PQDA:Policy-Aligned Q-Consistency Meets Decoupled Augmentation for Generalizable Visual RL

Yun Zhou[1,2,3], **Yuqiang Wu**[1], **Chunyu Tan**[1,3]

[1]School of Artificial Intelligence, Anhui University, Hefei, China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
[3]Anhui Provincial Key Laboratory of Security Artificial Intelligence, Anhui University, Hefei, China
zhouy@ahu.edu.cn, wuyq@stu.ahu.edu.cn, cytan@ahu.edu.cn

Abstract

A fundamental challenge in visual reinforcement learning (RL) is achieving robust generalization across environments with varying visual distractions. Current RL methods struggle with generalization due to their inability to differentiate foreground and background features during augmentation,while their Q-consistency mechanisms rely on outdated actions from replay buffers that drift from the current policy. In this paper, we present **PQDA**, a novel framework that addresses generalization challenges in RL through two key innovations: (1) Foreground-Background Decoupled Augmentation leverages Gaussian mixture model-based segmentation to efficiently generate and cache masks in replay buffers, applying differentiated augmentation strategies to foreground and background regions, thereby enhancing data diversity while maintaining task-relevant features. (2) Policy-Aligned Q-Consistency enforces policy alignment by sampling actions from the current policy for Q-regularization, achieving faster and more stable convergence. Notably, PQDA eliminates auxiliary tasks entirely through a unified architecture that co-optimizes the encoder and RL components directly. Extensive experiments on DMControl benchmarks (including our newly proposed CVDMC benchmark) and robotic manipulation tasks demonstrate PQDA's superior generalization performance, outperforming state-of-the-art methods.

Introduction

Deep reinforcement learning (RL) has achieved remarkable success in solving complex sequential decision-making problems, from game play (Lanctot et al. 2019; Ye et al. 2020) to robotic control (Han et al. 2023; Kargin and Kołota 2023). Despite these successes, deploying RL in real-world scenarios faces fundamental challenges, with *generalization* standing as the foremost bottleneck. Agents often fail to adapt to environmental variations unseen during training—such as visual distractions, physical parameter shifts, or task modifications—due to overfitting to narrow training distributions.

Recent efforts address this through data augmentation (e.g., RAD (Laskin et al. 2020), DrQ(Yarats, Kostrikov, and Fergus 2021)) to synthetically expand the training distribution, Q-consistency regularization (e.g., SVEA (Hansen, Su, and Wang 2021)) to enforce temporal coherence of value functions across perturbations, and auxiliary objectives (e.g., CURL (Laskin, Srinivas, and Abbeel 2020)) to learn invariant representations. Despite their empirical successes, these methods face four fundamental limitations as presented in Figure 1: (1) *Redundant Mask Generation for Identical Sampled States*: Existing approaches require the network to reprocess identical states sampled from the buffer and regenerate masks independently for each update, creating unnecessary computational overhead. (2) *Lack of Foreground-Specific Augmentation*: Existing augmentation methods typically adopt either global image augmentation (e.g., random overlay (Hansen, Su, and Wang 2021)) or background-only enhancement, failing to prioritize task-relevant foreground regions. (3) *Training Instability with Inconsistent Policy*: Standard Q-consistency regularization relies on outdated actions sampled from replay buffers, which were generated by historical policies. This policy lag introduces bias in gradient estimates and destabilizes training, as the Q-function learns from actions misaligned with the current policy. (4) *Excessive Learning Objectives with Multi-Loss*: While auxiliary tasks (e.g., reconstruction, contrastive learning) jointly train encoders with RL objectives to accelerate convergence, they often optimize features irrelevant to the control task. This not only wastes capacity but introduces competing gradients that increase training variance.

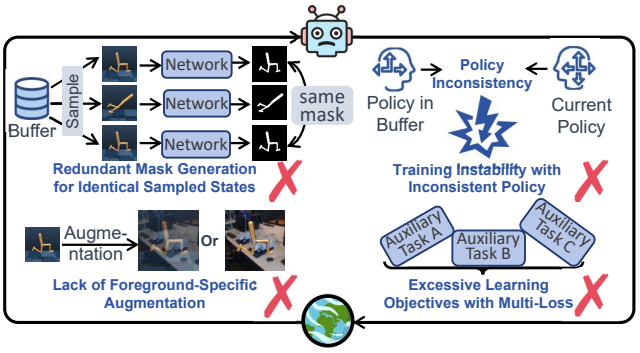

Figure 1: The primary limitations of existing methods. Our method in this paper is specifically designed to overcome these key limitations.

In this work, we present **PQDA** (**P**olicy-aligned **Q**-

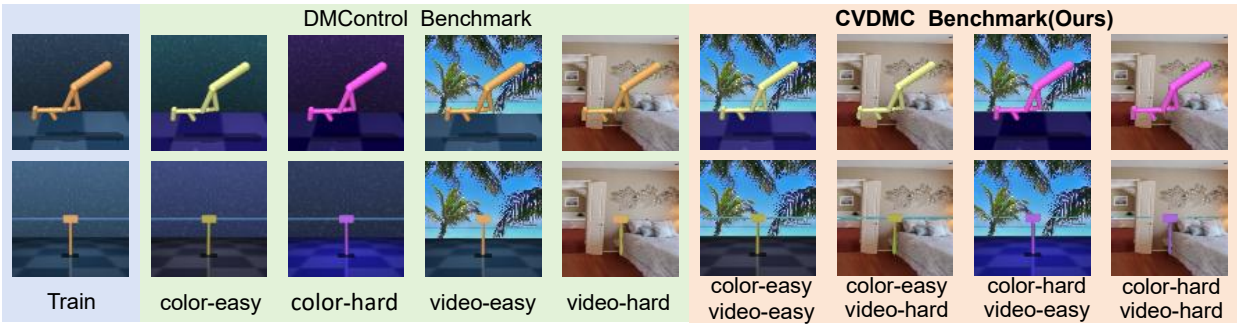

Figure 2: Examples of our proposed Color-Video DMControl (CVDMC) benchmark (*walker-stand* and *cartpole-swingup*), with independent foreground-background variations. More examples can be deferred to the *Appendix*.

consistency with **D**ecoupled **A**ugmentation), a novel framework that addresses these limitations through two synergistic innovations: (1)**Foreground-Background Decoupled Augmentation**: Unlike prior works that augment observations uniformly or only focus on the background, PQDA explicitly decouples foreground and background processing. We employ Gaussian Mixture Models (GMMs) (Zivkovic 2004) to efficiently segment observations into foreground (task-relevant) and background (distracting) regions. The computed segmentation masks are stored directly in the replay buffer, eliminating redundant recomputation when the same state is sampled multiple times. (2)**Policy-Aligned Q-Consistency**: While most off-policy algorithms (e.g., SGQN (Bertoin et al. 2022)) rely on buffer-sampled actions for Q-consistency, we introduce a novel on-policy-inspired loss, dynamically sampling actions from the latest policy for consistency computation. Our method eliminates policy-lag bias by ensuring Q-learning aligns with the current policy's state-action distribution while maintaining off-policy efficiency unlike pure on-policy methods, PQDA strategically leverages the replay buffer for sample-efficient training while avoiding its temporal inconsistency pitfalls through dynamic action resampling from the latest policy.

Our contributions can be summarized as follows:

- We present PQDA, a novel RL framework that eliminates dependency on auxiliary tasks while achieving superior sample efficiency and generalization.
- We propose the first augmentation strategy that explicitly decouples foreground/background processing while maintaining computational efficiency via GMM-based masking and buffer storage.
- We employ the policy-aligned Q-consistency mechanism, which sample actions from the current policy for consistency loss. Our approach is supported by a theoretical analysis that formalizes its robustness and generalization guarantees.
- We introduce the Color-Video DMControl(CVDMC) benchmark, comprising four environments with independent foreground-background variations for generalization evaluation (see Figure 2). Experiments across three platforms(DMControl, CVDMC, and robotic manipulation tasks) demonstrate PQDA's superior performance over sota RL algorithms with higher training efficiency.

Background and Related Work

Reinforcement Learning

Reinforcement learning (RL) is usually formulated as a Markov Decision Process (MDP) (Sutton, Barto et al. 1998) defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}\colon \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the state transition probability function, $\mathcal{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. At each time step t, the agent receives a state $s_t \in \mathcal{S}$ from the environment, selects an action $a_t \in \mathcal{A}$ from a policy $\pi_\theta(a|s)$ parameterized by θ, performs the action a_t in environment, then gains a reward $r_t \in \mathcal{R}$ and next state s_{t+1} from environment completing an interaction with the environment. The agent continuously interacts with the environment, while updating the policy $\pi(a|s)$ for the purpose of generating appropriate actions each time. The agent's goal of RL is to learn an optimal policy $\pi(a|s)$ that maps states to actions to maximize the expected return $J(\pi) = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t r_t]$, where $\tau = (s_0, a_0, s_1, a_1, \ldots, s_t, a_t)$ is the trajectory of the agent in the environment.

Representation Learning for Visual RL

Visual RL leverages high-dimensional image inputs but faces dual challenges: improving sample efficiency for faster convergence and learning generalizable representations for robustness. To accelerate convergence, prior work employs auxiliary tasks such as dynamics prediction (Shelhamer et al. 2016; Guo et al. 2020; Lee et al. 2020a,b; Schwarzer et al. 2021), contrastive learning (Laskin, Srinivas, and Abbeel 2020; Zhu et al. 2022; Yu et al. 2022), reconstruction objectives (Shelhamer et al. 2016; Yarats et al. 2021; Zhao et al. 2024; Zhou et al. 2026), and bisimulation(Zhang et al. 2020). These methods co-train the encoder with auxiliary task, enabling the encoder to learn more effective representations and thereby significantly accelerating the training speed of reinforcement learning. However, their ability to generalize to complex real-world scenarios remains limited—models often fail catastrophically in unseen environments due to overfitting to training-specific features. Recent breakthroughs address this mainly through four synergistic approaches: data augmentation (SODA (Hansen and Wang 2021),SRM(Huang et al. 2022)) diversifies inputs while preserving semantics, Q-value consistency

(SVEA(Hansen, Su, and Wang 2021),SGQN(Bertoin et al. 2022)) stabilizes learning across augmented views, attribute mask (MaDi(Grooten et al. 2023)) selectively obscures task-irrelevant features to force invariant representation learning and hybrid methods (SMG(Zhang et al. 2024)) combine all with auxiliary objectives. Building on these advances, we propose PQDA, a framework that achieves an optimal trade-off between generalization and training efficiency by introducing decoupled augmentation to maximize environmental variability without distorting task-relevant features and redesigning Q-consistency with policy alignment to improve the agent's performance.

Attribution Mask

In visual RL, the expectation for models with strong generalization capabilities is to successfully generalize from a single environment to unseen environments. In other words, the model should focus on task-relevant features, while other distracting information should be regarded as ephemeral and insubstantial. Consequently, numerous excellent approaches have been proposed to encourage the agent to focus on task-relevant features, such as segmentation models(Wang et al. 2023), separate models(Zhang et al. 2024) and saliency maps(Bertoin et al. 2022; Wang et al. 2024b; Song et al. 2024; Sun et al. 2025). In this work, we employ a Gaussian Mixture Model (GMM) for mask generation, achieving computational efficiency, stable high-precision mask generation, and simple implementation.

Data Augmentation

Data augmentation, a cornerstone technique in computer vision, has demonstrated exceptional value in visual reinforcement learning by directly improving model generalization. Through strategic transformations of input data, augmentation provides diversified training samples that enable models to learn robust, generalizable representations while preserving semantic consistency. This approach has been successfully adapted to visual RL through several paradigms: random cropping(Yarats, Kostrikov, and Fergus 2021), color transformation(Hansen, Su, and Wang 2021; Huang et al. 2022), overlay(Zhang et al. 2024; Wang et al. 2024a), and background replacement(Bertoin et al. 2022; Grooten et al. 2023; Zhang et al. 2024). Building upon these advances, we propose a Foreground-Background Decoupled Augmentation (FBDA) framework that fundamentally rethinks how perturbations should be applied in visual RL.

Regularization

Research on generalization in visual RL aims to develop policies that learn robust representations from a single training environment and successfully transfer to unseen test environments. This requires the model to extract invariant semantic features from observationally distinct but semantically equivalent inputs. While Q-consistency regularization has emerged as a dominant approach to stabilize learning, these methods (Yarats, Kostrikov, and Fergus 2021; Hansen, Su, and Wang 2021; Bertoin et al. 2022; Zhang et al. 2024) share a critical limitation: they enforce consistency using historical policy actions from replay buffers, leading to suboptimal gradient signals. To address this, we propose Policy-Aligned Q-Consistency (PAQC), which computes regularization targets using the current policy's actions, and simultaneously improves training stability and enhances final performance.

Approach

We present PQDA (Policy-aligned Q-consistency with Decoupled Augmentation), a novel framework for visual RL generalization that integrates three key innovations: (1) foreground mask extraction via Gaussian Mixture Models, (2) decoupled augmentation applying distinct transformations to foreground (color jitter) and background (natural image blending), and (3) policy-aligned Q-consistency regularization between original and augmented observations. As shown in Figure 3, unlike methods requiring pretrained segmentation model or foreground-agnostic augmentation, our approach simultaneously addresses foreground and background generalization within a unified framework. Through policy-aligned Q-consistency constraint, we eliminate auxiliary supervision needs while achieving maintained optimal performance and explicit training objectives that reduce instability.

Gaussian Mixture Model for Mask Generation

Our framework employs an online Gaussian Mixture Model (GMM) (Zivkovic 2004) to dynamically segment foreground objects from background scenes. As the agent interacts with the environment, we concurrently process each observation frame o_t to generate binary masks M_t, which are stored alongside the original observations in the replay buffer $\mathcal{B}$.

For each pixel x in the observation frame o_t, the foreground mask M_t is obtained through:

$$M_t(x) = \begin{cases} 1 & \text{if } p(x \mid \mathrm{BG}) < c_{\mathrm{thr}} \quad \text{(Foreground)}, \\ 0 & \text{otherwise} \quad \text{(Background)}, \end{cases} \tag{1}$$

where c_{thr} is the threshold. The background model is estimated via a M-component GMM:

$$p(x|\mathcal{X}_T, \mathrm{BG}) = \sum_{m=1}^{M} w_m \mathcal{N}(x; \mu_m, \sigma_m^2 I), \tag{2}$$

where $\mathcal{X}_T = \{x_t, ..., x_{t+T-1}\}$ with a length of T, I is the identity matrix, $\mu_1, , ..., \mu_M$ and $\sigma_1, ..., \sigma_M$ are the estimates of the means and variances of the corresponding Gaussian components. The mixing weights denoted by w_m are non-negative and add up to one. The detailed initialization and update process of the GMM is provided in *Appendix* .

Each generated mask M_t is integrated into the standard experience tuple τ_t, creating augmented transitions for storage in replay buffer $\mathcal{B}$. This design yields two key advantages over existing segmentation-based approaches: (1) *Computational Efficiency*: Each mask is generated once and reused throughout training, eliminating redundant computation required by methods that regenerate masks per sampling. (2) *Training Stability*: The GMM operates independently from RL parameter updates, preventing the propagation of learning instability into mask quality—a critical

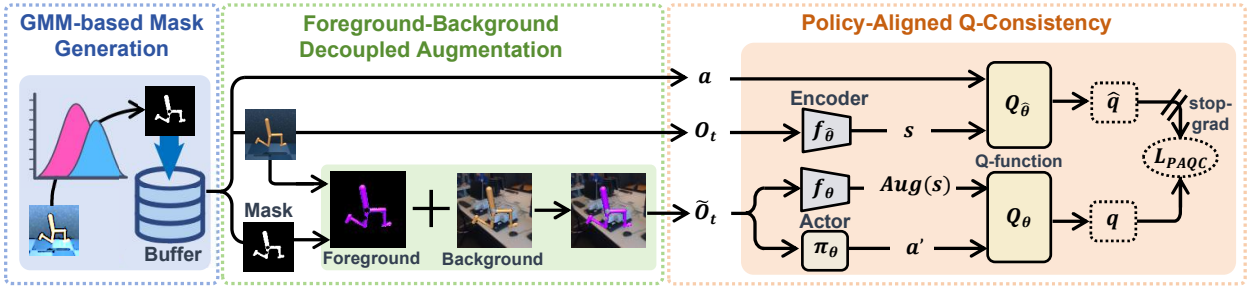

Figure 3: Overall framework of our proposed PQDA method, a novel framework for visual RL generalization that integrates GMM-based Mask Generation, Foreground-Background Decoupled Augmentation, and Policy-Aligned Q-Consistency.

limitation of neural network-based segmentation methods (Zhang et al. 2024; Sun et al. 2025).

Foreground-Background Decoupled Augmentation

Existing approaches for visual RL generalization typically rely on two types of augmentation. Random overlay augmentation usually blends observations o_t with random natural images $b_t \sim \mathcal{D}$ from Places dataset (Zhou et al. 2017) via linear interpolation: $\tilde{o}_t = \alpha b_t + (1-\alpha)o_t, b_t \sim \mathcal{D}$, where $\alpha = 0.5$ is a typical mixing coefficient. Background augmentation preserves foreground pixels (via binary mask M_t) while replacing the background: $\tilde{o}_t = o_t \odot M_t + b_t \odot (1 - M_t), b_t \sim \mathcal{D}$, where $\odot$ denotes the Hadamard product.

While these methods improve background generalization, they exhibit two critical shortcomings. Ineffective for foreground generalization: Neither method perturbs foreground appearance (e.g., object color/texture). Marginal gains: Overlay slightly outperforms mask-based methods, but both fail when both foreground and background vary.

To address these limitations, we propose a decoupled augmentation strategy that independently transforms foreground and background regions:

$$\begin{aligned} \mathcal{A}_{\text{fg}}(o_t) &= \mathcal{C}olor(o_t) \odot M_t, \\ \mathcal{A}_{\text{bg}}(o_t, b_t) &= b_t \odot (1 - M_t), b_t \sim \mathcal{D}. \end{aligned} \tag{3}$$

As shown in Figure 3, the foreground augmentation $\mathcal{A}_{\text{fg}}$ applies stochastic color-jitter (random hue, saturation, brightness, contrast) to foreground pixels ($M_t = 1$) by $\mathcal{C}olor(\cdot)$, enhancing robustness to appearance shifts. The background augmentation $\mathcal{A}_{\text{bg}}$ replaces the background ($M_t = 0$) with diverse natural images $b_t \sim \mathcal{D}$, preserving spatial structure. Through this dual-path augmentation paradigm, we generate the final augmented observation: $\tilde{o}_t = \mathcal{A}_{\text{fg}}(o_t) \oplus \mathcal{A}_{\text{bg}}(o_t, b_t)$, where $\oplus$ denotes the composition operator blending foreground and background.

Policy-Aligned Q-Consistency

Traditional Q-consistency methods enforce temporal consistency by minimizing the discrepancy between Q-values of augmented and original observations, using actions sampled from the replay buffer. Within the standard Soft Actor-Critic (SAC) framework, a approach regularize the Q-network by minimizing the squared error between the Q-values of original and augmented state–action pairs:

$$\mathcal{L}_{\text{Q-orig}} = \mathbb{E}_{(s,a)\sim\mathcal{B}}\left[\left(Q_\theta(s,a) - Q_\theta(\mathcal{A}\text{ug}(s), a)\right)^2\right], \tag{4}$$

where $\mathcal{A}\text{ug}(s)$ denotes the augmented state s'. Empirically, this simple consistency objective substantially boosts the agent's robustness when deployed across environments with unseen visual distractions, establishing it as a widely adopted baseline for improving generalization in visual RL.

However, this approach suffers from a critical limitation: the actions stored in the buffer are generated by historical policies that may significantly deviate from the current policy π_θ due to ongoing policy updates. This mismatch introduces stale-action bias, where the Q-value regularization is computed using outdated action distributions, leading to: *suboptimal gradient directions* that destabilize training and *inconsistent learning signals*, as the buffer actions fail to reflect the current policy's behavior.

To address this, we introduce the Policy-Aligned Q-Consistency (PAQC), which computes consistency targets using actions sampled from the current policy $\pi_\theta(\cdot)$. Unlike stochastic policies that introduce exploration noise, we generates actions deterministically through the policy mean $a' = \mu_\theta(\mathcal{A}\text{ug}(s))$, where μ_θ is the deterministic output of π_θ, eliminating exploratory noise ($\sigma = 0$) typically required in stochastic policies. The loss of our PAQC can be calculated as follows:

$$\mathcal{L}_{\text{PAQC}} = \mathbb{E}_{(s,a)\sim\mathcal{B}}\left[\left(Q_\theta(s,a) - Q_\theta(\mathcal{A}\text{ug}(s), a')\right)^2\right]. \tag{5}$$

The following theorem provides a theoretical guarantee for our PAQC design:

Theorem 1 (Explicit Upper Bound)**.** *Assume Q_θ is Lipschitz continuous in state and action with constant L_Q, and π_θ is Lipschitz continuous in state with constant L_π. For the PAQC objective, the Q-value discrepancy satisfies:*

$$|Q_\theta(s,a) - Q_\theta(\mathcal{A}ug(s), a')| \leq L_Q(1 + L_\pi) \cdot \epsilon_{\text{aug}}. \tag{6}$$

Proof. See *Appendix* for the detailed derivation. □

Theorem 1 demonstrates that PAQC typically leads to a *tighter effective bound* compared to the standard objective. Furthermore, our approach is compatible with trust region

DMControl (*color-easy*)	SAC	SODA	SVEA (overlay)	SRM	SGQN	SMG	**PQDA (Ours)**
cartpole, swingup	178±24	720±109	809±40	856±14	764±84	854±13	**858±8**
finger, spin	296±22	761±87	919±43	916±34	852±126	957±52	**979±7**
walker, stand	592±274	929±23	957±4	953±5	906±50	965±13	**975±9**
walker, walk	430±33	539±51	705±124	632±93	805±47	915±36	**933±8**
cheetah, run	253±27	219±46	289±43	272±24	312±34	346±27	**436±29**
Average	350	634	736	726	728	807	**836**

DMControl *video-easy*	SAC	SODA	SVEA (overlay)	SRM	SGQN	SMG	PQDA (**Ours**)
cartpole, swingup	175±23	617±76	718±101	645±108	717±77	839±16	**850±7**
finger, spin	171±37	615±56	817±94	642±101	860±82	952±48	**968±10**
walker, stand	484±185	924±28	928±50	947±14	949±10	961±19	**980±8**
walker, walk	325±26	518±92	691±120	662±75	830±58	904±34	**945±8**
cheetah, run	179±65	215±15	278±51	253±27	308±34	348±28	**414±21**
Average	267	578	686	630	733	801	**831**

DMControl (*color-hard*)	SAC	SODA	SVEA (overlay)	SRM	SGQN	SMG	**PQDA (Ours)**
cartpole, swingup	184±26	585±66	752±86	752±103	636±110	726±62	**852±10**
finger, spin	271±23	663±106	868±74	834±90	700±219	841±113	**980±8**
walker, stand	526±259	719±138	799±118	807±128	788±114	878±70	**970±17**
walker, walk	379±37	396±78	571±134	483±123	632±176	739±31	**903±23**
cheetah, run	208±54	199±38	238±69	203±30	210±18	299±22	**401±27**
Average	314	512	646	616	593	697	**821**

DMControl *video-hard*	SAC	SODA	SVEA (overlay)	SRM	SGQN	SMG	**PQDA (Ours)**
cartpole,swingup	156±16	346±59	510±177	254±69	599±112	764±32	**818±14**
finger,spin	22±10	310±72	353±71	131±89	710±159	910±61	**943±13**
walker,stand	212±41	406±68	814±57	558±139	870±78	955±9	**978±6**
walker,walk	132±26	175±31	348±80	165±99	634±136	814±51	**844±40**
cheetah,run	56±30	118±40	105±13	87±24	135±44	**303±46**	185±42
Average	116	271	426	239	590	749	**754**

Table 1: DMControl Results with color and video-based distractions.

policy optimization methods. When combined with KL divergence constraints, it can leverage established policy improvement guarantees while providing enhanced generalization through our proposed augmentation and consistency mechanisms.

The complete critic objective combines with the classical critic loss: $\mathcal{L}_{\text{critic}} = \mathcal{L}_{\text{Q}} + \lambda \mathcal{L}_{\text{PAQC}}$, where λ controls the Q-Consistency strength. Our experiments demonstrate that this Policy-Aligned Q-Consistency approach leads to more stable learning and improved generalization compared to the conventional implementation. Our PAQC strategy delivers three key advantages that address fundamental challenges in policy-value alignment: (1) *Gradient alignment*: The Q-value targets are consistent with the current policy's action distribution. (2) *Stability*: Eliminates stale-action bias and improves the training robustness. (3) *Compatibility*: Integrates seamlessly with standard off-policy RL frameworks.

Experimental Results

Setup

We evaluate PQDA on three challenging domains: (1) the standard DMControl Suite with visual distractions (background videos and color variations), (2) the simulated Robotic Manipulation tasks with varying background textures/colors during testing, and (3) four new DMControl-based environments we developed with independent foreground-background variations, termed the Color-Video DMControl (**CVDMC**) Benchmark in Figure 2. We compare our approach against the following methods, including SAC(Haarnoja et al. 2018), SODA(Hansen and Wang 2021), SVEA(Hansen, Su, and Wang 2021), SRM(Huang et al. 2022), SGQN(Bertoin et al. 2022), and SMG(Zhang et al. 2024). All experiments adopt identical backbone networks to SAC to ensure fair comparison. Throughout all experimental results, best performance values are highlighted in bold.

DMControl Results

Our comprehensive evaluation across five DMControl tasks demonstrates consistent superiority of the proposed method under four challenging generalization settings (as provided in Figure 2). The results reveal three key findings: (1) Our method outperforms the SAC baseline by at least **138.9**% in reward return (reaching up to **550.0**% gains in *video-hard* condition), establishing robust performance across all difficulty levels; (2) When compared to the current SOTA SMG algorithm, we achieve a consistent 6.5% average reward improvement; (3) This superior generalization stems from our dual innovation—FBDA combined with PAQC —which enables exceptional adaptability to both color variations (showing only **1.8**% performance drop from *color-easy* to *color-hard*, versus SMG's 13.6% decline) and dynamic video distractions (maintaining the top performance in *video-easy* and *video-hard*), validating our approach's effectiveness in handling diverse visual perturbations while preserving task-relevant features.

Robotic Manipulation Results

To demonstrate our method's generalization capability, we evaluate each task across five distinct test environments with varying visual conditions in simulated Robotic Manipulation tasks. As quantitatively shown in Table 2, our approach establishes SOTA performance in both training efficiency and testing robustness. In the precision-demanding *peg-in-box* task, our proposed PQDA achieves an average reward of 263, outperforming SMG's 234 by 12.4%. These results validate our framework's simulation-to-simulation generalization capacity, crucial for developing robust visuomotor policies in synthetic training environments. Additional results for *Reach* task are provided in the *Appendix*.

CVDMC Benchmark Results

To rigorously evaluate our method's robustness to complex visual distractions, we introduce the **CVDMC**, a novel evaluation framework that systematically combines foreground and background variations across four distinct environments (see Figure 2). As demonstrated in Table 3, our algorithm achieves SOTA performance under the most challenging conditions, particularly in video-hard settings, where dynamic background interference is severe. Under *color-easy video-hard*, our method attains an average reward of 725, outperforming SMG by **21.0**%. Under *color-hard video-hard*, the performance gap widens to **33.3**% (708

Robot-Manip (*peg-in-box*)	SAC	SODA	SVEA (overlay)	SRM	SGQN	SMG	**PQDA (Ours)**
train	31±73	232±20	212±39	227±15	232±19	237±16	**273±18**
test1	-33±25	34±143	-18±59	55±98	-67±28	237±18	**259±11**
test2	-42±31	76±119	85±68	11±54	194±51	219±37	**259±16**
test3	-8±46	66±147	67±73	147±114	198±34	237±15	**270±14**
test4	-42±51	80±122	109±98	112±123	-51±46	237±17	**259±14**
test5	-52±31	-104±51	-26±102	143±122	-108±24	237±15	**258±7**
Average	-24±28	64±98	72±80	116±69	66±143	234±7	**263±6**

Table 2: Robtic manipulation result in *peg-in-box*.

CVDMC Benchmark (*color-easy video-hard*)	SAC (*Baseline*)	SGQN	SMG	**PQDA (Ours)**
cartpole, swingup	179±31	409±66	504±100	**810±6**
finger, spin	40±29	683±50	862±81	**912±10**
walker, stand	402±115	678±38	724±78	**958±22**
walker, walk	81±21	611±38	644±12	**811±50**
cheetah, run	58±19	124±36	**261±43**	135±27
Average	152	501	599	**725**
CVDMC Benchmark (*color-easy video-hard*)	SAC (*Baseline*)	SGQN	SMG	**PQDA (Ours)**
cartpole, swingup	159±37	430±99	527±115	**809±12**
finger, spin	31±24	578±101	730±118	**903±19**
walker, stand	293±70	585±78	596±99	**932±25**
walker, walk	73±13	502±80	591±59	**772±63**
cheetah, run	54±10	77±20	**212±25**	125±39
Average	122	434	531	**708**

Table 3: Performance comparison on CVDMC Benchmark under video-hard conditions.

vs. SMG's 531), highlighting our method's resilience to simultaneous foreground and background distortions. By independently perturbing foreground (agent) and background regions, our method maintains task-relevant features while diversifying distractions—critical for handling video-hard scenarios. The alignment of Q-consistency with the current policy ensures stable learning despite extreme visual noise, reducing performance variance compared to replay-buffer-based approaches. The consistent gains in CVDMC's most difficult settings suggest our method is particularly suited for real-world deployments where visual conditions are unpredictable.

Efficiency Analysis

As demonstrated in Figure 4, our PQDA framework achieves a superior trade-off between training efficiency and final performance compared to other methods. While maintaining comparable training time to SAC (6.5h vs. 5.5h), PQDA delivers **3.7×** higher reward (810 vs. 217), highlighting its exceptional sample efficiency. More significantly, PQDA outperforms SGQN and SMG in both metrics: attaining 22.6%/6.2% higher rewards and reducing training time by **50%**/**61.8%**. This dual advantage stems from two design innovations: (1) Our *training-free* GMM segmentation with mask caching, which eliminating the bottleneck of *training-dependent* mask generation in SMG/SGQN. (2) The policy-aligned Q-consistency, which removes stale-action bias from replay buffers—reducing Q-value drift and directly improving reward performance.

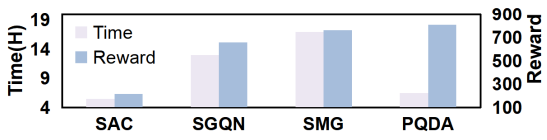

Figure 4: Comparison of training efficiency and average reward in DMControl with color and video-based distractions.

CVDMC Benchmark (*color-easy video-easy*)	SAC (Baseline)	SAC w/ PAQC	SAC * w/ PAQC+FBDA
cartpole, swingup	566±77	602±33	**847±7**
finger, spin	673±56	721±34	**978±8**
walker, stand	790±47	881±25	**968±7**
walker, walk	596±69	657±60	**925±20**
cheetah, run	115±47	167±47	**390±19**
Average	548	606	**822**

* SAC w/ PAQC+FBDA is equal to our full PQDA.

Table 4: Performance Comparison of PQDA Components on CVDMC Benchmark.

Ablation Study

To validate the contributions of each component in our framework, we conduct systematic ablation studies on the selected *color-easy video-easy* environment from our CVDMC Benchmark. This configuration represents a balanced test case with moderate visual complexity, allowing for clear isolation of component effects.

How Q-Consistency and Augmentation Strategy Jointly Drive PQDA Performance? With SAC as the *Baseline*, we compare the following configurations: (1) vanilla SAC; (2) SAC with PAQC; (3) SAC with PAQC and FBDA, with comparative results detailed in Table 4. The results demonstrate clear incremental improvements: (1) *Q-Consistency as a Foundation*: The PAQC alone (SAC w/ PAQC) boosts SAC's performance by 10.6% (548→606) by stabilizing training (lower variance). (2) *Decoupled Augmentation as an Amplifier*: Adding FBDA to the PAQC (full PQDA) yields synergistic effects, further increasing the reward by 35.6% (606→822), indicating its ability to preserve foreground semantics while diversifying backgrounds. This confirms that both components offer complementary benefits: PAQC enhances the policy training stability, while FBDA further improves the performance through foreground-background decoupled augmentation against appearance variations.

Does PAQC Outperform Existing Q-Consistency Method? To evaluate the effectiveness of our Policy-Aligned Q-Consistency (PAQC) mechanism, we conduct bidirectional replacement experiments: (1) integrating PAQC into SMG and SGQN by replacing their original Q-regularization method (Q-orig, as defined in Eq. (4)), and (2) substituting PAQC in our full method with Q-orig (denoted as PQDA w/ Q-orig). As quantitatively shown in Table 5, replacing the original Q-regularization in SMG with PAQC improves the average reward by 3.7% (from 763 to 791). Notably, when retrofitting SGQN with PAQC, we observe an 9.0% reward gain (691→753). Conversely,

CVDMC Benchmark (*color-easy video-easy*)	SGQN Variants		SGQN	SMG Variants		SMG	PQDA Variants		PQDA (Full)
	w/ PAQC	w/ GMM-Mask		w/ PAQC	w/ GMM-Mask		w/ Q-orig	w/ GT-Mask	
cartpole, swingup	781±20	838±11	742±27	828±6	837±16	799±29	833±8	**849±12**	847±7
finger, spin	959±18	971±10	854±46	965±13	979±4	957±3	974±3	**980±2**	978±8
walker, stand	960±10	967±16	905±35	955±4	967±6	918±10	947±11	950±16	**968±7**
walker, walk	803±29	861±37	747±20	895±17	899±3	850±22	901±20	915±16	**925±20**
cheetah, run	263±13	379±20	207±46	310±28	359±11	290±28	328±37	385±30	**390±19**
Average	753	803	691	791	808	763	797	816	**822**

Table 5: Ablation Study of our PAQC vs. Original Q-Consistency(Q-orig) and our GMM-based mask generation strategy vs. other segmentation methods on CVDMC Benchmark.

CVDMC Benchmark (*color-easy video-easy*)	SGQN		SMG	
	Time(h)	Reward	Time(h)	Reward
Original	13	691	17	763
w/ GMM-Mask	10.5	803	13.5	808
Performance Gap	-2.5	+112	-3.5	+45

Table 6: Training Efficiency of our GMM-based mask generation on SGQN and SMG.

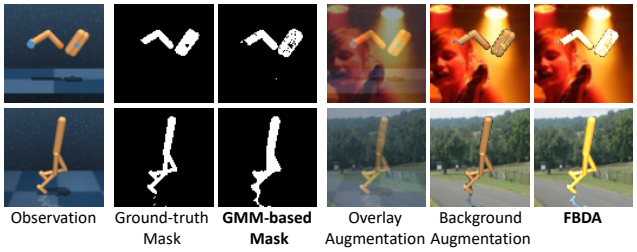

Figure 5: Examples of GMM-based mask and Foreground-Background Decoupled Augmentation (FBDA).

replacing PAQC in our full PQDA framework with Q-orig drops performance by 3.0%, confirming its critical role. The results validate that policy-aligned action sampling provides more stable gradient signals than replay-buffer-based alternatives (Q-orig), while maintaining better sample efficiency across all test cases.

Does GMM-Based Foreground-Background Decoupled Augmentation Improve Generalization? We address two fundamental questions to validate our GMM-Based FBDA design:

(1) **Why choose GMM for mask generation?** Our GMM eliminates both training overhead and runtime recomputation by caching masks in buffers (e.g., SMG's segmentation on reprocessing identical states), while maintaining ground truth (GT)—comparable segmentation as illustrated in Figure 5. Crucially, the reward performance (822 vs. 816) between GMM and GT mask (PQDA vs. PQDA w/GT-Mask) in Table 5 confirms GMM's adequacy for RL tasks. When integrated into SMG and SGQN, our GMM improves their average reward by 5.9% and 16.2% respectively, while achieving 20.6% and 19.2% faster convergence—attributable to the training-free nature of GMM and our buffer-cached mask strategy as shown in Table 6.

(2) **Does decoupled augmentation outperform other augmentation?** Figure 5 demonstrates the superior augmentation strategy of our FBDA framework compared to conventional approaches like background-only augmentation (e.g., SGQN) and random overlay methods (e.g., SMG), where GMM-based masks enable independent yet coordinated transformations: while applying diversified background texture variations, we simultaneously perform targeted color randomization exclusively to foreground agent pixels, preserving structural integrity.

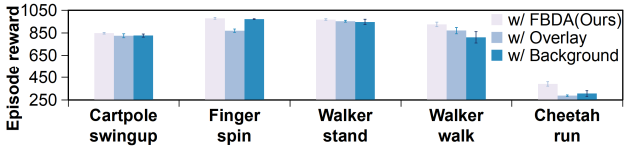

Figure 6: Results of our PQDA with different augmentation strategy in CVDMC Benchmark (*color-easy video-easy*).

This precise region-specific augmentation yields significant performance improvements, as evidenced in Figure 6 by a 6.5% higher average reward over background-only methods (822±12 vs. 772±23) and 8.0% improvement over overlay techniques (vs. 761±15) on CVDMC benchmark, confirming that decoupled augmentation uniquely maintains task-relevant foreground features while maximizing background diversity. The strategic separation inherently aligns with visual RL agents' processing priorities: foreground consistency for stable action determination and background variability for robust generalization, establishing FBDA as both theoretically grounded and empirically validated.

Conclusion

In this work, our PQDA presents a principled solution to the generalization challenge in visual RL by introducing two key innovations: foreground-background decoupled augmentation and policy-aligned Q-consistency. By explicitly differentiating task-relevant and distraction features during augmentation while enforcing Q-consistency with on-policy actions, PQDA achieves superior robustness against visual distractions compared to prior methods. The elimination of auxiliary tasks through a unified architecture further demonstrates that strong generalization can be attained without compromising simplicity or training stability. Extensive empirical validation across DMControl benchmarks (including our proposed CVDMC benchmark) and robotic manipulation tasks confirms PQDA's effectiveness, setting a new SOTA for visual RL in dynamic environments.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No. 62102387 and 62206005), and the University Synergy Innovation Program of Anhui Province (Project No. GXXT-2022-041).

References

Bertoin, D.; Zouitine, A.; Zouitine, M.; and Rachelson, E. 2022. Look where you look! Saliency-guided Q-networks for generalization in visual Reinforcement Learning. *Advances in neural information processing systems*, 35: 30693–30706.

Grooten, B.; Tomilin, T.; Vasan, G.; Taylor, M. E.; Mahmood, A. R.; Fang, M.; Pechenizkiy, M.; and Mocanu, D. C. 2023. Madi: Learning to mask distractions for generalization in visual deep reinforcement learning. *arXiv preprint arXiv:2312.15339*.

Guo, Z. D.; Pires, B. A.; Piot, B.; Grill, J.-B.; Altché, F.; Munos, R.; and Azar, M. G. 2020. Bootstrap latent-predictive representations for multitask reinforcement learning. In *International Conference on Machine Learning*, 3875–3886. PMLR.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. Pmlr.

Han, D.; Mulyana, B.; Stankovic, V.; and Cheng, S. 2023. A survey on deep reinforcement learning algorithms for robotic manipulation. *Sensors*, 23(7): 3762.

Hansen, N.; Su, H.; and Wang, X. 2021. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. *Advances in neural information processing systems*, 34: 3680–3693.

Hansen, N.; and Wang, X. 2021. Generalization in reinforcement learning by soft data augmentation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 13611–13617. IEEE.

Huang, Y.; Peng, P.; Zhao, Y.; Chen, G.; and Tian, Y. 2022. Spectrum random masking for generalization in image-based reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 20393–20406.

Kargin, T. C.; and Kołota, J. 2023. A reinforcement learning approach for continuum robot control. *Journal of Intelligent & Robotic Systems*, 109(4): 77.

Lanctot, M.; Lockhart, E.; Lespiau, J.-B.; Zambaldi, V. F.; Upadhyay, S.; Pérolat, J.; Srinivasan, S.; Timbers, F.; Tuyls, K.; Omidshafiei, S.; Hennes, D.; Morrill, D.; Muller, P.; Ewalds, T.; Faulkner, R.; Kramár, J.; Vylder, B. D.; Saeta, B.; Bradbury, J.; Ding, D.; Borgeaud, S.; Lai, M.; Schrittwieser, J.; Anthony, T. W.; Hughes, E.; Danihelka, I.; and Ryan-Davis, J. 2019. OpenSpiel: A Framework for Reinforcement Learning in Games. *ArXiv*, abs/1908.09453.

Laskin, M.; Lee, K.; Stooke, A.; Pinto, L.; Abbeel, P.; and Srinivas, A. 2020. Reinforcement learning with augmented data. *Advances in neural information processing systems*, 33: 19884–19895.

Laskin, M.; Srinivas, A.; and Abbeel, P. 2020. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, 5639–5650.

Lee, A. X.; Nagabandi, A.; Abbeel, P.; and Levine, S. 2020a. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. In *Advances in Neural Information Processing Systems*, volume 33, 741–752.

Lee, K.-H.; Fischer, I.; Liu, A.; Guo, Y.; Lee, H.; Canny, J.; and Guadarrama, S. 2020b. Predictive information accelerates learning in rl. In *Advances in Neural Information Processing Systems*, volume 33, 11890–11901.

Schwarzer, M.; Anand, A.; Goel, A.; Hjelm, R. D.; Courville, A.; and Bachman, P. 2021. Data-efficient reinforcement learning with self-predictive representations. In *International Conference on Learning Representations*, 1–18.

Shelhamer, E.; Mahmoudieh, P.; Argus, M.; and Darrell, T. 2016. Loss is its own Reward: Self-Supervision for Reinforcement Learning. *ArXiv*, abs/1612.07307.

Song, W.; Choi, H.; Sohn, K.; and Min, D. 2024. A simple framework for generalization in visual RL under dynamic scene perturbations. *Advances in Neural Information Processing Systems*, 37: 121790–121826.

Sun, J.; Tu, S.; Zhang, Q.; Chen, K.; and Zhao, D. 2025. Salience-Invariant Consistent Policy Learning for Generalization in Visual Reinforcement Learning. *arXiv preprint arXiv:2502.08336*.

Sutton, R. S.; Barto, A. G.; et al. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Wang, S.; Wu, Z.; Hu, X.; Wang, J.; Lin, Y.; and Lv, K. 2024a. What effects the generalization in visual reinforcement learning: policy consistency with truncated return prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 5590–5598.

Wang, S.; Wu, Z.; Wang, J.; Hu, X.; Lin, Y.; and Lv, K. 2024b. How to learn domain-invariant representations for visual reinforcement learning: an information-theoretical perspective. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 1389–1397.

Wang, Z.; Ze, Y.; Sun, Y.; Yuan, Z.; and Xu, H. 2023. Generalizable visual reinforcement learning with segment anything model. *arXiv preprint arXiv:2312.17116*.

Yarats, D.; Kostrikov, I.; and Fergus, R. 2021. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International conference on learning representations*.

Yarats, D.; Zhang, A.; Kostrikov, I.; Amos, B.; Pineau, J.; and Fergus, R. 2021. Improving sample efficiency in model-free reinforcement learning from images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12, 10674–10681.

Ye, D.; Chen, G.; Zhang, W.; Chen, S.; Yuan, B.; Liu, B.; Chen, J.; Liu, Z.; Qiu, F.; Yu, H.; et al. 2020. Towards playing full moba games with deep reinforcement learning. In

Advances in Neural Information Processing Systems, volume 33, 621–632.

Yu, T.; Zhang, Z.; Lan, C.; Lu, Y.; and Chen, Z. 2022. Mask-based latent reconstruction for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 25117–25131.

Zhang, A.; McAllister, R. T.; Calandra, R.; Gal, Y.; and Levine, S. 2020. Learning Invariant Representations for Reinforcement Learning without Reconstruction. *ArXiv*, abs/2006.10742.

Zhang, D.; Lv, B.; Zhang, H.; Yang, F.; Zhao, J.; Yu, H.; Huang, C.; Zhou, H.; Ye, C.; et al. 2024. Focus on what matters: Separated models for visual-based rl generalization. *Advances in Neural Information Processing Systems*, 37: 116960–116986.

Zhao, T.; Li, G.; Zhao, T.; Chen, Y.; Xie, N.; Niu, G.; and Sugiyama, M. 2024. Learning explainable task-relevant state representation for model-free deep reinforcement learning. *Neural Networks*, 180: 106741.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464.

Zhou, Y.; Wu, Y.; Wu, Q.; Tan, C.; Zhan, S.; and Hong, R. 2026. Dual-head prediction and reconstruction with coarse-to-fine masks for visual reinforcement learning. *Neural Networks*, 194: 108149.

Zhu, J.; Xia, Y.; Wu, L.; Deng, J.; Zhou, W.; Qin, T.; Liu, T.-Y.; and Li, H. 2022. Masked contrastive representation learning for reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3421–3433.

Zivkovic, Z. 2004. Improved adaptive Gaussian mixture model for background subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, 28–31. IEEE.