Salience-Invariant Consistent Policy Learning for Generalization in Visual Reinforcement Learning

Jingbo Sun
Institute of Automation, CASIA
Beijing, China
Pengcheng Laboratory
Shenzhen, China
School of Artificial Intelligence, UCAS
Beijing, China
sunjingbo2022@ia.ac.cn

Songjun Tu
Institute of Automation, CASIA
Beijing, China
Pengcheng Laboratory
Shenzhen, China
School of Artificial Intelligence, UCAS
Beijing, China
tusongjun2023@ia.ac.cn

Qichao Zhang
Institute of Automation, CASIA
Beijing, China
School of Artificial Intelligence, UCAS
Beijing, China
zhangqichao2014@ia.ac.cn

Ke Chen
Pengcheng Laboratory
Shenzhen, China
chenk02@pcl.ac.cn

Dongbin Zhao
Institute of Automation, CASIA
Beijing, China
School of Artificial Intelligence, UCAS
Beijing, China
dongbin.zhao@ia.ac.cn

ABSTRACT

Generalizing policies to unseen scenarios remains a critical challenge in visual reinforcement learning, where agents often overfit to the specific visual observations of the training environment. In unseen environments, distracting pixels may lead agents to extract representations containing task-irrelevant information. As a result, agents may deviate from the optimal behaviors learned during training, thereby hindering visual generalization. To address this issue, we propose the **S**alience-Invariant **C**onsistent **P**olicy **L**earning (SCPL) algorithm, an efficient framework for zero-shot generalization. Our approach introduces a novel value consistency module alongside a dynamics module to effectively capture task-relevant representations. The value consistency module, guided by saliency, ensures the agent focuses on task-relevant pixels in both original and perturbed observations, while the dynamics module uses augmented data to help the encoder capture dynamic- and reward-relevant representations. Additionally, our theoretical analysis highlights the importance of policy consistency for generalization. To strengthen this, we introduce a policy consistency module with a KL divergence constraint to maintain consistent policies across original and perturbed observations. Extensive experiments on the DMC-GB, Robotic Manipulation, and CARLA benchmarks demonstrate that SCPL significantly outperforms state-of-the-art methods in terms of generalization. Notably, SCPL achieves average performance improvements of 14%, 39%, and 69% in the challenging DMC video hard setting, the Robotic hard setting, and the CARLA benchmark, respectively. Project Page: https://sites.google.com/view/scpl-rl.

Corresponding author: Qichao Zhang, Dongbin Zhao.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

KEYWORDS

Visual Reinforcement Learning, Zero-shot Generalization, Salience-Invariant, Consistent Policy

ACM Reference Format:
Jingbo Sun, Songjun Tu, Qichao Zhang, Ke Chen, and Dongbin Zhao. 2025. Salience-Invariant Consistent Policy Learning for Generalization in Visual Reinforcement Learning. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

1 INTRODUCTION

In recent years, visual reinforcement learning (RL) [30] has achieved remarkable success across various domains, including video games [26, 31], robot control [16, 17, 33], and autonomous driving [29, 34, 35]. However, generalizing policies to novel scenarios remains a significant challenge. Small visual perturbations in observations can distract RL agents [7, 20], leading to representations containing task-irrelevant information and decisions that deviate from their training behavior, ultimately hindering visual generalization. In this paper, we aim to develop generalizable RL agents that can generate effective task-relevant representations and make consistent decisions across both original and perturbed observations.

Data augmentation (DA)-based methods [13, 15, 38] are widely used to enhance the representational ability of visual RL agents. Recent advances, such as SVEA [11] and SGQN [2], leverage augmented data for implicit regularization to improve generalization. Unfortunately, as illustrated in Fig.1 (left),these methods fail to maintain consistent task-relevant attention regions in perturbed observations, impeding the learning of task-relevant representations. Other studies [27, 42] employ dynamic models as auxiliary tasks to capture task-relevant representations. However, the encoder's primary design for original observations prevents it from generating task-relevant representations for perturbed observations. Furthermore, it is difficult to generate task-relevant representations with uncritical attention to task-relevant regions.

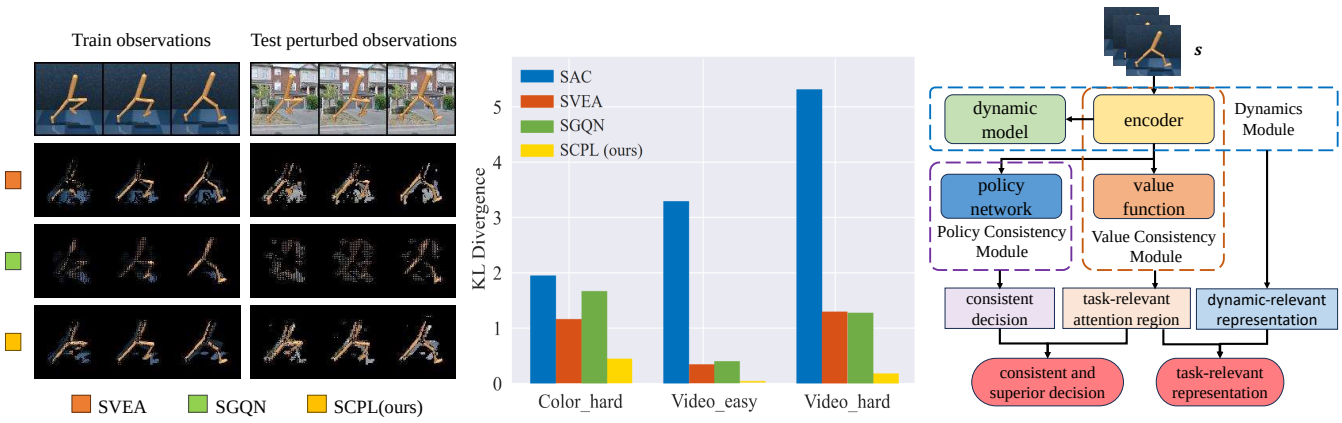

Figure 1: (Left) Saliency masked map of SVEA, SGQN, and SCPL (ours), which shows the attention regions of value functions on the DMC-GB benchmark. (Middle) The KL divergence of action distribution between training and test environments on DMC-GB, where our method holds the smallest KL divergence. (Right) Contribution overview of SCPL, which aims to improve visual generalization by achieving task-relevant representations and consistent and superior decisions.

Additionally, our theoretical analysis reveals that policy consistency between environments with original and perturbed observations is crucial for generalization. However, previous methods focus on improving representations, often neglecting policy consistency. As shown in Fig.1 (middle), both SVEA [11] and SGQN [2] exhibit high KL divergence in action distributions between training and test environments, indicating inconsistent decisions across original and perturbed observations. Improving the consistency of action distributions between original and perturbed observations is also essential to improve generalization. These insights prompt us to consider: **Can we develop generalizable agents that maintain consistent task-relevant representations and policies?**

To address these challenges, we propose Salience-Invariant Consistent Policy Learning (**SCPL**), which improves generalization by encouraging RL agents to capture consistent task-relevant representations and make consistent superior decisions across diverse observations. First, we introduce a novel value consistency module that encourages the encoder and value function to capture **task-relevant attention region** in observations. Meanwhile, we introduce a dynamics module to generate **dynamic-relevant representations** for observations. By combining the value consistency module and dynamics module, SCPL produces consistent task-relevant representations. Furthermore, SCPL regularizes the policy network using a KL divergence constraint between the policies for original and augmented observations, enabling agents to make **consistent decisions** in test environments. An overview of the motivation behind SCPL is illustrated in Fig.1 (right).

In summary, our contribution includes the following aspects:

- We propose an effective generalization framework, SCPL, in which a novel value consistency module with saliency guidance and a dynamics module with augmented data is introduced to generate task-relevant representations.
- Through theoretical analysis, we reveal that improved policy consistency leads to enhanced generalization. We further introduce a policy consistency module to regularize the policy network for consistent decisions to improve generalization.
- The proposed SCPL achieves state-of-the-art (SOTA) performances in 3 popular visual generalization benchmarks, with an average boost of 14%, 39%, and 69% on the *video hard* setting in DMC-GB, the Robotic hard setting, and the CARLA benchmark, respectively.

2 RELATED WORKS

2.1 Data augmentation for visual RL

Data augmentation is widely used to enhance the generalization of visual reinforcement learning (RL) [8, 15, 37]. DrQ [38] employs image transformation strategies to augment observations through implicit regularization. SVEA [11] enhances generalization by updating the value function with both original and augmented data. CG2A [21] improves generalization by combining various data augmentations and alleviating the gradient conflict bias caused by these augmentations. CNSN [18] employs normalization techniques to improve visual generalization. SGQN [2] utilizes saliency guidance to focus agents' attention on task-relevant areas in original observations while aligning their attention across original and augmented data using a trainable network. MaDi [7] improves generalization by incorporating a mask network before the value function to filter out task-irrelevant regions in observations. While these methods can identify effective task-relevant regions in original observations, they often struggle with perturbed observations. In this paper, SCPL focuses on maintaining consistent task-relevant regions in both original and perturbed data with a value consistency module.

2.2 Representation learning in visual RL

Numerous methods [5, 23] improve generalization by learning task-relevant representations through auxiliary tasks. Some approaches [19] improve representation effectiveness by employing observation reconstruction as auxiliary tasks. DBC [42] minimizes the bisimulation metric in latent spaces to learn invariant representations without task-irrelevant information. PAD [10] utilizes an inverse dynamic model to predict actions based on current and next states. Dr.G [8] trains the encoder and world model using

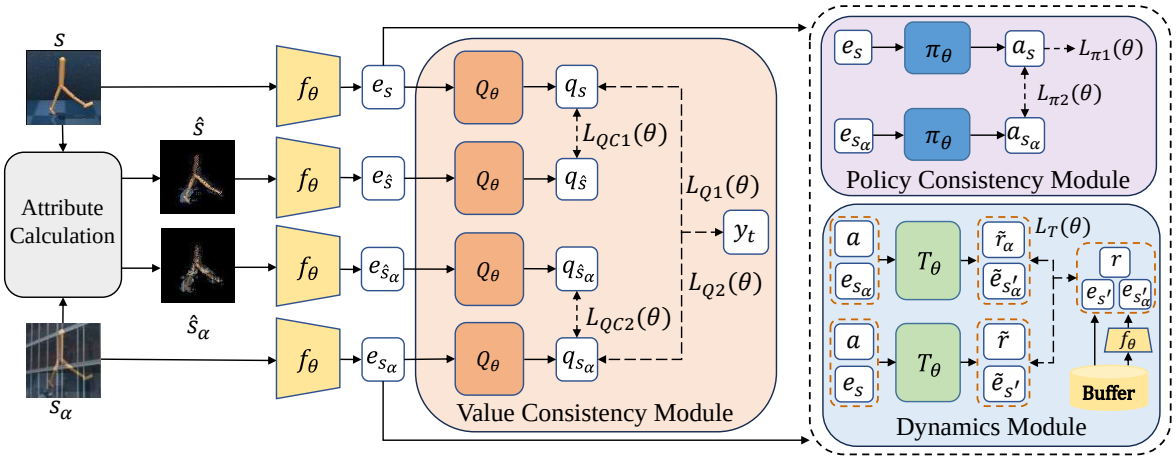

Figure 2: Overview of SCPL. The value consistency module is trained using the original and augmented observations s and s_α, along with their saliency attribute maps $\hat{s}$ and $\hat{s}_\alpha$. The dynamics module aids the encoder f_θ in acquiring task-relevant representations, while the policy consistency module introduces a constraint to maintain consistency in action distributions.

contrastive learning and introduces an inverse dynamics model to capture temporal structure. These methods learn task-relevant information through the guidance of rewards and dynamic consistency. However, they struggle to extract task-relevant information for perturbed observations due to their exclusive training on original data and uncritical task-relevant attention. SCPL achieves task-relevant representations for both original and perturbed observations by training a dynamics module using both original and augmented data while focusing on task-relevant regions.

2.3 Policy learning for RL

Some studies explore the decoupling of value functions and policy networks to learn effective policies or obtain invariant representations [1, 39]. PPG [3] mitigates the interference between policy and value optimization by distilling the value function while constraining the policy. IDAAC [25] models both the policy and value function while introducing an auxiliary loss to obtain representations that remain invariant to task-irrelevant properties. DCPG [22] implicitly penalizes value estimates by optimizing the value network less frequently, using more training data than the policy network. However, prior studies that focus on learning invariant representations often overlook the consistency of policies across both original and perturbed observations. In contrast, SCPL learns task-relevant representations while maintaining a consistent superior policy for perturbed observations, similar to that for original observations. To the best of our knowledge, we are the first to highlight the importance of policy consistency between original and perturbed observations for generalization ability.

3 PROBLEM FORMULATION

Visual reinforcement learning (RL) is considered a partially observable Markov decision process (POMDP) because only partial states are observed from images. A POMDP is defined as a tuple $M = \langle S, O, A, P, r, \gamma \rangle$, where S is the state space, O is the observation space, A is the action space, $P : S \times A \times S \to \mathbb{R}$ is the transition probability distribution, $r : S \to \mathbb{R}$ is the reward function, and $\gamma \in (0, 1)$ is the discount factor. Let π denote a stochastic policy and $\eta(\pi)$ denote its expected cumulative reward: $\eta(\pi) = \mathbb{E}_{s_0, a_0, \ldots}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right]$, where π is the policy, a is the action, and s_t is the state in the t step. The purpose of visual RL is to find a policy π^* to maximize the expected cumulative reward. The state-action value function Q_π, the value function V_π, and the advantage function A_π are defined as: $Q_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1}, a_{t+1}, \ldots}\left[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})\right]$, $V_\pi(s_t) = \mathbb{E}_{a_t, s_{t+1}, \ldots}\left[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})\right]$, and $A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$.

4 METHODOLOGY

We propose a salience-invariant consistent policy learning (SCPL) framework to improve the zero-shot generalization of visual RL. As shown in Fig.2, SCPL mainly consists of three modules: the value consistency module, the policy consistency module, and the dynamics module. In this paper, θ, ζ, ϕ, and ψ represent the parameters of the encoder, the value function, the policy network, and the dynamic model, respectively.

4.1 Value Consistency Module

To extract task-relevant representations from both original and perturbed observations, it is essential for the encoder and value function to consistently focus on task-relevant regions. We introduce a value consistency module with a novel loss function for the encoder and value function, leveraging augmented data and their saliency maps. To improve the consistency of attention regions for original and perturbed observations, we update the value function with both original and augmented observations. The value loss for the original data L_{Q1} is:

$$L_{Q1}(\theta, \zeta) = \mathbb{E}_{s,a}[(Q_\zeta(f_\theta(s), a) - y_t)^2]. \tag{1}$$

The value loss for the augmented data L_{Q2} is:

$$L_{Q2}(\theta, \zeta) = \mathbb{E}_{s_\alpha,a}[(Q_\zeta(f_\theta(s_\alpha), a) - y_t)^2], \tag{2}$$

where y_t represents the target of Q-value. These losses encourage the value function to estimate the same values for both original and augmented observations, thereby promoting consistency in the attention regions. However, maintaining consistency in value

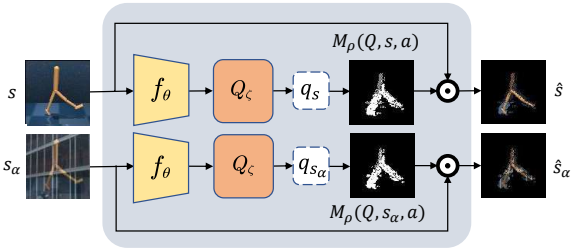

Figure 3: The generation of saliency attribute masked maps.

estimation alone may be insufficient to ensure that agents focus on task-relevant regions amidst increasing perturbations. Therefore, additional guidance is necessary to help agents remain attentive to task-relevant regions in the observations.

SCPL utilizes saliency attribute masked maps to guide the encoder and the value function to focus on task-relevant regions for observations. As shown in Fig.3, we generate the saliency attribute masked maps ($\hat{s}$ and $\hat{s}_\alpha$) for original and augmented observations (s and s_α) using the vanilla gradient method [28]. We use guided backpropagation to compute the gradient map $M(Q,s,a)$ of the Q-network, represented as $M(Q,s,a) = \partial Q(s,a)/\partial s$. Let $M_\rho(Q,s,a)$ be the binarized ρ-quantile saliency attribute map. Specifically, if the gradient pixel $M(Q,s,a)_j$ belongs to the top $1-\rho$ quantile of gradient values, then $M_\rho(Q,s,a)_j$ will be set to 1, otherwise 0. The ρ-quantile saliency attribute map $\hat{s}$ and $\hat{s}_\alpha$ represents the attention of the value function towards input observations, where white areas indicate regions of high attention and less attended regions are masked out. Then, the saliency attribute masked maps are generated by multiplying the observations with their saliency attribute maps to show the attended pixels. $\odot$ denotes the Hadamard product.

To guide agents to focus on task-relevant pixels, we introduce a saliency consistency term between original observations and their respective saliency attribute maps. The saliency consistency loss for the original data is:

$$L_{QC1}(\theta,\zeta) = \mathbb{E}_{s,\hat{s},a}[(Q_\zeta(f_\theta(\hat{s}),a) - Q_\zeta(f_\theta(s),a))^2]. \quad (3)$$

To ensure that agents focus on task-relevant regions in both original and perturbed observations, we extend saliency guidance to augmented data while updating the value function with this augmented data. The saliency consistency loss for the augmented data is:

$$L_{QC2}(\theta,\zeta) = \mathbb{E}_{s_\alpha,\hat{s}_\alpha,a}[(Q_\zeta(f_\theta(\hat{s}_\alpha),a) - Q_\zeta(f_\theta(s_\alpha),a))^2]. \quad (4)$$

With the saliency guidance from Eq.(3) and Eq.(4), agents are able to focus on task-relevant regions in both original and perturbed observations.

We combine the value loss with the saliency consistency loss to form the training objective. The value function's objective is:

$$L_Q(\theta,\zeta) = L_{Q1}(\theta,\zeta) + L_{Q2}(\theta,\zeta) + \lambda(L_{QC1}(\theta,\zeta) + L_{QC2}(\theta,\zeta)), \quad (5)$$

where λ is the value consistency coefficient. $L_{Q1}(\theta,\zeta)$ and $L_{Q2}(\theta,\zeta)$ ensure the encoder and value function attend to consistent regions in both the training and test environments, while $L_{QC1}(\theta,\zeta)$ and $L_{QC2}(\theta,\zeta)$ ensure agent focus on task-relevant regions within observations. This value loss enables the encoder and value function to capture consistent task-relevant pixels from both original and perturbed observations.

4.2 Dynamics Module

To enable the encoder to effectively provide task-relevant representations, we develop a dynamic model to ensure representations meet the conditions of rewards and dynamics. Specifically, we construct this dynamic model by predicting rewards and next-state representations for both the original and augmented observations. The loss of the dynamics module for embedding e is:

$$L_{Te}(\theta,\psi) = \mathbb{E}_{s,a}[(e' - P_\psi(f_\theta(s),a))^2 + (r - R_\psi(f_\theta(s),a))^2], \quad (6)$$

where e and e' are the latent representations of the current observation and the next observation. Dynamics module T consists of dynamic head P and reward head R. The training objective of the dynamics module of the embedding for augmented data e_α, is:

$$L_{Te_\alpha}(\theta,\psi) = \mathbb{E}_{s_\alpha,a}[(e'_\alpha - P_\psi(f_\theta(s_\alpha),a))^2 + (r - R_\psi(f_\theta(s_\alpha),a))^2]. \quad (7)$$

The training objective for the dynamics module is:

$$L_T(\theta,\psi) = L_{Te}(\theta,\psi) + L_{Te_\alpha}(\theta,\psi). \quad (8)$$

In SCPL, the value consistency module ensures attention to task-relevant regions, while the dynamics module guides representations that align with reward and dynamics conditions. With the combination of the value consistency module and the dynamics module, the encoder can generate task-relevant representations.

4.3 Policy Consistency Module

As illustrated in Fig.1 (middle), previous RL agents frequently exhibit poor policy consistency, reflected in the substantial KL divergence, between training and test environments. In this section, our theoretical analysis reveals that the policy consistency of agents contributes to enhanced generalization capability. Furthermore, we propose a policy consistency module that improves generalization by enhancing agents' policy consistency across both original and perturbed observations.

Relationship between policy consistency and generalization ability. We utilize the KL divergence of action distributions between training and test environments to assess the policy consistency of agents. Additionally, the divergence in cumulative rewards between these environments reflects the agents' generalization capabilities. In the context of visual generalization, the training and test environments are identical except for visual observations. Consequently, the agent's policies in both environments can be regarded as two equivalent policies within the training environment. We prove that there is a positive correlation between the upper bound of the divergence in cumulative rewards of the two policies and their KL divergence.

Initially, we utilize the total variation divergence to measure the distance between two policies' distributions. The divergence is defined as : $D_{TV}(p\|q) = \frac{1}{2}\sum_i |p_i - q_i|$ for probability distributions p and q. Define $D_{\text{TV}}^{\max}(\pi_o,\pi_p)$ as :

$$D_{\text{TV}}^{\max}(\pi_o,\pi_p) = \max_s D_{TV}(\pi_o(\cdot \mid s)\|\pi_p(\cdot \mid s)), \quad (9)$$

Algorithm 1 SCPL (changes to SAC in blue)

Parameter: $\mathcal{B}$: replay buffer, N_A: dynamics module update frequency, τ: data augmentation function, α: learning rate, λ: value consistency coefficient, β: policy consistency coefficient.

1: **for** each iteration **do**
2: Sample a transition: $a, s^{'} \sim \pi_\phi(\cdot|s), P(\cdot|s,a)$
3: Add transition to replay buffer: $\mathcal{B} \leftarrow \mathcal{B} \cup \{(s, a, \mathcal{R}(s,a), s')\}$
4: Sample a batch of transition: $\{s, a, r, s^{'}\} \sim \mathcal{B}$
5: Generate augmented data: $s_\alpha \leftarrow \tau(s)$
6: Update value consistency module: $\{\theta, \zeta\} \leftarrow \{\theta, \zeta\} - \alpha\nabla_{\{\theta,\zeta\}}(L_{Q1}(\theta,\zeta)+L_{Q2}(\theta,\zeta) + \lambda L_{QC1}(\theta,\zeta) + \lambda L_{QC2}(\theta,\zeta))$
7: Update dynamics module: $\{\theta, \psi\} \leftarrow \{\theta, \psi\} - \alpha\nabla_{\{\theta,\psi\}} L_T(\theta,\psi)$
8: Update policy consistency module: $\phi \leftarrow \phi - \alpha\nabla_\phi(L_{\pi o}(\phi)+\beta L_{\pi c}(\phi))$
9: **end for**

where π_o and π_p are policies in training and test environments, respectively. With this measure in hand, we can now state the following theorem:

Theorem 1. *Let* $\alpha = D_{\mathrm{TV}}^{\max}(\pi_o, \pi_p)$*, the following bound holds:*

$$\eta(\pi_o) - \eta(\pi_p) \leq \frac{2\epsilon\gamma}{(1-\gamma)^2}\alpha^2, \tag{10}$$

where η is expected return, $\epsilon = \max_{s,a}|A_\pi(s,a)|$. According to [24], the relationship between the total variation divergence and the KL divergence is: $D_{TV}(p\|q)^2 \leq D_{\mathrm{KL}}(p\|q)$. Let $D_{\mathrm{KL}}^{\max}(\pi,\tilde{\pi}) = \max_s D_{\mathrm{KL}}(\pi(\cdot \mid s)\|\tilde{\pi}(\cdot \mid s))$. With Theorem 1, the following bound holds:

$$\eta(\pi_o) - \eta(\pi_p) \leq CD_{\mathrm{KL}}^{\max}(\pi_o, \pi_p), \tag{11}$$

where $C = \frac{2\epsilon\gamma}{(1-\gamma)^2}$. Hence, a smaller KL divergence of action distributions between training and test environments pushes a tighter upper bound on the disparity of cumulative rewards in these environments. This implies that policy consistency contributes to the generalization performance of agents.

Policy consistency module. To enhance the generalization of visual RL algorithms, agents should produce consistent policies for both original and perturbed observations. Therefore, we design a policy consistency loss with the KL divergence of action distributions of both the original and augmented data for the policy network. The policy loss for the original observation is:

$$L_{\pi 1}(\phi) = \mathbb{E}_{s,a\sim\pi_\phi(\cdot|e_s)}[\alpha \log \pi_\phi(a|e_s) - Q(e_s,a)], \tag{12}$$

where e_s, e_{s_α} are embeddings for observations s and s_α. The policy consistency loss is:

$$L_{\pi 2}(\phi) = \mathbb{E}_{s,s_\alpha}[D_{KL}\left(\pi_\phi(\cdot|e_s)||\pi_\phi(\cdot|e_{s_\alpha})\right)]. \tag{13}$$

With the policy consistency loss, SCPL improves generalization by encouraging agents to generate consistent policies for original and perturbed observations. In summary, the loss of the policy consistency module is:

$$L_\pi(\phi) = L_{\pi 1}(\phi) + \beta L_{\pi 2}(\phi), \tag{14}$$

where β is the policy consistency coefficient. By minimizing the total loss, the policy consistency module attains a consistently superior policy in test environments, similar to that in training environments. Algorithm 1 presents the pseudocode for SCPL.

5 EXPERIMENTS

In this section, we conduct experiments to investigate the following questions: (1) Does SCPL exhibit superior visual generalization capability compared to current state-of-the-art methods? (2) Can SCPL focus on consistent task-relevant pixels in both original and perturbed observations? (3) Does SCPL possess consistent representations and policies? (4) What is the contribution of various modules to generalization performance? (5) Can SCPL demonstrate advanced generalization in challenging robotic and autonomous driving environments?

5.1 Experimental Settings

We evaluate the zero-shot generalization performance of our method in DeepMind Control Suite (DMC) [12, 32], Robotic Manipulation tasks [14], and CARLA [4]. All methods are trained in the default environment and evaluated with visual perturbations. In the DMC experiment, we compare the generalization ability of SCPL with SOTA methods including SAC [9], SVEA [11], SIM [36], TLDA [40], PIE-G [41], SGQN [2], CG2A [21], MaDi [7], and CNSN [18].

5.2 Evaluation on the DeepMind Control Suite

We evaluate the agent's generalization ability on five tasks in DMC-GB [12]. The agent is trained with default backgrounds and evaluated on test environments: *Color hard*, *Video easy*, and *Video hard*.

Does SCPL exhibit superior visual generalization capability? We evaluate the visual generalization performance of SCPL across 15 visual perturbed control tasks in the DMC. As shown in Table 1, we report the mean and standard deviation of episode returns over three seeds. SCPL agents are trained using two data augmentation techniques from [12]: *random convolution* and *random overlay*. The SCPL results in Table 1 are based on random convolution for the *Color hard* task, and random overlay for both the *Video easy* and *Video hard* tasks. Table 1 shows that SCPL outperforms other baselines in 13 out of 15 tasks within unseen test environments. Notably, SCPL achieves performance improvements of 12% in walker stand, 11% in walker walk, 28% in cartpole swing-up, 18% in ball in cup, and 9% in finger spin tasks in the challenging *video hard* setting. Overall, SCPL achieves an average performance improvement of 14% across all tasks in the *video hard* environments. Fig.4 presents the test curves for SCPL, SGQN, SVEA, and SAC in these environments, where SCPL demonstrates faster convergence due to its effective task-relevant representations and consistent policies. The experimental results demonstrate that SCPL exhibits superior visual generalization ability in various perturbed environments.

Can SCPL focus on consistent task-relevant pixels? To evaluate the SCPL agent's attention to task-relevant regions, we visualized the saliency maps of agents in both original and perturbed *video hard* observations across five DMC tasks. Fig.5 presents a

Table 1: DMC-GB Generalization Performance

Setting	Task	SAC[9]	SVEA[11]	SIM[36]	TLDA[40]	PIE-G[41]	SGQN[2]	CG2A[21]	MaDi[7]	CNSN[18]	**SCPL**
Color hard	Walker stand	423 ± 155	942 ± 26	940 ± 2	947 ± 26	941 ± 35	948 ± 25	**972 ± 23**	–	942 ± 19	960 ± 11
	Walker walk	255 ± 61	760 ± 145	803 ± 33	823 ± 58	884 ± 20	810 ± 43	902 ± 46	–	815 ± 65	**939 ± 19**
	Cartpole	615 ± 29	837 ± 23	841 ± 13	760 ± 60	749 ± 46	806 ± 6	856 ± 40	–	679 ± 35	**857 ± 12**
	Ball in cup	391 ± 245	961 ± 7	953 ± 7	932 ± 32	964 ± 7	887 ± 10	**972 ± 10**	–	894 ± 78	966 ± 9
	Finger spin	373 ± 70	**977 ± 5**	960 ± 6	–	–	899 ± 27	928 ± 43	–	–	929 ± 24
	Average	411	895	899	865	884	870	926	–	833	**930(+1%)**
Video easy	Walker stand	351 ± 245	961 ± 8	963 ± 5	**973 ± 6**	957 ± 12	955 ± 9	968 ± 6	967 ± 3	967 ± 6	968 ± 8
	Walker walk	228 ± 65	819 ± 71	861 ± 33	873 ± 34	870 ± 22	910 ± 24	918 ± 20	895 ± 24	842 ± 58	**941 ± 9**
	Cartpole	359 ± 80	782 ± 27	770 ± 13	671 ± 57	597 ± 61	761 ± 28	788 ± 24	**848 ± 6**	752 ± 26	814 ± 21
	Ball in cup	338 ± 201	871 ± 106	820 ± 135	887 ± 58	922 ± 20	950 ± 24	**963 ± 28**	807 ± 144	913 ± 45	**963 ± 10**
	Finger spin	260 ± 98	808 ± 33	815 ± 38	744 ± 18	837 ± 107	956 ± 26	912 ± 69	679 ± 17	–	**963 ± 8**
	Average	300	848	845	830	837	906	909	839	869	**930(+2%)**
Video hard	Walker stand	225 ± 58	747 ± 43	827 ± 24	602 ± 51	852 ± 56	851 ± 24	895 ± 35	920 ± 14	871 ± 23	**953 ± 15**
	Walker walk	104 ± 18	385 ± 63	459 ± 67	271 ± 55	600 ± 28	739 ± 21	687 ± 18	504 ± 33	480 ± 46	**818 ± 32**
	Cartpole	174 ± 24	401 ± 38	367 ± 47	286 ± 47	401 ± 21	544 ± 43	472 ± 24	619 ± 24	417 ± 31	**675 ± 3**
	Ball in cup	196 ± 82	498 ± 147	287 ± 39	257 ± 57	786 ± 47	782 ± 57	806 ± 44	758 ± 135	691 ± 72	**924 ± 7**
	Finger spin	26 ± 21	307 ± 24	362 ± 9	241 ± 29	762 ± 59	822 ± 24	819 ± 38	358 ± 25	–	**897 ± 22**
	Average	145	467	460	331	680	747	736	632	615	**853(+14%)**

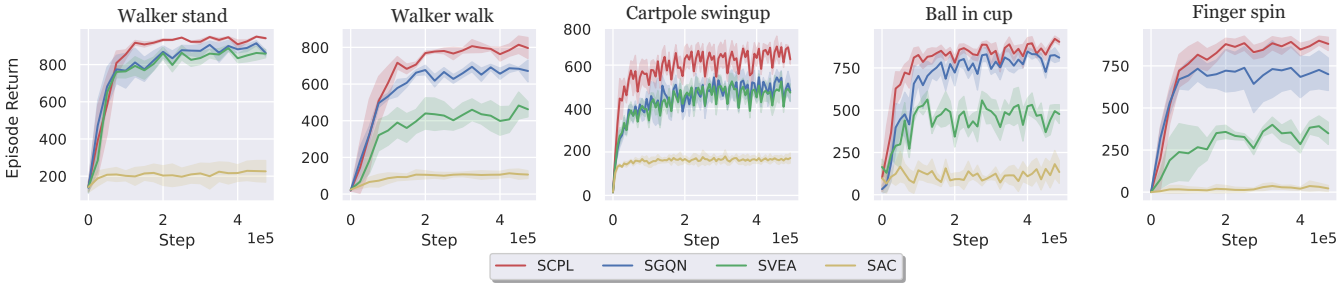

Figure 4: The performance of SAC, SVEA, SGQN, and SCPL in *Video hard* setting. SCPL (red line) shows better generalization.

comparison of the saliency attribute maps for SAC, SVEA, SGQN, and SCPL in the *video hard* setting. By comparing the saliency maps of the agents in original and perturbed observations, SCPL exhibits similar areas of focus for both types of observations across all tasks, while other baselines typically focus on different regions between the original and perturbed observations. The comparison indicates that SCPL demonstrates more consistent attention regions.

By comparing the saliency maps of different methods in original and perturbed observations, we find that SCPL consistently focuses on significant task-relevant regions across both types of observations. In contrast, other baselines typically capture only approximate task-relevant regions in the original observations and struggle to maintain consistent attention to significant task-relevant areas in the perturbed observations. Furthermore, Table 2 evaluates the accuracy of the agent's attended task-relevant regions in perturbed observations using the following metrics: ACC, measuring overall prediction correctness for pixels; AUC, representing the area under the Receiver Operating Characteristic (ROC) curve; and F1 score, considering both precision and recall to compute a unified score. These metrics are averaged across five *Video hard* tasks within the DMC tasks. According to the comparison of saliency maps and statistical results, it's evident that SCPL captures more critical task-relevant regions within various perturbed observations compared to the other methods. Thus, SCPL can consistently focus on task-relevant pixels in both original and perturbed observations.

Table 2: Metrics for attention region of RL agents

	SAC	SVEA	SGQN	**SCPL**
ACC	0.889	0.926	0.932	**0.942**
AUC	0.811	0.833	0.862	**0.908**
F1	0.341	0.462	0.463	**0.566**

Does SCPL possess consistent representations and policies? To further evaluate the task relevance of SCPL's representations and the consistency of its policies, we employed principal component analysis (PCA) to project the representations and actions onto a two-dimensional plane. We plotted the t-SNE visualization of the agent's representations and actions for 800 observations, composed of 20 motions, each featuring various *video hard* backgrounds. Dots of the same color represent representations or actions corresponding to observations with the same motion but different backgrounds. The first row of Fig. 6 displays the t-SNE maps of embeddings learned using SAC, SVEA, SGQN, and SCPL. In the t-SNE map for SCPL, dots of the same color cluster closely together, while clusters of different colors are distinctly separated compared to baseline

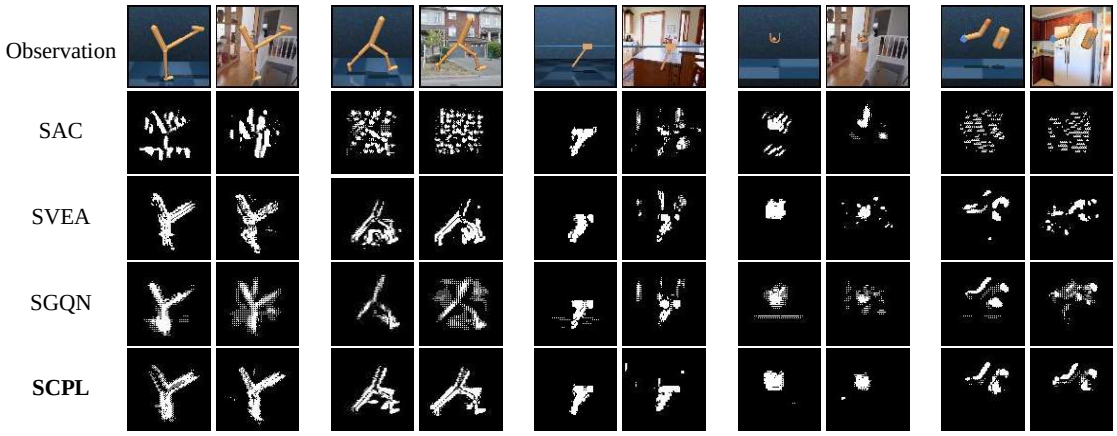

Figure 5: Saliency attribute maps for SAC, SVEA, SGQN, and SCPL in *Training* and *Video hard* setting. In observations of each task, the first column is the original observation, and the second column is the perturbed observation.

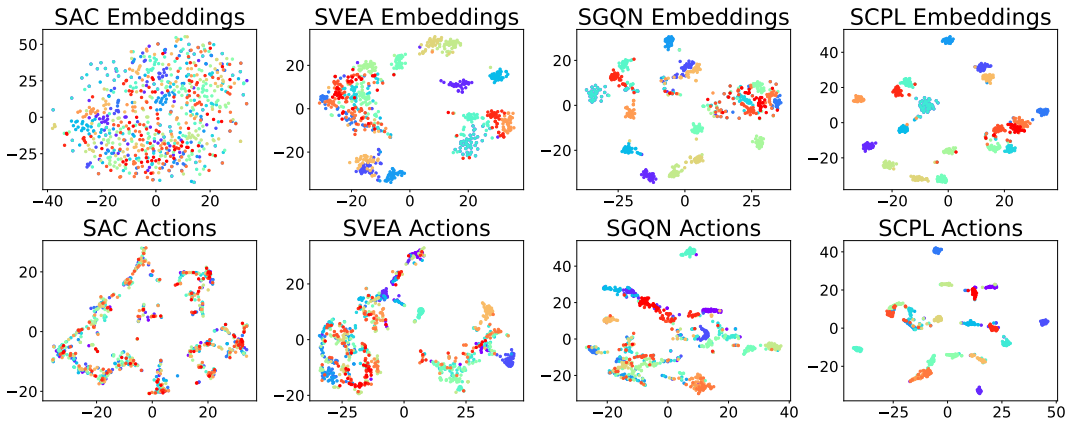

Figure 6: t-SNE maps of embeddings and actions learned with SVEA, SGQN, and SCPL for 20 motion situations, generated by randomly selecting 40 backgrounds from *Video hard*. Different motion situations are represented by different colors, and dots represent representations or actions.

methods. This demonstrates that SCPL generates consistent task-relevant representations for perturbed observations, similar to those for original observations. The second row of Fig. 6 presents t-SNE maps of actions in perturbed observations. In the t-SNE maps of baseline methods, actions for observations with different motions cluster together, indicating inconsistency in their policies. In contrast, actions of SCPL for various perturbed observations with the same motion tend to cluster together, while actions for observations with different motions are clearly separated. The t-SNE maps indicate that SCPL is capable of generating consistent task-relevant representations and policies.

5.3 Ablation Study

What is the contribution of various modules? SCPL leverages the value consistency module, policy consistency module, and dynamics module to enhance generalization. To assess the contribution of each component, we evaluate the generalization performance of SAC with the different modules and analyze their respective and combined effects. The results are demonstrated in Table 3. *SAC + dynamics module* and *SAC + value consistency* refer to the application of the dynamics module and the value consistency module to SAC, respectively. *SAC + value + policy consistency* represents applying both the value consistency module and the policy consistency module to SAC. The percentages denote the enhanced performance of modules within SCPL compared to the performance of vanilla SAC. Specifically, the dynamics module yields improvements of 101% in *color hard* environments, 166% in *video easy*, and 178% in *video hard* environments. The value consistency module achieves impressive gains of 97% in *color hard*, 191% in *video easy*, and 405% in *video hard*. When combining both the value and policy consistency modules, generalization improves further, resulting in performance gains of 112%, 206%, and 465% across the three environments, respectively. In SCPL, the integration of these modules significantly boosts performance across all modes. The ablation

Table 3: Ablation study of three significant components in SCPL

Benchmark	Environment	SAC	SAC + dynamics module	SAC + value consistency	SAC + value + policy consistency	**SCPL**
Video hard	Walker stand	225 ± 58	630 ± 26	918 ± 29	949 ± 9	**953 ± 15**
	Walker walk	104 ± 18	336 ± 14	673 ± 9	812 ± 20	**818 ± 32**
	Cartpole	174 ± 24	351 ± 15	567 ± 64	624 ± 52	**675 ± 3**
	Ball in cup	196 ± 82	405 ± 29	805 ± 67	909 ± 12	**924 ± 7**
	Finger spin	26 ± 21	292 ± 6	703 ± 15	802 ± 6	**897 ± 22**
	Average	145	403(**+178**%)	733(**+405**%)	819(**+465**%)	**853**(**+488**%)

results demonstrate that each component plays a crucial role in enhancing SCPL's visual generalization.

5.4 Generalization in robotic and autonomous driving environments?

Evaluation on Vision-based Robotic Manipulation. To further evaluate the generalization ability of the proposed SCPL, we consider three robot manipulation tasks based on third-person visual input introduced in [14]: *Reach*, *Push*, and *Peg in Box*. All agents are trained using the default settings and evaluated in two modes. The easy mode substitutes the default environment with five different background colors and desktop textures, while the hard mode further replaces the desktop textures with complex images. We compare SCPL with baseline algorithms SAC, SVEA, and SGQN. The results, presented in Table 4, demonstrate that SCPL outperforms the best prior methods in terms of generalization, achieving an average improvement of +7% on the training set, +52% on the easy set, and +39% on the hard set. The experimental results indicate that SCPL outperforms previous methods in robotic environments.

Table 4: Performance comparison on Robotic Manipulation

Setting	Task	SAC	SVEA	SGQN	**SCPL(ours)**
Train	Reach	1.5 ± 6.7	33.6 ± 0.6	33.6 ± 0.7	**33.8 ± 0.3**
	Push	−25.3 ± 13.4	10.8 ± 7.0	18.8 ± 6.4	**19.2 ± 6.1**
	Peg	−12.6 ± 13.2	152.6 ± 21.1	179.8 ± 23.1	**194.6 ± 12.1**
	Average	−12.1	65.7	77.4	**82.6(+7%)**
Test easy	Reach	−22.7 ± 6.4	32.2 ± 1.0	28.2 ± 5.9	**33.3 ± 0.3**
	Push	−23.6 ± 11.2	2.9 ± 10.4	−12.6 ± 12.6	**6.4 ± 6.5**
	Peg	−33.6 ± 20.2	110.4 ± 44.3	94.6 ± 12.0	**181.0 ± 14.0**
	Average	−26.6	48.5	36.7	**73.6(+52%)**
Test hard	Reach	−19.9 ± 4.8	27.8 ± 1.2	18.9 ± 4.6	**31.9 ± 2.0**
	Push	−24.2 ± 11.0	**-1.7 ± 13.5**	−17.7 ± 11.3	−3.1 ± 5.1
	Peg	−25.1 ± 5.2	114.0 ± 43.6	124.8 ± 28.8	**166.4 ± 15.0**
	Average	−23.1	46.7	42.0	**65.1(+39%)**

Evaluation on CARLA autonomous driving environments. CARLA [4] is a widely used simulator for autonomous driving. In our generalization experiments [6], the agents aim to navigate along the road in the *Highway Town04* map, striving to travel as far as possible without colliding within 1000 time steps. The agent is trained under clear noon weather conditions and evaluated in five different weather scenarios, which include varying lighting conditions, realistic rain, and slippery surfaces. We adapted the reward function to align with the settings used in prior work [42]. In Table 5, we present the average driven distance without collisions for vehicles across different weather conditions. Averaged over 10 episodes per weather condition and three training runs, SCPL is able to drive, on average, 69% farther than previous baselines during tests. Notably, in the *sunset* weather scenario, where all other methods struggle, SCPL demonstrates exceptional generalization capabilities. These experimental results indicate that SCPL achieves superior visual generalization performance in CARLA's autonomous driving environments.

Table 5: Performance comparison on CARLA

Setting	SAC	SVEA	SGQN	**SCPL(ours)**
Train	472 ± 110	297 ± 14	614 ± 41	**643 ± 87**
Wet noon	468 ± 68	353 ± 112	473 ± 187	**564 ± 123**
Hard rain noon	306 ± 114	268 ± 89	406 ± 63	**442 ± 199**
Wet sunset	23 ± 16	125 ± 36	39 ± 17	**271 ± 28**
Soft rain sunset	45 ± 25	22 ± 5	59 ± 44	**243 ± 29**
Mid rain sunset	44 ± 24	42 ± 31	63 ± 46	**242 ± 11**
Test Average	177	162	208	**352(+69%)**

6 CONCLUSION

This paper proposes a Salience-Invariant Consistent Policy Learning (SCPL) algorithm for generalization in visual RL. SCPL improves visual generalization by promoting task-relevant representations through its value consistency module, which ensures consistent focus on critical regions in both original and perturbed observations, and its dynamics module, which learns dynamics-relevant features. Additionally, our theoretical analysis reveals that maintaining policy consistency between original and perturbed observations is crucial for visual generalization. Therefore, we propose a policy consistency module to enhance generalization performance by reinforcing policy consistency. Through the extensive experiment results, SCPL demonstrates superior zero-shot generalization performance compared to prior SOTA methods. In this study, SCPL employs fixed saliency quantiles during training. Exploring adaptive quantiles for saliency maps to enhance task-relevant attention regions presents a promising direction for future research.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China under Grants 2022YFA1004000, the Beijing Natural Science Foundation under No. 4242052, the National Natural Science Foundation of China under Grants 62173325, and the CAS for Grand Challenges under Grants 104GJHZ2022013GC.

REFERENCES

[1] Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, et al. 2021. What matters for on-policy deep actor-critic methods? a large-scale study. In *International Conference on Learning Representations (ICLR)*.

[2] David Bertoin, Adil Zouitine, Mehdi Zouitine, and Emmanuel Rachelson. 2022. Look where you look! Saliency-guided Q-networks for generalization in visual Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 30693–30706.

[3] Karl Cobbe, Jacob Hilton, Oleg Klimov, and John Schulman. 2021. Phasic Policy Gradient. In *International Conference on Machine Learning (ICML)*, Vol. 139. PMLR, 2020–2027.

[4] Alexey Dosovitskiy, Germán Ros, Felipe Codevilla, Antonio M. López, and Vladlen Koltun. 2017. CARLA: An Open Urban Driving Simulator. In *Conference on Robot Learning (CoRL)*, Vol. 78. PMLR, 1–16.

[5] Mhairi Dunion, Trevor McInroe, Kevin Sebastian Luck, Josiah P. Hanna, and Stefano V. Albrecht. 2023. Temporal Disentanglement of Representations for Improved Generalisation in Reinforcement Learning. In *International Conference on Learning Representations (ICLR)*.

[6] Linxi Fan, Guanzhi Wang, De-An Huang, Zhiding Yu, Li Fei-Fei, Yuke Zhu, and Animashree Anandkumar. 2021. SECANT: Self-Expert Cloning for Zero-Shot Generalization of Visual Policies. In *International Conference on Machine Learning (ICML)*. PMLR, 3088–3099.

[7] Bram Grooten, Tristan Tomilin, Gautham Vasan, Matthew E. Taylor, A. Rupam Mahmood, Meng Fang, Mykola Pechenizkiy, and Decebal Constantin Mocanu. 2024. MaDi: Learning to Mask Distractions for Generalization in Visual Deep Reinforcement Learning.. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 733–742.

[8] Jeongsoo Ha, Kyungsoo Kim, and Yusung Kim. 2023. Dream to generalize: zero-shot model-based reinforcement learning for unseen visual distractions. In *Association for the Advancement of Artificial Intelligence (AAAI)*, Vol. 37. 7802–7810.

[9] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *International Conference on Machine Learning (ICML)*, Vol. 80. PMLR, 1856–1865.

[10] Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A. Efros, Lerrel Pinto, and Xiaolong Wang. 2021. Self-Supervised Policy Adaptation during Deployment. In *International Conference on Learning Representations (ICLR)*.

[11] Nicklas Hansen, Hao Su, and Xiaolong Wang. 2021. Stabilizing Deep Q-Learning with ConvNets and Vision Transformers under Data Augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*. 3680–3693.

[12] Nicklas Hansen and Xiaolong Wang. 2021. Generalization in Reinforcement Learning by Soft Data Augmentation. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 13611–13617.

[13] Yangru Huang, Peixi Peng, Yifan Zhao, Guangyao Chen, and Yonghong Tian. 2022. Spectrum Random Masking for Generalization in Image-based Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 20393–20406.

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

[15] Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. 2020. Reinforcement learning with augmented data. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. 19884–19895.

[16] Haoran Li, Zhennan Jiang, Yuhui Chen, and Dongbin Zhao. 2024. Generalizing Consistency Policy to Visual RL with Prioritized Proximal Experience Regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

[18] Lu Li, Jiafei Lyu, Guozheng Ma, Zilin Wang, Zhenjie Yang, Xiu Li, and Zhiheng Li. 2024. Normalization Enhances Generalization in Visual Reinforcement Learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 1137–1146.

[19] Xiang Li, Jinghuan Shang, Srijan Das, and Michael Ryoo. 2022. Does self-supervised learning really improve reinforcement learning from pixels?. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 30865–30881.

[20] Minsong Liu, Luntong Li, Shuai Hao, Yuanheng Zhu, and Dongbin Zhao. 2023. Soft Contrastive Learning with Q-irrelevance Abstraction for Reinforcement Learning. *IEEE Transactions on Cognitive and Developmental Systems (TCDS)* 15, 3, 1463–1473.

[21] Siao Liu, Zhaoyu Chen, Yang Liu, Yuzheng Wang, Dingkang Yang, Zhile Zhao, Ziqing Zhou, Xie Yi, Wei Li, Wenqiang Zhang, et al. 2023. Improving generalization in visual reinforcement learning via conflict-aware gradient agreement augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 23436–23446.

[22] Seungyong Moon, JunYeong Lee, and Hyun Oh Song. 2022. Rethinking Value Function Learning for Generalization in Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 34846–34858.

[23] Minting Pan, Xiangming Zhu, Yunbo Wang, and Xiaokang Yang. 2022. Iso-dream: Isolating and leveraging noncontrollable visual dynamics in world models. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 23178–23191.

[24] David Pollard. 2000. Asymptopia: an exposition of statistical asymptotic theory. http://www.stat.yale.edu/Ëœpollard/Books/Asymptopia

[25] Roberta Raileanu and Rob Fergus. 2021. Decoupling value and policy for generalization in reinforcement learning. In *International Conference on Machine Learning (ICML)*. PMLR, 8787–8798.

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

[27] Max Schwarzer, Ankesh Anand, Rishab Goel, R. Devon Hjelm, Aaron C. Courville, and Philip Bachman. 2021. Data-Efficient Reinforcement Learning with Self-Predictive Representations. In *International Conference on Learning Representations (ICLR)*.

[28] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations (ICLR)*.

[29] Jingbo Sun, Xing Fang, and Qichao Zhang. 2023. Reinforcement learning driving strategy based on auxiliary task for multi-scenarios autonomous driving. In *2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS)*. IEEE, 1337–1342.

[30] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[31] Zhentao Tang, Yuanheng Zhu, Dongbin Zhao, and Simon M Lucas. 2023. Enhanced rolling horizon evolution algorithm with opponent model learning: results for the fighting game AI competition. *IEEE Transactions on Games (TOG)* 15, 1, 5–15.

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

[33] Songjun Tu, Jingbo Sun, Qichao Zhang, Yaocheng Zhang, Jia Liu, Ke Chen, and Dongbin Zhao. 2025. In-Dataset Trajectory Return Regularization for Offline Preference-based Reinforcement Learning. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

[34] Junjie Wang, Qichao Zhang, and Dongbin Zhao. 2022. Highway lane change decision-making via attention-based deep reinforcement learning. *IEEE/CAA Journal of Automatica Sinica* 9, 3, 567–569.

[35] Jingda Wu, Zhiyu Huang, and Chen Lv. 2022. Uncertainty-aware model-based reinforcement learning: Methodology and application in autonomous driving. *IEEE Transactions on Intelligent Vehicles (TIV)* 8, 1, 194–203.

[36] Keyu Wu, Min Wu, Zhenghua Chen, Yuecong Xu, and Xiaoli Li. 2022. Generalizing Reinforcement Learning through Fusing Self-Supervised Learning into Intrinsic Motivation. In *Association for the Advancement of Artificial Intelligence (AAAI)*. 8683–8690.

[37] Denis Yarats, Ilya Kostrikov, and Rob Fergus. 2020. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations (ICLR)*.

[38] Denis Yarats, Ilya Kostrikov, and Rob Fergus. 2021. Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels. In *International Conference on Learning Representations (ICLR)*.

[39] Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. 2021. Improving sample efficiency in model-free reinforcement learning from images. In *Association for the Advancement of Artificial Intelligence (AAAI)*, Vol. 35. 10674–10681.

[40] Zhecheng Yuan, Guozheng Ma, Yao Mu, Bo Xia, Bo Yuan, Xueqian Wang, Ping Luo, and Huazhe Xu. 2022. Don't Touch What Matters: Task-Aware Lipschitz Data Augmentation for Visual Reinforcement Learning. In *International Joint Conferences on Artificial Intelligence (IJCAI)*. 3702–3708.

[41] Zhecheng Yuan, Zhengrong Xue, Bo Yuan, Xueqian Wang, Yi Wu, Yang Gao, and Huazhe Xu. 2022. Pre-trained image encoder for generalizable visual reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 13022–13037.

[42] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. 2021. Learning Invariant Representations for Reinforcement Learning without Reconstruction. In *International Conference on Learning Representations (ICLR)*.