Reinforcing Good Experiences: AWAC–DreamerV3 with Confidence-Aware Policy Gating

Group: B559 Advisor: 吳尚鴻 教授
Member: 黃祥陞 (Huang, Hsiang-Sheng)、江承瀚 (Chiang, Cheng-Han)

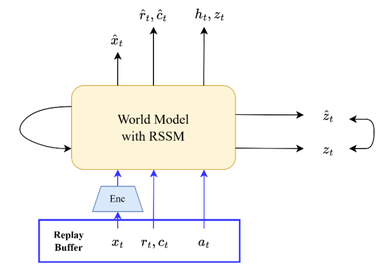
Abstract

We aim to collect and reinforce good experiences by integrating Advantage-Weighted Actor-Critic[1] (AWAC) into DreamerV3[2]. **Our approach adds a second policy network trained using AWAC to emphasize high-value transitions from replay data.** Because DreamerV3 relies heavily on imagined rollouts that can drift when the learned dynamics are imperfect, the AWAC head serves as a grounding mechanism that amplifies reliable behavior directly from real replay. To intelligently choose between the original Dreamer policy and the AWAC policy, **we introduce a gating mechanism that evaluates policy confidence through normalized entropy and measures distributional agreement via symmetric KL divergence.**

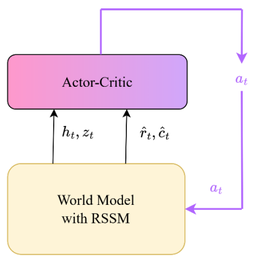
The gate dynamically selects the more confident policy while being conservative when policies disagree significantly. This dual-policy architecture with intelligent gating enables the system to effectively leverage both the robust world-model learning of DreamerV3 and the experience-focused learning of AWAC. Our method provides formal analysis of gate properties including boundedness and monotonicity. Experiments on DMC[3] proprioceptive control and Crafter[4] demonstrate stable gating behavior and competitive sample efficiency, showing effective reinforcement of good experiences while maintaining robust performance.

Introduction

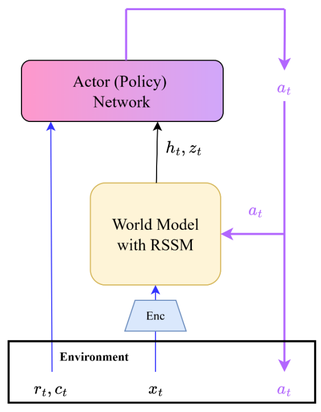
World Model DreamerV3



World Model: RSSM component learning
Off Policy, Training with replay Buffer



World Model: Actor-Critic component learning
Training on imagined Trajectories (On model, Off-env)

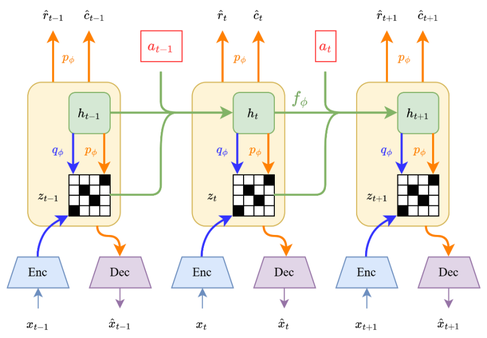


DreamerV3 Inference Flow

DreamerV3's RSSM

$$\text{RSSM} \begin{cases} h_t = f_\phi(h_{t-1}, z_{t-1}, a_{t-1}) & (Sequence) \\ z_t \sim q_\phi(z_t|h_t, x_t) & (Encoder) \\ \hat{z}_t \sim p_\phi(\hat{z}_t|h_t) & (Dynamics) \end{cases}$$

$$\begin{aligned} \hat{r}_t &\sim p_\phi(\hat{r}_t|h_t, z_t) && (Reward) \\ \hat{c}_t &\sim p_\phi(\hat{c}_t|h_t, z_t) && (Continue) \\ \hat{x}_t &\sim p_\phi(\hat{x}_t|h_t, z_t) && (Decoder) \end{aligned}$$



DreamerV3's Actor-Critic

Actor Network Learning:

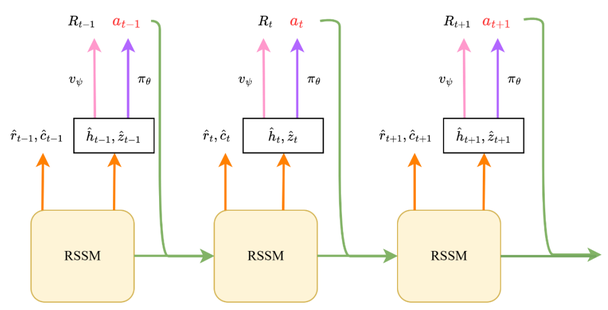
$$\max_\theta E_{(s,a)\sim\mathbb{D}} [\log \pi_\theta (a|s)] \quad (1)$$

Critic Network Learning:

$$\min_\theta \mathrm{L}_{(s,a,r,s')\sim\mathbb{D}} \quad (2)$$

$$\mathrm{L}_{(s,a,r,s')\sim\mathbb{D}} = E_{(s,a,r,s')\sim\mathbb{D}} \left[\left(v_\psi(s,a) - R_t^{0.95} \right)^2 \right] \quad (3)$$

$$R_t^{0.95} = r_t + \gamma c_t \left(0.05\, v_\phi(s_t) + 0.95\, R_{t+1}^{0.95} \right) \quad (4)$$



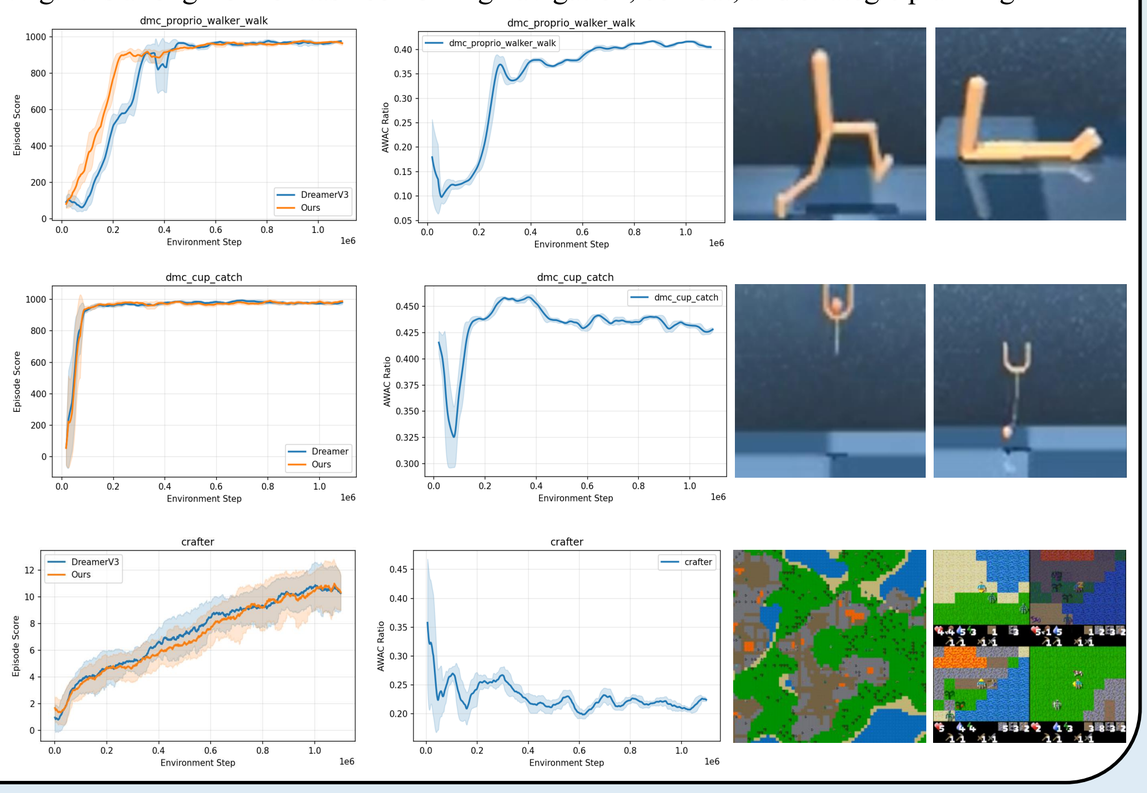
Results

We've test multiple experiments on 3 different tasks: *DeepMind Control Suite(DMC): proprio walker walk*, *DeepMind Control Suite(DMC): catch cup* and *Crafter*.

DMC: proprio walker walk is a locomotion task where an agent controls a bipedal robot using proprioceptive inputs. It is a challenge focused on maintaining balance and producing stable forward walking.

DMC: catch cup is a continuous-control task where an agent moves a cup to catch a falling ball. It is a test of precise tracking and fast reactive motion.

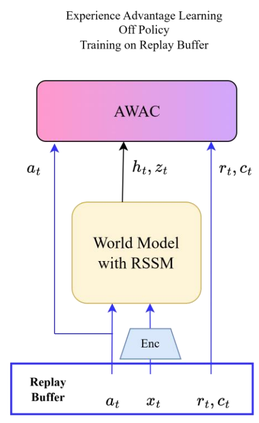
Crafter is a survival-style environment requiring exploration, resource gathering, and tool crafting. It is a long-horizon task combining navigation, combat, and strategic planning.



Methodology

A. Advantaged Weighted Actor-Critic (AWAC)

AWAC[1] is first used for solving the on policy and off policy learning gap. AWAC uses a dataset that collects all experience. We modified the dataset as the training buffer, using catastrophic forgetting to make model learn new good experiences.



The following is the AWAC's actor learning rule:

$$\max_\theta E_{(s,a)\sim\mathbb{D}} [w(s,a) \log \pi_\theta (a|s)] \quad (5)$$

$$w(s,a) = \exp\left(\frac{\max\left(A^\pi \mathrm{slow}(s,a), 0\right)}{\beta} \right) \quad (6)$$

B. Confidence-Based Policy-Gating Method

Now, we have 2 policy nets: DreamerV3's[2] Actor-Critic Network, learning from RSSM's imagination and Experience-Based AWAC Network.

We developed an action transition selection method that based on the confidence and the policy disagreement, each of them using Normalized (Bounded) Entropy (Shannon Entropy for discrete task, Differential Entropy for continuous task) and mean Symmetric KL divergence.

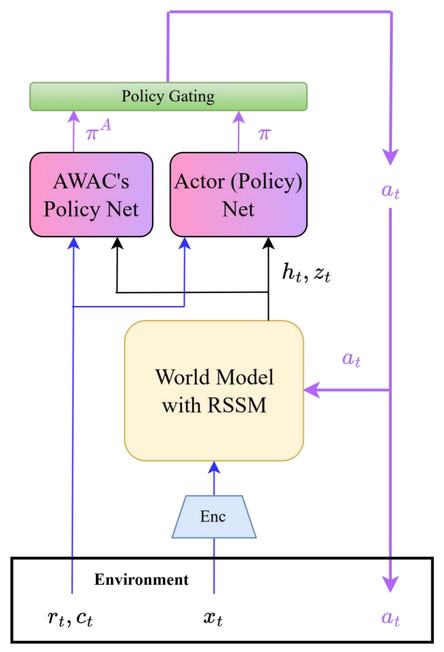
To prevent the scale difference of mean symmetric KL-divergence and Normalized Entropy, we designed a clipped hyperbolic tangent function for mean symmetric KL-divergence, and give it a β_{div} weight.

Finally, after calculating p_{gate} , we use a Bernoulli sampling to select the policy from AWAC and DreamerV3's Actor Network.

Algorithm 1 AWAC Gate at Policy Time

Input: features x_t, policies $\pi(\cdot \mid x_t)$, $\pi^A(\cdot \mid x_t)$
Sample $a \sim \pi$, $a^A \sim \pi^A$;
for each head k **do**
  Compute $\hat{H}_k(\pi_k)$ and $\hat{H}_k(\pi_k^A)$ using bounded entropy.
  Compute $\mathcal{D}_k = \frac{1}{2}\big(\mathrm{KL}(\pi_k \| \pi_k^A) + \mathrm{KL}(\pi_k^A \| \pi_k)\big)$.
end for
$\Delta\hat{H} \leftarrow \mathrm{mean}_k\big(\hat{H}_k(\pi_k) - \hat{H}_k(\pi_k^A)\big)$; $\mathcal{D} \leftarrow \mathrm{mean}_k(\mathcal{D}_k)$.
penalty $\leftarrow \tanh\big(\max(\mathcal{D} - \tau, 0)/s\big)$.
score $\leftarrow \Delta\hat{H} - \beta_{div} \cdot$ penalty $-$ bias.
$p^*_{gate} \leftarrow \sigma(\alpha \cdot \text{score})$; $p_{gate} \leftarrow \mathrm{clip}(p_{gate}, p_{\min}, p_{\max})$.
$m \sim \mathrm{Bernoulli}(p^*_{gate})$ independently per environment.
action $\leftarrow \mathrm{where}(m, a^A, a)$.
Return action and log metrics $\{p_{gate}, m, \Delta\hat{H}, \mathcal{D}\}$.

Policy Gating		
α	5.0	Logistic slope for sigmoid transformation
bias	0.05	Gate bias (shifts toward Dreamer when positive)
τ	0.0	Divergence tolerance threshold
s	150.0	Divergence scale for tanh saturation
β_{div}	0.3	Disagreement weight in gate computation
$p_{\min}$	0.0	Minimum gate probability
$p_{\max}$	0.7	Maximum gate probability



Conclusion

We propose an intelligent gating mechanism that integrates AWAC[1] with DreamerV3[2] by using entropy-based confidence and KL divergence to dynamically choose between policies. Across DMC[3] proprioceptive control, DMC cup-catch, and Crafter[4] tasks, the method shows domain-dependent strengths: improved sample efficiency on walker-walk, Dreamer-level performance on cup-catch, and stable, conservative mixing in exploration-heavy Crafter.

The core finding is that the gate autonomously determines when advantage-weighted learning is beneficial and when the world model should dominate, reducing the need for manual tuning and improving robustness across tasks.

Our future work may extend this adaptive selection framework to other policy combinations and explore richer confidence measures to further test its generality.

Reference

[1] *Nair, Ashvin, Murtaza Dalal, Abhishek Gupta and Sergey Levine. "Accelerating Online Reinforcement Learning with Offline Datasets." ArXiv abs/2006.09359 (2020): n. pag.*
[2] *Hafner, Danijar, J. Pašukonis, Jimmy Ba and Timothy P. Lillicrap. "Mastering diverse control tasks through world models." Nature 640 (2025): 647 - 653.*
[3] *Tassa, Yuval, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy P. Lillicrap and Martin A. Riedmiller. "DeepMind Control Suite." ArXiv abs/1801.00690 (2018): n. pag.*
[4] Hafner, Danijar. 2022. "Benchmarking the Spectrum of Agent Capabilities." Paper presented at the *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2109.06780
[5] *Henrich, Allison, Noel MacNaughton, Sneha Narayan, Oliver Pechenik, Robert Silversmith and Jennifer Townsend. "A Midsummer Knot's Dream." The College Mathematics Journal 42 (2010): 126 - 134.*
[6] Astley, Rick. 1987. *Never Gonna Give You Up*. Single. EMI Records.