Learning What Matters: A Problem in Robotic Reinforcement Learning

Baran Özer
03783103
Technical University of Munich

Serife Damla Konur
03780211
Technical University of Munich

I. Introduction

Reinforcement learning (RL) has shown great promise in enabling autonomous agents to learn complex behaviors through interaction with their environment. Traditionally, much of this progress has been driven by training in simulated environments, where large-scale data collection and safe experimentation are possible. However, recent trends point toward a growing interest in training RL agents directly on physical robots, a paradigm known as on-robot reinforcement learning. This approach eliminates the need for carefully designed simulators and addresses the longstanding simulation-to-reality (sim-to-real) gap, thereby enabling more robust real-world policy deployment.

Despite these advantages, on-robot RL presents unique challenges. In particular, agents must process high-dimensional, noisy, and redundant sensor data in real time. Without the benefit of hand-crafted features or idealized simulation inputs, agents risk being overwhelmed by irrelevant information, leading to inefficient learning or unstable policies.

To address these challenges, we explore the integration of self-attention mechanisms into Proximal Policy Optimization (PPO) to enhance the agent's ability to dynamically filter out noisy or irrelevant features. We systematically design and evaluate several attention-based network architectures, including Feature-Attention, Selective-Attention, Hard-Gated Attention, and our proposed Frame Attention architecture, each targeting different aspects of noise suppression and relevant feature selection.

Our experimental strategy follows a curriculum starting from a baseline with vanilla PPO, establishing upper-bound performance, incorporating frame stacking, and incrementally integrating attention mechanisms. We evaluate all architectures on two standard continuous control environments, AntBullet and LunarLanderContinuous, under a range of observation noise conditions.

Our contributions are:

- We propose Frame Attention, a new attention module for PPO that improves robustness to observation noise by prioritizing relevant temporal features
- We demonstrate that attention-enhanced PPO policies improve learning robustness in noisy, high-dimensional settings relevant to real-world robot learning

II. Related Work

In recent years, attention mechanisms such as self-attention have gained significant attraction in deep reinforcement learning, particularly in the context of vision-based tasks [1], [2], [3]. For instance, self-attention mechanisms have been successfully applied to enable agents to focus on relevant parts of an image, reducing the computational load and improving decision-making [1]. Additionally, the method in [2] employs an attention-based hierarchical policy to guide lane-change behaviors by focusing on the most relevant regions of camera inputs by enhancing decision accuracy. Similarly, the method in [3] applies attention over visual features in both the policy and value branches, enabling interpretable decision-making in vision-based deep reinforcement learning tasks such as Atari games and robotic manipulation. These works demonstrate how attention can filter and prioritize high-dimensional sensory data and they outperform traditional methods in terms of efficiency and generalization.

However, the application of self-attention mechanisms to filter out noise in robotic reinforcement learning is still an under-explored area. This gap in the literature motivates the investigation of self-attention-based approaches in robotic RL to address the problem of information overload and noise filtering.

III. Method

In this section, we present the Frame Attention mechanism, a lightweight self-attention module designed to filter noisy observation features in reinforcement learning. The core idea is to reorganize stacked observations into feature tokens and apply self-attention across features to dynamically suppress irrelevant or inconsistent inputs.

A. Feature History as Tokens

Let the observation at time t be $\mathbf{o}_t \in \mathbb{R}^d$. We maintain a framestack of F recent observations:

$$\{\mathbf{o}_{t-F+1}, \dots, \mathbf{o}_t\} \in \mathbb{R}^{F \times d}.$$

For each feature index $i \in \{1, \dots, d\}$, we collect its past F values to form a feature history vector:

$$\mathbf{h}_i = \begin{bmatrix} o_{t-F+1}^{(i)} & \dots & o_t^{(i)} \end{bmatrix} \in \mathbb{R}^F.$$

Stacking all feature history vectors yields a feature-by-time matrix:

$$H = \begin{bmatrix} \mathbf{h}_1^\top \\ \vdots \\ \mathbf{h}_d^\top \end{bmatrix} \in \mathbb{R}^{d \times F}.$$

A learned positional encoding over frames $P \in \mathbb{R}^{d \times F}$ is added elementwise:

$$\tilde{H} = H + P.$$

B. Feature-Token Self-Attention

Each feature history (row of $\tilde{H}$) is projected from F to d_{model}:

$$Z = \tilde{H}W + \mathbf{1}b^\top \quad \in \mathbb{R}^{d \times d_{\text{model}}}, \qquad W \in \mathbb{R}^{F \times d_{\text{model}}},\ b \in \mathbb{R}^{d_{\text{model}}}.$$

We treat the d rows $\{\mathbf{z}_i\}_{i=1}^d$ of Z as a sequence of *feature tokens* and apply multi-head self-attention *across features* (sequence length $= d$). For head $h = 1, \ldots, H$:

$$Q_h = ZW_h^Q, \quad K_h = ZW_h^K, \quad V_h = ZW_h^V,$$

$$W_h^Q,\ W_h^K,\ W_h^V \in \mathbb{R}^{d_{\text{model}} \times d_k}.$$

$$A_h = \operatorname{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_k}}\right), \qquad U_h = A_h V_h \ \in\ \mathbb{R}^{d \times d_k}.$$

Heads are concatenated and linearly mapped:

$$\hat{Z} = \big[U_1 \mid \cdots \mid U_H\big] W^O \in \mathbb{R}^{d \times d_{\text{model}}}, \quad W^O \in \mathbb{R}^{(Hd_k) \times d_{\text{model}}}.$$

A pointwise nonlinearity is applied:

$$Y = \operatorname{ReLU}(\hat{Z}) \in \mathbb{R}^{d \times d_{\text{model}}},$$

then tokens are flattened and passed through a shallow MLP to produce the final latent:

$$\mathbf{y} = \operatorname{vec}(Y) \in \mathbb{R}^{d \cdot d_{\text{model}}}, \qquad \boldsymbol{\phi} = g(\mathbf{y}) \in \mathbb{R}^m,$$

where g is a shallow MLP (single hidden layer of size `out_dim`) and m is the latent dimension fed to the policy/value heads. In our implementation, attention can be enabled for the actor (and optionally the critic); batching follows the same shapes with an extra leading dimension B.

C. Integration with PPO

As shown in fig. 1, the Frame Attention module is implemented as a pre-processing layer before the policy and value networks in a Proximal Policy Optimization (PPO) agent. During training, it operates in an end-to-end differentiable manner. No architectural modifications are made to PPO's policy or value heads beyond replacing the raw input vector with the attended one.

This design ensures compatibility with standard RL pipelines and introduces minimal overhead.

D. Other Attention Variants Investigated

In addition to our proposed Frame Attention, we experimented with several alternative self-attention designs to assess their robustness under observation noise:

- **Feature-Attention**: Self-attention applied within each observation, learning correlations among raw features.
- **Selective-Attention**: Similar to Feature-Attention, but outputs are compressed through a bottleneck to force feature selection.
- **Hard-Gated Attention**: The attended value overwrites the original feature, forcing the network to replace noisy dimensions rather than reweight them.
- **Frame-Attention (ours)**: Treats each stacked frame as a token and applies attention across features within the framestack.

All variants replaced the default MLP in the policy and/or value networks. Ablation studies compared their performance under identical noise settings to determine which design preserved performance most effectively. Across environments and noise types, **Frame Attention consistently achieved the best balance of robustness and policy quality**, which motivated us to adopt it as the final architecture proposed in this work.

IV. Results

A. Experimental Setup

Our experimental evaluation is conducted on two standard continuous control benchmarks with vector observations: AntBullet and LunarLanderContinuous. All models were implemented using the PPO algorithm as provided in the Stable Baselines3 library. To ensure fair and reliable comparisons, we adopted the set of hyper-parameters recommended by the official Stable Baselines3 [4] benchmarks for each environment.

We structured our experiments in a curriculum, beginning with a noise-free setting to establish upper-bound performance for both vanilla PPO and our attention-based variants. First, we trained baseline PPO agents without any observation noise. Next, we introduced frame stacking as a form of temporal augmentation, experimenting with different stack sizes (ranging from 1 to 6 frames) to quantify its effect on policy performance and stability. Then we decided to use stack size of four, this configuration was used for all subsequent experiments. After establishing the optimal frame-stacking configuration, we systematically introduced noise into the observation space to evaluate policy robustness. For each noise experiment, both the vanilla PPO and Frame Attention architectures were trained under identical conditions for direct comparison. Every experiment was repeated with three random seeds to account for variance due to initialization.

B. Noise Injection and Ablation Protocol

To study the effectiveness of attention in filtering out irrelevant or misleading information, we injected three distinct types of noise into the observation vectors:

- **Ramp Noise**: Additive noise that increases linearly during each episode (with a slope of 0.001 per step) for 1000 steps, and resets to zero at the start of each new episode.

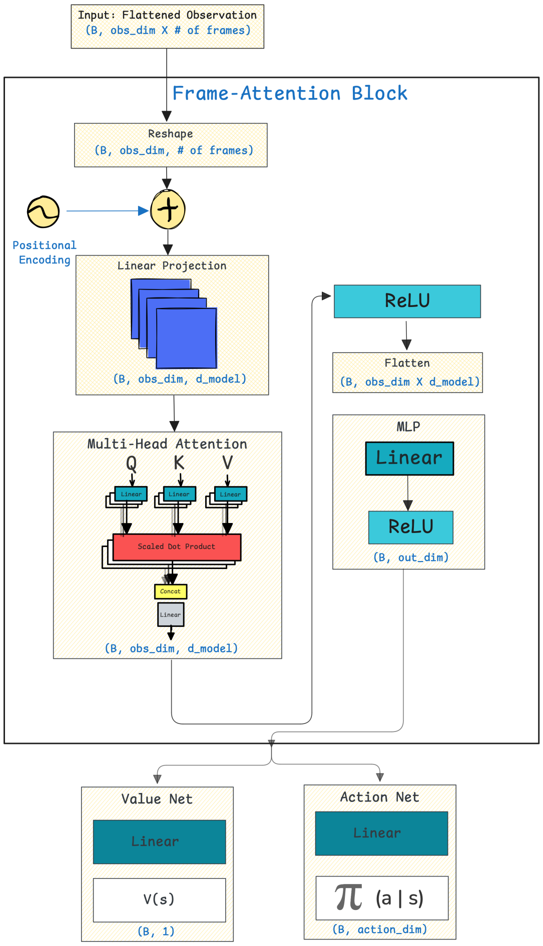

Fig. 1. Proposed Frame Attention architecture: stacked observations are reshaped into feature tokens, encoded with positional information, processed with multi-head self-attention across features, flattened, and passed through a shallow MLP before feeding to the policy and value heads.

- **Uniform Noise**: For each feature, noise is sampled independently from a uniform distribution in the interval [-10, 10].
- **Gaussian Noise**: For each feature, noise is sampled independently from a normal distribution with zero mean and unit standard deviation.

To investigate how noise intensity and scope affect learning, we performed ablations on the number of noisy dimensions. For each noise type, we ran experiments where 10, 20, 40 randomly chosen observation dimensions were corrupted. This allowed us to assess both moderate and extreme scenarios of observation degradation.

C. Training and Evaluation Procedure

Each model was trained for a fixed budget of environment steps (e.g., up to 4 million timesteps for AntBullet, 1 million for LunarLanderContinuous). Episode reward (mean over three seeds) was tracked throughout training to assess learning speed, final policy performance, and robustness to noise. All runs used identical random seeds and the same set of hyperparameters for PPO to isolate the effect of architectural modifications. Following this protocol, we compared vanilla PPO, PPO with frame stacking only, and PPO with both frame stacking and our Frame Attention mechanism under each noise scenario and ablation setting.

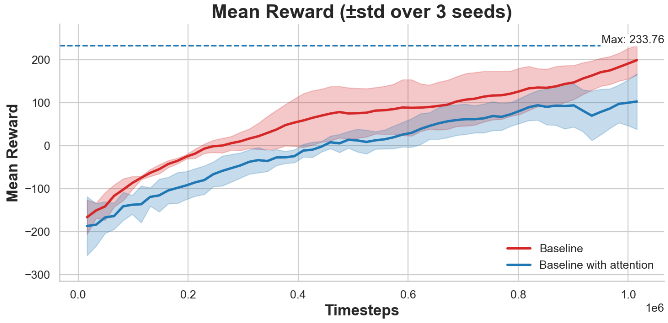

Fig. 2. The red curve represents the baseline performance of the default PPO agent trained with benchmark hyperparameters for **LunarLanderContinuous** environment. The blue curve shows the performance of Frame Attention under the same conditions.

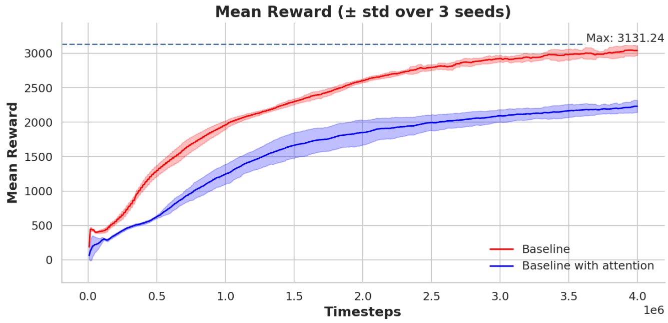

Fig. 3. The red curve represents the baseline performance of the default PPO agent trained with benchmark hyperparameters for **AntBullet** environment. The blue curve shows the performance of Frame Attention under the same conditions.

D. Evaluation Results

As shown in fig. 2 and fig. 3, under clean observation settings, PPO achieves slightly higher rewards compared to the attention-based policy. This performance gap is observed when no noise is present in the environment.

The results are summarized in fig. 4, which presents a comprehensive comparison of PPO and PPO with Frame Attention under different types and intensities of observation noise. Each subplot shows the mean episodic reward (averaged over three seeds) as a function of training timesteps for AntBullet (left) and LunarLanderContinuous (right). Results are reported for ramp noise (top), uniform noise (middle), and Gaussian noise (bottom). The orange curves correspond to the default PPO baseline, while the purple curves represent PPO augmented with Frame Attention. The effect of increasing the number of noisy observation dimensions (10, 20, 40, 100) is indicated by different line styles. Across all noise types and both environments, a clear trend emerges: as the number of

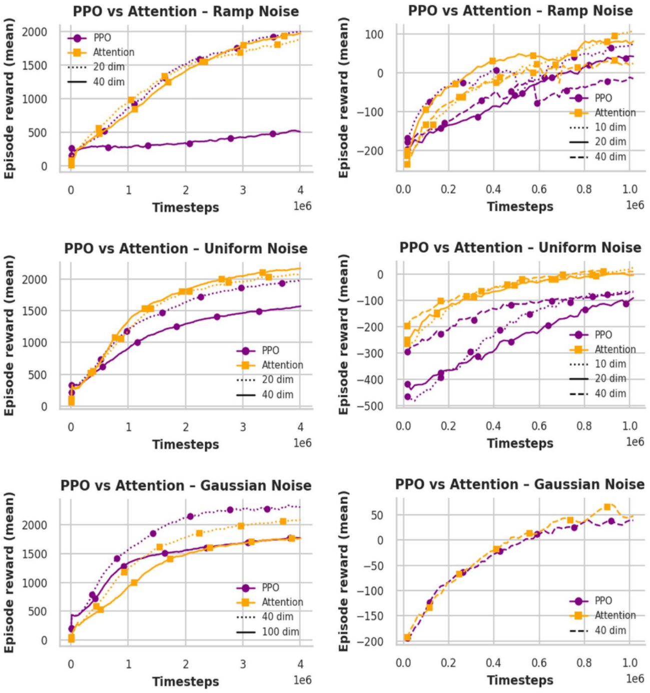

Fig. 4. **Performance comparison of PPO and PPO with Frame Attention under various observation noise conditions.** Each subplot shows the mean episodic reward (averaged over 3 seeds) as a function of training timesteps for AntBullet (left column) and LunarLanderContinuous (right column) environments, under three types of injected noise: ramp (top), uniform (middle), and Gaussian (bottom). The orange curves represent the experiments that use default PPO baseline, while the purple curves show the performance of Frame Attention, using an identical training setup. Dotted and dashed lines correspond to different numbers of noisy dimensions in the observation vector, reflecting increasing levels of input corruption (10, 20, 40, 100).

noisy dimensions increases, the performance of both PPO and Frame Attention drops gradually, confirming that higher levels of observation corruption degrade policy learning. Notably, Frame Attention consistently achieves higher or comparable reward to the baseline PPO under moderate to severe noise, except Gaussian noise with 40 dimensions .

A particularly interesting phenomenon occurs with ramp noise in AntBullet. Although the maximum reward curves for PPO and Frame Attention are similar, qualitative evaluation by rendering the trained policies reveals a critical difference: the baseline PPO agent, despite achieving high episodic reward, fails to produce meaningful movement and remains largely stationary, likely exploiting the reward structure without learning the intended walking behavior. In contrast, the Frame Attention agent demonstrates effective and robust walking under the same conditions, indicating genuinely improved policy robustness and task performance.

For LunarLanderContinuous, Frame Attention also outperforms PPO in terms of mean episodic reward under all noise types. However, video renderings suggest that the behavioral difference between the two agents is less pronounced, with Frame Attention showing only slight qualitative improvement over PPO.

Overall, these results indicate that attention-based architectures not only improve quantitative robustness to observation noise but, especially in the case of AntBullet, can lead to qualitatively better and more robust policies, even when standard reward curves alone might obscure meaningful differences in agent behavior.

V. Conclusion

As shown in fig. 2 and fig. 3, PPO slightly outperforms the attention-based architecture under clean (noise-free) observations. This can be attributed to the additional complexity introduced by the attention mechanism, combined with the use of baseline PPO hyperparameters without further tuning. Such performance drops in ideal settings are not uncommon, as architectural changes often require tailored optimization to reach their full potential.

In this project, we systematically investigated the performance and robustness of a novel Frame Attention module integrated within PPO. Our evaluation covered a range of observation noise types and intensities across two continuous control benchmarks. We compared the proposed attention-augmented agent to a strong PPO baseline under identical training protocols, incorporating frame stacking and varying the number of noisy input dimensions.

Our experiments show that the Frame Attention architecture consistently improves robustness and policy performance under noisy observation conditions. Notably, in the AntBullet environment, Frame Attention led to qualitatively superior behaviors, even when reward curves appeared similar. These findings suggest that attention mechanisms can provide substantial benefits for RL agents operating in realistic, noisy settings.

However, further improvements could likely be achieved through targeted hyperparameter tuning and more extensive architecture search. Future work should explore optimizing learning rates, attention-specific parameters, and regularization schemes to enhance both the performance and generalization of attention-augmented policies across a broader range of environments.

References

[1] Y. Tang, D. Nguyen, and D. Ha, “Neuroevolution of self-interpretable agents,” in *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, ser. GECCO ’20. ACM, Jun. 2020, p. 414–424. [Online]. Available: http://dx.doi.org/10.1145/3377930.3389847

[2] Y. Chen, C. Dong, P. Palanisamy, P. Mudalige, K. Muelling, and J. M. Dolan, “Attention-based hierarchical deep reinforcement learning for lane change behaviors in autonomous driving,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE Press, 2019, p. 3697–3703. [Online]. Available: https://doi.org/10.1109/IROS40897.2019.8968565

[3] H. Itaya, T. Hirakawa, T. Yamashita, H. Fujiyoshi, and K. Sugiura, “Mask-attention a3c: Visual explanation of action–state value in deep reinforcement learning,” *IEEE Access*, vol. 12, pp. 86 553–86 571, 2024.

[4] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, “Stable-baselines3: Reliable reinforcement learning implementations,” *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1364.html