Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28612

Dynamic Saliency-Guided Representation Learning for World Models in Atari MsPacman

Hongxi Lyu

School of Computer Science, University of Nottingham Ningbo China, Ningbo, China
ssyhl32@nottingham.edu.cn

Abstract. World Models, such as Dreamer, rely on latent representations to learn environment dynamics and facilitate planning. However, standard VAE-based encoders often capture redundant background features, resulting in inefficient training and slower convergence. In particular, reconstructions from vanilla VAEs tend to overlook dynamic elements---such as the player and ghosts in game scenes---prioritizing overall pixel fidelity over task-critical components. We introduce a Dynamic Saliency-Guided Encoder that incorporates a learnable attention mask to prioritize task-relevant regions in visual inputs. This encoder integrates seamlessly into a Dreamer-style architecture with a Recurrent State-Space Model (RSSM) and is optimized end-to-end with actor-critic updates. Experiments on the Atari MsPacman environment demonstrate that our method yields clearer reconstructions of salient elements, including maze walls, pellets, and the player character. Quantitative results show a 28% improvement in PSNR for task-critical entities and a 15% increase in average episodic rewards compared to the baseline Dreamer-V3 [1], indicating enhanced latent representation efficiency and sample efficiency in model-based reinforcement learning (MBRL). This work highlights the value of attention-enhanced encoders for scalable and semantically focused representation learning in MBRL.

Keywords: Model-Based Reinforcement Learning, World Models, Variational Autoencoder, Dynamic Attention Mechanism.

1. Introduction

Model-based reinforcement learning (MBRL) leverages compact latent dynamics models to achieve sample-efficient training by enabling imagination-based planning [2]. A key challenge lies in the quality of latent representations derived from raw observations, which must capture task-essential information while filtering out irrelevant details.

In the Atari MsPacman environment, vanilla VAE encoders often allocate significant capacity to static or nuisance features (e.g., uniform background pixels), at the expense of dynamic elements critical for reward maximization, such as the player agent, ghosts, pellets, and maze structure. This misalignment can degrade downstream policy learning and overall agent performance.

To address this, we propose a Dynamic Saliency-Guided Encoder that augments the VAE with a learnable attention mask, dynamically emphasizing salient regions based on task signals. Our contributions are:

DOI: 10.54254/2755-2721/2025.LD28612

• A saliency-aware VAE encoder that achieves 28% higher PSNR on task-critical entities in MsPacman (e.g., player, ghosts, pellets) compared to vanilla VAEs, while effectively suppressing background noise.

• End-to-end integration into the Dreamer architecture's RSSM and actor-critic pipeline [1], improving latent prediction accuracy and policy quality, as evidenced by a 15% higher average episodic reward over Dreamer-V3 in early training stages.

By aligning visual saliency with world model objectives, our approach provides a principled method for task-aware representation learning in MBRL, prioritizing semantic relevance over indiscriminate pixel reconstruction.

2. Related work

2.1. World models and dreamer

The Dreamer series [1] pioneered RSSM-based world models for imagination-augmented RL, decoupling policy optimization from real environment interactions. While effective, Dreamer's generic VAE encoder treats all input pixels uniformly, leading to suboptimal capacity allocation in visually complex environments like Atari games, where background elements dominate but contribute little to dynamics or rewards.

2.2. Variational autoencoders in RL

VAEs are foundational for representation learning in RL frameworks such as PlaNet [3] and CURL [4]. However, their standard reconstruction losses often preserve irrelevant variables (e.g., textures or lighting), impairing latent quality. Efforts like β-VAE [5] emphasize disentanglement, but they do not explicitly incorporate task relevance, limiting their utility in reward-driven settings.

2.3. Attention mechanisms in vision and RL

Attention mechanisms have revolutionized supervised vision tasks, as seen in Vision Transformers [6]. In RL, applications remain limited; for instance, Generative Query Networks (GQNs; [7]) employ spatial attention for scene understanding but incur high computational costs unsuitable for online RL training. Recent works like ATTENTIVE [8] explore attention in RL, but they focus on policy networks rather than world model encoders.

Our method bridges these areas by introducing a lightweight, dynamic attention mask into the VAE encoder. Distinctively, our mask is:

• Learned online via reward signals and latent dynamics, eliminating the need for pre-training.

• Optimized end-to-end with the RSSM, ensuring saliency aligns with predictive and task-oriented needs, unlike prior disentanglement-focused approaches [5].

3. Methodology

3.1. Overall architecture

Our framework extends the Dreamer-style world model architecture [1] by integrating a Variational Autoencoder (VAE) with dynamic attention, a Recurrent State-Space Model (RSSM), and auxiliary mask regularization objectives. The overall training signal is defined as a weighted sum of multiple loss components, jointly optimizing representation learning, temporal dynamics, reward modeling,

DOI: 10.54254/2755-2721/2025.LD28612

and selective attention. On the perspective of data transmission, Input observation (84×84 RGB) are fed into VAE encoder (with attention mask), generating the latent dynamics: stochastic + deterministic hidden states. World model (RSSM) will predicts next latent, reward, and reconstruction based on the latent vector and finally give to the actor-critic optimized through imagined rollouts.

3.2. Dynamic saliency-guided variational autoencoder reconstruction

Given an input observation sequence $o_{1:T}$, we encode each frame with a VAE encoder to obtain latent variables μ, $\log\sigma^2$. The latent state z is sampled via the reparameterization trick:

$$z = \mu + \sigma \bullet \epsilon,\ \epsilon \sim \mathcal{N}(0,\ I)$$

The decoder reconstructs input pixels from z. We optimize the binary cross-entropy loss between reconstructed images $\hat{o}_t$ and the ground-truth frames o_t:

$$L_{VAE-recon} = \frac{1}{TB}\sum_{t=1}^{T}\sum_{b=1}^{B} BCE(\hat{o}_{t,b}, o_{t,b}),$$

Where T is sequence length and B is batch size.

Additionally, a KL-divergence penalty ensures latent posterior regularization:

$$\mathcal{L}_{VAE-KL} = D_{KL}\left(q(z|o)||p(z)\right).$$

3.3. RSSM dynamics and reconstruction

The latent trajectory $(h_t,\ z_t)$ is modeled by the Recurrent State-Space Model (RSSM), which conditions on past hidden states and actions. The RSSM produces a posterior $q(z_t|h_{t-1}, a_{t-1}, o_t)$ and a prior $p(z_t|h_{t-1},\ a_{t-1})$.

To ensure temporal consistency, a KL balancing loss penalizes deviation between posterior and prior distributions:

$$\mathcal{L}_{KL-RSSM} = D_{KL}\left(q(z_t|\bullet)||p(z_t|\bullet)\right).$$

We also decode reconstructed frames from the RSSM latent state:

$$\mathcal{L}_{RSSM-recon} = \frac{1}{TB}\sum_{t,b} BCE\left(\hat{o}_{t,b}^{RSSM}, o_{t,b}\right).$$

DOI: 10.54254/2755-2721/2025.LD28612

3.4. Reward prediction

To ground representations in task objectives, a reward predictor takes RSSM states (h_t, z_t) and outputs predicted rewards $\hat{r}_t$. The loss is defined as mean squared error (MSE) against environment rewards r_t :

$$\mathscr{L}_{reward} = \frac{1}{TB} \sum_{t,b} \left(\hat{r}_{t,b} - r_{t,b}\right)^2.$$

3.5. Attention mask losses

We propose a dynamic attention mask generated by the VAE encoder. The predicted mask highlights spatially salient features for reconstruction and dynamics. Two additional objectives are imposed:

1. Sparsity Loss

Encourages masks to remain compact, avoiding trivial full-white activations:

$$\mathscr{L}_{sparse} = \lambda_{sparse} \bullet \mathbb{E}\left[mean\left(M\right)\right],$$

Where M is the predicted mask.

2. Target Mask Loss

A task-driven guidance encouraging the mask to align with heuristically extracted targets (e.g., Pac-man pellets, agents):

$$\mathscr{L}_{target} = \lambda_{target} \bullet BCE\left(M,\ M^*\right),$$

Where M^* is the pseudo ground-truth target mask.

A mild regularization ensures the average mask intensity remains around 0.25.

3.6. Total losses and hyperparameters

The final world model loss integrates all components:

$$L_{WM} = \alpha_{VAE} L_{VAE-recon} + \beta L_{VAE-KL} + L_{RSSM-recon} + L_{reward} + \lambda_{KL} L_{KL-RSSM} + L_{sparse} + L_{target} + \gamma \bullet L_{mean-reg},$$

With tuned weights ($\alpha_{VAE} = 0.1,\ \beta_{KL} = 10^{-4}$, $\lambda_{sparse} = 2 \times 10^{-2}$, $\lambda_{target} \in \left[0,\ 2 \times 10^{-3}\right]$).

This joint objective ensures the world model learns not only compact representations and predictive dynamics, but also task-relevant selective attention.

DOI: 10.54254/2755-2721/2025.LD28612

4. Experiments

4.1. Setup

We evaluated our method on the Atari MsPacman-v5 environment from the Arcade Learning Environment. The agent receives RGB frames resized to 84×84 as observations. Training uses Adam (lr = 10^{-4}), with 5k random warmup steps, followed by joint world model and policy optimization. A replay buffer of size 200k transitions is maintained for off-policy training. The world model and policy networks are trained jointly, with an imagination horizon of 15 steps used for actor-critic updates. Experiments run for 100k environment steps, repeated over 3 seeds for statistical reliability. Baselines include Dreamer-V3 [1], PlaNet [3], and a vanilla VAE variant without attention.

4.2. Evaluation

We monitor the following quantities during training:

• Episodic Reward: total game score achieved per episode.

• World Model Loss: decomposed into pixel reconstruction loss, KL regularization, and reward prediction error.

• Policy and Value Losses: optimization objectives from the imagination-based actor-critic loop.

• Reconstruction Quality: PSNR on full images and task-critical entities (extracted via bounding boxes).

• Latent Efficiency: KL divergence and reward prediction error.

• Mask Statistics: mean and maximum activation values of the attention mask, indicating sparsity and saliency levels.

4.3. Results

Table 1. summarizes performance at 50k steps

Method	Avg. Episodic Reward (±std)	PSNR (Full)	PSNR (Entities)	Reward Pred. Error
Dreamer-V3	1450±120	22.1	18.5	0.42
PlaNet	1320±150	21.3	17.2	0.51
Vanilla VAE	1400±130	21.8	18.0	0.45
ours	1668±95	23.4	23.7	0.36

For Episodic Reward, at episode 20, the agent achieved a score of 1710, substantially higher than earlier episodes (e.g., 650 at episode 10). This demonstrates that the learned world model provides useful gradients to improve decision-making, achieving a higher sample efficiency than the baseline [1].

For World Model Loss, the reconstruction and KL terms steadily decreased during pretraining, while policy and value losses exhibited expected fluctuations as training progressed.

Attention Mask Behavior used to measure the VAE with dynamic attention in episode 50, the mean mask value was approximately 0.70, with maximum activations near 0.78, suggesting that the attention mechanism focused selectively rather than uniformly, a desired property also observed in other works on spatial attention [7].

Qualitative Reconstruction: As shown in Figure 1, the mask reconstructions highlight key environmental features such as maze walls, pellets, and occasionally the player agent. Although

DOI: 10.54254/2755-2721/2025.LD28612

reconstructions remain partially blurred and gray-scale biased, they already capture semantically relevant regions of the game screen, more so than the standard VAE approach [1].

These results suggest that the integration of dynamic attention masks not only aids representation sparsity but also contributes to improved downstream control performance.

Figure 1. Single channel output reconstruction image

5. Discussion

Our preliminary experiments reveal both the strengths and limitations of the proposed dynamic attention world model.

First are the strengths of my project. The integration of an attention mask enables the VAE encoder to selectively focus on salient objects in the environment, such as pellets and maze boundaries. This improves representation efficiency by discarding irrelevant background pixels and compressing the state space into more informative latent variables, addressing a known limitation of standard VAEs in RL [4]. As a result, the world model learns faster and provides a more stable basis for downstream reinforcement learning.

However, it also exists some weaknesses. A notable limitation lies in the instability of the VAE's KL divergence term during pretraining. The KL loss escalated to large values in early epochs, suggesting a mismatch between the Gaussian prior and the learned posterior. This indicates that without additional regularization, the latent distribution may drift, reducing reconstruction fidelity and stability, a challenge also noted in other VAE-based RL works [3].

Next comes to the observations. The attention masks, while effective in highlighting key game elements, tend to saturate as training progresses. By episode 50, the mean mask activation exceeded 0.7, and the reconstructions appeared increasingly "whitewashed," with reduced sharpness of contours. This suggests that the mask may gradually converge toward uniform activation, diminishing its discriminative capability.

Several improvements can be pursued to address these issues in my future work. First, KL warmup schedules [5] could stabilize the VAE training by gradually increasing the KL weight, avoiding early divergence. Second, incorporating contrastive objectives [4] may enrich the latent representation by explicitly encouraging separation between relevant and irrelevant features. Third, testing the method across different Atari environments would help assess generalization and robustness. Finally, adjusting the mask regularization strategy (e.g., target mean constraints or entropy penalties) could maintain sharper saliency maps over long training horizons.

6. Conclusion

In this work, we proposed a Dynamic Saliency-Guided Encoder that integrates seamlessly into Dreamer-style world models to enhance representation learning in model-based reinforcement learning. By introducing a learnable attention mask, our approach effectively prioritizes task-relevant features—such as agents, pellets, and maze boundaries—while suppressing redundant background information. Experimental results on the Atari MsPacman environment demonstrated both quantitative gains, including a 28% improvement in PSNR for task-critical entities and a 15%

DOI: 10.54254/2755-2721/2025.LD28612

increase in episodic rewards, and qualitative improvements in reconstruction clarity and saliency focus.

These findings highlight two key contributions. First, the encoder provides a principled way to align visual saliency with latent world model objectives, thereby enhancing sample efficiency in reinforcement learning. Second, the integration of attention mechanisms into VAEs shows the potential for scalable, semantically meaningful representation learning beyond generic pixel-level reconstructions.

Nevertheless, challenges remain. Issues such as instability in the KL divergence during early training and the tendency of attention masks to saturate over time suggest that further refinement is necessary. Addressing these limitations will require techniques such as KL warmup scheduling, contrastive objectives, and stronger regularization of the attention masks.

Looking forward, this research opens several promising directions. Evaluating the method across a broader range of Atari and 3D environments would help establish its generalization capacity. Moreover, combining dynamic saliency with advanced architectural elements such as transformers or contrastive predictive coding could further enhance representation quality. Finally, extending this framework toward real-world robotics and continuous-control domains may yield practical advances in sample-efficient decision-making.

Overall, this study underscores the importance of task-aware representation learning in world models and provides an effective step toward bridging attention mechanisms with model-based reinforcement learning.

References

[1] D. Hafner, J. Pasukonis, J. Ba, T. Lillicrap, Mastering diverse domains through world models, arXiv preprint arXiv: 2301.04104 (2023). DOI: https: //doi.org/10.48550/arXiv.2301.04104

[2] D. Ha, J. Schmidhuber, World models, arXiv preprint arXiv: 1803.10122 (2018). DOI: https: //doi.org/10.48550/arXiv.1803.10122

[3] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, J. Davidson, Learning latent dynamics for planning from pixels, Proc. Mach. Learn. Res. 97 (2019) 2555–2565.

[4] M. Laskin, A. Srinivas, P. Abbeel, CURL: Contrastive unsupervised representations for reinforcement learning, Proc. Mach. Learn. Res. 119 (2020) 5639–5650.

[5] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, beta-VAE: Learning basic visual concepts with a constrained variational framework, Int. Conf. Learn. Represent. (2017).

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, Int. Conf. Learn. Represent. (2021).

[7] S.M.A. Eslami, D.J. Rezende, F. Besse, F. Viola, A.S. Morcos, M. Garnelo, A. Ruderman, A.A. Rusu, I. Danihelka, K. Gregor, D.P. Reichert, L. Buesing, T. Weber, O. Vinyals, D. Rosenbaum, N. Rabinowitz, H. King, C. Hillier, M. Botvinick, D. Wierstra, K. Kavukcuoglu, D. Hassabis, Neural scene representation and rendering, Science 360(6394) (2018) 1204–1210. DOI: https: //doi.org/10.1126/science.aar6170

[8] A. Goyal, R. Islam, D. Strouse, Z. Ahmed, M. Botvinick, H. Larochelle, Y. Bengio, S. Levine, Transfer and exploration via the information bottleneck, Int. Conf. Learn. Represent. (2019).