DR2: Revisiting Visual Reinforcement Learning from the Dimensional Analysis Perspective

Chuxiong Sun*
National Key Laboratory of Space Integrated Information System, Institute of Software Chinese Academy of Sciences
Beijing, China
chuxiong2016@iscas.ac.cn

Jinli Chen*
Institute of Software Chinese Academy of Sciences
Beijing, China
University of Chinese Academy of Sciences
Beijing, China
chenjinli24@mails.ucas.ac.cn

Zehua Zang
Institute of Software Chinese Academy of Sciences
Beijing, China
University of Chinese Academy of Sciences
Beijing, China
zehua2020@iscas.ac.cn

Jiangmeng Li
National Key Laboratory of Space Integrated Information System, Institute of Software Chinese Academy of Sciences
Beijing, China
jiangmeng2019@iscas.ac.cn

Rui Wang
National Key Laboratory of Space Integrated Information System, Institute of Software Chinese Academy of Sciences
Beijing, China
wangrui@iscas.ac.cn

Changwen Zheng†
National Key Laboratory of Space Integrated Information System, Institute of Software Chinese Academy of Sciences
Beijing, China
changwen@iscas.ac.cn

ABSTRACT

Despite impressive progress on visual control challenges, visual reinforcement learning (VRL) remains sample-inefficient. Existing work commonly leverages auxiliary objectives and data augmentation to learn discriminative representations from observation space containing redundant and task-irrelevant information. However, our analysis shows that the learned representation space still contain dimensional redundancy and dimensional confounders, impeding policy learning. To address these problems, we introduce DR2, a simple plug-and-play module that first constructs a redundancy-reduced representation space and then identifies dimensions most critical for decision making. Concretely, DR2 first applies a redundancy-reduction regularizer to decorrelate latent dimensions, then learns a dimensional mask that models each dimension's gradient contribution to policy learning, dynamically down-weighting task-irrelevant confounders during training. Across diverse visual control benchmarks, DR2 consistently improves sample efficiency and generalization over state-of-the-art baselines. These results indicate that addressing redundancy and confounding at the representation level provides a complementary—rather than substitutive—benefit to existing augmentation and self-supervised strategies.

KEYWORDS

Reinforcement Learning; Visual Reinforcement Learning

*Equal contribution.
†Corresponding author.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/PHPN6008

ACM Reference Format:
Chuxiong Sun, Jinli Chen, Zehua Zang, Jiangmeng Li, Rui Wang, and Changwen Zheng. 2026. DR2: Revisiting Visual Reinforcement Learning from the Dimensional Analysis Perspective. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/PHPN6008

1 INTRODUCTION

Visual reinforcement learning (VRL) enables agents to perceive and make decisions directly from high-dimensional visual inputs, demonstrating significant potential in real-world applications such as robotic manipulation [16, 25] and autonomous driving [3, 53], where manually obtaining compact, low-dimensional observations can be prohibitively expensive or infeasible. However, visual inputs typically contain substantial redundant and task-irrelevant information, severely limiting policy learning. To address this problem, recent work explores auxiliary objectives for learning more effective representations—spanning reconstruction[44, 46], dynamic prediction[15, 32, 45], and contrastive learning[18, 20, 37, 52]. In parallel, data augmentation techniques[14, 19, 28, 42] significantly improve sample efficiency and generalization by synthetically expanding the training data, increasing diversity and quality without requiring additional environment interactions.

Essentially, existing methods based on auxiliary tasks and data augmentation aim to construct a discriminative representation space that mitigate redundant and task-irrelevant information within observation space. However, through explicit dimensional analysis, we demonstrate that the learned representations still retain redundant and task-irrelevant components, which we formally define as **dimensional redundancy** and **dimensional confounders**, respectively.

From the information theory perspective, each dimension holds a subset of the representation's information entropy. Dimensional

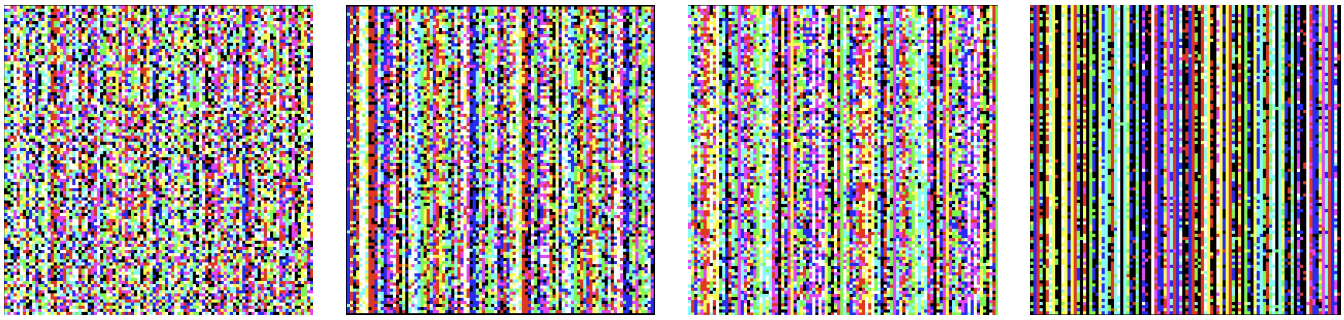

Figure 1: We visualize learned embeddings produced by CURL, DrQv2, PlayVirtual and our proposed method DR2 on a challenging DMC task, FingerSpin. Each visualization encodes the representations into RGB-colored images, where columns represent individual representation dimensions and rows correspond to samples from different trajectories. Different colors indicate different types of features, with greater color contrast reflecting lower dimensional feature similarity. In contrast to existing approaches, DR2 employs a redundancy-reduction technique that efficiently decouples the information into distinct dimensions. Each of these dimensions represents a unique part of the information's entropy. Hence, the learned representation is more informative, and less redundant.

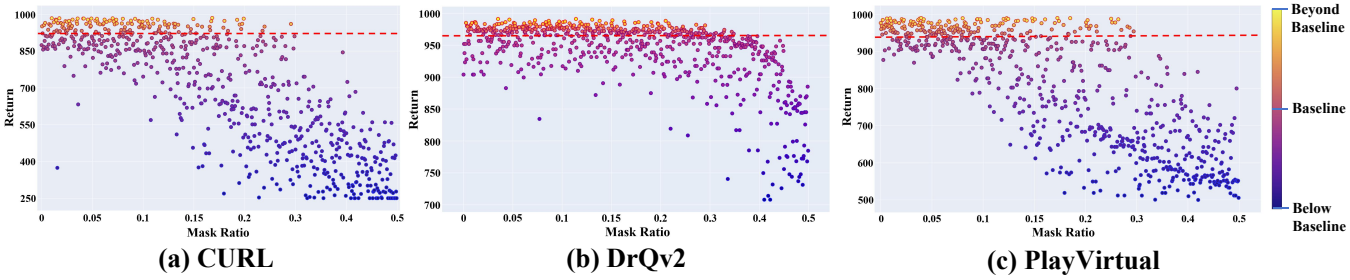

Figure 2: Performance scatter plots demonstrating dimensional confounders in representations learned by CURL, DrQv2, and PlayVirtual on the challenging finger_spin task from the DMC benchmark. Each point corresponds to an independent evaluation with randomly masked representation dimensions at different mask ratios. Notably, certain masked representations achieve higher performance than their unmasked counterparts, empirically validating the pervasive presence of dimensional confounders—latent dimensions containing irrelevant or harmful information—in learned embeddings.

redundancy occurs when multiple dimensions contain overlapping information entropy. To demonstrate the existence of such dimensional redundancy, we visualize the learned representations produced by both CURL[20], DrQv2[14] and PlayVitrual[45]. As illustrated in Fig.1, even with data augmentation and self-supervised objectives, substantial redundancy persists in the learned representations. Therefore, explicit redundancy reduction remains necessary to further improve performance in VRL.

The dimensional confounder refers to a set of dimensions containing "harmful" information, which ultimately degrades the model's performance. To demonstrate the existence of dimensional confounders, we conduct a set of motivating experiments using trained models from CURL, DrQv2, and PlayVirtual. Specifically, for each method, we randomly mask a subset of dimensions in the learned representation at test time by setting their values to zero, and then evaluate the resulting policy performance. The experimental results are shown in Fig.2, where each point corresponds to the average return over multiple trajectories at a given masking ratio. Interestingly, in many cases, partially masked representations yield better performance than their unmasked counterparts. This empirical finding provides strong evidence for the presence of dimensional confounders, suggesting that even after applying data augmentation and auxiliary objectives, harmful task-irrelevant features remain embedded in the learned representations.

With this vision, we propose Dimensional Rational Representation (DR2) to address both dimensional redundancy and confounders in VRL. Specifically, we introduce a redundancy-reduction regularization term that encourages the cross-correlation matrix between two augmented views of the representation to approximate an identity matrix. This simple yet effective constraint recodes highly redundant visual inputs into a decoupled representation space, where components are statistically independent.

To further mitigate the impact of dimensional confounders, we incorporate a dimensional mask mechanism that amplifies the gradient flow of task-relevant dimensions while suppressing the influence of confounding ones during training. The mask is learned via a meta-learning framework, which optimizes the masking weights by maximizing the performance of the masked representation on downstream VRL tasks. As training progresses, the dimensional mask dynamically adjusts based on each dimension's gradient contribution, thereby guiding the model to focus on decision-critical information and improving policy learning efficiency. Empirically, we evaluate DR2 on two widely used VRL benchmarks—Atari[29] and the DeepMind Control Suite(DMC)[39]—covering 38 diverse tasks ranging from discrete-action video games to continuous robot control. Compared to state-of-the-art VRL algorithms, DR2 consistently demonstrates significantly higher sample efficiency and generalization. Moreover, as a plug-and-play module, we demonstrate that DR2 is compatible with a variety of existing methods and serves as a complementary enhancement rather than a replacement. Our key contributions are threefold.

- We perform a dimensional analysis of the representations learned by state-of-the-art VRL algorithms and reveal that, similar to the raw visual observation space, the learned representation space also contains redundant and task-irrelevant information.
- We propose DR2, a unified framework that incorporates a redundancy reduction regularization term and a learnable dimensional mask to explicitly address the challenges of dimensional redundancy and confounders, respectively.
- Empirically, we show that DR2 not only improves sample efficiency and generalization, but also serves as a complementary module to existing VRL methods, enhancing their performance without requiring architectural changes.

2 RELATED WORKS

Reinforcement learning has recently achieved significant breakthroughs in complex decision-making scenarios[7, 8, 23, 24, 30, 31, 35, 36, 38]. Within this domain, VRL has demonstrated remarkable progress in enabling agents to make decisions directly from high-dimensional visual inputs. Existing studies mainly fall into two broad categories: auxiliary learning objectives and data augmentation. Beyond these, we further discuss recent advances in redundancy reduction and disentangled representation learning, which together form the broader context of our work.

Auxiliary Learning Objectives. A large body of research improves VRL by introducing auxiliary objectives to enhance representation quality. CURL[20] pioneers contrastive learning, aligning embeddings from augmentations. ATC[34] maximizes mutual information between current and future representations. ADAT[18] introduces an action-driven contrastive objective, while SPR[32] leverages temporal consistency by predicting future latent features. PSRL[5] proposes a local-guided global contrastive scheme capturing dynamic continuity. TACO[52] uses a temporal action-driven contrastive loss, while CoIT[27] focuses on invariance under image transformations. Recently, CoCo[50] introduces a sequential consistency preserved policy contrast. While enhancing discriminability, these methods do not explicitly address redundancy or confounding at the dimensional level.

Data Augmentation. Another direction focuses on augmentation to improve sample efficiency. RAD[19] first demonstrates that pixel-based augmentations like random cropping stabilize Q-learning. DrQ[43] and DrQv2[14] exploit multi-view averaging for better data utilization. PlayVirtual[45] performs trajectory-level augmentation via cycle-consistent virtual rollouts. Subsequent works such as CycAug[28], SODA[13], SRM[17], CG2A [26], and FGA3[22] explore domain-specific or frequency-domain augmentations to enhance generalization. While increasing data diversity, they typically do not alter latent structure or address dimensional redundancy.

Redundancy Reduction in Self-supervised Learning. Inspired by neuroscience[2], several self-supervised methods introduce decorrelation-based regularizers. Barlow Twins[51] enforces the cross-correlation matrix to approximate an identity matrix, while VICReg[1] and Whitening-based approaches[6] regularize variance and covariance for non-redundant features. The MCR^2 principle[47] emphasizes information-theoretic compression to achieve diverse and discriminative embeddings. Despite advances, their efficacy in sequential decision-making remains unclear. DR2 tackles the noisy-TV issue from a dimensional-analysis perspective. To alleviate difficulties in VRL, DR2 applies redundancy reduction to decouple the feature space, reducing the complexity of discovering the optimal dimensional rational that aligns representations with decision making.

In summary, while prior research improved sample efficiency and generalizatio via auxiliary objectives and augmentation, most methods overlook dimensional redundancy and confounders. DR2 addresses these by combining redundancy reduction with a meta-learned dimensional mask, producing compact, decision-relevant representations. This mechanism complements existing VRL approaches and provides a new dimension-wise perspective on improving sample efficiency and generalization.

3 BACKGROUND

Markov Decision Process. We analyze VRL within the framework of a standard Markov Decision Process (MDP), which models the interaction between an agent and its environment over episodic trajectories. Formally, an MDP is defined as a tuple $(\mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, d_0)$, where $\mathcal{O}$ represents the observation space, consisting of pixel observations o; $\mathcal{A}$ denotes the action space, from which actions a are sampled; $\mathcal{P} : \mathcal{O} \times \mathcal{A} \to \triangle(\mathcal{O})$ specifies the transition dynamics, with $\triangle(\mathcal{O})$ denoting the probability distribution over subsequent observations; $\mathcal{R} : \mathcal{O} \times \mathcal{A} \to \mathbb{R}$ defines the reward function; $\gamma \in (0, 1)$ is the discount factor that governs the importance of future rewards; and $d_0 \in \triangle(\mathcal{O})$ represents the initial observation distribution, defining the probability distribution of the initial observation o_0.

Deep Reinforcement Learning from Pixels. In VRL, learning a policy directly from the high-dimensional observation o_t is intractable. Therefore, algorithms typically employ a deep encoder network, f_{θ_E}, to map the pixel input to a low-dimensional latent representation, $z_t = f_{\theta_E}(o_t)$. This learned representation is intended to capture the essential information required for decision-making.

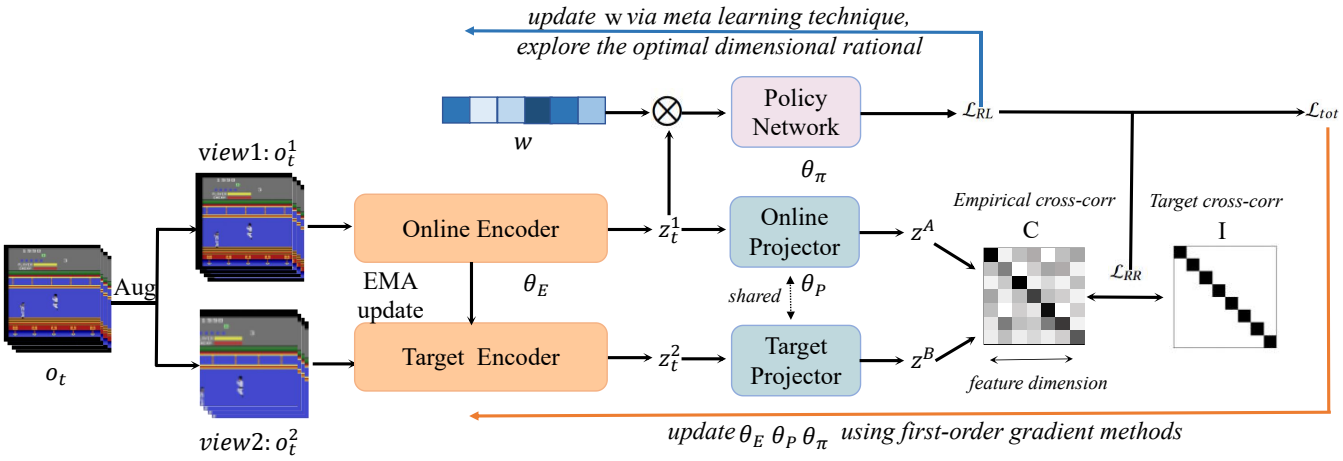

Figure 3: Framework of DR2.

The latent representation z_t serves as the input for a policy network and a value network, which are typically trained by RL and representation learning objectives.

4 METHODS

4.1 Dimensional Redundancy Reduction

As illustrated in Fig.1, representations learned by existing VRL methods still exhibit dimensional redundancy, where multiple dimensions within the embedding space capture overlapping information entropy. To address this issue, we draw inspiration from redundancy reduction theory in neuroscience[2], which posits that a key ability of sensory processing is to transform highly redundant input signals into a factorial code—a representation composed of statistically independent components. Building upon its successful application in self-supervised learning [51], we introduce dimensional redundancy reduction into the VRL setting to explicitly suppress overlapping information across dimensions. Concretely, given a visual observation, we apply data augmentation followed by an encoder and projector to generate two embedding views. To simplify notations, we utilize z^A and z^B to denote the two embeddings after online projector and target projector respectively. The redundancy reduction objective is formulated as follows:

$$\mathcal{L}_{\text{RR}} \triangleq \underbrace{\sum_i (1 - C_{ii})^2}_{\text{invariance term}} + \underbrace{\lambda \sum_i \sum_{j \neq i} C_{ij}^2}_{\text{redundancy reduction term}} \tag{1}$$

where λ is a hyper-parameter, and C is computed between the outputs of the online and target projectors:

$$C_{ij} \triangleq \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b \left(z_{b,i}^A\right)^2} \sqrt{\sum_b \left(z_{b,j}^B\right)^2}} \tag{2}$$

where b refers to batch sample index, and i, j refers to the dimensional index of the projector network outputs. The matrix C is a square matrix with a size equal to the dimensionality of the network's output, and its values range in [-1,1], where -1 denotes perfect anti-correlation and 1 denotes perfect correlation.

Intuitively, the invariance term drives the diagonal entries of the cross-correlation matrix to 1, enforcing that embeddings remain invariant under the applied augmentations. In parallel, the redundancy-reduction term pushes the off-diagonal entries toward 0, decorrelating the embedding components. This decorrelation suppresses redundancy across output units, ensuring that each unit captures distinct, non-overlapping information. With the redundancy-reduction objective $\mathcal{L}_{RR}$ in place, the online encoder efficiently produces well-decoupled and redundancy-reduced representations.

4.2 Dimensional Mask: Addressing Dimensional Confounders

After optimizing the encoder with $\mathcal{L}_{\text{RR}}$, we obtain a redundancy-reduced representation space where each representation space encodes a distinct portion of the information entropy. Yet not every information is beneficial for decision-making tasks: as shown in Fig. 2, masking certain dimensions can even improve policy performance. This observation raises a fundamental question:

Can we reliably identify the contribution of each latent dimension to downstream decision-making?

To answer this question, we introduce a learnable dimensional mask $\boldsymbol{\omega} = \{\omega_k \mid k \in [\![1, D]\!]\}$ to model each dimension's gradient contribution to the RL loss $\mathcal{L}_{\text{RL}}$. The dimensional mask is then used to assign a weight to each dimension of the representation.

$$\tilde{z}_t^1 = z_t^1 \otimes \omega \tag{3}$$

where $\tilde{z}_t^1$ refers to the masked representation capturing task-important information, ω represents the dimensional mask and $\otimes$ is an element-wise Hadamard product function. This process enables to extract only the information from the decoupled representation that is most relevant to decision-making context, while ignoring dimensions that serve as confounders.

Table 1: Performance comparison of DR2 and baseline algorithms on the Atari. The best score for each game is highlighted in bold, and the second-best score is underlined.

Game	Human	Random	SimPLe	CURL	DrQ	SPR	PlayVirtual	CCLF	PSRL	CoIT	Ours
Alien	7127.7	227.8	616.9	558.2	771.2	801.5	947.8	920.0	1030.1	**1206.7**	1144.3
Amidar	1719.5	5.8	88.0	142.1	102.8	176.3	165.3	154.7	114.3	**182.3**	179.3
Assault	742.0	222.4	527.2	600.6	452.4	571.0	702.3	612.4	708.3	635.7	**751.2**
Asterix	8503.3	210.0	**1128.3**	734.5	603.5	977.8	933.3	708.8	959.3	709.0	1057.7
BankHeist	753.1	14.2	34.2	131.6	168.9	**380.9**	245.9	36.0	95.8	124.8	334.3
BattleZone	37187.5	2360.0	5184.4	14870.0	12954.0	16651.0	13260.0	5775.0	16688.0	13760.0	**17322.0**
Boxing	12.1	0.1	9.1	1.2	6.0	35.8	**38.3**	7.4	35.9	23.6	34.3
Breakout	30.5	1.7	16.4	4.9	16.1	17.1	20.6	2.7	17.5	16.1	**21.8**
ChopperCommand	7387.8	811.0	1246.9	1058.5	780.3	974.8	922.4	765.0	1251.2	1338.0	**1783.0**
CrazyClimber	35829.4	10780.5	**62583.6**	12146.5	20516.5	42923.6	23176.7	7845.0	42544.0	17538.0	32091.0
DemonAttack	1971.0	152.1	208.1	817.6	1113.4	545.2	1131.7	**1360.9**	884.0	846.4	1089.0
Freeway	29.6	0.0	20.3	26.7	9.8	24.4	16.1	22.6	24.8	**29.6**	25.4
Frostbite	4334.7	65.2	254.7	1181.3	331.1	1821.5	1984.7	1401.0	776.9	2069.8	**2253.7**
Gopher	2412.5	257.6	771.0	669.3	636.3	715.2	684.3	814.7	**920.3**	746.8	747.3
Hero	30826.4	1027.0	2656.6	6279.3	3736.3	7019.2	**8597.5**	6944.5	3977.3	7572.8	8113.0
Jamesbond	302.8	29.0	125.3	471.0	236.0	365.4	394.7	308.8	**471.4**	336.0	455.7
Kangaroo	3035.0	52.0	323.1	872.5	940.6	3276.4	2384.7	650.0	1580.0	**4116.6**	2652.3
Krull	2665.5	1598.0	4539.9	4229.6	4018.1	3688.9	3880.7	3975.0	**4958.3**	3426.2	4599.3
KungFuMaster	22736.3	258.5	17257.2	14307.8	9111.0	13192.7	14259.0	12605.0	**17759.5**	9250.0	16778.0
MsPacman	6951.6	307.3	1480.0	1465.5	960.5	1313.2	1335.4	1397.5	**1597.3**	1509.6	1329.3
Pong	14.6	-20.7	**12.8**	-16.5	-8.5	-5.9	-3.0	-17.3	-8.2	1.5	4.8
PrivateEye	69571.3	24.9	58.3	**218.4**	-13.6	124.0	93.9	100.0	158.0	145.7	146.7
Qbert	13455.0	163.9	1288.8	1042.4	854.4	669.1	**3620.1**	953.8	1290.3	2117.5	2668.3
RoadRunner	7845.0	11.5	5640.6	5661.0	8895.1	**14220.5**	13534.0	11730.0	3175.7	11758.5	14166.3
Seaquest	42054.7	68.4	683.3	384.5	301.2	583.1	527.7	550.5	**734.9**	554.0	653.7
UpNDown	11693.2	533.4	3350.3	2955.2	3180.8	**28138.5**	10225.2	3376.3	4263.8	4734.2	14258.4
Mean HNS	1.000	0.000	0.443	0.381	0.357	0.703	0.637	0.382	0.610	0.543	**0.739**
Median HNS	1.000	0.000	0.144	0.175	0.268	0.415	0.472	0.181	0.344	0.352	**0.514**

4.3 How to Obtain the Optimal ω?

The dimensional mask is introduced to identify dimensional confounders in the redundancy-reduced representation space, thereby tailoring the learned features to the demands of the decision-making task. The main difficulty stems from both the task's intrinsic complexity and the high dimensionality of the observation space. As a result, attributing each latent dimension's contribution to policy performance is intractable with conventional dimension–importance modeling techniques such as attention mechanisms[40]. Furthermore, estimating the contribution of $\boldsymbol{\omega}$ to $\mathcal{L}_{RR}$ via first-order optimization alone is unreliable, as it is easily trapped by noisy, non-stationary gradients and local minima. Hence, we update the dimensional mask via a meta-learning strategy. Meta-learning mines task-specific knowledge from previous training episodes and, in doing so, endows the dimensional mask with a learn-to-learn capability: it quickly adapts its weighting scheme to track the importance of each dimension, ensuring that the policy concentrates on truly decision-relevant information.

In DR2' training framework, only the ω are refined through the meta-learning process as the complexity of acquiring relationship between representation space and downstream decision-making task, while the rest of the parameters, such as encoder, projector and policy network are updated via standard first-order gradient methods. Specifically, in the first regular training step, we optimize the parameter set $\theta = (\theta_E, \theta_P, \theta_\pi)$ by jointly minimizing the redundancy-reduction loss and the RL objective (with $\boldsymbol{\omega}$ frozen):

$$\arg\min_{\theta} \mathcal{L}_{tot}(\theta, \omega), \tag{4}$$

where $\mathcal{L}_{tot}(\theta, \omega) = \mathcal{L}_{RL} + \beta\mathcal{L}_{RR}$, and β is a weighting coefficient to balance the RL objective with the redundancy reduction objective. It is worth noting that our DR2 framework is compatible with any VRL learning objectives. The experimental results presented in **Section 5.4** substantiate this assertion.

In the second meta-learning-based step, ω is updated using a meta-learning approach that leverages second-order gradient. This method is crucial for adapting ω to properly recognize the importance of each information dimension. The update involves calculating the gradient of ω based on the combined performance metric $\mathcal{L}_{RL}$, and is formulated as follows:

$$\arg\min_{\omega} \mathcal{L}_{RL}(\theta_{trial}, \omega), \tag{5}$$

where $\theta_{trial} = (\theta_E^{trial}, \theta_P^{trial}, \theta_\pi^{trial})$ are the trial weights obtained after a single gradient update on $\mathcal{L}_{RL}$. This update of trial weights is computed as:

$$\theta_{trial} = \theta - \ell_\theta \nabla_\theta \mathcal{L}_{RL}, \tag{6}$$

Table 2: Performance comparison of DR2 and baseline algorithms on the DMC-100k and DMC-500k.

100K Step Score	DrQ	SPR	PlayVirtual	SVEA	CCLF	PSRL	TACO	MaDi	Ours
FingerSpin	901±42	868±143	915±49	859±77	**944±42**	882±132	876±67	810±95	927±32
CartpoleSwingup	759±92	799±42	816±36	727±86	799±61	**849±63**	782±51	704±54	804±54
ReacherEasy	601±213	638±269	785±142	811±115	738±99	621±202	**821±97**	766±101	806±64
CheetahRun	344±67	467±36	474±50	375±54	317±38	398±71	402±62	432±44	**481±55**
WalkerWalk	612±164	398±165	460±173	**747±65**	648±165	595±104	601±103	574±94	547±77
BallInCupCatch	913±53	861±233	929±31	915±71	861±233	922±60	902±54	884±36	**931±21**
Mean Score	688.3	671.8	729.8	739.0	726.7	711.1	730.7	695.0	**749.2**
500K Step Score	**DrQ**	**SPR**	**PlayVirtual**	**SVEA**	**CCLF**	**PSRL**	**TACO**	**MaDi**	**Ours**
FingerSpin	938±103	924±132	963±40	924±93	974±6	961±121	972±89	951±47	**976±6**
CartpoleSwingup	868±10	870±12	865±11	865±10	869±9	**895±39**	870±21	849±6	887±20
ReacherEasy	942±71	925±79	942±66	944±52	941±48	932±41	944±50	955±31	**966±15**
CheetahRun	660±96	716±47	719±51	682±65	588±22	686±80	663±30	732±45	**738±27**
WalkerWalk	921±45	916±75	928±30	919±24	936±23	930±75	914±87	912±26	**944±17**
BallInCupCatch	863±9	963±8	967±5	960±19	961±9	**988±54**	960±22	912±62	974±7
Mean Score	865.3	885.7	897.3	882.3	878.2	894.1	887.1	885.1	**914.3**

with ℓ_θ representing the learning rate. Notably, during the calculation of these trial weights, back-propagation of gradients is excluded to ensure computational efficiency. Consequently, ω is refined through second-order gradient optimization based on θ. This process allows ω to be continuously fine-tuned by considering the gradient contributions from $\mathcal{L}_{RL}$, enabling it to effectively discern the relevance of each information dimension and adapt accordingly. Under this meta-learning scheme, ω dynamically prioritizes decision-critical dimensions, thereby enhancing policy performance.

We train the dimensional mask ω solely from RL objective for two reasons. First, optimizing ω against multiple objectives introduces conflicting gradients and increases optimization difficulty, hindering stable convergence. Second, redundancy reduction is task-agnostic, whereas our goal is to probe the link between representation dimensions and downstream decision making to extract decision-critical information. By updating ω only with respect to the RL objective $\mathcal{L}_{RL}$, the agent learns to highlight task-relevant dimensions and suppress dimensional confounders that are irrelevant to control.

5 EXPERIMENTS

In this section, our experimental design is meticulously structured to address three fundamental questions:

- **RQ1.** How does DR2 compare with state-of-the-art VRL methods in terms of sample efficiency and generalization?
- **RQ2.** Which components of DR2 are critical to its effectiveness?
- **RQ3.** Can DR2 be integrated as a plug-and-play module into diverse VRL algorithms to consistently enhance sample efficiency?

5.1 Setup

Benchmarks. Our evaluation spans data-efficient bench—Atari-100K, DMC-100K, DMC-500K—and a generalization benchmark, DMC-GB, providing coverage of both sample efficiency and robustness. All results are reported under identical training budgets and evaluation protocols to assess sample efficiency and generalization fairly. For each task, we run five random seeds and report the mean and standard deviation of return for evaluation.

Baselines. We compare DR2 against leading VRL algorithms:

- Atari-100K: SimPLe[54], CURL[20], DrQ[43], SPR[32], PlayVirtual[45], CCLF[37], PSRL[5], and CoIT[27];
- DMC-100K and DMC-500K: DrQ[43], SPR[32], PlayVirtual[45], SVEA[12], CCLF[37], PSRL[5], TACO[52] and MaDi[9].
- DMC-GB: SAC[10], DrQ[43], DrQv2[14], RAD[19], PAD[11], SODA[13], SVEA[12], TLDA[48], PIE-G[49], EAR[4].

The results of the baseline algorithms are directly cited and kept consistent with their original papers or subsequent works that follow those baselines.

5.2 Performance(RQ1)

Although DR2 is a plug-and-play module applicable to any VRL algorithm, we instantiate it on PlayVirtual [45] for the sample-efficient benchmarks and on SVEA[12] for the generalization benchmark, and compare against strong baselines to evaluate both sample efficiency and generalization.

Atari-100k. For Atari-100K—a discrete-action benchmark—we evaluate 26 games. As shown in Table 1, DR2 achieves the best overall sample efficiency and asymptotic performance, with a mean human-normalized score (HNS) of 0.739 and a median HNS of 0.514.

Table 3: Performance comparison of DR2 and baseline algorithms on the DMC-GB.

	DMCGB	SAC	DrQ	DrQv2	RAD	PAD	SODA	SVEA	TLDA	PIE-G	EAR	Ours
Video Easy	Walker,Walk	245±165	682±8	175±117	608±92	717±79	771±66	819±71	873±83	871±22	**913±38**	845±17
	Walker,Stand	389±131	873±83	560±48	879±64	935±20	965±7	961±8	946±6	957±12	970±23	**971±8**
	Ball_in_cup,Catch	192±157	318±157	453±60	363±158	436±55	**939±10**	871±106	892±68	922±20	911±40	919±22
	Finger,Spin	152±8	533±119	456±15	334±54	691±80	535±52	808±23	744±18	837±107	717±51	**860±47**
	Cartpole,Swingup	472±26	485±105	267±41	391±66	521±76	678±120	702±80	671±57	587±61	762±88	**797±35**
	Cheetah,Run	87±21	102±30	64±22	43±21	206±34	184±64	249±20	308±57	287±20	334±56	**336±41**
	DMCGB	**SAC**	**DrQ**	**DrQv2**	**RAD**	**PAD**	**SODA**	**SVEA**	**TLDA**	**PIE-G**	**EAR**	**Ours**
Video Hard	Walker,Walk	122±47	104±22	34±11	80±10	189±54	312±32	385±63	271±55	600±28	383±59	**641±22**
	Walker,Stand	231±57	289±49	151±13	229±45	411±36	736±132	747±43	602±51	852±56	744±62	**871±29**
	Ball_in_cup,Catch	101±37	92±23	97±27	98±40	174±71	381±163	403±174	257±57	**786±47**	320±48	779±50
	Finger,Spin	25±6	71±45	21±4	15±6	144±19	221±48	335±58	241±29	762±59	277±62	**790±43**
	Cartpole,Swingup	153±22	138±9	130±3	117±22	255±60	339±87	393±45	286±47	401±21	375±37	**421±18**
	Cheetah,Run	28±6	32±13	23±5	21±7	35±22	94±75	105±37	90±27	134±17	91±46	**138±21**

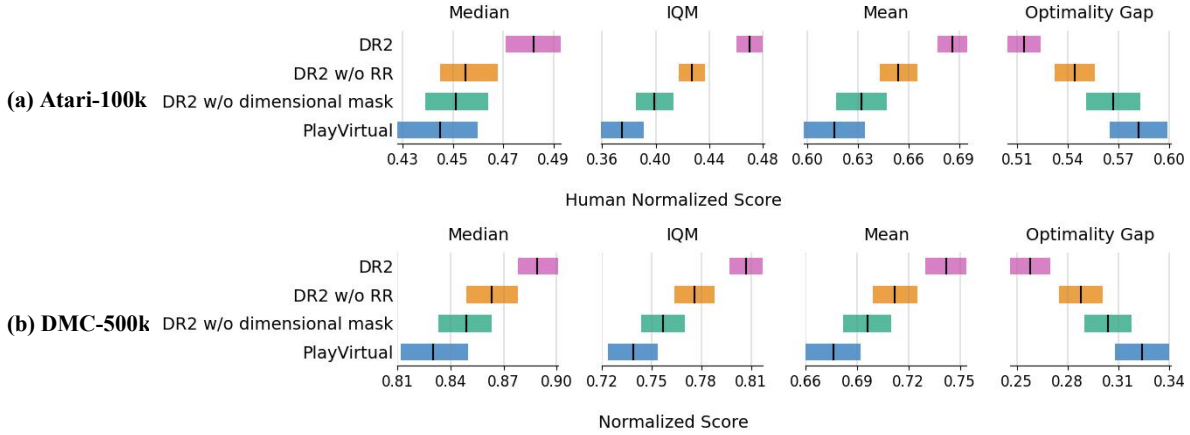

Figure 4: Ablation studies of different DR2 variants.

DMC-100k and DMC-500k. We then evaluate DR2's effectiveness in improving sample efficiency on six challenging continuous-control tasks from the DMC, a setting widely known to exhibit severe sample inefficiency under pixel observations due to high-dimensional inputs and complex, nonstationary dynamics. As shown in Table 2, DR2 achieves the best sample efficiency at both 100K and 500K training steps, with mean scores of 749.2 (DMC-100K) and 914.3 (DMC-500K), respectively.

DMC-GB. At last, we assess DR2's generalization. In this benchmark, agents are trained in a fixed environment and evaluated on two distribution shifts: Video-Easy and Video-Hard. We report results on six tasks for each test distribution. As shown in Table 3, DR2 achieves the top performance on 4/6 Video-Easy tasks and 5/6 Video-Hard tasks. These consistent gains suggest that the redundancy-reduced and de-confounded representations learned by DR2 transfer robustly to out-of-distribution visuals.

These results demonstrate that DR2 effectively mitigates dimensional redundancy and dimensional confounders, achieving state-of-the-art performance in VRL.

5.3 Ablation Study(RQ2)

To evaluate the specific contributions of different components in DR2, we perform ablation studies on five Atari games—Alien, BattleZone, Frostbite, KungFuMaster, RoadRunner—and three DMC tasks—FingerSpin, ReacherEasy, BallInCupCatch. We evaluate four configurations: **DR2** refers to the complete method proposed in our work, incorporating all components. **PlayVirtual** represents the baseline algorithm without any DR2 enhancements. **DR2 w/o RR** is a variant of DR2 which removes the redundancy-reduction regularizer, enabling us to quantify the benefit of decoupled representation space. **DR2 w/o dimensional mask** omits dimensional mask and treats every dimension as equally important. This allows us to assess the significance of mitigating dimensional confounders.

As shown in Fig. 4, omitting either component of DR2 leads to a clear drop in performance. Removing the dimensional mask (DR2 w/o dimensional mask) causes a larger performance drop, because redundancy reduction alone merely produces a decoupled representation, whereas the dimensional mask can align the representation with the decision-critical information, which is essential for decision-making performance. Furthmore, the performance

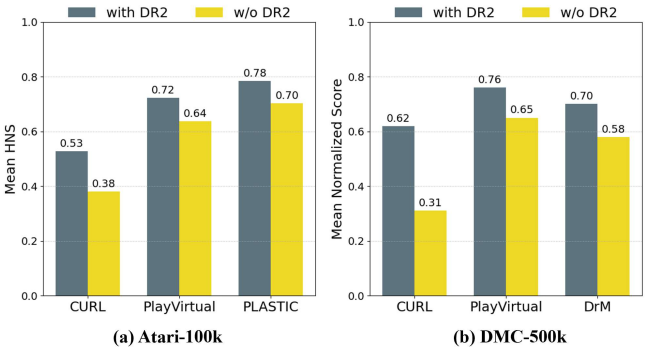

Figure 5: Performance of DR2 integrated with various VRL baselines.

drop of DR2 w/o Redundancy Reduction also highlight the importance of the redundancy-reduction regularizer. Without this step, dimensional redundancy substantially enlarges the search space and hinders the discovery of an optimal dimensional mask. In summary, redundancy reduction and the dimensional mask are mutually reinforcing: the former produces a disentangled representation, and the latter selects the most informative subspace, with both being essential for achieving state-of-the-art performance.

5.4 Performance as a Plug-and-play Module(RQ3)

To assess the plug-and-play ability of DR2, we integrate it into four representative state-of-the-art VRL algorithms: CURL[20], PlayVirtual[45], PLASTIC[21] and DrM[41]. Among these methods, CURL represents the first attempt to combine contrastive learning and data augmentation in VRL, PlayVirtual performs trajectory-level augmentation via cycle consistency, PLASTIC and DrM tackle the plasticity-loss phenomenon in neural networks. We compare each baseline with and without DR2 under identical training budgets and report sample efficiency on the same set of five Atari games and three DMC tasks described in **Section 5.3**. As shown in Fig. 5, we observe that DR2, as a plug-and-play module, consistently enhances the performance and sample efficiency of these leading methods. These results (i) further confirm the presence of dimensional redundancy and dimensional confounders in existing representations and their negative effect on sample efficiency, and (ii) demonstrate that DR2 effectively removes redundancy and suppresses confounders. In summary, dimensional analysis with DR2 serves as a complementary enhancement that reliably improves the efficacy of diverse VRL approaches.

5.5 Visualization

To better understand the effectiveness of DR2, we visualize its learned representations. As shown in Fig. 1, DR2 produces a decoupled embedding, in which different latent dimensions encode distinct portions of the information entropy. This visualization clearly confirms that DR2 effectively removes redundant factors and yields a more disentangled representation than existing methods. Furthermore, we employ Grad-CAM [33], which highlights input regions that most influence the model's decisions, to visualize

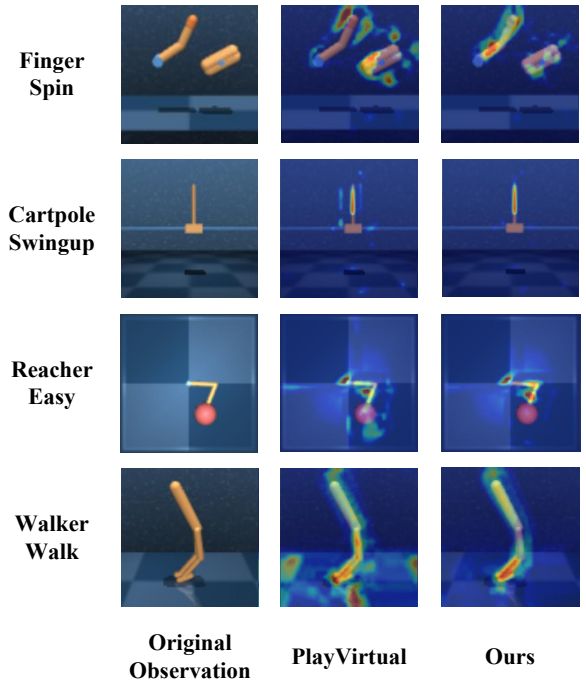

Figure 6: Grad-CAM visualizations on DMC. Brighter regions indicate stronger model attention. Columns (left → right): original pixel observations; features learned by PlayVirtual; features learned by DR2.

what the dimensional mask has learned. The resulting heat maps expose the specific areas of the visual input that DR2 attends to after redundancy reduction and masking. As illustrated in Fig.6, after training on the DMC tasks, the baseline, PlayVirtual, still attends to several task-irrelevant background regions. In contrast, DR2 consistently concentrates on task-relevant areas, effectively filtering out these distracting visual elements.

6 CONCLUSION

In this work, we tackle the longstanding sample-efficiency and generalization bottleneck in VRL. A dimensional analysis of existing methods reveals that dimensional redundancy and dimensional confounders remain in their learned representations, substantially impairing policy learning. To address these issues, we introduce DR2—a simple yet powerful, plug-and-play module that first constructs a redundancy-reduced representation space and then explores the dimensions most critical to decision-making. Concretely, DR2 combines (i) a redundancy-reduction regularizer that encourages statistically decoupled embeddings, and (ii) a learnable dimensional mask that weights each dimension according to its gradient contribution to policy improvement. Comprehensive experiments on Atari and DMC demonstrate that DR2 not only delivers state-of-the-art sample efficiency and generalization but also consistently boosts the performance of a wide range of VRL baselines without requiring any architectural changes. These results underscore DR2's effectiveness and highlight the importance of explicitly handling redundancy and confounders in visual representations.

REFERENCES

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. 2022. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=xm6YD62D1Ub

[2] Horace B Barlow et al. 1961. Possible principles underlying the transformation of sensory messages. *Sensory communication* 1, 01 (1961), 217–233.

[3] Zhong Cao, Kun Jiang, Weitao Zhou, Shaobing Xu, Huei Peng, and Diange Yang. 2023. Continuous improvement of self-driving cars using dynamic confidence-aware reinforcement learning. *Nat. Mac. Intell.* 5, 2 (2023), 145–158. https://doi.org/10.1038/S42256-023-00610-Y

[4] Hyesong Choi, Hunsang Lee, Seongwon Jeong, and Dongbo Min. 2023. Environment Agnostic Representation for Visual Reinforcement Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 263–273.

[5] Hyesong Choi, Hunsang Lee, Wonil Song, Sangryul Jeon, Kwanghoon Sohn, and Dongbo Min. 2023. Local-guided global: Paired similarity representation for visual reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15072–15082.

[6] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. 2021. Whitening for Self-Supervised Representation Learning. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 3015–3024. https://proceedings.mlr.press/v139/ermolov21a.html

[7] Yuqian Fu, Yuanheng Zhu, Haoran Li, Zijie Zhao, Jiajun Chai, and Dongbin Zhao. 2025. CPIG: Leveraging Consistency Policy with Intention Guidance for Multi-agent Exploration. *IEEE Transactions on Cognitive and Developmental Systems* (2025).

[8] Yuqian Fu, Yuanheng Zhu, Jian Zhao, Jiajun Chai, and Dongbin Zhao. [n.d.]. INS: Interaction-aware Synthesis to Enhance Offline Multi-agent Reinforcement Learning. In *The Thirteenth International Conference on Learning Representations*.

[9] 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 2024, Auckland, New Zealand, May 6-10, 2024*. International Foundation for Autonomous Agents and Multiagent Systems / ACM, 733–742. https://doi.org/10.5555/3635637.3662926

[10] 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 *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 1861–1870. https://proceedings.mlr.press/v80/haarnoja18b.html

[11] 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*. https://openreview.net/forum?id=o_V-MjyyGV_

[12] Nicklas Hansen, Hao Su, and Xiaolong Wang. 2021. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. *Advances in neural information processing systems* 34 (2021), 3680–3693.

[13] Nicklas Hansen and Xiaolong Wang. 2021. Generalization in reinforcement learning by soft data augmentation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 13611–13617.

[14] Jianshu Hu, Yunpeng Jiang, and Paul Weng. 2024. Revisiting Data Augmentation in Deep Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=EGQBpkIEuu

[15] Tao Huang, Jiachen Wang, and Xiao Chen. 2022. Accelerating representation learning with view-consistent dynamics in data-efficient reinforcement learning. *arXiv preprint arXiv:2201.07016* (2022).

[16] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. 2023. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. In *Proceedings of The 7th Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 229)*, Jie Tan, Marc Toussaint, and Kourosh Darvish (Eds.). PMLR, 540–562. https://proceedings.mlr.press/v229/huang23b.html

[17] Yangru Huang, Peixi Peng, Yifan Zhao, Guangyao Chen, and Yonghong Tian. 2022. Spectrum random masking for generalization in image-based reinforcement learning. *Advances in Neural Information Processing Systems* 35 (2022), 20393–20406.

[18] Minbeom Kim, Kyeongha Rho, Yong-duk Kim, and Kyomin Jung. 2022. Action-driven contrastive representation for reinforcement learning. *Plos one* 17, 3 (2022), e0265456.

[19] Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. 2020. Reinforcement learning with augmented data. *Advances in neural information processing systems* 33 (2020), 19884–19895.

[20] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. 2020. Curl: Contrastive unsupervised representations for reinforcement learning. In *International conference on machine learning*. PMLR, 5639–5650.

[21] Hojoon Lee, Hanseul Cho, Hyunseung Kim, Daehoon Gwak, Joonkee Kim, Jaegul Choo, Se-Young Yun, and Chulhee Yun. 2023. Plastic: Improving input and label plasticity for sample efficient reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2023), 62270–62295.

[22] Jeong Woon Lee and Hyoseok Hwang. 2025. Fourier Guided Adaptive Adversarial Augmentation for Generalization in Visual Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 18110–18118.

[23] Zhiyuan Li, Wenshuai Zhao, Lijun Wu, and Joni Pajarinen. 2024. Backpropagation through agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 13718–13726.

[24] Zhiyuan Li, Wenshuai Zhao, Lijun Wu, and Joni Pajarinen. 2025. AgentMixer: Multi-Agent Correlated Policy Factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 18611–18619.

[25] Qiyuan Liu, Qi Zhou, Rui Yang, and Jie Wang. 2023. Robust representation learning by clustering with bisimulation metrics for visual reinforcement learning with distractions. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 8843–8851.

[26] 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*. 23436–23446.

[27] Sicong Liu, Xi Sheryl Zhang, Yushuo Li, Yifan Zhang, and Jian Cheng. 2023. On the data-efficiency with contrastive image transformation in reinforcement learning. In *The Eleventh International Conference on Learning Representations*.

[28] Guozheng Ma, Linrui Zhang, Haoyu Wang, Lu Li, Zilin Wang, Zhen Wang, Li Shen, Xueqian Wang, and Dacheng Tao. 2023. Learning better with less: Effective augmentation for sample-efficient visual reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2023), 59832–59859.

[29] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

[30] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

[31] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

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

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[34] Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. 2021. Decoupling representation learning from reinforcement learning. In *International conference on machine learning*. PMLR, 9870–9879.

[35] Chuxiong Sun, Peng He, Qirui Ji, Zehua Zang, Jiangmeng Li, Rui Wang, and Wei Wang. 2024. M2i2: Learning efficient multi-agent communication via masked state modeling and intention inference. *arXiv preprint arXiv:2501.00312* (2024).

[36] Chuxiong Sun, Peng He, Rui Wang, and Changwen Zheng. 2025. Revisiting Communication Efficiency in Multi-Agent Reinforcement Learning from the Dimensional Analysis Perspective. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems* (Detroit, MI, USA) *(AAMAS '25)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1977–1986.

[37] Chenyu Sun, Hangwei Qian, and Chunyan Miao. 2022. CCLF: A Contrastive-Curiosity-Driven Learning Framework for Sample-Efficient Reinforcement Learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3444–3450. https://doi.org/10.24963/ijcai.2022/478 Main Track.

[38] Chuxiong Sun, Zehua Zang, Jiabao Li, Jiangmeng Li, Xiao Xu, Rui Wang, and Changwen Zheng. 2024. T2mac: Targeted and trusted multi-agent communication through selective engagement and evidence-driven integration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 15154–15163.

[39] 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* (2018).

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[41] 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. 2024. DrM: Mastering Visual Reinforcement Learning through Dormant Ratio Minimization. In *International Conference on Learning Representations*, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (Eds.), Vol. 2024. 56219–56243.

[42] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. 2021. Mastering Visual Continuous Control: Improved Data-Augmented Reinforcement Learning. In *Deep RL Workshop NeurIPS 2021*. https://openreview.net/forum?id=L5HKN-IsdSE

[43] 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*.

[44] 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 *Proceedings of the aaai conference on artificial intelligence*, Vol. 35. 10674–10681.

[45] Tao Yu, Cuiling Lan, Wenjun Zeng, Mingxiao Feng, Zhizheng Zhang, and Zhibo Chen. 2021. Playvirtual: Augmenting cycle-consistent virtual trajectories for reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 5276–5289.

[46] Tao Yu, Zhizheng Zhang, Cuiling Lan, Yan Lu, and Zhibo Chen. 2022. Mask-based latent reconstruction for reinforcement learning. *Advances in Neural Information Processing Systems* 35 (2022), 25117–25131.

[47] Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. 2020. Learning Diverse and Discriminative Representations via the Principle of Maximal Coding Rate Reduction. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9422–9434. https://proceedings.neurips.cc/paper_files/paper/2020/file/6ad4174eba19ecb5fed17411a34ff5e6-Paper.pdf

[48] 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 *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3702–3708. https://doi.org/10.24963/ijcai.2022/514 Main Track.

[49] 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*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=FQtku8rkp3

[50] Zehua Zang, Jiangmeng Li, Chuxiong Sun, Rui Wang, Lixiang Liu, and Fuchun Sun. 2026. Visual reinforcement learning via sequential consistency preserved policy contrast from optimal transport view. *Neural Networks* 193 (2026), 108019. https://doi.org/10.1016/j.neunet.2025.108019

[51] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*. PMLR, 12310–12320.

[52] Ruijie Zheng, Xiyao Wang, Yanchao Sun, Shuang Ma, Jieyu Zhao, Huazhe Xu, Hal Daumé III, and Furong Huang. 2023. TACO: Temporal Latent Action-Driven Contrastive Loss for Visual Reinforcement Learning. *Advances in Neural Information Processing Systems* 36 (2023), 48203–48225.

[53] Weitao Zhou, Zhong Cao, Nanshan Deng, Kun Jiang, and Diange Yang. 2023. Identify, estimate and bound the uncertainty of reinforcement learning for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems* 24, 8 (2023), 7932–7942.

[54] Łukasz Kaiser, Mohammad Babaeizadeh, Piotr Miłos, Błażej Osiński, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. 2020. Model Based Reinforcement Learning for Atari. In *International Conference on Learning Representations*.