Efficient Training of Generalizable Visuomotor Policies via Control-Aware Augmentation

Extended Abstract

Yinuo Zhao
Beijing Institute of Technology
Beijing, China
ynzhao@bit.edu.cn

Kun Wu
Beijing Innovation Center of Humanoid Robotics
Beijing, China
gongda.wu@x-humanoid.com

Tianjiao Yi
Beijing Institute of Technology
Beijing, China
tjyi@bit.edu.cn

Zhiyuan Xu
Beijing Innovation Center of Humanoid Robotics
Beijing, China
eric.xu@x-humanoid.com

Zhengping Che
Beijing Innovation Center of Humanoid Robotics
Beijing, China
z.che@x-humanoid.com

Chi Harold Liu
Beijing Institute of Technology
Beijing, China
liuchi02@gmail.com

Jian Tang
Beijing Innovation Center of Humanoid Robotics
Beijing, China
jian.tang@x-humanoid.com

ABSTRACT

Improving generalization is a key challenge in Embodied AI, where obtaining large-scale datasets from diverse scenarios is costly. Visuomotor policies trained with weak augmentations provide only marginal improvements when applied to new environments. Strong augmentations, such as random overlay, can disrupt task-relevant information and degrade performance. To overcome these challenges, we introduce **EAGLE**—an **E**fficient tr**A**ining framework for **G**enera**L**izabl**E** visuomotor policies. EAGLE enhances generalization by applying augmentation only to control-related regions using a self-supervised, control-aware mask. It also boosts training efficiency and stability by transferring knowledge from an expert to a student policy, enabling deployment in new environments without further fine-tuning. Experiments on the DMControl Generalization Benchmark (DMC-GB) demonstrate the effectiveness of our approach. Project website at https://vrl-eagle.github.io/

KEYWORDS

Zero-shot Generalization; Visuomotor Policies; Data Augmentation

ACM Reference Format:
Yinuo Zhao, Kun Wu, Tianjiao Yi, Zhiyuan Xu, Zhengping Che, Chi Harold Liu, and Jian Tang. 2025. Efficient Training of Generalizable Visuomotor Policies via Control-Aware Augmentation: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

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

1 INTRODUCTION

End-to-end visuomotor policies learn low-level controls directly from high-dimensional visual inputs, yielding promising results in tasks like robot manipulation [3, 8], autonomous navigation [1], and locomotion [13, 15]. However, visuomotor policies heavily rely on visual inputs for decision-making and control, making them susceptible to performance degradation when faced with changes in background, distractors, or viewpoints. This deficiency cannot be mitigated through reinforcement nor imitation learning alone.

One promising technique to reduce the impact of these visual discrepancies is Data Augmentation [4, 6, 11, 13, 14]. *Weak augmentations*, like random cropping and flipping, consistently enhance generalization, but with modest improvements. In contrast, *strong augmentations* such as random conv [10] and random overlay [7], boost generalization capabilities through significantly diversifying the data. Nevertheless, they can indiscriminately distort the entire observation space, disrupting the control-related environmental structures and dynamics captured in the data. This often complicates training and destabilizes both learning and testing phases. While previous research [2, 5] has focused on augmenting specific areas within the observation space, they are limited to identifying dynamic objects [12] without considering their task relevance. Recently, vision foundation models like SAM [9] have shown strong generalization abilities. But they still require fine-tuning or human-given priors to identify task-relevant regions in the observation space. Therefore, automatically identifying control-related pixels for generalizable visuomotor policies still remains challenging.

To improve generalization ability, we propose **EAGLE**—an efficient framework for generating generalizable visuomotor policies. EAGLE consists of two modules: 1) A control-aware augmentation module, which identifies control-related pixels using self-supervised reconstruction, and 2) A privilege-guided distillation

module, which extracts control knowledge from an expert agent trained with deep reinforcement learning. This approach enables zero-shot deployment in unseen environments, requiring no additional labels or reward signals. We evaluate **EAGLE**'s zero-shot generalization ability on the DMC-GB [7]. Experimental results demonstrate that **EAGLE** significantly improvements the generalization ability against challenging visual changes.

2 METHOD

We introduce **EAGLE**, an efficient training framework for generalizable visuomotor policies. The overall goal of **EAGLE** is to learn visuomotor policies that are invariant and capable of zero-shot generalization. **EAGLE** consists of two simultaneously optimized modules: a control-aware augmentation module and a privilege-guided distillation module. The former module retrieves temporal data from the replay buffer and conducts a self-supervised reconstruction task, accompanied by three auxiliary losses, to identify control-related pixels. The latter module augments the observation input and distills knowledge from a pretrained DRL expert (which processes only environment states) into the visuomotor student network (which processes only image observations). After training is completed, the visuomotor policy can be reliable deployed in complex environments with visual variations, without the need for fine-tuning or additional supervision.

The control-aware augmentation module learns an attention mask $\boldsymbol{m}$ through four networks: an encoder f_c, an attention block f_a, a decoder f_d and a control predictor f_{ctl}. Given consecutive observations $\boldsymbol{o}_t$ and $\boldsymbol{o}_{t+1}$, we first derive the latent features $\boldsymbol{z}_t = f_e(\boldsymbol{o}_t)$ and the attention mask $\boldsymbol{m}_t = f_a(\boldsymbol{z}_t)$ from the source image. Similarly, we obtain $\boldsymbol{z}_{t+1}$ and $\boldsymbol{m}_{t+1}$ from the target image. Thus, $\boldsymbol{z}_{t+1} \otimes \boldsymbol{m}_{t+1}$ represent the control-related features extracted by the control-aware masks from target images. Then, following [12], we synthesize reconstructed latent features $\hat{\boldsymbol{z}}_{t+1}$ by:

$$\hat{\boldsymbol{z}}_{t+1} = \boldsymbol{z}_{t+1} \otimes \boldsymbol{m}_{t+1} + \boldsymbol{z}_t \otimes (1 - \boldsymbol{m}_t) \otimes (1 - \boldsymbol{m}_{t+1}),$$

where $\otimes$ denotes element-wise multiplication. We decode the target image by processing $\hat{\boldsymbol{z}}_{t+1}$ through the decoder $f_d(\cdot)$ and computing the reconstruction loss as $\mathcal{L}_{rec}(\boldsymbol{o}_t, \boldsymbol{o}_{t+1}) = \|f_d(\hat{\boldsymbol{z}}_{t+1}) - \boldsymbol{o}_{t+1}\|_2^2$.

Then, we introduce three auxiliary losses to facilitate learning a clear control-aware mask. The auto-encoder loss $\mathcal{L}_{ae}$ is utilized to capture essential latent information, which is computed as: $\mathcal{L}_{ae}(\boldsymbol{o}_{t+1}) = \|f_d(\boldsymbol{z}_{t+1}) - \boldsymbol{o}_{t+1}\|_2^2$. The control prediction loss $\mathcal{L}_{ctl}$ is introduced to extract accurate control-related regions, computed as $\mathcal{L}_{ctl}(\boldsymbol{o}_t, \boldsymbol{o}_{t+1}) = \|f_{ctl}(\boldsymbol{z}_t \otimes \boldsymbol{m}_t, \boldsymbol{z}_{t+1} \otimes \boldsymbol{m}_{t+1}) - a_t\|_2^2$. And the sparsity penalty loss $\mathcal{L}_{sps}$ is added to flexibly control the generated attention mask as $\mathcal{L}_{sps} = \|m_j\|_1$. The overall optimization objective in control-aware augmentation module is defined as $\mathcal{L}_{att} = \mathcal{L}_{rec} + \mathcal{L}_{ae} + \beta\mathcal{L}_{ctl} + \lambda\mathcal{L}_{sps}$. We directly upsample the control-aware attention masks $\boldsymbol{m}$ to the observation scale and the augmented image is computed by $\boldsymbol{o}_{aug} = \boldsymbol{o} \otimes \boldsymbol{m} + aug(\boldsymbol{o}) \otimes (1 - \boldsymbol{m})$.

The privilege-guided distillation module receives the augmented image and distills a visuomotor policy from a state-based expert policy π_e, trained using DrQv2 [13]. The student π_θ is updated through minimizing the following objective:

$$\mathcal{L}(\pi_\theta) = \mathbb{E}_{(\boldsymbol{o},\boldsymbol{s})\sim\mathcal{D}}\left[\|\pi_\theta(\boldsymbol{o}_{aug}) - \pi_e(\boldsymbol{s})\|_2^2\right],$$

Settings	SVEA	TLDA	VAI	SAM+E	SGQN	EAGLE
Easy	654	671	738	768	745	**833**
Hard	435	261	616	548	647	**761**

Table 1: Generalization Performance on DMC-GB. We report the average episode returns over 7 tasks with 5 seeds.

	Q.	Aug.	Att.	Exp.	Train	Easy	Hard
Q-only	✓				628.3	318.6	104.7
Q+Aug	✓	✓			468.0	430.60	257.3
Q+Mask	✓	✓	✓		702.9	613.0	509.3
E+Aug		✓		✓	826.9	718.9	440.5
EAGLE		✓	✓	✓	888.8	833.3	761.3

Table 2: Ablation study of our control-aware attention module (Att.) , privilege-guided distillation module (Exp.) and the *random overlay* augmentation (Aug.) on DMC-GB.

where $\boldsymbol{o}_{aug}$ combines observation and augmentation images via $\boldsymbol{m}$. This distillation approach stabilizes the training process by constraining the action space complexity.

3 EXPERIMENTS

Experiment settings. We evaluate **EAGLE** using the DMC-GB [7], which tests an agent's generalization ability from simple to complex environments (Easy and Hard) with background changes. In Hard settings, the background features real-world videos that differ significantly from the training environment. Each method undergoes 500k training iterations, with evaluations based only on visual inputs. We compare **EAGLE** to several SOTA algorithms, including SVEA [6], TLDA [16], VAI [12], SGQN [2] in generalization ability. Besides, we develope a strong baseline SAM+E that combined SAM [9] with our privilege expert.

Comparison results. As shown in Tab. 1, **EAGLE** achieves an average return of 761 in Hard settings, which is 17.6% higher than previous state-of-the-art method SGQN. **EAGLE** overcomes visual distraction limitations via control-aware masks that preserves task-critical regions while augmenting all irrelevant areas.

Ablation studies. We investigate the effect of the proposed control-aware augmentation and privilege-guided distillation modules on training and generalization performance in Tab. 2. We can observe that indiscriminate use of *strong augmentations* degrades training performance, with Q+Aug achieving 160 lower average returns than Q-only. Adding the mask or the Expert alone can boost performance, with Q+Mask and E+Aug improving training results by 12% and 32%, respectively, and achieving 93% and 126% gains in Easy settings. In Hard settings, **EAGLE** achieves an average return of 761, with a 50% and 73% improvement over Q+Maks and E+Aug, respectively. This underscores the joint effect of two modules in enhancing the efficient generalization of visuomotor policies.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (2023YFE0209100) and NSFC, China (U23A20310).

REFERENCES

[1] Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. 2018. Vector-based navigation using grid-like representations in artificial agents. *Nature* 557, 7705 (2018), 429–433.

[2] David Bertoin, Adil Zouitine, Mehdi Zouitine, and Emmanuel Rachelson. 2022. Look where you look! Saliency-guided Q-networks for visual RL tasks. *Advances in neural information processing systems* (2022).

[3] Xi Chen, Ali Ghadirzadeh, Mårten Björkman, and Patric Jensfelt. 2020. Adversarial feature training for generalizable robotic visuomotor control. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1142–1148.

[4] Linxi Fan, Guanzhi Wang, De-An Huang, Zhiding Yu, Li Fei-Fei, Yuke Zhu, and Animashree Anandkumar. 2021. SECANT: Self-expert cloning for zero-shot generalization of visual policies. In *International Conference on Machine Learning*. PMLR, 3088–3099.

[5] Xiang Fu, Ge Yang, Pulkit Agrawal, and Tommi Jaakkola. 2021. Learning task informed abstractions. In *International Conference on Machine Learning*. PMLR, 3480–3491.

[6] 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.

[7] 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.

[8] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. 2018. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*. PMLR, 651–673.

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).

[10] Kimin Lee, Kibok Lee, Jinwoo Shin, and Honglak Lee. 2019. Network randomization: A simple technique for generalization in deep reinforcement learning. *International Conference on Learning Representations* (2019).

[11] Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. 2020. Data-efficient reinforcement learning with self-predictive representations. *International Conference on Learning Representations* (2020).

[12] Xudong Wang, Long Lian, and Stella X Yu. 2021. Unsupervised visual attention and invariance for reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6677–6687.

[13] 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*.

[14] 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*.

[15] Wenhao Yu, Deepali Jain, Alejandro Escontrela, Atil Iscen, Peng Xu, Erwin Coumans, Sehoon Ha, Jie Tan, and Tingnan Zhang. 2021. Visual-locomotion: Learning to walk on complex terrains with vision. In *5th Annual Conference on Robot Learning*.

[16] Z Yuan, G Ma, Y Mu, B Xia, B Yuan, X Wang, P Luo, and H Xu. 2022. Don't touch what matters: Task-aware lipschitz data augmentationfor visual reinforcement learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, Vienna, 23-29 July 2022*. International Joint Conferences on Artificial Intelligence.