowei Wen, Yuanheng Zhu, and Dongbin Zhao. Stabilizing diffusion model for robotic control with dynamic programming and transition feasibility. *IEEE Transactions on Artificial Intelligence*, 5(9):4585–4594, 2024. doi: 10.1109/TAI.2024.3387401.

[41] Long Yang, Zhixiong Huang, Fenghao Lei, Yucun Zhong, Yiming Yang, Cong Fang, Shiting Wen, Binbin Zhou, and Zhouchen Lin. Policy representation via diffusion probability model for reinforcement learning. *CoRR*, abs/2305.13122, 2023.

[42] Marc Rigter, Jun Yamada, and Ingmar Posner. World models via policy-guided trajectory diffusion. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

[43] Michael Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[44] 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, May 3-7, 2021*. OpenReview.net, 2021.

[45] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[46] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: contrastive unsupervised representations for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5639–5650. PMLR, 2020.

[47] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[48] Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9870–9879. PMLR, 2021.

[49] Ruijie Zheng, Xiyao Wang, Yanchao Sun, Shuang Ma, Jieyu Zhao, Huazhe Xu, Hal Daumé III, and Furong Huang. TACO: temporal latent action-driven contrastive loss for visual reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[50] Max Schwarzer, Ankesh Anand, Rishab Goel, R. Devon Hjelm, Aaron C. Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[51] Edoardo Cetin, Philip J. Ball, Stephen J. Roberts, and Oya Çeliktutan. Stabilizing off-policy deep reinforcement learning from pixels. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2784–2810. PMLR, 2022.

[52] Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G. Bellemare, and Aaron C. Courville. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[53] Guowei Xu, Ruijie Zheng, Yongyuan Liang, Xiyao Wang, Zhecheng Yuan, Tianying Ji, Yu Luo, Xiaoyu Liu, Jiaxin Yuan, Pu Hua, Shuzhen Li, Yanjie Ze, Hal Daumé III, Furong Huang, and Huazhe Xu. Drm: Mastering visual reinforcement learning through dormant ratio minimization. In *The Twelfth International Conference on Learning Representations*, 2024.

[54] Johan Obando-Ceron, Aaron C. Courville, and Pablo Samuel Castro. In value-based deep reinforcement learning, a pruned network is a good network. In *International Conference on Machine Learning*, 2024.

[55] Emlyn Williams and Athanasios Polydoros. Pretrained visual representations in reinforcement learning. *CoRR*, abs/2407.17238, 2024.

[56] Bobak Shahriari, Abbas Abdolmaleki, Arunkumar Byravan, Abe Friesen, Siqi Liu, Jost Tobias Springenberg, Nicolas Heess, Matt Hoffman, and Martin A. Riedmiller. Revisiting gaussian mixture critics in off-policy reinforcement learning: a sample-based approach. *CoRR*, abs/2204.10256, 2022.

[57] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[58] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy P. Lillicrap, and Martin A. Riedmiller. Deepmind control suite. *CoRR*, abs/1801.00690, 2018.

[59] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, October 30 - November 1, 2019, Proceedings*, volume 100 of *Proceedings of Machine Learning Research*, pages 1094–1100. PMLR, 2019.

[60] Edoardo Cetin, Philip J. Ball, Stephen J. Roberts, and Oya Çeliktutan. Stabilizing off-policy deep reinforcement learning from pixels. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2784–2810. PMLR, 2022.

[61] Ruijie Zheng, Xiyao Wang, Yanchao Sun, Shuang Ma, Jieyu Zhao, Huazhe Xu, Hal Daumé III, and Furong Huang. TACO: temporal latent action-driven contrastive loss for visual reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

A   More Results on Dormant Ratios

To demonstrate the phenomenon of policy degradation in Consistency-AC, we conduct experiments with different settings on more tasks and analyze the effect on the dormant ratio. For analyzing the impact of non-stationary data distribution on consistency model training in online RL, we employ 3 classic tasks from D4RL. The results are shown in Figure 10. It can be seen that the non-stationary distribution caused by online training does not significantly affect the dormant ratio of the policy network.

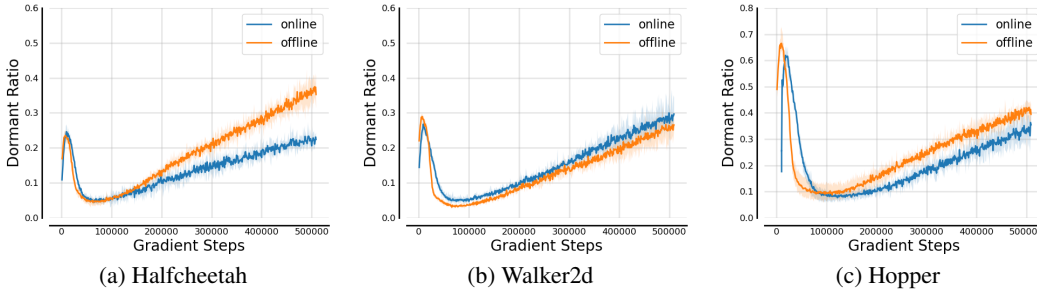|		
(a) Halfcheetah	(b) Walker2d	(c) Hopper

Figure 10: The dormant ratios of the policy under the online and offline training. All results are averaged over 4 random seeds, and the shaded region stands for standard deviation across different random seeds.

For analyzing the impact of the loss function and high-dimensional state input on the consistency policy, we conduct experiments on 5 tasks separately. Among them, the results of 2 tasks are presented in the main part, and the results of the remaining 3 tasks are shown here.

Figure 11 shows the impact of different loss functions on the performance and the dormant ratio of consistency policy. Compared to SAC policy, consistency policy with Q-loss achieves the higher dormant ratio and the worse performance. Especially on the Finger-turn-hard task, the policy under the Consistency-AC framework hardly learn any meaningful behavior, and the dormant ratios remain at a high value. Figure 12 shows the impact of different observation inputs on Consistency-AC. Compared to state-based settings, image-based policy is more difficult to learn meaningful behavior, and its dormant ratio also maintains a higher value.

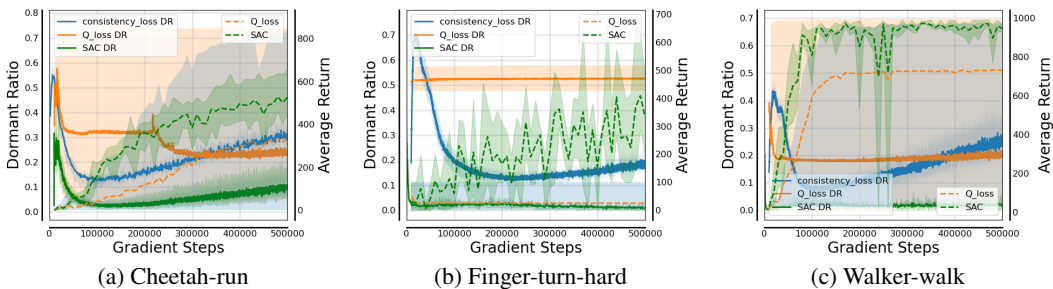|		
(a) Cheetah-run	(b) Finger-turn-hard	(c) Walker-walk

Figure 11: The dormant ratios of the policy with different training loss. All results are averaged over 4 random seeds, and the shaded region stands for standard deviation across different random seeds.

B   Implementation Details

In this paper, we propose Consistency Policy with Prioritized Proximal Experience Regularization (CP3ER), which is built on the basis of DrQ-v2 and its framework is shown in Figure 13. Compared to DrQ-v2, its difference lies in the use of Prioritized Proximal Experience (PPE) when sampling data from the replay buffer, and the use of consistency policy instead of Gaussian policy in the

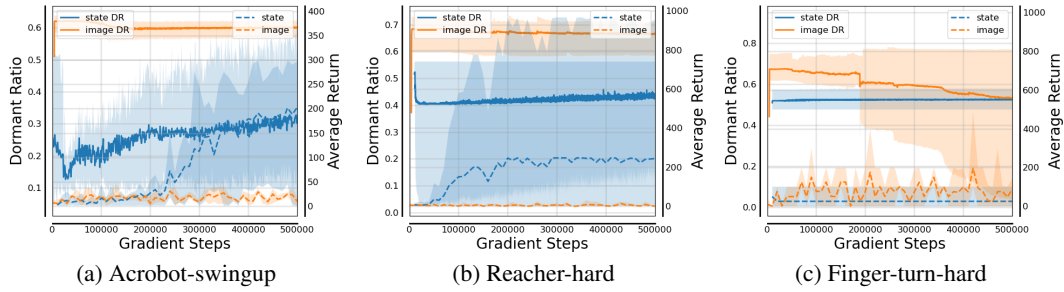| (a) Acrobot-swingup | (b) Reacher-hard | (c) Finger-turn-hard |

Figure 12: The dormant ratios of the policy with different observations. All results are averaged over 4 random seeds, and the shaded region stands for standard deviation across different random seeds.

actor. In addition, it employs a mixture of Gaussians to model the value distribution rather than the deterministic double Q-networks in DrQ-v2.

(a) DrQ-v2

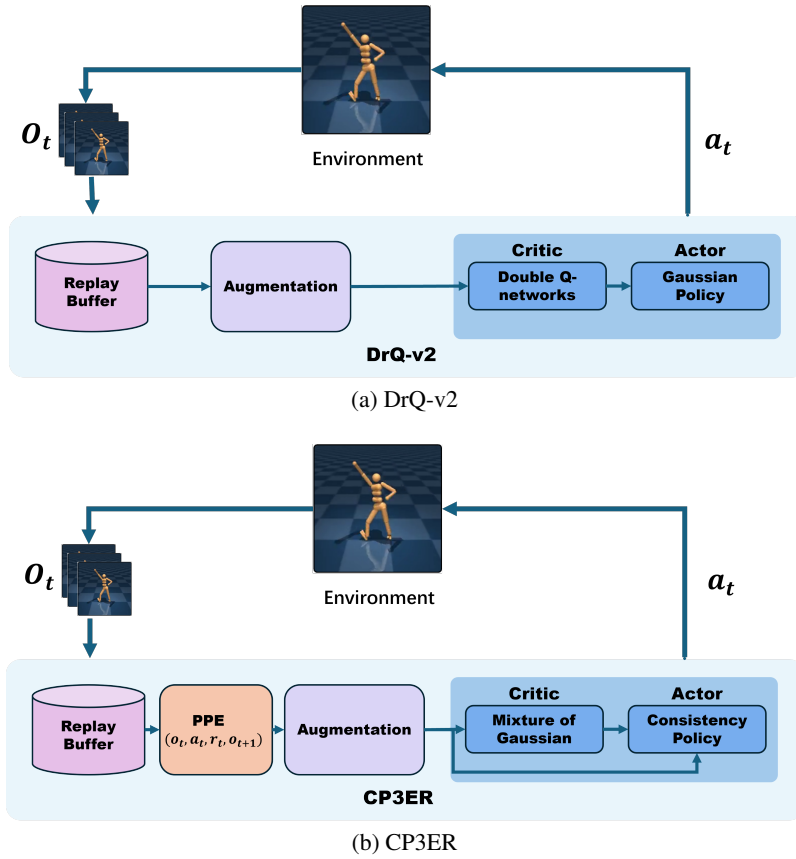(b) CP3ER

Figure 13: Comparison between DrQ-v2 and CP3ER.

B.1 Procedure of the proposed algorithm

We have demonstrated the complete procedure of CP3ER in Algorithm 1, 2 and 3.

B.2 Hyperparameters

We present a summary of all the hyperparameters for CP3ER in Table 1, where DMC is the abbreviation of DeepMind control suite. It is worth noting that for tasks in different domains, only the

---
Algorithm 1 Algorithm of CP3ER

Input:
$f_\xi, \pi_\theta, Q_\phi$: parametric networks for encoder, policy and Q-function respectively.
aug: random shifts image augmentation.
T, B, α, τ: training steps, mini-batch size, learning rate, target update rate.
Training routine:
    for each timestep $t = 1..T$ **do**
        $a_t \leftarrow \pi_\theta(f_\xi(o_t))$
        $o_{t+1} \sim P(\cdot|o_t, a_t)$
        $\mathcal{B} \leftarrow \mathcal{B} \cup (o_t, a_t, R(o_t, a_t), o_{t+1})$
        $(o_t, a_t, r_t, o_{t+1}) \sim \mathcal{B}$ with PPE sampling method
        UPDATECRITIC(o_t, a_t, r_t, o_{t+1})
        UPDATEACTOR(o_t, a_t)
    end for

Algorithm 2 Training critic
---
    procedure UPDATECRITIC(o_t, a_t, r_t, o_{t+1})
        $h_t, h_{t+1} \leftarrow f_\xi(aug(o_t)), f_\xi(aug(o_{t+1}))$
        $a_{t+1} \leftarrow \pi_\theta(o_{t+1})$
        Compute $\mathcal{L}_q(\phi)$ using Equation (9)
        $\xi \leftarrow \xi - \alpha\nabla_\xi \mathcal{L}_{q,\xi}(\phi)$
        $\phi \leftarrow \phi - \alpha\nabla_\phi \mathcal{L}_{q,\xi}(\phi)$
        $\bar{\phi} \leftarrow (1-\tau)\bar{\phi} + \tau\bar{\phi}$
    end procedure
---

learning rate and feature dimension are different, while other parameters are the same for all tasks. The parameters of our proposed method are not task-sensitive, which helps it be applied to a wider range of visual control tasks without the need for fine-tuning of parameters.

C   More Results

In this section, we present more detailed experimental results on tasks in DeepMind control suite and Meta-world, including performance curves during the training, performance profiles, and probability of performance improvement. We compare CP3ER to 4 baselines including DrQ-v2 [57], ALIX [60], TACO [61] and DrM [53]. All evaluations are based on a single NVIDIA GeForce RTX 2080 Ti. For CP3ER, training a run with 2M frames on this device will take about 13 hours.

- DrQ-v2: `https://github.com/facebookresearch/drqv2`
- ALIX: `https://github.com/Aladoro/Stabilizing-Off-Policy-RL`
- TACO: `https://github.com/FrankZheng2022/TACO`
- DrM: `https://github.com/XuGW-Kevin/DrM`

Algorithm 3 Training actor
---
    procedure UPDATEACTOR(o_t, a_t)
        $h_t \leftarrow f_\xi(o_t)$
        $\hat{a}_t \leftarrow \pi_\theta(o_t)$
        Compute $\mathcal{L}_a^r(\theta)$ using Equation (6)
        $\theta \leftarrow \theta - \alpha\nabla_\theta \mathcal{L}_a^r(\theta)$
        $\bar{\theta} \leftarrow (1-\tau)\bar{\theta} + \tau\bar{\theta}$
    end procedure

Parameter	Setting
Replay buffer capacity	10^6
Action repeat	2
Seed frames	4000
Exploration steps	10000
n-step returns	3
Mini-batch size	256
Discount γ	0.99
Optimizer	Adam
Learning rate	8×10^{-5} (Hard-level tasks in DMC)
	10^{-4} (Medium-level tasks in DMC & Meta-world)
Soft update rate	0.01
Features dimension	100 (Hard-level tasks in DMC)
	50 (Medium-level tasks in DMC & Meta-world)
Hidden dimension	1024
Number of Gaussian mixtures for Critic	3
Number of samples M for update Critic	20
Parameter α for PPE	2.0
Coefficient for consistency loss η	0.05
τ-Dormant ratio η	0.025

Table 1: The hyper-parameters for CP3ER.

C.1 Results on Medium-level Tasks in DeepMind Control Suite

In this subsection, we show the detail results on the 8 medium-level tasks [57] in DeepMind control suite. These tasks include: acrobot-swingup, cheetah-run, finger-turn-hard, hopper-hop, quadruped-run, quadruped-walk, reacher-hard and walker-walk. From the results in Figure 14, it can be seen that our proposed CP3ER has better sample efficiency and performance on almost all tasks, even though it does not use any loss for representation learning. In addition to the performance curve during the training, we also demonstrate the performance profiles at different checkpoints and the probability of performance improvement. According to the results in Figure 15, CP3ER is significantly superior to other baseline methods.

(a) Acrobot-swingup   (b) Cheetah-run   (c) Finger-turn-hard   (d) Hopper-hop

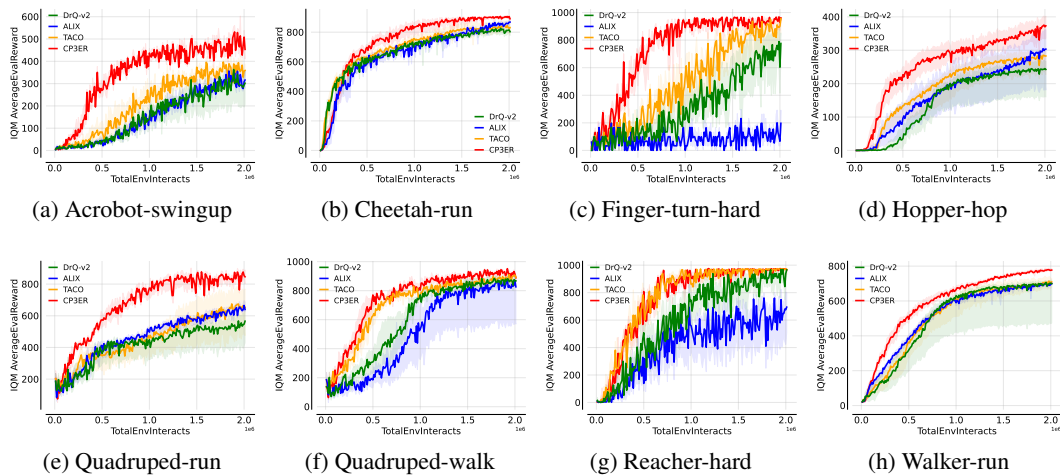(e) Quadruped-run   (f) Quadruped-walk   (g) Reacher-hard   (h) Walker-run

Figure 14: Performance of CP3ER against baseline algorithms DrQ-v2, ALIX, and TACO on the medium-level tasks in DeepMind control suite. All results are averaged over 5 random seeds, and the shaded region stands for standard deviation across different random seeds.
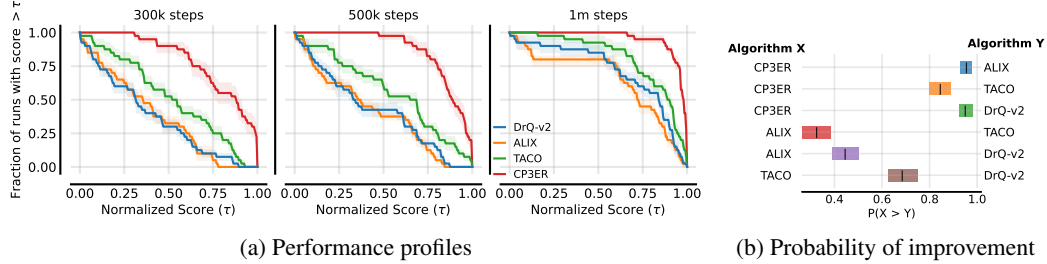
(a) Performance profiles   (b) Probability of improvement

Figure 15: Performance profiles and probabilities of improvement of different methods.

C.2 Results on Hard-level Tasks in DeepMind Control Suite

In this subsection, we show the detail results on the 7 hard-level tasks [53, 57] in DeepMind control suite. These tasks include: dog-run, dog-stand, dog-trot, dog-walk, humanoid-run, humanoid-stand and humanoid-walk. In addition to the 3 comparison baselines considered in the previous subsection, we also consider DrM which achieves SOTA performance on hard-level tasks in DeepMind control suite by weight perturbation and exploration strategies. Figure 16 shows the results CP3ER against several baselines. It is difficult for DrQ-v2, ALIX, and TACO to learn meaningful policy within the 5M framework on these tasks. CP3ER achieves comparable performance to DrM on 4 tasks: dog-run, dog-stand, dog-trot, and dog-walk without any exploration strategy. In addition, CP3ER significantly outperforms DrM in another 3 more challenging tasks. From the results in Figure 16, CP3ER also significantly outperforms other baselines on the tasks. Compared to the SOTA method DrM for difficult tasks, CP3ER also has a performance improvement probability of over 70%.

(a) Dog-run   (b) Dog-stand   (c) Dog-trot   (d) Dog-walk

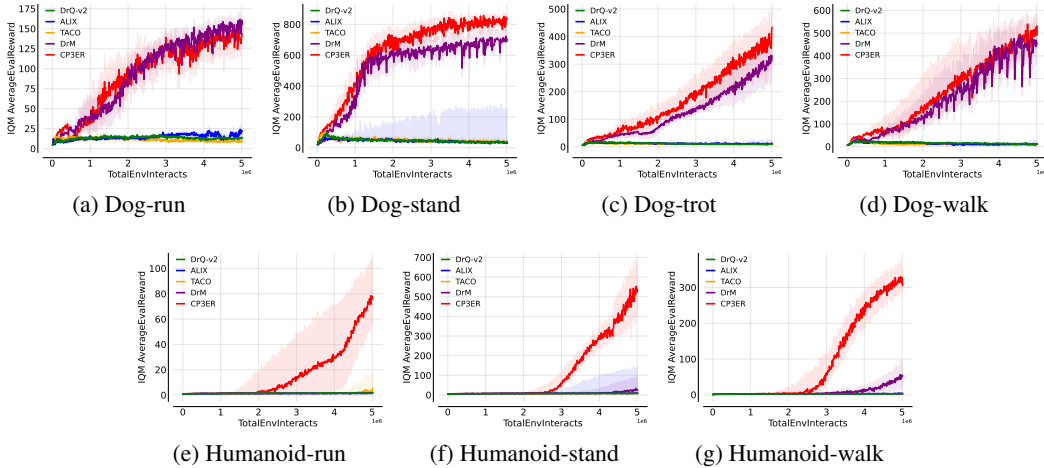(e) Humanoid-run   (f) Humanoid-stand   (g) Humanoid-walk

Figure 16: Performance of CP3ER against baseline algorithms DrQ-v2, ALIX, TACO and DrM on hard-level tasks in DeepMind control suite. All results are averaged over 5 random seeds, and the shaded region stands for standard deviation across different random seeds.

C.3 Results on Tasks in Meta-world

In this subsection, we show the detail results on the 6 hard tasks [53] in Metat-world. These tasks include: assembly, disassemble, hammer, hand insert, pick place wall and stick pull. In these tasks, we compared 4 baselines, including DrM. It should be noted that in [53], DrM achieves SOTA performance on these hard tasks in Meta-world, but we cannot reproduce the corresponding performance based on the official codes. Therefore, this result is used for reference and may cannot represent the true performance of DrM. According to the results in Figure 18 and Figure 19, CP3ER achieves SOTA performance on all 6 hard-level tasks in Meta-world, and compared to the baselines, CP3ER has a significant performance advantage.
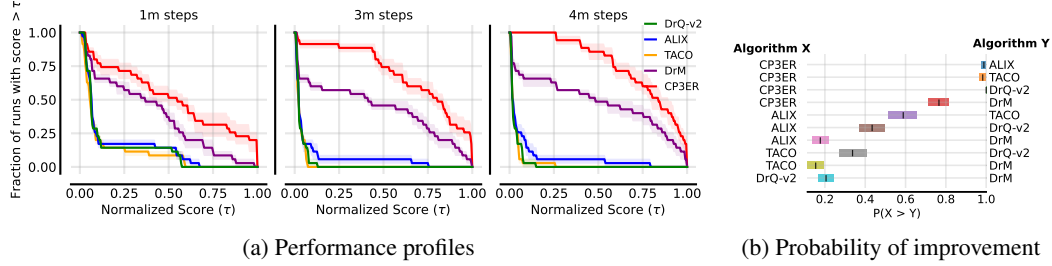
(a) Performance profiles

(b) Probability of improvement

Figure 17: Performance profiles and probabilities of improvement of different methods.

(a) Assembly

(b) Disassemble

(c) Hammer

(d) Hand insert
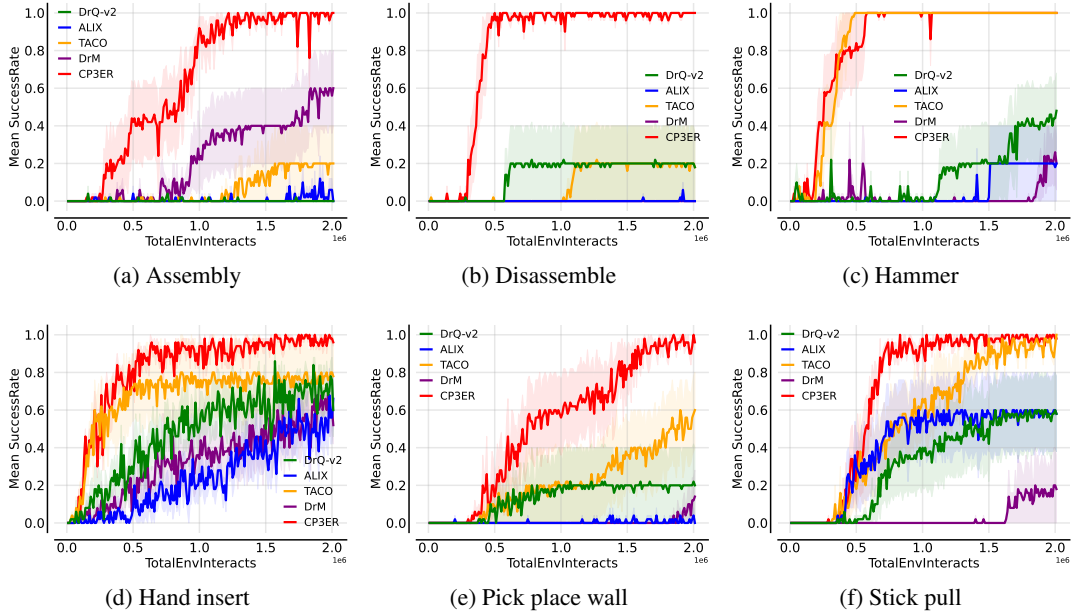
(e) Pick place wall

(f) Stick pull

Figure 18: Performance of CP3ER against baseline algorithms DrQ-v2, ALIX, TACO and DrM on tasks in Meta-world. All results are averaged over 5 random seeds, and the shaded region stands for standard deviation across different random seeds.

(a) Performance profiles

(b) Probability of improvement
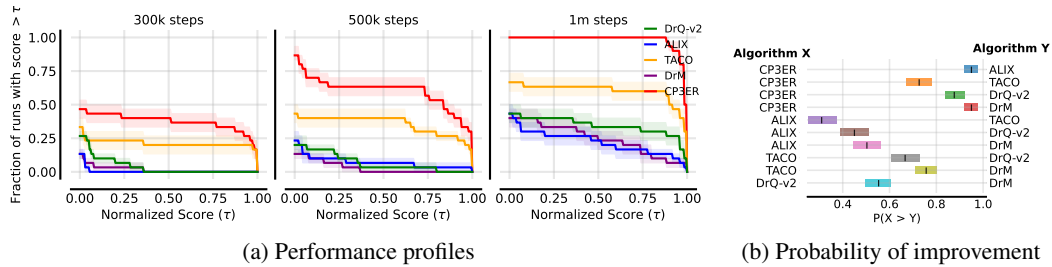
Figure 19: Performance profiles and probabilities of improvement of different methods.

C.4 Results on State-based Tasks

Sample efficiency has always been a highly concerned issue in visual RL, and high expressive power of the policy can improve exploration efficiency and thus increase sample efficiency. After introducing the diffusion model into visual RL, we find a serious phenomenon of policy degradation, which results in disastrous performance. Therefore, we propose CP3ER to stabilize the training of consistent policy. In addition to visual RL tasks, CP3ER theoretically has the potential to be applied to state-based RL tasks. According to the experimental setup in [34], we conduct a comparison to

state-based RL SOTA methods, and the experimental results are shown in the Table 2. It can be seen that compared to the current SOTA methods, CP3ER also has significant advantages in the tasks, which also proves the generalization of CP3ER on different observation tasks.

Tasks	TD3	SAC	PPO	MPO	DMPO	D4PG	DreamerV3	CPQL	CP3ER
Acrobot Swingup	46.8	33.2	34.4	80.6	98.5	125.5	154.5	183.1	**362.9**
Finger Turn Easy	337.6	371.4	275.2	430.4	593.8	524.5	745.4	**874.1**	867.45
Finger Turn Hard	334.4	344.8	5.06	250.8	384.5	379.2	841.0	864.6	**912.65**
Hopper Hop	40.0	41.7	0.0	37.5	71.5	67.5	111.0	130.1	**299.3**
Hopper Stand	322.7	270.9	2.2	279.3	519.5	755.4	573.2	902.1	**904.1**
Walker Run	274.5	445.9	131.7	539.5	462.9	593.1	632.7	**683.8**	646.5
Average	226	251.3	74.8	269.7	355.1	407.5	509.6	606.3	**665.5**

Table 2: Comparison of CP3ER and other methods on state-based RL tasks in DeepMind control suite.

In addition, we also compare the existing methods based on diffusion/consistency models following the settings in [33], and the results are shown in Table 3. It can be seen that compared to existing methods based on diffusion/consistency models, our method CP3ER, although aimed at visual RL problems, still has significant advantages when extended to state-based RL tasks.

Online tasks	Diffusion-QL	Consistency-AC	CP3ER
Halfcheetah-m	5745.9 ± 388.5	6725.2 ± 944.4	$\mathbf{10699.1} \pm 1054.7$
Hopper-m	$\mathbf{3675.5} \pm 47.7$	3589.7 ± 163.4	3309.6 ± 133.2
Walker2d-m	4316.2 ± 612.1	3790.9 ± 1677.5	$\mathbf{5201.0} \pm 111.3$
Average	4579.2	4701.9	**6403.2**

Table 3: Comparison of CP3ER with diffusion/consistency based RL methods.

D   Limitations

Although CP3ER enhances the training stability of consistency policy in the actor-critic framework and achieves excellent performance in visual control tasks, the policy diversity of CP3ER has not been thoroughly explored. In CP3ER, this diversity only depends on the multimodal actions in the replay buffer, which gradually disappears as the training steps increases. Therefore, CP3ER may face the risk of losing diversity in consistent policy, thereby weakening its exploration ability to a certain extent. In addition, there is a lack of theoretical analysis on CP3ER. Although it is based on the actor-critic framework, its policy improvement and convergence property require more rigorous theoretical analysis.

E   Broader Impacts

This work mainly focuses on the field of visual RL and proposes a new method that may significantly improve the efficiency of visual RL. This method may improve the efficiency of robot skill learning and have a wide impact on visual control fields such as robots, but it will not involve ethical and safety issues. Therefore, this work should not have negative social impacts.

NeurIPS Paper Checklist

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims stated in the abstract and introduction are consistent with the contributions and scope of our paper as detailed in Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper acknowledges its limitations in appendix D.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: There are no assumptions or proofs presented in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully discloses all necessary details for reproducing the main experimental results, including comprehensive descriptions of the algorithm in section 5 and appendix B.1, experiments setup in section 6, and parameters used in appendix B.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   We have submitted the code for training as the supplementary material, and once the paper is accepted, the code will be fully open source.

   Guidelines:
   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper provides specific details about hyperparameters settings in appendix B.2

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper comprehensively presents statistical significance in the results in sections 6 and appendix C

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper specifies the type of compute workers (a single NVIDIA GeForce RTX 2080 Ti) in appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics. We confirm that the work involved in the paper complies with Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper acknowledges its broader impacts in appendix E.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our research is mainly focused on visual RL, with the main potential application direction being robot skill learning. Therefore, this paper should not have the aforementioned risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have cited relevant papers and provided explanations for the referenced code in our codes. The majority of our code is licensed under the MIT license, however portions of the project are available under separate license terms: DeepMind is licensed under the Apache 2.0 license.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We have included a README file in the provided code, which provides detailed instructions on deployment, training, and license.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.