

# SEBA: Sample-Efficient Black-Box Attacks on Visual Reinforcement Learning

## Supplementary Material

### 1. Victim Agents

This section summarizes the three visual RL agents used in our experiments, providing a concise overview of the DrQ-SAC and A-LIX algorithms for MuJoCo control and the Rainbow agent for Atari games.

#### 1.1. DrQ-SAC (MuJoCo)

We adopt SAC [5] combined with DrQ [9], which stabilizes pixel-based MuJoCo training through data augmentation, as our primary victim agent.

SAC learns a stochastic policy  $\pi(a_t|s_t)$  via entropy-regularized actor-critic updates.

**Critic.** The critic minimizes the temporal-difference objective:

$$J_Q(\phi_c) = \mathbb{E} \left[ \frac{1}{2} (Q_{\phi_c}(s_t, a_t) - y)^2 \right], \quad (1)$$

$$y = r_t + \gamma \mathbb{E}_{s_{t+1}} V_{\bar{\phi}_c}(s_{t+1}),$$

with the soft value estimate

$$V_{\bar{\phi}_c}(s) = \mathbb{E}_{a \sim \pi} [Q_{\bar{\phi}_c}(s, a) - \alpha \log \pi(a|s)]. \quad (2)$$

**Actor.** The policy maximizes:

$$J_\pi(\phi_a) = \mathbb{E}_{s_t, a_t \sim \pi} [\alpha \log \pi(a_t|s_t) - Q_{\phi_c}(s_t, a_t)]. \quad (3)$$

**Temperature.** The entropy weight  $\alpha$  is adapted by minimizing:

$$J_\alpha(\alpha) = \mathbb{E}_{a_t \sim \pi} [-\alpha \log \pi(a_t|s_t) - \alpha \mathcal{H}]. \quad (4)$$

**Data-Regularized Q (DrQ).** DrQ stabilizes SAC from pixels using random augmentations. For augmentation  $f^{(k)}$ , the target is:

$$y_t = r_t + \gamma \frac{1}{K} \sum_{k=1}^K Q(f^{(k)}(s_{t+1}), a_{t+1}), \quad (5)$$

and the Q-loss averages over  $M$  independently augmented views:

$$\mathcal{L}_Q(\theta) = \frac{1}{M} \sum_{m=1}^M \mathbb{E} \left[ \left( Q_\theta(f^{(m)}(s_t), a_t) - y_t \right)^2 \right]. \quad (6)$$

#### 1.2. Rainbow DQN (Atari)

For Atari experiments, we adopt Rainbow [6], an integrated value-based agent combining six complementary improvements to DQN: (a) double Q-learning, (b) prioritized replay, (c) dueling network architecture, (d) multi-step returns, (e) distributional RL, (f) NoisyNets for exploration.

Rainbow learns a categorical return distribution with 51 atoms and minimizes the KL divergence between predicted and target value distributions. These components significantly enhance stability and sample efficiency in discrete-action visual RL.

#### 1.3. A-LIX (MuJoCo)

To assess cross-algorithm generality, we also attack A-LIX [3], a recent pixel-based RL method that stabilizes TD learning from images. A-LIX introduces a feature-mixing operator (LIX) that smooths encoder feature maps by interpolating each latent activation with its spatial neighbors:

$$\hat{z}_{cij} = (1 - \alpha)(1 - \beta) z_{c[\tilde{i}][\tilde{j}]} + (1 - \alpha)\beta z_{c[\tilde{i}][\tilde{j}+1]} \\ + \alpha(1 - \beta) z_{c[\tilde{i}+1][\tilde{j}]} + \alpha\beta z_{c[\tilde{i}+1][\tilde{j}+1]}. \quad (7)$$

A-LIX adaptively controls the interpolation radius  $S$  using a dual objective that enforces a target level of smoothness in the feature-gradient field. As discussed in Sec. 3, SEBA remains effective against A-LIX agents, demonstrating strong cross-algorithm generality.

## 2. Experimental Settings

### 2.1. General Setup

All results are reported as the mean and standard deviation over 10 random seeds (seed IDs: 1 through 10). Unless otherwise specified, all adversarial perturbations use an  $\epsilon = 8/255 L_\infty$  bound.

### 2.2. SEBA Configuration

**World Model.** We train the tokenizer and the transformer using a replay buffer of 200,000 transitions. The batch size is 64. The tokenizer is optimized using Adam with learning rate  $2 \times 10^{-4}$ , and the transformer uses the same optimizer and learning rate. We set the rollout horizon to  $H = 4$ .

**Shadow Q and GAN Training.** Both the shadow Q model and the GAN modules use the Adam optimizer with learning rate  $10^{-3}$  and batch size 128. Each of the two training phases runs for  $T_1 = 5K$  and  $T_2 = 5K$  iterations, respectively. The alternating two-stage optimization executes

Table 1. **SEBA evaluated on A-LIX visual RL agents.** Lower reward and FID indicate stronger attacks. SEBA maintains strong performance across all tasks, highlighting its generality.

Task (Reward ↓)	No Attack	White-box		Black-box		
	Clean	MAD	C&W	OPTIMAL	Square	SEBA
Cheetah Run	896.95 $\pm$ 21.45	49.53 $\pm$ 16.13	164.36 $\pm$ 39.82	327.92 $\pm$ 49.5	212.02 $\pm$ 34.37	<b>1.88<math>\pm</math>2.05</b>
Walker Walk	923.51 $\pm$ 31.28	243.25 $\pm$ 36.58	377.45 $\pm$ 55.99	544.05 $\pm$ 110.69	664.65 $\pm$ 27.9	<b>25.62<math>\pm</math>7.04</b>
Walker Run	760.47 $\pm$ 31.57	13.84 $\pm$ 4.16	100.61 $\pm$ 38.24	327.12 $\pm$ 7.66	230.53 $\pm$ 34.26	<b>8.73<math>\pm</math>5.38</b>
Reacher Hard	901.33 $\pm$ 44.15	36.11 $\pm$ 22.71	354.86 $\pm$ 78.06	641.39 $\pm$ 82.75	385.63 $\pm$ 63.1	<b>0.92<math>\pm</math>0.46</b>
Hopper Stand	862.9 $\pm$ 83.82	16.47 $\pm$ 11.8	12.6 $\pm$ 16.46	315.53 $\pm$ 19.97	574.89 $\pm$ 22.34	<b>2.72<math>\pm</math>2.28</b>
FID ↓	/	103.22	115.72	87.74	117.46	<b>69.5</b>
Train Env (total) ↓	/	/	/	4M	/	160K
Train Vic (total) ↓	/	/	/	4M	/	800K
Atk. Vic (per-step) ↓	/	11	20	0	202	0

Table 2. Influence of the trade-off parameter  $\lambda$  on SEBA.

Task	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$
Cheetah Run	1.62 $\pm$ 3.26	1.61 $\pm$ 3.97	1.50 $\pm$ 3.58
Walker Walk	37.07 $\pm$ 4.79	35.74 $\pm$ 6.67	34.35 $\pm$ 6.57
Walker Run	16.84 $\pm$ 4.68	17.09 $\pm$ 4.73	16.52 $\pm$ 4.16
FID ↓	61.15	62.43	67.88

Table 3. Influence of the perturbation budget  $\epsilon$  on SEBA.

Task	$\epsilon = 4/255$	$\epsilon = 8/255$	$\epsilon = 12/255$
Cheetah Run	34.22 $\pm$ 8.48	1.61 $\pm$ 3.97	0.81 $\pm$ 1.50
Walker Walk	108.19 $\pm$ 36.94	35.74 $\pm$ 6.67	26.18 $\pm$ 9.01
Walker Run	76.41 $\pm$ 14.32	17.09 $\pm$ 4.73	13.52 $\pm$ 6.86
FID ↓	30.07	62.43	87.24

for a total of  $N_{\text{iter}} = 20$  outer iterations. The generator uses a balancing weight  $\lambda = 1$  to trade off attack strength and imperceptibility.

### 2.3. Baseline Configurations

**Image-based Attacks.** PGD [7]: step size  $\alpha = 2$ , number of update steps = 10. C&W [2]: step size  $\alpha = 1$ , number of update steps = 10, confidence parameter  $c = 10^{-4}$ . SimBA [4]: 100 update steps. Square [1]: 100 update steps, block size = 12.

**Vector-state RL Attack Methods.** Critic-Based [10]:  $\alpha = 2$ , 10 steps. MAD [10]:  $\alpha = 2$ , 10 steps,  $k_{\text{samples}} = 16$ .

**OPTIMAL.** OPTIMAL [11] employs the same generator architecture as SEBA for fairness. The training procedure

follows standard RL: `collect_batch`, `compute_gae`, and `ppo_update`. We use Adam with learning rate  $10^{-4}$ , clip ratio = 0.2, entropy coefficient = 0.0,  $\gamma = 0.99$ , GAE parameter  $\lambda = 0.95$ , PPO epochs = 4, and minibatch size = 64.

**PA-AD.** PA-AD [8] extends OPTIMAL with an additional non-RL actor that proposes  $K = 4$  candidate perturbations, selecting the one with highest estimated attack score.

**Code Reference.** Further implementation details for SEBA and all baselines can be found in our released code-base.

### 3. SEBA on A-LIX Agents

A-LIX [3] is a visual reinforcement learning algorithm that improves training stability by adaptively regularizing latent feature maps and maintaining consistent visual representations. Although A-LIX adopts a different encoder structure and learning dynamics compared with DrQ [9], the results in Tab. 1 show that SEBA still achieves strong attack performance, including substantial reward reduction, low FID, and efficient query usage across all tasks. All reported results are averaged over ten random seeds with mean and standard deviation. These results indicate that SEBA generalizes well across different visual RL algorithms.

### 4. Additional Ablation Studies

This section examines how SEBA behaves under different hyperparameter choices. We study (1) the trade-off parameter  $\lambda$  in the generator loss and (2) the perturbation budget  $\epsilon$  for  $L_{\infty}$ -bounded attacks.

#### 4.1. Effect of the Trade-off Parameter $\lambda$

We evaluate SEBA under three values of the generator weight  $\lambda$ . As shown in Table 2, larger  $\lambda$  mildly strengthens the attack (lower reward) and slightly increases FID. However, all differences remain small, indicating that SEBA is not sensitive to the choice of  $\lambda$ .

#### 4.2. Effect of the Perturbation Budget $\epsilon$

We also test three perturbation budgets for the  $L_\infty$  attack bound. Table 3 reports results under perturbation levels  $\epsilon \in \{4/255, 8/255, 12/255\}$ . A larger budget increases attack strength (lower reward) but leads to higher FID, as expected. Overall, the results vary gradually across different  $\epsilon$  values, indicating that SEBA behaves consistently when the perturbation budget changes.

## References

- [1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII*, pages 484–501. Springer, 2020. 2
- [2] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society, 2017. 2
- [3] Edoardo Ceting, Philip J. Ball, Stephen J. Roberts, and Oya Çelikütan. Stabilizing off-policy deep reinforcement learning from pixels. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 2784–2810. PMLR, 2022. 1, 2
- [4] Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 2484–2493. PMLR, 2019. 2
- [5] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1856–1865. PMLR, 2018. 1
- [6] Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Gheshlaghi Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 3215–3222. AAAI Press, 2018. 1
- [7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2
- [8] Yanchao Sun, Ruijie Zheng, Yongyuan Liang, and Furong Huang. Who is the strongest enemy? towards optimal and efficient evasion attacks in deep RL. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 2
- [9] Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1, 2
- [10] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane S. Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2
- [11] Huan Zhang, Hongge Chen, Duane S. Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2