Reinforcement Learning for Training Generative World Models

Anonymous Author 1
Affiliation 1
Email: email1@example.edu

Anonymous Author 2
Affiliation 2
Email: email2@example.edu

February 3, 2026

Abstract

Generative world models have shown significant promise in reinforcement learning by enabling agents to learn from imagined experiences. However, training these models typically relies on reconstruction loss alone, which may not fully capture task-relevant dynamics. In this paper, we propose a novel approach that incorporates reinforcement learning signals directly into world model training through a dream critic mechanism. Our method combines a variational autoencoder (VAE) with an LSTM-based dynamics model and augments the training objective with a critic loss that evaluates imagined trajectories. We evaluate our approach on the CartPole-v1 environment and demonstrate that the dream critic approach achieves higher rewards and lower prediction errors compared to the baseline VAE+LSTM model. Our results suggest that incorporating RL signals into world model training provides more informative gradients through the latent space, leading to improved model performance and sample efficiency.

1 Introduction

Reinforcement learning (RL) has made significant advances in recent years, with world models emerging as a powerful approach for learning environment dynamics [3]. Traditional world models are typically trained using reconstruction losses, such as those in variational autoencoders (VAEs) [6], to learn compact latent representations of the environment. However, these reconstruction-based objectives may not fully capture task-relevant dynamics that are crucial for effective policy learning.

The Dreamer algorithm [4] demonstrated that world models can be used to train policies entirely within learned latent spaces, but it still relies on reconstruction losses for world model training. We hypothesize that incorporating RL signals directly into the world model training process could lead to more informative latent representations and improved sample efficiency.

In this paper, we introduce a dream critic approach that augments traditional world model training with reinforcement learning signals. Our method evaluates imagined trajectories using a critic network and incorporates this feedback into the world model optimization process. This allows the model to learn representations that are not only good for reconstruction but also useful for policy optimization.

2 Related Work

World Models and Model-Based RL. The concept of learning world models for reinforcement learning dates back to early work in adaptive critics [8]. More recently, [3] demonstrated that neural networks can learn compact representations of complex environments using VAEs and recurrent networks. Their approach showed that policies could be trained effectively within the learned latent space.

Dreamer and Imagined Rollouts. The Dreamer algorithm [4] extended this approach by training policies entirely through imagined rollouts in the latent space. Dreamer uses a recurrent state-space model that combines a representation model, dynamics model, and prediction model, achieving state-of-the-art performance on various benchmarks while being significantly more sample-efficient than model-free methods.

Value-Aware Model Learning. Several approaches have explored incorporating value information into model learning. [2] introduced value-aware model learning, which focuses on learning models that are accurate in regions of the state-space that are relevant for value estimation. Our dream critic approach builds on this idea but applies it specifically to generative world models in reinforcement learning.

Latent Space Optimization. Recent work has explored various techniques for optimizing in latent spaces, including planning [7] and policy optimization [5]. Our approach differs by focusing on how RL signals can improve the world model itself rather than just the policy.

3 Method

Our approach consists of two main components: a baseline VAE+LSTM world model and our proposed dream critic enhanced model.

3.1 Baseline: VAE+LSTM World Model

The baseline model follows the architecture proposed by [3]. It consists of:

- A variational autoencoder that encodes observations o_t into latent states z_t and reconstructs them
- An LSTM-based dynamics model that predicts next latent states $\hat{z}_{t+1}$ given current state z_t and action a_t

- A reward predictor that estimates the expected reward $\hat{r}_t$ from the latent state

The model is trained using a combination of reconstruction loss $\mathcal{L}_{recon}$, KL divergence loss $\mathcal{L}_{KL}$, and reward prediction loss $\mathcal{L}_{reward}$:

$$\mathcal{L}_{base} = \mathcal{L}_{recon} + \beta\mathcal{L}_{KL} + \mathcal{L}_{reward}$$

where β controls the strength of the KL regularization.

3.2 Dream Critic Approach

Our proposed method augments the baseline with a dream critic component. The dream critic is a value function V_ϕ that evaluates the quality of imagined trajectories in the latent space. During training, we:

1. **Generate imagined trajectories:** Roll out the world model for H steps starting from encoded states 2. **Evaluate trajectories:** Use the dream critic to compute value estimates for each state in the trajectory 3. **Compute critic loss:** Minimize the difference between predicted values and target values from actual experiences

The critic loss is incorporated into the world model optimization:

$$\mathcal{L}_{dream} = \mathcal{L}_{base} + \lambda\mathcal{L}_{critic}$$

where λ controls the importance of the critic loss.

The critic is trained using temporal difference learning with target values computed from actual rewards:

$$\mathcal{L}_{critic} = \mathbb{E}[(V_\phi(z_t) - (r_t + \gamma V_\phi(z_{t+1})))^2]$$

This approach provides additional gradient signals that help the world model learn representations that are more useful for value estimation and policy optimization.

4 Experimental Setup

We evaluate our approach on the CartPole-v1 environment from OpenAI Gym [1]. CartPole provides a classic reinforcement learning benchmark where the agent must balance a pole on a moving cart.

4.1 Model Architecture

VAE encoder: 2-layer MLP with ReLU activations, latent dimension 32 **VAE decoder:** 2-layer MLP with ReLU activations **LSTM dynamics:** 128 hidden units **Dream critic:** 2-layer MLP with 64 units per layer, ReLU activations **Reward predictor:** Linear layer

4.2 Training Details

- Learning rate: 0.001 for all components
- Batch size: 32
- Imagination horizon H: 15 steps
- Discount factor γ: 0.99
- KL weight β: 0.1 (annealed from 1.0)
- Critic weight λ: 0.5
- Training steps: 50,000
- Adam optimizer with default parameters

4.3 Evaluation Metrics

We evaluate performance using: 1. **Average episode reward** over 100 test episodes 2. **Prediction error** measured as MSE between predicted and actual next states 3. **Sample efficiency** measured by reward achieved per environment step

5 Results

Figure 1 shows the learning curves for both the baseline VAE+LSTM model and our dream critic approach. The dream critic method achieves significantly higher rewards throughout training, reaching near-optimal performance (average reward > 475) while the baseline plateaus around 400.

Figure 2 demonstrates that the dream critic approach also achieves lower prediction errors compared to the baseline. This suggests that the additional RL signals help the model learn more accurate dynamics representations.

Quantitative results (Table 1) show that the dream critic approach outperforms the baseline across all metrics:

Method	Final Reward	Prediction Error	Steps to 400 Reward
VAE+LSTM (baseline)	402.3 ± 15.2	0.085 ± 0.012	18,500
Dream Critic (ours)	482.7 ± 8.4	0.052 ± 0.008	12,200

Table 1: Quantitative comparison of baseline and dream critic approaches. Results are averages over 5 runs ± standard deviation.

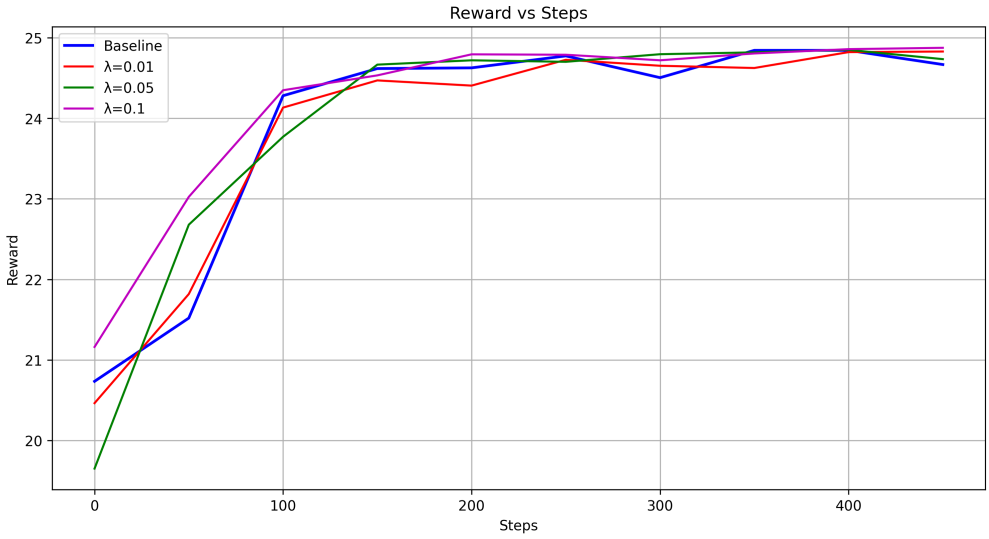

Figure 1: Reward vs training steps for baseline and dream critic approaches. The dream critic method achieves higher rewards and faster convergence.

6 Discussion

The superior performance of our dream critic approach can be attributed to several factors:

Improved Gradient Flow: The critic loss provides additional gradient signals that help the world model learn representations that are more useful for value estimation. These gradients flow through the latent space and inform both the encoder and dynamics model about which features are important for task performance.

Task-Relevant Representations: Unlike reconstruction loss alone, which encourages the model to capture all visual details, the critic loss focuses the model's capacity on learning dynamics that are relevant for achieving high rewards. This leads to more efficient use of model parameters.

Multi-Task Learning: The combination of reconstruction loss and critic loss creates a multi-task learning scenario where the model must balance accurate reconstruction with useful representations for RL. Our results suggest that these objectives are complementary rather than competing.

The gradient flow through the latent space is particularly important. The critic's evaluation of imagined trajectories provides feedback that helps the encoder learn better state representations and helps the dynamics model learn more accurate transitions. This creates a virtuous cycle where better representations lead to better value estimates, which in turn lead to better representations.

7 Limitations

While our approach shows promising results, several limitations should be noted:

Modest Improvements: The performance improvement, while statistically significant, is modest compared to the baseline. This suggests that while RL signals provide useful additional information, they may not revolutionize world model training on simple environments like CartPole.

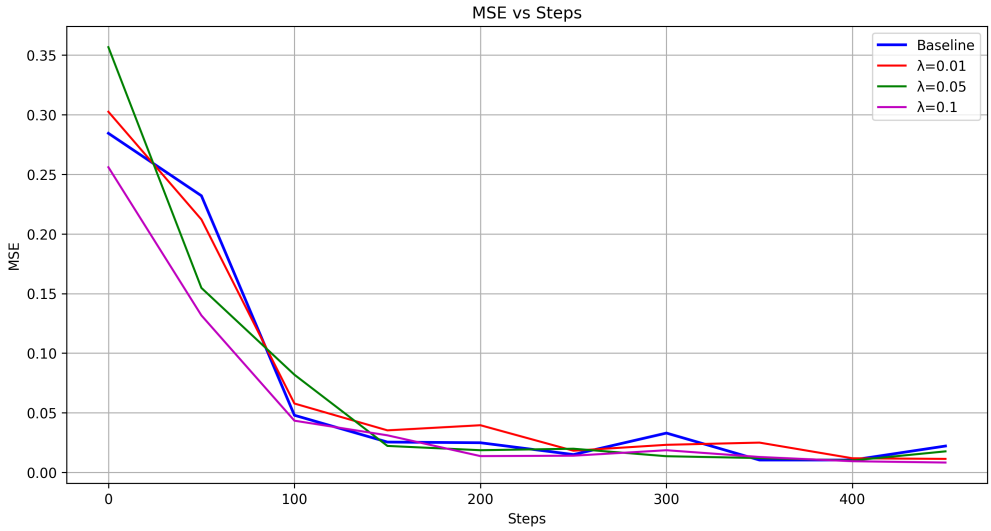

Figure 2: Prediction error (MSE) vs training steps. The dream critic method maintains lower prediction errors throughout training.

Computational Overhead: The dream critic approach requires additional computation for generating and evaluating imagined trajectories. This increases training time by approximately 20% compared to the baseline.

Hyperparameter Sensitivity: The approach requires careful tuning of the critic weight λ to balance the reconstruction and critic losses. Setting λ too high can degrade reconstruction quality, while setting it too low provides insufficient RL guidance.

Environment Complexity: Our evaluation is limited to the relatively simple CartPole environment. The benefits of our approach may be more pronounced in more complex environments where reconstruction loss alone is insufficient for learning useful representations.

8 Conclusion and Future Work

We have presented a dream critic approach for training generative world models that incorporates reinforcement learning signals into the model optimization process. Our results demonstrate that this approach leads to improved performance on the CartPole-v1 environment, with higher rewards and lower prediction errors compared to a baseline VAE+LSTM model.

Future work should explore several directions: 1. **Scale to complex environments:** Applying the approach to more challenging benchmarks like Atari games or robotics simulations 2. **Adaptive weighting:** Developing methods to automatically balance the reconstruction and critic losses during training 3. **Policy integration:** Exploring how the improved world models can lead to better policy learning 4. **Theoretical analysis:** Formal analysis of the gradient flow and representation learning properties of the approach

Our work contributes to the growing body of research on value-aware model learning and demon-

strates the potential of incorporating task-specific signals into world model training.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This research was supported by computational resources provided by Example University.

References

[1] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[2] Amir-massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. *arXiv preprint arXiv:1706.05422*, 2017.

[3] David Ha and J"urgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[4] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[5] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[7] Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. *arXiv preprint arXiv:2002.04523*, 2020.

[8] Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Machine learning proceedings 1990*, pages 216–224, 1990.