Learning Important Subtle Signals in World Model Training

By

YUKUAN LU

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Electrical and Computer Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Yubei Chen, Chair

Anthony Thomas

Zhi Ding

Committee in Charge

2025

Contents

List of Figures

List of Tables

Abstract

In deep learning and reinforcement learning, world models are models that learn an inner representation of the exterior world. In particular, they learn to compress high-dimensional visual observations into a low-dimensional latent space, enabling efficient planning and decision-making. However, this compression often leads to the loss of important but subtle signals—features that are critical for downstream policy learning yet have low magnitude in certain statistical measures (e.g., pixel intensity, area) in their original perception space, i.e., the pixel space. In this work, we propose a regularization method to explicitly retain such important signals in world model training, preventing their loss during compression. Our method introduces an auxiliary loss based on a segmentation function $m(\cdot)$ that enforces the preservation of critical information in the latent space. Experiments on Atari Pong, with additional evaluations on Breakout and Ms. Pac-Man, demonstrate that our approach accelerates the learning of important signals, enhances their retention, and significantly improves policy learning performance, particularly in environments where these signals are subtle and difficult to learn.

Acknowledgments

I sincerely thank Prof. Yubei Chen for inspiring my research interests and guiding me through severe eye pain during my 18 months at UC Davis. His support shaped my academic path, research aesthetic, and methodological approach—focusing on solving specific problems.

CHAPTER 1

Introduction

In this thesis, we introduce a novel regularization method to preserve important subtle signals in world model learning. We show that the improvement is substantial, evident in video reconstruction quality, signal probing, and performance gains in downstream tasks. This section sets the stage by discussing the motivations, objectives of this research, our approaches, and the contributions of the thesis.

1.1. Motivation

World models learn compressed representations of the environment to enable planning and decision-making, playing an essential role in model-based reinforcement learning. While effective, these models often suffer from information loss, particularly when encoding important but subtle signals. In this thesis, we focus on vision-based world models that encode visual observations into low-dimensional latent spaces and perform temporal prediction within these compressed representations. During encoding, the model must capture visual features relevant for downstream tasks. However, not all visual signals contribute equally—some features are critical for decision-making while others may be irrelevant.

A key challenge arises when important visual signals are subtle—characterized by small spatial footprint, low pixel intensity variation, or short temporal persistence. While prominent visual features (large, high-contrast objects) are typically preserved during compression due to their statistical dominance, subtle signals face higher risk of being discarded or poorly represented in the latent space. This selective information loss can significantly degrade both predictive accuracy and downstream policy learning performance.

Consider Atari Pong: the ball represents the most critical visual element for gameplay, yet occupies minimal spatial footprint compared to paddles and boundaries. In noisy environments, the ball's statistical significance diminishes further, making it vulnerable to being filtered out during encoding.

When the world model fails to represent the ball adequately, the policy cannot learn effective strategies, resulting in poor performance.

Conversely, non-subtle signals—such as large, high-contrast paddles—are inherently easier to learn and retain due to their statistical prominence. While also task-relevant, their robustness to compression requires no additional preservation mechanisms. Therefore, this work specifically targets preserving important subtle signals, as they represent the primary bottleneck in world model learning.

1.2. Our Approach

We propose a regularization method that enforces the preservation of important subtle signals in world model training. Our method:

- Uses SAM2, a segmentation model, to define a function $m(\cdot)$ that identify important subtle signals in pixel space.
- Introduces a regularization loss that prevents the world model from discarding these signals.
- Evaluates our approach on Atari Pong, with additional experiments on Breakout and Ms. Pac-Man, confirming its generalizability.

Our results demonstrate that regularization accelerates signal learning, enhances retention, and improves policy performance, especially when signals are subtle and prone to being ignored.

1.3. Contributions

This work addresses the challenge of preserving important but subtle signals in world model training. Our key contributions are:

- Regularization for Subtle Signal Retention: We introduce a segmentation-based regularization method that explicitly prevents the loss of subtle yet critical signals during world model training.
- Segmentation-Guided Loss Function: We incorporate SAM2, a state-of-the-art segmentation model, to define a function $m(\cdot)$ that identifies important signals and enforces their preservation in the latent space.

- Comprehensive Evaluation: We validate our approach on Atari Pong, where the ball signal is challenging to retain, and extend our analysis to Breakout and Ms. Pac-Man to demonstrate generalizability.

1.4. Notation and Symbol Definitions

TABLE 1.1. Notation and Symbol Definitions

Symbol	**Type**	**Explanation**
l_{wm}	$\in \mathbb{R}$	Total world model loss
$l_{\text{mask predictor}}$	$\in \mathbb{R}$	Training loss of the mask predictor
$\text{MSE}(\cdot, \cdot)$	Function	Mean Squared Error function
$\mathbf{x}$	$\in \mathbb{R}^{h \times w \times c}$	Input image (observation) with height h, width w, and channels c
$\hat{\mathbf{x}}$	$\in \mathbb{R}^{h \times w \times c}$	Reconstructed image
l_{others}	$\in \mathbb{R}$	Other loss terms (e.g., reward prediction, sequence modeling)
ω, ω'	$\in \mathbb{R}$	Regularization weight
$m(\mathbf{x})$	$\in \{0, 1\}^{h \times w}$	Boolean segmentation mask
$m(\mathbf{x}) \cdot \mathbf{x}$	$\in \mathbb{R}^{h \times w \times c}$	Masked input image; the mask is broadcasted across channels c to match the shape of $\mathbf{x}$ during element-wise multiplication (as in PyTorch broadcasting)
$m(\mathbf{x}) \cdot \hat{\mathbf{x}}$	$\in \mathbb{R}^{h \times w \times c}$	Masked reconstructed image
t	$\in \mathbb{N}$	Training step index
z_t	$\in \mathbb{R}^d$	Discrete latent representation at time step t with hidden dimension d
h_t	$\in \mathbb{R}^d$	Recurrent state at time step t
a_t	$\in \mathbb{R}^a$	Action taken at time step t
v_t	$\in \mathbb{R}$	Value estimate at time step t
$\text{Enc}(\cdot)$	Function	Observations encoder of world model (including visual signals, rewards, termination flags, etc.)
$\text{Dec}(\cdot)$	Function	Image decoder of world model (including visual signals, rewards, termination flags, etc.)
$f(\cdot, \cdot)$	Function	Dynamics model that predicts future states given current state and action. This component learns the "physics" of the environment—how the game world evolves over time. It is often stateful, meaning it maintains memory (e.g., RNN hidden state) to track temporal dependencies

1.5. Abbreviations and Terms

Table 1.2. Abbreviations and Terms

Term	**Explanation**
env.	Short for environment. In reinforcement learning, this refers to the external system or simulator with which the agent interacts. It provides observations and rewards in response to the agent's actions.
MSE	Mean Squared Error. A common loss function used to measure the average squared difference between predicted and true values.
Mask Pred. Loss	Abbreviation for Mask Prediction Loss. This refers to a loss function associated with predicting segmentation masks from world model states.
Image Recon. Loss	Abbreviation for Image Reconstruction Loss. This term denotes the loss between the input image and the reconstructed image, often calculated using MSE. It evaluates how well the model can reconstruct visual observations.

CHAPTER 2

Background and Problem Formulation

2.1. World Models in Reinforcement Learning

World models [1] are predictive models that learn to simulate environment dynamics in a latent state space, enabling reinforcement learning agents to plan and make decisions through simulated rollouts rather than direct interaction with the real environment. This approach is particularly valuable when the actual environment is unavailable, computationally expensive, or unsafe to interact with extensively.

The core functionality of a world model operates as a complete system that learns environment dynamics from high-dimensional observations [2]. In vision-based reinforcement learning, a world model typically consists of three key components:

(1) **Encoder**: $s_t = \text{Encode}(o_t)$ converts observations (including visual signals, rewards, termination flags, etc.) to latent states

(2) **Dynamics Model**: $s_{t+1} = f(s_t, a_t)$ predicts future latent states given current states and actions

(3) **Decoder** (optional): $\hat{o}_{t+1} = \text{Decode}(s_{t+1})$ reconstructs observations from predicted latent states

The complete world model can thus be represented as:

$$\hat{o}_{t+1} = \text{WorldModel}(o_t, a_t) = \text{Decode}(f(\text{Encode}(o_t), a_t)) \tag{2.1}$$

where, in practice, observations o_t may include visual frames, reward signals, termination indicators, and other environmental information. This encode-predict-decode framework allows agents to simulate potential futures in compressed latent space while maintaining the ability to reconstruct observations for evaluation.

Common encoder-decoder implementations include Variational Autoencoders (VAEs) [3], Vector Quantized VAEs (VQ-VAEs) [4, 5], and diffusion models [6, 7]. Recently, GAIA-1 [8] equips its

world model with a diffusion model as video decoder for autonomous driving applications. Building on these foundations, several approaches have advanced world model learning:

- **PlaNet** [9] introduced a latent dynamics model based on a Recurrent State-Space Model (RSSM) [10], enabling long-horizon predictions without pixel-space reconstructions. However, it required external planning methods as it did not explicitly learn a policy.
- **DreamerV1** [11] extended PlaNet by incorporating actor-critic policy learning, allowing agents to optimize actions directly in the learned latent space and enabling end-to-end learning without external planning.
- **DreamerV2** [12] improved upon DreamerV1 by introducing discrete and continuous latent variables, enhancing expressiveness and stability in world model learning.
- **DreamerV3** [13] further refined the framework by improving scalability, robustness, and generalization, making it applicable to a broader range of reinforcement learning tasks, including large-scale continuous control environments.

Despite these advancements, world models remain susceptible to information loss during latent space compression, particularly for subtle but important signals. Our work addresses this issue by introducing a regularization method that explicitly enforces the retention of critical signals in the latent representation.

2.2. Latent Space Representations in RL

Efficient latent space learning [14] is critical for representation learning in reinforcement learning (RL). One prominent approach is contrastive learning [15, 16], which aims to extract task-relevant features while discarding irrelevant information.

CURL [17] introduced a contrastive learning objective that maximizes agreement between representations of augmented views of the same observation while pushing apart representations of different observations. This method significantly improved sample efficiency [18] in vision-based RL tasks.

While contrastive learning has been widely used to improve latent representations, it does not explicitly enforce the retention of specific task-critical signals. Our work builds on these insights by introducing a regularization technique that directly preserves important environmental signals, ensuring better world model learning and improved policy performance.

2.3. Segmentation-Based Learning in RL

Segmentation techniques have been explored in robotics and vision-based reinforcement learning (RL) to identify task-relevant regions in observations. Notable approaches include:

- **Object-centric learning** [19], which enhances generalization by segmenting objects for improved policy adaptation. For instance, the Slot-Attention for Object-centric Latent Dynamics (SOLD) algorithm learns object-centric dynamics models in an unsupervised manner from pixel inputs, facilitating better policy learning.
- **Attention-based models** [20], which prioritize salient features for reinforcement learning. A self-supervised approach combining self-attention with reinforcement learning has produced significant improvements, including new state-of-the-art results in the Arcade Learning Environment.

Our method integrates SAM2 [21] [22], a state-of-the-art segmentation model developed by Meta, to extract important signals for world model learning.

2.4. World Models Training

We use DreamerV3, a model-based reinforcement learning algorithm, as our baseline.
DreamerV3 learns a compressed representation of the environment to help its policy agent make better decisions. The world model is a reconstruction-based model trained with the following objective:

$$l_{\text{wm}} = \text{MSE}(\mathbf{x}, \hat{\mathbf{x}}) + l_{\text{others}} \tag{2.2}$$

where l_{others} includes reward prediction, sequence modeling, and other auxiliary losses. A detailed notation table listing all mathematical symbols used in this paper can be found in 1.1.
Despite achieving strong policy performance across diverse tasks, DreamerV3's world model does not explicitly enforce the preservation of important signals. As a result, small but crucial features such as the ball in the popular Atari Pong task [23] [24] [25] may be lost during compression, potentially leading to suboptimal policy learning.

CHAPTER 3

Methodology and Results

3.1. Identifying Important Signals

A key step in our approach is to define a function $m(\cdot)$ that identifies important signals in the observation space. Given an image $\mathbf{x}$, this function produces a Boolean mask image $m(\mathbf{x})$, where each pixel value is 1 if it belongs to an important region and 0 otherwise:
To validate that the selected signals are indeed important, we later empirically show in Section 3.4.3 that policies trained with world models preserving these signals significantly outperform those that ignore them.

$$m(\mathbf{x}) = \text{SAM2}(\mathbf{x}) \tag{3.1}$$

where SAM2 is a segmentation model used to extract task-relevant regions. An ideal $m(\cdot)$ should accurately highlight all important but subtle signals in the observation space.
In our implementation, we use SAM2, which generates high-quality segmentation masks guided by human-defined prompts. The integration and usage of SAM2 within our framework are detailed in Section 3.3.3.
Our primary benchmark is Atari Pong, where the ball signal is a key feature for policy learning. To assess the generalizability of our method, we also conduct additional experiments on Breakout and Ms. Pac-Man, which involve different types of critical signals (e.g., paddle and bricks in Breakout, player and ghosts in Ms. Pac-Man).

3.2. Enhancing the Learning of Important Signals

To prevent the loss of important subtle signals during compression, we introduce a regularization term into the world model objective:

$$l_{\text{wm}} = \text{MSE}(\mathbf{x}, \hat{\mathbf{x}}) + l_{\text{others}} + \omega \cdot \text{MSE}(m(\mathbf{x}) \cdot \mathbf{x}, m(\mathbf{x}) \cdot \hat{\mathbf{x}}) \tag{3.2}$$

where:

- $\mathbf{x}$ is the input image,
- $\hat{\mathbf{x}}$ is the reconstructed image,
- ω is the weight assigned to the regularization term, controlling the emphasis on preserving important signals. A higher ω prioritizes signal retention over general reconstruction, and we set $\omega = 10$ in our experiments.
- $m(\cdot)$ is the masking function. Notably, an all-zero mask indicates that no important signals were selected for retention.
- $m(\mathbf{x}) \cdot \mathbf{x}$ is the Hadamard (element-wise) product, ensuring that the regularization loss is only applied to important regions identified by the segmentation mask.

We establish the following three key points and validate them through experiments:

(1) Regularization accelerates the learning of important subtle signals, allowing the world model to capture them earlier in training.

(2) Regularization improves the retention of important signals, especially when they are highly subtle and prone to being overlooked.

(3) The identified signals, i.e., $m(\mathbf{x}) \cdot \mathbf{x}$, are indeed critical for policy learning, as demonstrated by their direct impact on downstream performance. In this sense, learning them leads to better world model representations.

Our primary analysis focuses on Atari Pong, with additional results for Breakout and Ms. Pac-Man provided in the Supplements.

3.3. Experiment Setup

This section describes our experimental methodology for evaluating the proposed regularization approach. We begin with the evaluation protocol, present our results on Atari Pong, and then detail the technical implementation.

3.3.1. Evaluation Protocol. Benchmark and Training Setup: We evaluate our method on Atari Pong following the Atari100k [26] benchmark established by DreamerV3 [13]. To better assess long-term signal retention capabilities, we extend the training duration from the standard 400K steps to 1M steps (approximately 1M frames). Additional experiments on Breakout and Ms. Pac-Man are provided in Section 3.5.

Game Environment: In Atari Pong, the player controls the right paddle and competes against a computer-controlled left paddle to deflect a ball into the opponent's goal. The first player to reach 21 points wins the game. Since the ball is the central game element that determines success, ensuring that the world model accurately captures its position and movement is critical for effective policy learning.

Experimental Conditions: We conduct experiments under two conditions:

- **Standard environment**: The original Atari Pong without modifications
- **Noisy environment**: Global Gaussian noise is applied to input images, reducing the signal-to-noise ratio of critical visual elements and making signal preservation more challenging

The noisy condition allows us to assess whether regularization provides enhanced benefits when preserving subtle signals becomes a limiting factor for representation learning.

Evaluation Methodology: We employ a comprehensive evaluation framework with both quantitative and qualitative methods. Depending on the specific research question being addressed, we selectively apply appropriate subsets of these evaluation approaches:

Quantitative Methods:

- **Image Reconstruction Loss**: We measure Mean Squared Error (MSE) loss by sampling unseen data and performing forward passes on converged world models without backward propagation. Lower reconstruction error indicates better learned representations.
- **Mask Prediction Loss**: We train a mask predictor CNN to reconstruct Boolean mask images from the world model's latent representations. Given a trained world model checkpoint at step t, we freeze the world model parameters and train the mask predictor from scratch using:

$$l_{\text{mask predictor}} = \omega' \cdot \text{MSE}(m(\mathbf{x}) \cdot \mathbf{x}, m(\mathbf{x}) \cdot \hat{\mathbf{x}}) \tag{3.3}$$

where $\omega' = 10$ aligns with our regularization weight in Equation (3.2). This loss trains only the mask predictor without affecting the world model's learned representations. During evaluation, we compare mask prediction loss to quantitatively measure how well the world model retains the ball signal in its latent space.

- **Policy Performance**: We compare reinforcement learning scores between standard DreamerV3 and our regularized version to demonstrate that preserved signals are critical for effective decision-making.

Qualitative Methods:

- **Mask Prediction Visual Analysis**: We conduct visual examination of predicted masks to observe signal preservation behavior, including consistency, stability, and trajectory tracking of the ball across frame sequences.

For each experiment presented in this work, we apply the most relevant subset of these evaluation methods based on the specific hypotheses being tested, ensuring focused and appropriate assessment of our regularization approach.

3.3.2. Model Architecture. Our reinforcement learning algorithm, including both world model and policy model, is based on DreamerV3. Architectures of DreamerV3 world model and policy model are illustrated in Figures 3.1 and 3.2, respectively.

The world model consists of the following components:

- Encoder: A CNN that extracts latent features from input image observations.
- Latent Dynamics Model: A discrete latent space model with a Recurrent State-Space Model (RSSM), enabling long-horizon rollouts by predicting future latent states.
- Decoder: A transposed CNN that reconstructs image observations from latent representations.
- Reward Model: An MLP that predicts expected rewards from latent states.
- Termination Model: A MLP that predicts episode termination signals from latent states.

The policy model consists of:

- Actor: Selects actions based on the learned latent representations to maximize expected rewards.
- Critic: Evaluates state values to guide policy optimization.

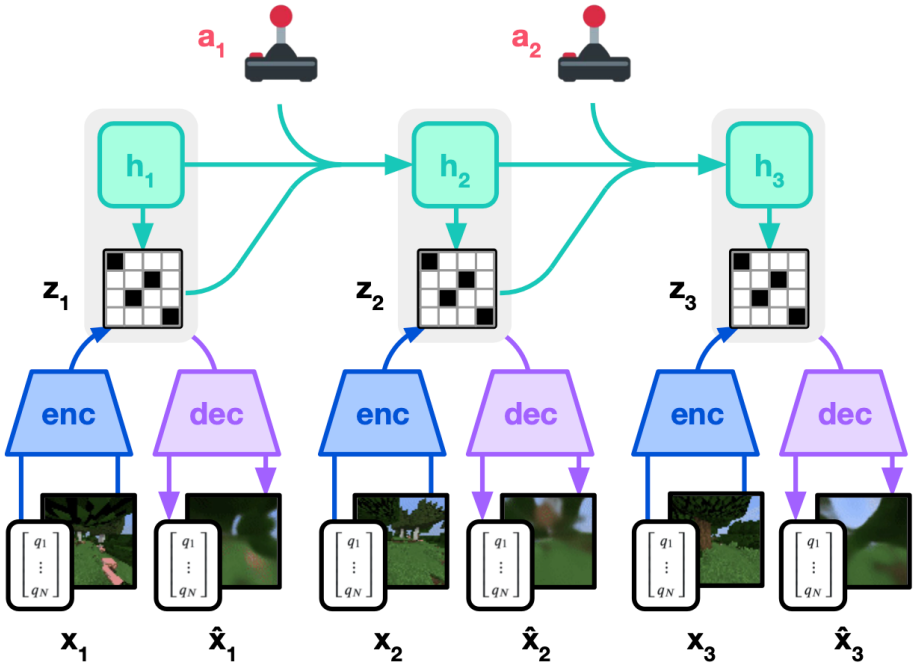

FIGURE 3.1. DreamerV3 world model architecture

During training step t, the DreamerV3 world model encodes sensory image inputs $\mathbf{x}_t$ into discrete representations z_t, which are predicted by a sequence model with recurrent state h_t given actions a_t. The model reconstructs the input as $\hat{\mathbf{x}}_t$, ensuring latent consistency. The actor and critic learn from trajectories of abstract representations predicted by the world model, optimizing actions a_t and value estimates v_t.

Additional Evaluation Component: For evaluation purposes only, we introduce a mask predictor that shares the same architectural structure as DreamerV3's image decoder but maintains separate, independently trained parameters. This component takes the world model's latent representations as input and predicts segmentation masks, allowing us to assess whether important visual signals are preserved in the learned representations. The mask predictor is trained only during evaluation and does not affect the core DreamerV3 training process.

3.3.3. SAM2 Implementation and Integration. To enforce preservation of important visual signals, we integrate a segmentation function $m(\cdot)$ based on SAM2 (Segment Anything Model 2), developed by Meta. This integration operates through a client-server architecture where the **backend server** hosts the SAM2 segmentation model and provides segmentation services through

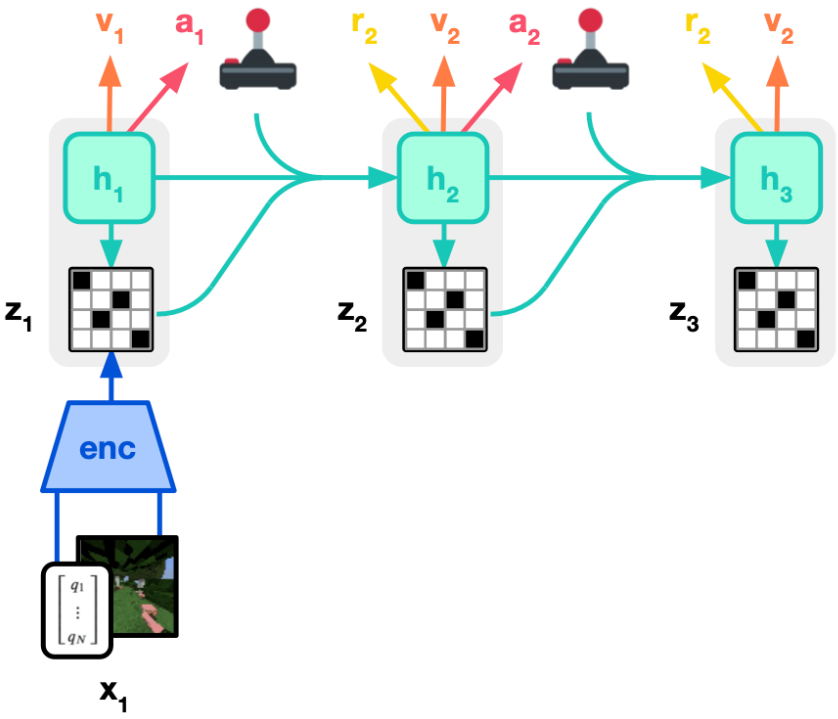

FIGURE 3.2. DreamerV3 policy model architecture

API calls, while the **client interface** is integrated within DreamerV3 to request and retrieve segmentation results via GraphQL API. Segmentation masks are returned in Run-Length Encoding (RLE) format and converted to binary mask images for training.

Implementation Workflow: The integration follows a three-step process:

(1) Segmentation Prompt Definition: We begin by extracting representative gameplay videos for each target task (Pong, Breakout, Ms. Pac-Man) and manually annotating key frames (e.g., 15 frames for Pong) by marking locations of critical visual elements such as the ball in Pong. These annotations are then stored as segmentation prompts for SAM2.

(2) Prompt Validation: This quality assurance step ensures segmentation accuracy before deployment. We insert the annotated frames into the original video sequence and send the augmented video (original plus annotated frames) with prompts to SAM2. The system retrieves segmentation masks in RLE format, decodes them to binary images, removes masks corresponding to inserted annotation frames, and manually evaluates mask quality by comparing extracted regions against ground truth. For example, a 64-frame video with

10 annotated frames creates a processed sequence of 74 frames. After SAM2 generates 74 masks, we remove the 10 masks corresponding to inserted frames, yielding 64 validated segmentation masks.

(3) Training Integration: During regularized world model training, we apply the same validated process to training batches of 1024 frames (16 sequences $\times$ 64 frames each). We append the 10 pre-validated annotation frames, creating 1034 total frames, send the augmented batch to SAM2 for segmentation, decode the 1034 returned masks, remove the 10 annotation-related masks, and use the remaining 1024 masks for regularization loss computation.

3.3.4. Hyperparameter Settings. We use the same hyperparameters as DreamerV3 to ensure consistency with prior work. The key settings include:

- Batch Size: 16
- Batch Length (Sequence Length per Sample): 64
- Regularization Weight ω in l_{wm}: 10.0
- Regularization Weight ω' in $l_{\text{mask predictor}}$: 10.0

These values maintain stable training across both standard and noisy environments.

3.4. Experimental Results

This section presents our experimental findings validating three key propositions:

(1) regularization accelerates the learning of important signals,

(2) regularization enhances the retention of important signals in converged models, and

(3) the identified signals are critical for policy learning.

We evaluate these claims using both standard Atari Pong environments and noisy variants, as the preservation of subtle signals becomes even more critical under challenging conditions.
Additionally, we discover and analyze a special phenomenon: regularization induces local minimum in noise reconstruction in Section 3.5.1.

3.4.1. Regularization Accelerates Learning of Important Signals. To demonstrate that our regularization approach accelerates the learning of important visual signals, we evaluate both

models at 60K training steps—well before convergence—using quantitative metrics and qualitative observations in the standard environment.

Quantitative Methods:

TABLE 3.1. Quantitative Assessment at 60K Steps (Standard Environment)

Model Variant	**Mask Pred. Loss ↓**	**Image Recon. Loss ↓**	**Policy Score ↑**
Baseline	7.185	0.2974	-19.5
Regularized	**0.4694**	**0.2599**	**-3.0**

The regularized model achieves dramatically lower mask prediction loss (0.4694 vs 7.185), reduced image reconstruction loss (0.2599 vs 0.2974), and substantially improved policy performance (-3.0 vs -19.5), demonstrating that regularization enables the world model to learn important signal representation, i.e., the representation of the ball, much earlier in training with direct benefits for downstream policy learning.

Qualitative Methods - Mask Prediction Analysis:

TABLE 3.2. Mask Prediction Performance Comparison at 60K Steps

Model Variant	**Observed Issues**	**Figure**
Baseline	Ball completely disappears in long sequences. From frames 6 to 17, the mask predictor fails to localize the ball.	Figure 3.3
Baseline	The prediction sometimes includes two balls simultaneously, indicating severe instability in representation learning.	Figure 3.4
Regularized	Ball remains consistently visible and follows the correct motion trajectory throughout all sequences.	Figure 3.5

These results provide strong evidence that regularization not only improves final performance but significantly accelerates the learning process for critical visual signals during early training stages, with measurable improvements in both representation quality and policy effectiveness.

3.4.2. Regularization Enhances Retention of Important Signals. To validate that regularization enhances signal retention in fully trained models, we evaluate converged models (1M steps) using both quantitative metrics and qualitative observations across standard and noisy environments. We specifically include noisy environment evaluation because environmental noise makes the ball signal more subtle and harder for the world model to learn, creating conditions where our

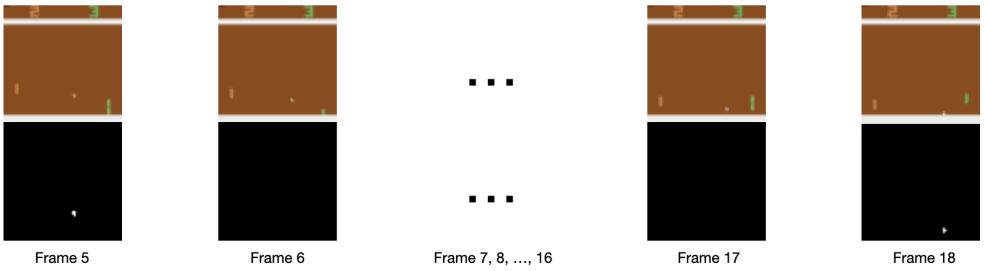

FIGURE 3.3. Baseline model fails to track during extended sequences. Each figure shows original game frames (top row) and corresponding predicted masks (bottom row). The mask predictor fails to maintain consistent ball localization, with the ball completely disappearing from frames 6-17, demonstrating poor signal retention in early training stages.

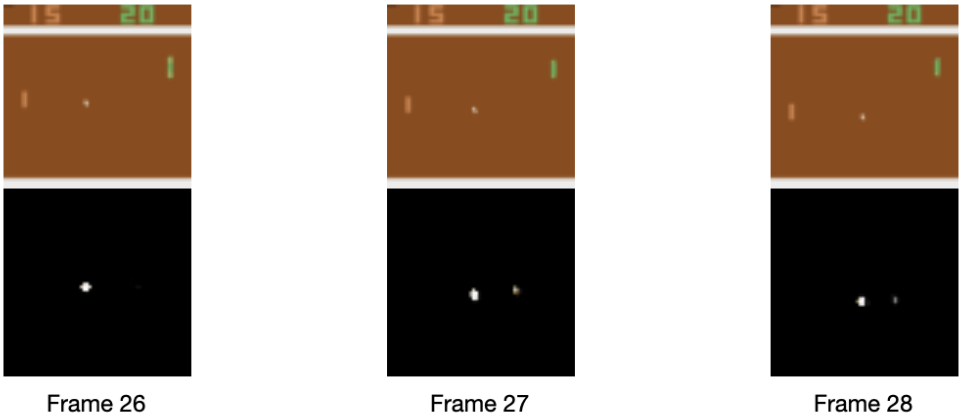

FIGURE 3.4. Dual ball prediction artifacts. Original game frames (top row) show single ball positions, while predicted masks (bottom row) generate multiple simultaneous ball predictions, indicating unstable representation learning and poor signal-noise discrimination at 60K training steps.

regularization method is expected to demonstrate greater advantages in preserving critical signals against interference.

Quantitative Methods:

TABLE 3.3. Quantitative Assessment at 1M Steps

Environment	**Model Variant**	**Mask Pred. Loss ↓**	**Image Recon. Loss ↓**
Standard	Baseline	0.4024	0.0156
	Regularized	**0.204**	**0.0142**
Noisy	Baseline	10.78	0.0389
	Regularized	**0.250**	**0.0198**

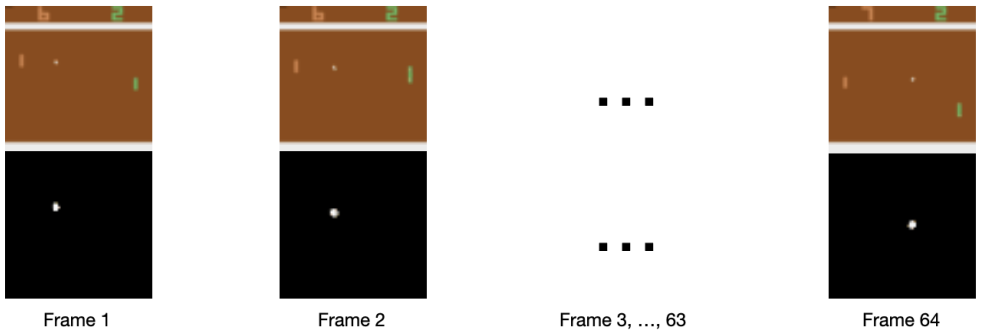

FIGURE 3.5. Regularized model achieves consistent ball tracking throughout sequences. Original game frames (top row) and predicted masks (bottom row) demonstrate stable, accurate ball localization across all time steps, with masks correctly following the ball's motion trajectory.

The regularized model demonstrates superior performance in both environments, with benefits becoming more pronounced under challenging noisy conditions where the baseline model's performance degrades dramatically (mask prediction loss: 0.4024→10.78) while the regularized model maintains robust performance (0.204→0.250).

Qualitative Methods - Mask Prediction Analysis:

TABLE 3.4. Mask Prediction Performance Comparison at 1M Steps

Environment	Model Variant	Observed Issues	Figure
Standard	Baseline	Ball shows intermittent tracking instabilities and occasional signal loss during complex motion patterns.	Figure 3.6
	Regularized	Ball remains consistently visible with accurate trajectory tracking throughout all sequences.	Figure 3.7
Noisy	Baseline	Severe tracking failures with frequent ball disappearance and multiple false positive detections due to noise interference.	Figure 3.8
	Regularized	Robust ball tracking maintained despite noise, with minimal impact on localization accuracy.	Figure 3.9

These comprehensive evaluations demonstrate that regularization fundamentally improves the world model's ability to retain important signals, with benefits becoming more pronounced under challenging noisy conditions.

3.4.3. Identified Signals Are Critical for Policy Learning. To establish that the preserved signals are indeed critical for effective decision-making, we compare policy performance

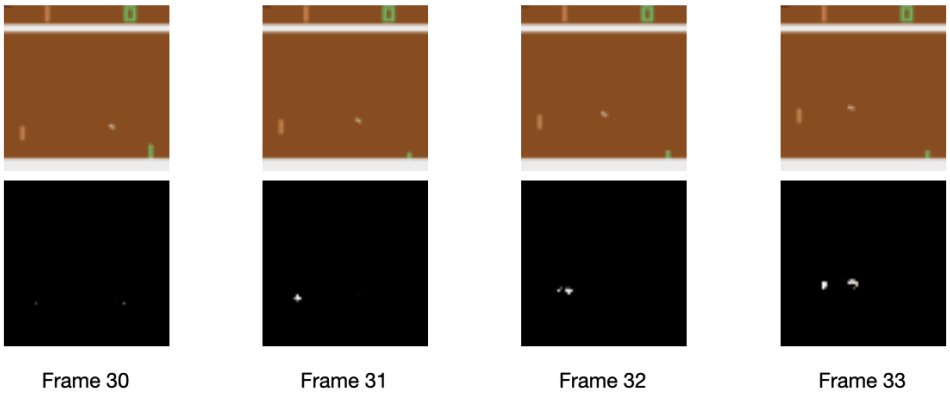

FIGURE 3.6. Baseline model at 1M steps (standard environment): Intermittent tracking instabilities. Original game frames (top row) and predicted masks (bottom row) show periodic signal loss and inconsistent ball localization during complex motion sequences despite full convergence.

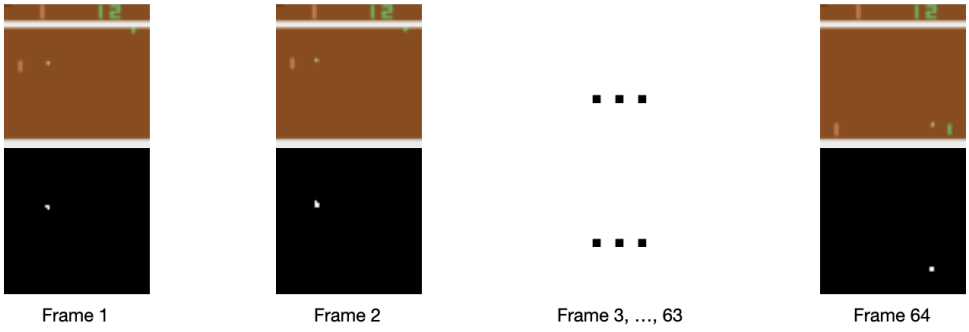

FIGURE 3.7. Regularized model at 1M steps (standard environment): Consistent tracking performance. Original game frames (top row) and predicted masks (bottom row) demonstrate stable, accurate ball localization throughout all motion patterns, showing superior signal retention.

between baseline and regularized models across different training durations and environmental conditions.

Quantitative Methods:

TABLE 3.5. Policy Performance Comparison Across Training Duration and Environments

Steps	Model Variant	Standard Env. Score ↑	Noisy Env. Score ↑
400K	Baseline	-4.0	-20.1
	Regularized	**3.0**	**4.2**
1M	Baseline	-3.0	-18.5
	Regularized	**15.0**	**10.4**

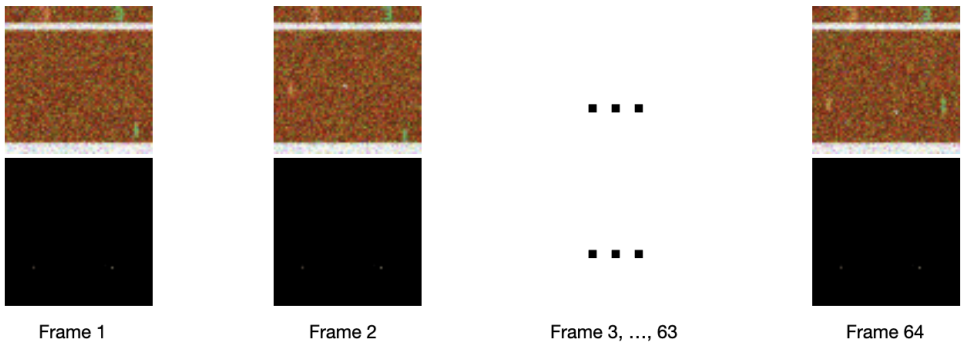

FIGURE 3.8. Baseline model at 1M steps (noisy environment): Severe degradation under noise. Original game frames (top row) show noisy conditions while predicted masks (bottom row) exhibit frequent tracking failures, ball disappearance, and multiple false positive detections, highlighting poor noise robustness.

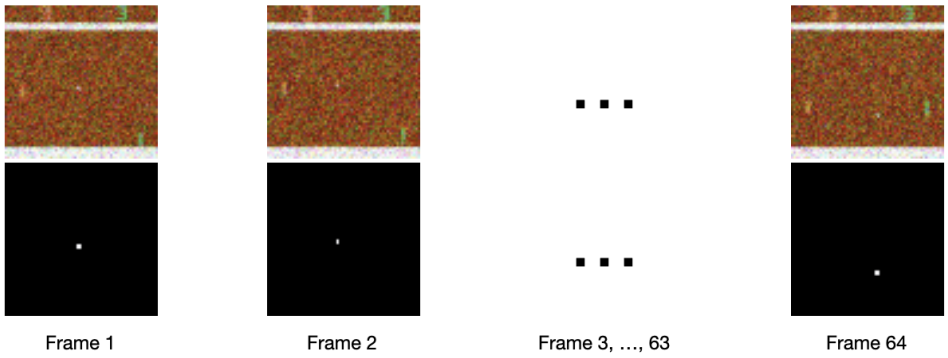

FIGURE 3.9. Regularized model at 1M steps (noisy environment): Robust performance under noise. Original game frames (top row) show the same noisy conditions while predicted masks (bottom row) maintain accurate ball tracking despite environmental noise, demonstrating enhanced signal retention capabilities.

TABLE 3.6. Overall Performance Comparison

Task	**Random Policy**	**Human Record**	**DreamerV3 (1M)**	**Our Method (1M)**
Pong	-21	21	-3	**15**

The regularized model demonstrates substantial improvements across all conditions: transforming negative performance to positive at 400K steps (-4.0→-3.0), achieving dramatically higher scores at convergence (3.0→15.0), and maintaining robust performance in noisy environments where the baseline fails catastrophically (-18.5 vs 10.4). The policy learning curves in Figures 3.10 and 3.11

illustrate the dramatic training dynamics differences between baseline and regularized models across both environmental conditions.

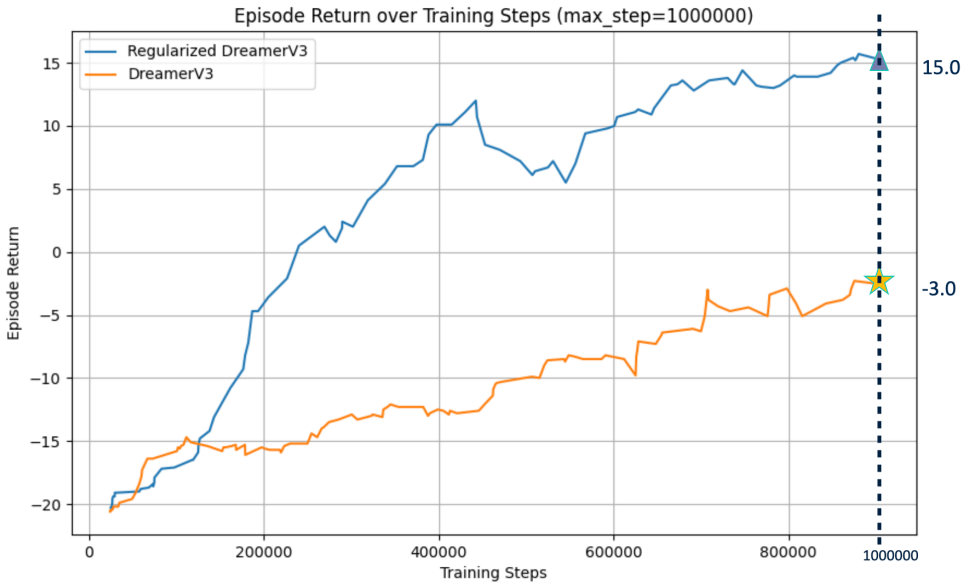

FIGURE 3.10. Policy learning curves in standard Pong environment. The regularized model (blue) achieves higher peak performance compared to the baseline model (red). The enhanced ball signal representation enables consistent performance improvement throughout training, reaching positive scores by 400K steps and achieving 15.0 at convergence versus -3.0 for the baseline.

Since enhancing the ball signal representation leads to these significant policy improvements across both standard and challenging noisy conditions, this provides strong evidence that the ball signal is indeed a critical component for effective Pong gameplay.

3.5. Additional Experiment Results on Other Tasks

To evaluate the generalizability of our method, we conduct additional experiments on Atari Breakout and Ms. Pac-Man, using a 400K training step budget due to computational constraints.

TABLE 3.7. Performance Comparison on Additional Atari Tasks

Task	**Random**	**Human Record**	**DreamerV3 (400K)**	**Our Method (400K)**
Breakout	2	864	3	**16**
Ms. Pac-Man	307	290090	1521	**4220**

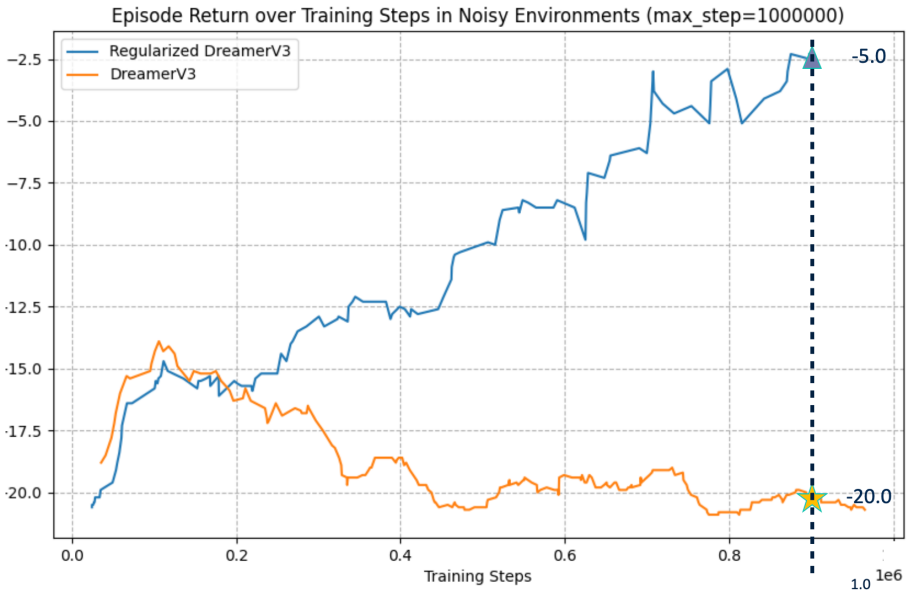

FIGURE 3.11. Policy learning curves in noisy Pong environment. The regularized model (blue) maintains robust learning performance despite environmental noise. In contrast, the baseline model (red) shows severe performance degradation under noise.

The regularization method demonstrates consistent improvements across different Atari tasks, achieving higher scores in Breakout (3→16) and Ms. Pac-Man (1521→4220), confirming the generalizability of our approach beyond Pong. The policy learning curves in Figures 3.12 and 3.13 illustrate the training dynamics for these additional tasks.

3.5.1. Regularization-Induced Local Minimum in Noise Reconstruction. During our experiments, we observe an intriguing phenomenon where regularization guides the world model toward a beneficial local minimum [27] that prioritizes signal preservation over complete reconstruction fidelity, leading to improved policy performance despite higher image reconstruction loss.

Discrepancy Between Image Reconstruction and Regularization Loss

We observe counterintuitive optimization behavior in noisy environments where loss magnitudes become disproportionate. As shown in Figures 3.14 and 3.15, at 400K steps the image reconstruction loss reaches 10.74 while the regularization loss remains at 3.606—a significant magnitude difference that theoretically should drive the optimizer to prioritize reconstruction error minimization.

Visual Evidence of Selective Reconstruction

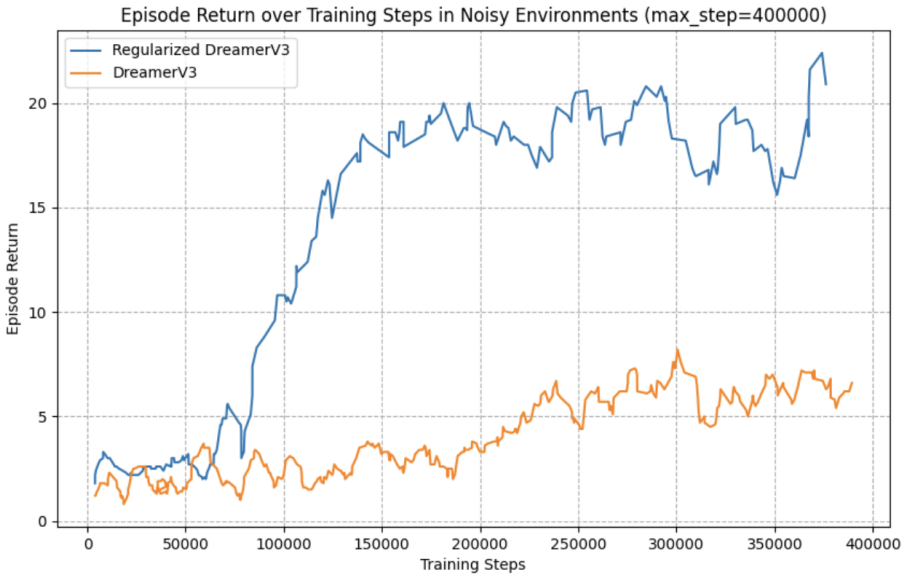

FIGURE 3.12. Policy learning curves for Atari Breakout. The regularized model (blue) shows improved policy performance compared to the baseline DreamerV3 (orange).

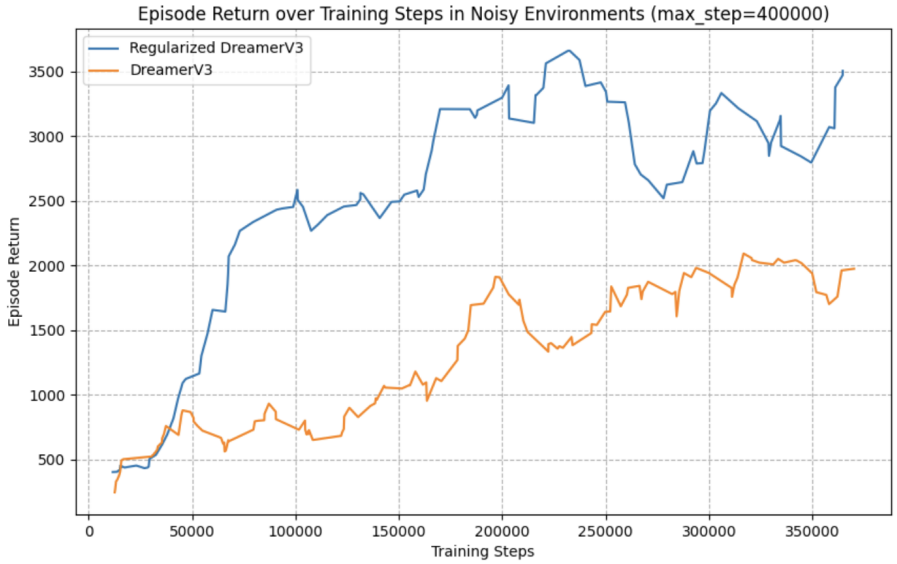

FIGURE 3.13. Policy learning curves for Ms. Pac-Man. The regularized model (blue) shows improved policy performance compared to the baseline DreamerV3 (orange).

Most remarkably, Figure 3.16 demonstrates that the regularized world model completely fails to reconstruct global Gaussian noise while maintaining perfect ball signal preservation. This selective

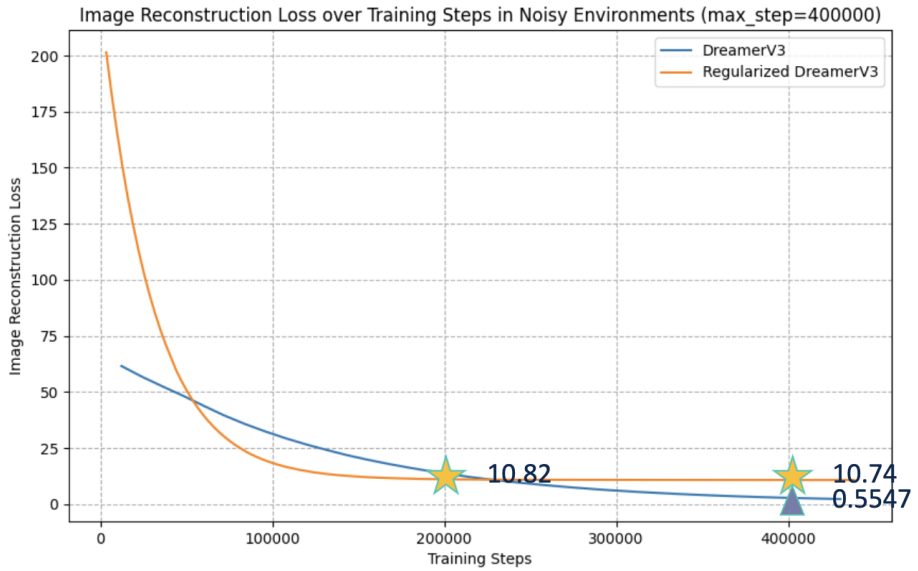

FIGURE 3.14. Image reconstruction loss in noisy Pong environment. The regularized model (blue) maintains higher reconstruction loss compared to baseline (orange), indicating deliberate neglect of noise reconstruction in favor of signal preservation.

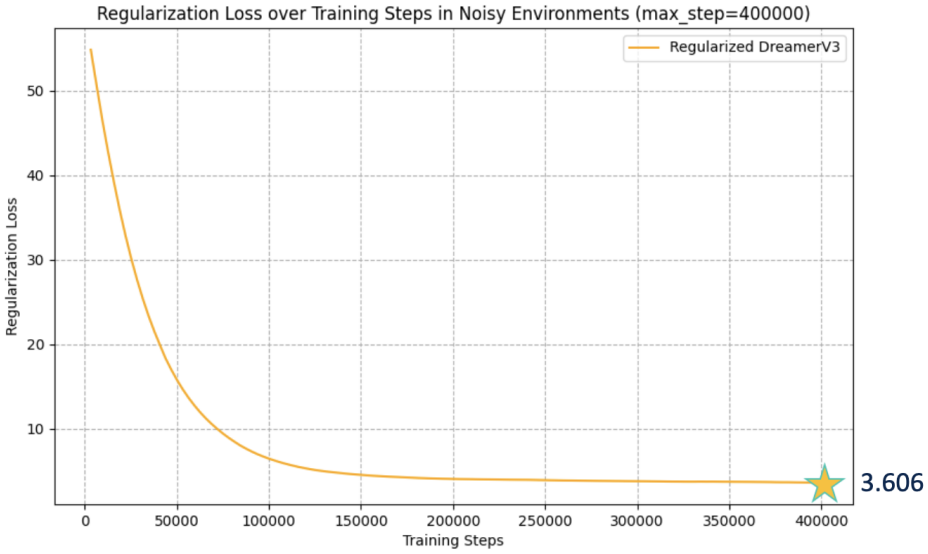

FIGURE 3.15. Regularization loss in noisy Pong environment. The regularized model achieves consistently lower regularization loss throughout training, demonstrating successful preservation of important signal components despite environmental noise.

reconstruction behavior contradicts standard optimization theory, where MSE loss should drive the model to learn all image components equally.

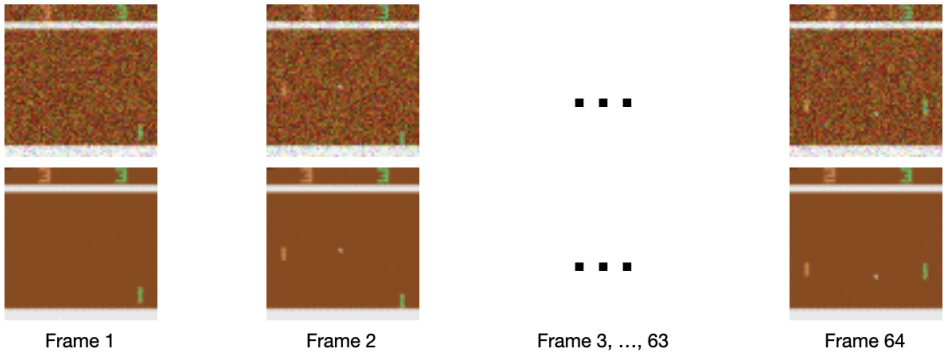

FIGURE 3.16. Selective reconstruction before reheating. Original images with global Gaussian noise (top row) and corresponding reconstructions (bottom row). At 400K steps, the regularized model exhibits counter-intuitive selective reconstruction, completely ignoring background noise while perfectly preserving ball signals, despite the noise reconstruction error being larger than the regularization loss magnitude.

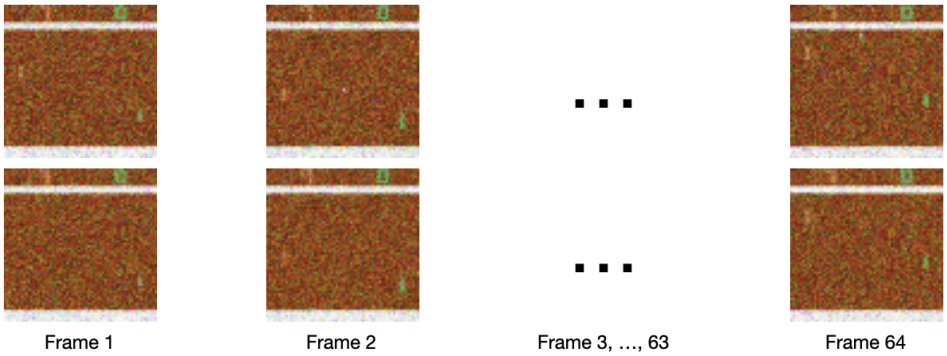

FIGURE 3.17. Complete reconstruction after reheating. Original images with global Gaussian noise (top row) and corresponding reconstructions (bottom row). After learning rate reheating, the model escapes the local minimum and begins properly reconstructing noise components, validating the local minimum hypothesis while maintaining signal preservation benefits.

Theoretical Insights & Practical Implications

From a practical perspective, this behavior is desirable—the world model avoids wasting representational capacity on irrelevant noise and focuses on task-critical signals. From an optimization

perspective, this reveals that regularization fundamentally alters the optimization landscape, guiding the optimizer toward beneficial local minima that prioritize signal preservation over complete reconstruction fidelity.

Learning Rate Reheating Validates Local Minimum Hypothesis

To confirm our local minimum hypothesis, we conduct learning rate reheating experiments by resuming training beyond 400K steps with a 5× learning rate increase. The results in Figures 3.18, 3.17, 3.19, and 3.20 demonstrate successful escape from the local minimum with resumed learning of noise reconstruction while maintaining signal preservation benefits.

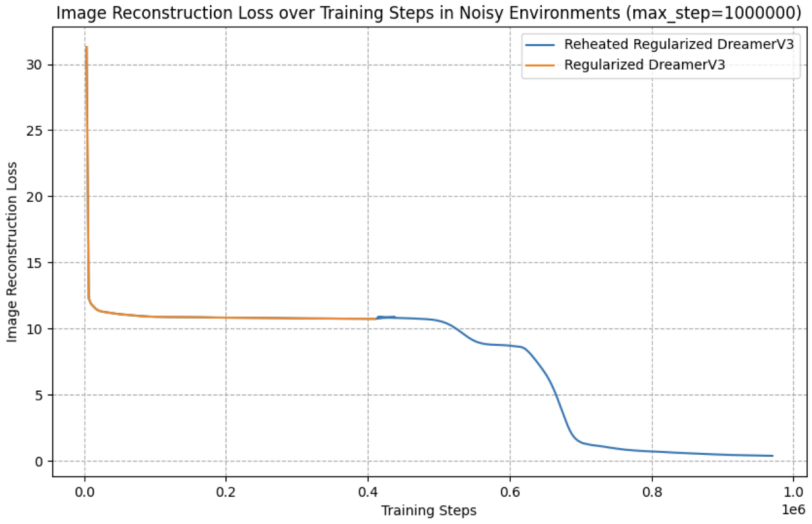

FIGURE 3.18. Image reconstruction loss after learning rate reheating. The loss resumes decreasing, confirming escape from the local minimum and renewed learning of previously ignored noise components.

These results validate that the regularization-induced local minimum represents a practically beneficial optimization state that balances signal preservation with reconstruction quality, providing both theoretical insights into regularization effects and practical evidence for improved learning dynamics.

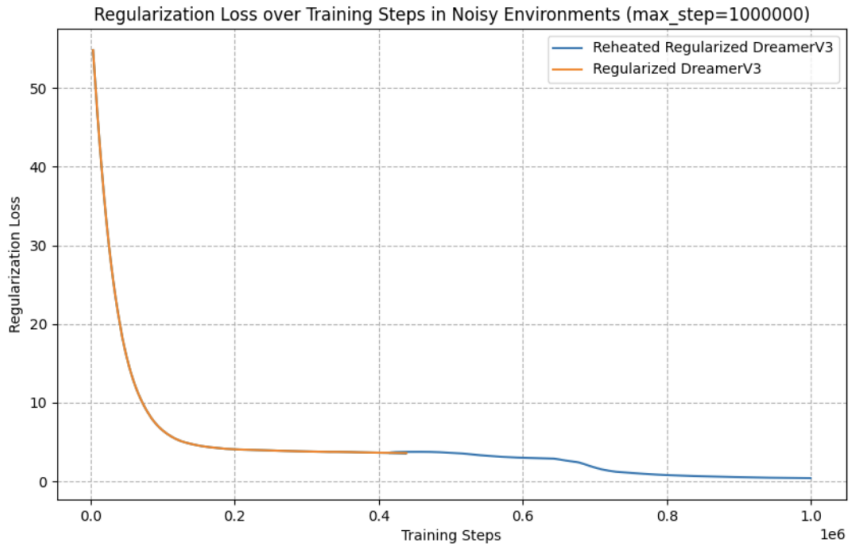

FIGURE 3.19. Regularization loss after learning rate reheating. The regularization loss continues decreasing, confirming that important signal preservation benefits are maintained during the transition out of the local minimum.

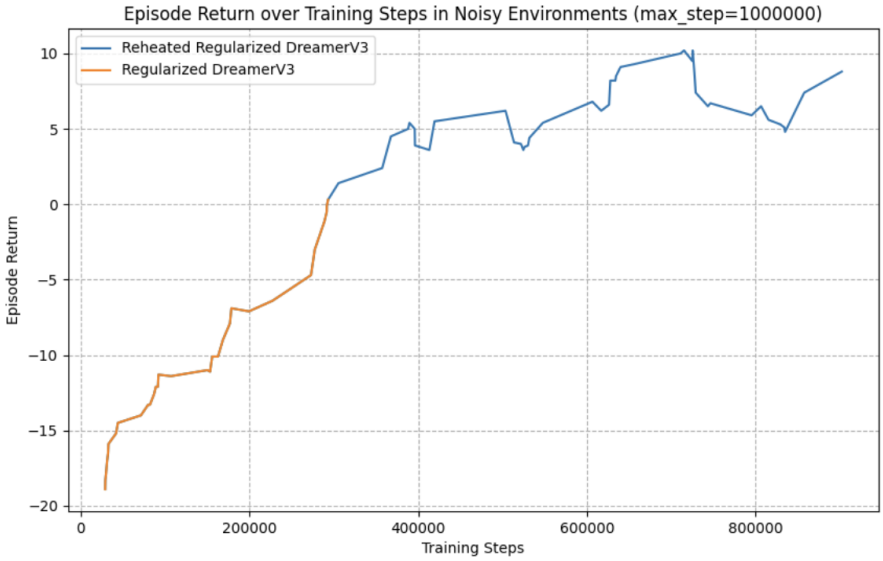

FIGURE 3.20. Policy performance after learning rate reheating. Performance continues improving, demonstrating that escaping the local minimum maintains the practical benefits of signal preservation while enabling more complete reconstruction learning.

CHAPTER 4

Conclusion

We introduced a regularization method to preserve important subtle signals in world model learning, addressing a key limitation in latent space compression. Our approach ensures that world models retain subtle task-relevant signals, preventing critical information loss.

Through extensive experiments on Atari Pong, Breakout, and Ms. Pac-Man across both standard and noisy environments, we demonstrated that:

- **Regularization accelerates the learning of important subtle signals**, allowing world models to capture them earlier in training, with particularly pronounced benefits in challenging noisy conditions.
- **Regularization enhances the retention of these signals** in converged models, providing robust performance improvements that become more significant when signal preservation is critical.
- **The identified signals are critical for policy learning**, leading to significantly improved performance in reinforcement learning tasks, with our method achieving remarkable improvements especially in noisy environments where baseline models fail catastrophically.
- **Enhanced robustness under challenging conditions**: Our method provides substantially greater improvements in noisy environments, where baseline models show dramatic performance degradation while our regularized approach maintains strong performance, confirming that signal preservation becomes increasingly important when distinguishing critical features from noise.
- **Discovery of beneficial local minimum**: We identified and analyzed a novel phenomenon where regularization guides the optimizer toward a beneficial local minimum that prioritizes signal preservation over complete reconstruction. This selective optimization behavior, while counterintuitive from a theoretical standpoint, proves practically advantageous for downstream tasks. Our learning rate reheating experiments further validate this understanding and demonstrate controllable escape from these local minima when desired.

Bibliography

[1] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[2] Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chong Qi, John Wright, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in Neural Information Processing Systems*, 33:9422–9434, 2020.

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[4] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 6306–6315, 2017.

[5] Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. 32:14866–14876, 2019.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[7] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[8] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[9] Daniil Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pages 2555–2565. PMLR, 2019.

[10] Daniil Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Recurrent state space models for reinforcement learning. In *International Conference on Machine Learning*, pages 2477–2486. PMLR, 2019.

[11] Daniil Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

[12] Daniil Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021.

[13] Daniil Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[14] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Tobias Pohlen, et al. Mastering atari, go, chess and shogi by planning with a learned model. volume 588, pages 604–609. Nature Publishing Group, 2020.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[17] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pages 5639–5650. PMLR, 2020.

[18] Piotr Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. Sample efficiency in deep reinforcement learning: A survey. *arXiv preprint arXiv:2206.04670*, 2022.

[19] Max Kipf, Maurits van Steenkiste, Aäron van den Oord, and Thomas Kipf. Sold: Reinforcement learning with slot object-centric latent dynamics. *arXiv preprint arXiv:2410.08822*, 2024.

[20] Soroush Rabiee, Farhad Zandi, and Hassan Ghasemzadeh. Attention-driven multi-agent reinforcement learning: Enhancing decisions with expertise-informed tasks. *arXiv preprint arXiv:2404.05840*, 2024.

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[22] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[23] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[24] Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

[25] Jesse Farebrother and Pablo Samuel Castro. Cale: Continuous arcade learning environment. In *Advances in Neural Information Processing Systems*, 2024. arXiv preprint arXiv:2410.23810.

[26] Denis Yarats, Ilya Kostrikov, and Rob Fergus. Improving sample efficiency in model-free reinforcement learning from images. 35:10674–10681, 2021.

[27] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in Neural Information Processing Systems*, 27, 2014.