Unbounded Memory and Consistent Imagination via Unified Diffusion–SSM World Models

Jia-Hua Lee [1] **Bor-Jiun Lin** [1] **Wei-Fang Sun** [2] **Chun-Yi Lee** [3]

Abstract

World models represent a promising approach for training reinforcement learning agents with significantly improved sample efficiency. While most world model methods primarily rely on sequences of discrete latent variables to model environment dynamics, this compression often neglects critical visual details essential for reinforcement learning. Recent diffusion-based world models condition generation on a fixed context length of frames to predict the next observation, using separate recurrent neural networks to model rewards and termination signals. Although this architecture effectively enhances visual fidelity, the fixed context length approach inherently limits memory capacity. In this paper, we introduce EDELINE, a unified world model architecture that integrates state space models with diffusion models. Our approach demonstrates superior performance on the memory-demanding Crafter benchmark.

1. Introduction

World models (Ha & Schmidhuber, 2018) constitute a foundational element of modern reinforcement learning (RL) by simulating environment dynamics for agent planning and reasoning. The capacity to learn environment representations (Hafner et al., 2024; Schrittwieser et al., 2020) facilitates policy optimization through imagined trajectories, which substantially enhances sample efficiency (Ye et al., 2021) relative to conventional RL approaches. This capability is especially valuable for real-world applications in robotics and autonomous systems.

Existing world models fall into two principal paradigms: *latent-space models* and *generative models*. Latent-space approaches (Hafner et al., 2020; 2021; 2024) employ recurrent neural networks (RNNs) or variants to predict future states within a compressed latent space for efficient policy optimization. This compression, however, introduces information loss that compromises generality and reconstruction quality. Generative models, particularly diffusion-based approaches (Alonso et al., 2024), have transformed world modeling through high-fidelity visual predictions via noise-reversal processes. Nevertheless, prior generative models depend on fixed-length observation-action windows that truncate historical context and fail to capture extended temporal dependencies. This limitation presents a challenge especially in partially observable environments where agents must retain and reason over prolonged observation sequences for informed decisions. Moreover, the architectural segregation of reward prediction, termination signals, and observation modeling in existing frameworks can potentially lead to suboptimal representation sharing and optimization conflicts that further impair performance.

In order to mitigate long sequence dependency issues, recent state space models (SSMs) (Gu et al., 2022a;b; Smith et al., 2023; Gu & Dao, 2024b;a) provide a complementary advantage through their capacity to model long-term dependencies efficiently. With linear-time complexity and selective state updates (Gu & Dao, 2024b), SSMs can process theoretically unbounded sequences while preserving critical historical information. This capability is particularly valuable for world modeling, where accurate trajectory prediction often necessitates retention and reasoning across extended observation-action histories.

Based on these considerations, we introduce EDELINE (**E**nhancing **D**iffusion-bas**E**d World Models via **LINE**ar-Time Sequence Modeling), a unified framework that integrates the advantages of diffusion models and SSMs. EDELINE advances the state-of-the-art (SOTA) through three key innovations: (1) **Memory Enhancement**: A recurrent embedding module (REM) based on Mamba SSMs that processes unbounded observation-action sequences to enable adaptive memory retention beyond fixed-context limitations, (2) **Unified Framework**: Direct conditioning of reward and termination prediction on REM hidden states that eliminates separate networks for efficient representation sharing, and (3) **Dynamic Loss Harmonization**: Adaptive weighting of observation and reward losses that addresses scale disparities in multi-task optimization. To validate EDELINE's

[1]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan [2]NVIDIA AI Technology Center (NVAITC), NVIDIA Corporation, Santa Clara, CA, USA [3]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan. Correspondence to: Chun-Yi Lee <cylee@csie.ntu.edu.tw>.

ICML 2025 Workshop on Assessing World Models. Copyright 2025 by the author(s).

effectiveness, we conduct evaluation on Crafter (Hafner, 2022), a procedurally generated survival environment specifically designed to evaluate "wide and deep exploration, long-term reasoning and credit assignment, and generalization". The complete technical details and extended experimental analysis are available in the full paper (Lee et al., 2025).

2. Background

In this Section, we focus on the essential concepts necessary for understanding our EDELINE framework. We provide additional background material on score-based diffusion models and multi-task world model learning in Appendix A.

2.1. Reinforcement Learning and World Models

The problem considered in this study focuses on image-based reinforcement learning (RL), formulated as a Partially Observable Markov Decision Process (POMDP) (Åström, 1965) defined by tuple $(S, A, \mathcal{O}, P, R, O, \gamma)$. Our formulation specifically considers high-dimensional image observations as inputs, as described in Section 1. The state space S comprises states $s_t \in S$, while the action space A can be either discrete or continuous with actions $a_t \in A$. The observation space $\mathcal{O}$ contains image observations $o_t \in \mathbb{R}^{3\times H\times W}$. A transition function $P : S \times A \times S \to [0, 1]$ characterizes the environment dynamics $p(s_{t+1}|s_t, a_t)$, while the reward function $R : S \times A \times S \to \mathbb{R}$ maps transitions to scalar rewards $r_t \in \mathbb{R}$. The observation function $O : S \times \mathcal{O} \to [0, 1]$ establishes observation probabilities $p(o_t|s_t)$. The objective centers on learning a policy π that maximizes the expected discounted return $\mathbb{E}_\pi[\sum_{t\geq 0} \gamma^t r_t]$, with discount factor $\gamma \in [0, 1]$. Model-based Reinforcement Learning (MBRL) (Sutton, 1988) achieves this objective by learning a world model that encapsulates the environment dynamics $p(o_{t+1}, r_t|o_{\leq t}, a_{\leq t})$. MBRL enables learning in imagination through three systematic stages: (1) collecting real environment interactions, (2) updating the world model, and (3) training the policy through world model interactions.

2.2. Linear-Time Sequence Modeling with Mamba

SSMs (Gu et al., 2022a) provide an alternative paradigm to attention-based architectures for sequence modeling. The Mamba architecture (Gu & Dao, 2024b) introduces a selective state space model that offers linear time complexity and efficient parallel processing, which employs variable-dependent projection matrices to implement its selective mechanism, thus overcoming the inherent limitations of computational inefficiency and quadratic complexity in conventional SSMs (Gu et al., 2020; 2022a; Smith et al., 2023; Gu & Dao, 2024b). The foundational mechanism of Mamba is characterized by a linear continuous-time state space formulation via first-order differential equations as follows:

$$\begin{aligned} \frac{\partial x(t)}{\partial t} &= Ax(t) + B(u(t))u(t), \\ y(t) &= C(u(t))x(t), \end{aligned} \tag{1}$$

where $x(t)$ represents the latent state, $u(t)$ denotes the input, and $y(t)$ indicates the output. The matrix A adheres to specifications from (Gu et al., 2022b). The primary innovation compared to traditional SSMs lies in $B(u(t))$ and $C(u(t))$, which function as state-dependent linear operators to enable selective state updates based on input content. For discretization, the system employs the zero-order-hold (ZOH) rule (Chifu et al., 2018) to transform the A and B matrices into $\tilde{A} = \exp(\Delta A)$ and $\tilde{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B$, where the step size Δ serves as a variable-dependent parameter. This transformation enables SSMs to process continuous inputs as discrete signals and converts the original Linear Time-Invariant (LTI) equation into a recurrence format.

2.3. Diffusion-based World Model Learning

To adapt diffusion models for world modeling, which offers superior sample quality and tractable likelihood estimation, a key requirement is modeling the conditional distribution $p(o_{t+1}|o_{\leq t}, a_{\leq t})$, where o_t and a_t represent observations and actions at time step t. The denoising process incorporates both the noised next observation and the conditioning context as input: $D_\theta(o^\tau_{t+1}, \tau, o_{\leq t}, a_{\leq t})$. While diffusion-based world models (Alonso et al., 2024) have shown promise, the state-of-the-art approach DIAMOND (Alonso et al., 2024) exhibits limitations although it achieves superior performance on the Atari 100k benchmark (Łukasz Kaiser et al., 2020). These models face two critical limitations. The first limitation stems from their constrained conditioning context, which typically considers only the most recent observations and actions. For instance, DIAMOND restricts its context to the last four observations and actions in the sequence. This constraint impairs the model's capacity to capture long-term dependencies and leads to inaccurate predictions in scenarios that require extensive historical context. The second limitation in current diffusion-based world models lies in their architectural separation of predictive tasks. For example, DIAMOND implements a separate recurrent neural network for reward and termination prediction. This separation prevents the sharing of learned representations between the diffusion model and these predictive tasks and results in reduced overall learning efficiency of the system.

3. Methodology

Conventional diffusion-based world models (Alonso et al., 2024) demonstrate promise in learning environment dynamics yet face fundamental limitations in memory capacity and horizon prediction consistency. To address these challenges, this paper presents EDELINE, as illustrated in Fig 1, a unified architecture that integrates state space models (SSMs) with diffusion-based world models. EDELINE's core innovation lies in its integration of SSMs for encoding sequential observations and actions into hidden embeddings, which a diffusion model then processes for future frame predic-

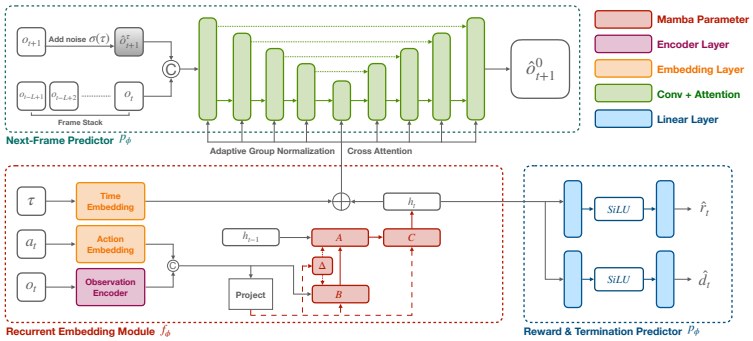

Figure 1: **Framework Overview of EDELINE.** The model integrates three principal components: (1) An U-Net-like *Next-Frame Predictor* enhanced by adaptive group normalization and cross-attention mechanisms, (2) A *Recurrent Embedding Module* built on Mamba architecture for temporal sequence processing, and (3) A *Reward/Termination Predictor* implemented through linear layers. The EDELINE framework uses shared hidden representations across the components for efficient world model learning.

tion. This hybrid design maintains temporal consistency while generating high-quality visual predictions. A Convolutional Neural Network based actor processes these predicted frames to determine actions, thus enabling autoregressive generation of imagined trajectories for policy optimization.

3.1. World Model Learning

The core architecture of EDELINE consists of a *Recurrent Embedding Module (REM)* f_ϕ that processes the history of observations and actions $(o_0, a_0, o_1, a_1, ..., o_t, a_t)$ to generate a hidden embedding h_t through recursive computation. This embedding enables the *Next-Frame Predictor* p_ϕ to generate predictions of the subsequent observation $\hat{o}_{t+1}$. The architecture further incorporates dedicated *Reward and Termination Predictors* to estimate the reward $\hat{r}_t$ and episode termination signal $\hat{d}_t$ respectively. The trainable components of EDELINE's world model are formalized as:

- Recurrent Embedding Module: $h_t = f_\phi(h_{t-1}, o_t, a_t)$
- Next-Frame Predictor: $\hat{o}_{t+1} \sim p_\phi(\hat{o}_{t+1}|h_t)$
- Reward Predictor: $\hat{r}_t \sim p_\phi(\hat{r}_t|h_t)$
- Termination Predictor: $\hat{d}_t \sim p_\phi(\hat{d}_t|h_t)$

3.1.1. Recurrent Embedding Module

While DIAMOND, the current state-of-the-art in diffusion-based world models, relies on a fixed context window of four previous observations and actions sequence, the proposed EDELINE architecture advances beyond this limitation through a recurrent architecture for extended temporal sequence processing. At each timestep t, the Recurrent Embedding Module processes the current observation-action pair (o_t, a_t) to update a hidden embedding $h_t = f_\phi(h_{t-1}, o_t, a_t)$. The implementation of REM utilizes Mamba (Gu & Dao, 2024b), an SSM architecture that offers distinct advantages for world modeling.

3.1.2. Next-Frame Predictor

While motivated by DIAMOND's success in diffusion-based world modeling, EDELINE introduces significant architectural innovations in its Next-Frame Predictor to enhance temporal consistency and feature integration. At time step t, the model conditions on both the last L frames and the hidden embedding h_t from the Recurrent Embedding Module to predict the next frame $\hat{o}_{t+1}$. The predictive distribution $p_\phi(o^0_{t+1}|h_t)$ is implemented through a denoising diffusion process, where D_ϕ functions as the denoising network. Let $y^\tau_t = (\tau, o^0_{t-L+1}, ..., o^0_t, h_t)$ represent the conditioning information, where τ represents the diffusion time. The denoising process can be formulated as $o^0_{t+1} = D_\phi(o^\tau_{t+1}, y^\tau_t)$. To effectively integrate both visual and hidden information, D_ϕ employs two complementary conditioning mechanisms. First, the architecture incorporates Adaptive Group Normalization (AGN) (Zheng et al., 2020) layers within each residual block to condition normalization parameters on the hidden embedding h_t and diffusion time τ, which establishes context-aware feature normalization (Zheng et al., 2020). This design significantly extends DIAMOND's implementation, which limits AGN conditioning to τ and action embeddings only. The second key innovation introduces cross-attention blocks inspired by Latent Diffusion Models (LDMs), which utilize h_t and τ as context vectors. The UNet's feature maps generate the query, while h_t and τ project to keys and values. This novel attention mechanism, which is absent in DIAMOND, facilitates the fusion of spatial-temporal features with abstract dynamics encoded in h_t. The observation modeling loss $\mathcal{L}_{\text{obs}}(\phi)$ is defined based on Eq. (7), and can be formulated as follows:

$$\mathcal{L}_{\text{obs}}(\phi) = \mathbb{E}\left[\|D_\phi(o^\tau_{t+1}, y^\tau_t) - o^0_{t+1}\|^2\right]. \tag{2}$$

3.1.3. Reward / Termination Predictor

EDELINE advances beyond DIAMOND's architectural limitations through an integrated approach to reward and termination prediction. Rather than employing separate neural networks, EDELINE leverages the rich representations from its REM. The reward and termination predictors are implemented as multilayer perceptrons (MLPs) that utilize the deterministic hidden embedding h_t as their conditioning input. This architectural unification enables efficient representation sharing across all predictive tasks. EDELINE processes both reward and termination signals as probability distributions conditioned on the hidden embedding: $p_\phi(\hat{r}_t|h_t)$ and $p_\phi(\hat{d}_t|h_t)$ respectively. The predictors are

optimized via negative log-likelihood losses, expressed as:

$$\mathcal{L}_{\text{rew}}(\phi) = -\ln p_\phi(r_t|h_t), \mathcal{L}_{\text{end}}(\phi) = -\ln p_\phi(d_t|h_t). \quad (3)$$

This unified architectural design represents an improvement over DIAMOND's separate network approach, where reward and termination predictions require independent representation learning from the world model. The integration of these predictive tasks with shared representations enables REM to learn dynamics that encompass all relevant aspects of the environment. The architectural efficiency facilitates enhanced learning effectiveness and better performance.

3.1.4. EDELINE World Model Training

The world model integrates an innovative end-to-end training strategy with a self-supervised approach. EDELINE extends the harmonization technique from HarmonyDream (Ma et al., 2024) through the adoption of harmonizers w_o and w_r, which dynamically balance the observation modeling loss $\mathcal{L}_{\text{obs}}(\phi)$ and reward modeling loss $\mathcal{L}_{\text{rew}}(\phi)$. This adaptive mechanism results in the total loss function $\mathcal{L}(\phi)$:

$$\begin{aligned}\mathcal{L}(\phi) = w_0\mathcal{L}_{\text{obs}}(\phi) + w_r\mathcal{L}_{\text{rew}}(\phi) + \mathcal{L}_{\text{end}}(\phi) \\ + \log(w_o^{-1}) + \log(w_r^{-1})\end{aligned} \quad (4)$$

3.2. Agent Behavior Learning

To enable fair comparison and demonstrate the effectiveness of EDELINE's world model architecture, the agent architecture adopts the same optimization framework as DIAMOND. Specifically, the agent integrates policy π_θ and value V_θ networks with REINFORCE value baseline and Bellman error optimization using λ-returns (Alonso et al., 2024). The training framework executes a procedure with three key phases: experience collection, world model updates, and policy optimization. This method follows the established paradigms in model-based RL literature (Łukasz Kaiser et al., 2020; Hafner et al., 2020; Micheli et al., 2023; Alonso et al., 2024). To ensure reproducibility, we provide extensive details in the Appendix, with documentation of objective functions in Appendix B.

4. Experiments

This section presents our experimental results of EDELINE on the Crafter benchmark.

4.1. Crafter Experiments

To evaluate EDELINE's memory enhancement capabilities, we conducted experiments on Crafter (Hafner, 2022), a procedurally generated survival environment that presents complex memory challenges. Crafter was specifically designed to assess "wide and deep exploration, long-term reasoning and credit assignment, and generalization" (Hafner et al., 2024), which establishes it as an ideal benchmark for the evaluation of an agent's long-term memory utilization capabilities.

Table 1: Comparison of different methods on Crafter in terms of average return and world model parameter count.

Method	Avg Return	#World Model Params
EDELINE	**11.5 ± 0.9**	**11M**
DreamerV3 XL	9.2 ± 0.3	200M
Δ-IRIS	7.7 ± 0.5	25M
DreamerV3 M	6.2 ± 0.5	37M
IRIS	5.5 ± 0.7	48M
DIAMOND	2.8 ± 0.5	**10.4M**

Crafter requires substantial memory capabilities due to its demand for agents to retain information about previously collected resources, crafted items, and explored territories for optimal decision-making, which establishes it as an ideal testbed for the evaluation of our memory-enhanced architecture. Within a 1M environment step budget, EDELINE achieves superior performance compared to state-of-the-art baselines including DIAMOND (Alonso et al., 2024), DreamerV3 (Hafner et al., 2024), Δ-IRIS (Alonso et al., 2023), and IRIS (Micheli et al., 2023), despite its relatively modest parameter count of 11M. These results highlight the significant advantages resulting from the integration of Mamba's memory capabilities with diffusion's generative abilities.

Table 1 presents our experimental results with a 1M environment step budget. The results reveal that EDELINE significantly outperforms all baselines with 25% higher returns than DreamerV3 XL despite the utilization of 18× fewer parameters. Most importantly, EDELINE delivers a 4.1× improvement over DIAMOND with a comparable parameter count, which demonstrates the substantial benefits of our enhanced memory mechanism.

5. Conclusions

In this work, we addressed the limitations of current diffusion-based world models in handling long-term dependencies and maintaining prediction consistency. Through the integration of Mamba SSMs, EDELINE effectively processed extended observation-action sequences through its recurrent embedding module, which enabled adaptive memory retention beyond fixed-context approaches. The unified framework eliminated architectural separation between observation, reward, and termination prediction, which fostered efficient representation sharing. Dynamic loss harmonization further mitigated optimization conflicts arising from multi-task learning. Our evaluation on Crafter demonstrates EDELINE's superior performance in memory-demanding environments. The results validate EDELINE's ability to maintain long-term spatial awareness and environmental consistency, highlighting the effectiveness of combining state space models with diffusion-based world modeling for reinforcement learning applications.

Acknowledgements

The authors gratefully acknowledge support from the National Science and Technology Council (NSTC) in Taiwan under grant numbers MOST 111-2223-E-002-011-MY3, NSTC 113-2221-E-002-212-MY3, and NSTC 113-2640-E-002-003. We also express our sincere appreciation to NVIDIA Corporation and the NVIDIA AI Technology Center (NVAITC) for the donation of GPUs and access to the Taipei-1 supercomputer. Furthermore, we thank the National Center for High-Performance Computing (NCHC) for providing computational and storage resources.

References

Alonso, E., Micheli, V., and Fleuret, F. Towards efficient world models. In *Workshop on Efficient Systems for Foundation Models @ ICML2023*, 2023.

Alonso, E., Jelley, A., Micheli, V., Kanervisto, A., Storkey, A., and et al. Diffusion for world modeling: Visual details matter in atari. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, 2024.

Åström, K. J. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965.

Caruana, R. Multitask learning. *Machine Learning*, 28: 41–75, 1997.

Chao, C.-H., Sun, W.-F., Cheng, B.-W., Lo, Y.-C., Chang, C.-C., Liu, Y.-L., Chang, Y.-L., Chen, C.-P., and Lee, C.-Y. Denoising likelihood score matching for conditional score-based data generation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=LcF-EEt8cCC.

Chifu, Y., Shuang, G., and Zhu, X. Improving the closed-loop tracking performance using the first-order hold sensing technique with experiments. arXiv:1801.01263, 2018.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Gu, A. and Dao, T. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2024a.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *Proc. Int. Conf. on Language Modeling (CoLM)*, 2024b.

Gu, A., Dao, T., Ermon, S., Rudra, A., and Ré, C. Hippo: Recurrent memory with optimal polynomial projections. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, 2020.

Gu, A., Goel, K., and Ré, C. Efficiently modeling long sequences with structured state spaces. In *Int. Conf. on Learning Representations (ICLR)*, 2022a.

Gu, A., Gupta, A., Goel, K., and Ré, C. On the parameterization and initialization of diagonal state space models. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, 2022b.

Ha, D. and Schmidhuber, J. Recurrent world models facilitate policy evolution. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, 2018.

Hafner, D. Benchmarking the spectrum of agent capabilities. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=1W0z96MFEoH.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020.

Hafner, D., Lillicrap, T. P., Norouzi, M., and Ba, J. Mastering atari with discrete world models. In *International Conference on Learning Representations (ICLR)*, 2021.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models. arXiv:2301.04104, 2024.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022.

Lee, J.-H., Lin, B.-J., Sun, W.-F., and Lee, C.-Y. Edeline: Enhancing memory in diffusion-based world models via linear-time sequence modeling, 2025. URL https://arxiv.org/abs/2502.00466.

Ma, H., Wu, J., Feng, N., Xiao, C., Li, D., HAO, J., Wang, J., and Long, M. Harmonydream: Task harmonization inside world models. In *Forty-first International Conference on Machine Learning (ICML)*, 2024.

Micheli, V., Alonso, E., and Fleuret, F. Transformers are sample-efficient world models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., and et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 2020.

Smith, J. T., Warrington, A., and Linderman, S. Simplified state space layers for sequence modeling. In *Int. Conf. on Learning Representations (ICLR)*, 2023.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and et al. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. The MIT Press, 2018.

Ye, W., Liu, S., Kurutach, T., Abbeel, P., and Gao, Y. Mastering atari games with limited data. In *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, 2021.

Zheng, H., Fu, J., Zeng, Y., Luo, J., and Zha, Z.-J. Learning semantic-aware normalization for generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Łukasz Kaiser, Babaeizadeh, M., Miłos, P., Osiński, B., Campbell, R. H., and et al. Model based reinforcement learning for atari. In *International Conference on Learning Representations (ICLR)*, 2020.

A. Additional Background Material Section

A.1. Score-based Diffusion Generative Models

Diffusion probabilistic modeling (Sohl-Dickstein et al., 2015; Ho et al., 2020; Dhariwal & Nichol, 2021) and score-based generative modeling (Song & Ermon, 2019; 2020; Chao et al., 2022) can be unified through a forward stochastic differential equation (SDE) formulation (Song et al., 2021). The forward diffusion process $\{\mathbf{x}^\tau\}$ with continuous time variable τ transforms the data distribution $p^0 = p^{\text{data}}$ to prior distribution $p^T = p^{\text{prior}}$, expressed as:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, \tau)\mathrm{d}\tau + g(\tau)\mathrm{d}\mathbf{w}, \tag{5}$$

where $\mathbf{f}(\mathbf{x}, \tau)$ represents the drift coefficient, $g(\tau)$ denotes the diffusion coefficient, and $\mathbf{w}$ is the Wiener process. The corresponding reverse-time SDE can then be formulated as:

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, \tau) - g(\tau)^2 \nabla_{\mathbf{x}} \log p^\tau(\mathbf{x})\right] \mathrm{d}\tau + g(\tau)\mathrm{d}\bar{\mathbf{w}}, \tag{6}$$

where $\bar{\mathbf{w}}$ is the reverse-time Wiener process. Eq. (6) enables sampling from p^0 when the (Stein) score function $\nabla_{\mathbf{x}} \log p^\tau(\mathbf{x})$ is available. A common approach to estimate the score function is through the introduction of a denoiser D_θ, which is trained to minimize the following objective:

$$\mathbb{E}_{\sigma \sim p^{\text{train}}} \mathbb{E}_{\mathbf{x}^0 \sim p^{\text{data}}} \mathbb{E}_{\mathbf{n} \sim \mathcal{N}(0, \sigma^2 I)} \left[\|D_\theta(\mathbf{x}^0 + \mathbf{n}; \sigma) - \mathbf{x}^0\|_2^2\right], \tag{7}$$

where $\mathbf{n}$ is Gaussian noise with zero mean and variance determined by a variance scheduler $\sigma(\tau)$ that follows a noise distribution p^{train}, and $(\mathbf{x}^0 + \mathbf{n})$ corresponds to the perturbed data $\mathbf{x}^\tau$. The score function can then be estimated through: $\nabla_{\mathbf{x}} \log p^\tau(\mathbf{x}) = \frac{1}{\sigma^2}(D_\theta(\mathbf{x}; \sigma) - \mathbf{x})$. In practice, modeling the denoiser D_θ directly can be challenging due to the wide range of noise scales. To address this, EDM (Karras et al., 2022) introduces a design space that isolates key design choices, including preconditioning functions $\{c_{\text{skip}}, c_{\text{out}}, c_{\text{in}}, c_{\text{noise}}\}$ to modulate the unconditioned neural network F_θ to represent D_θ, which can be formulated as:

$$D_\theta(\mathbf{x}; \sigma) = c_{\text{skip}}(\sigma)\mathbf{x} + c_{\text{out}}(\sigma) F_\theta(c_{\text{in}}(\sigma)\mathbf{x}; c_{\text{noise}}(\sigma)). \tag{8}$$

The preconditioners serve distinct purposes: $c_{\text{in}}(\sigma)$ and $c_{\text{out}}(\sigma)$ maintain unit variance for network inputs and outputs across noise levels, $c_{\text{noise}}(\sigma)$ provides transformed noise level conditioning, and $c_{\text{skip}}(\sigma)$ adaptively balances signal mixing. This principled framework improves the robustness and efficiency of diffusion models, enabling state-of-the-art performance across various generative tasks.

A.2. Multi-task Essence of World Model Learning

Modern world models (Łukasz Kaiser et al., 2020; Hafner et al., 2021; 2024; Alonso et al., 2024) typically address two fundamental prediction tasks: the modeling of environment dynamics through observations and the prediction of reward signals. The learning of these tasks requires distinct considerations based on the complexity of the environment. In simple low-dimensional settings, separate learning approaches suffice for each task. However, the introduction of high-dimensional visual inputs fundamentally alters this paradigm, as partial observability creates an inherent coupling between state estimation and reward prediction. This coupling necessitates joint learning through shared representations, an approach that aligns with established multi-task learning principles (Caruana, 1997). The implementation of such joint learning through shared representations introduces several technical challenges. The integration of multiple learning objectives requires careful consideration of their relative importance and interactions. A fundamental difficulty stems from the inherent scale disparity between high-dimensional visual observations and scalar reward signals. This disparity manifests in the world model learning objective, which combines observation modeling $\mathcal{L}_o(\theta)$, reward modeling $\mathcal{L}_r(\theta)$, and dynamics modeling $\mathcal{L}_d(\theta)$ losses with weights w_o, w_r, w_d to control relative contributions:

$$\mathcal{L}(\theta) = w_o \mathcal{L}_o(\theta) + w_r \mathcal{L}_r(\theta) + w_d \mathcal{L}_d(\theta). \tag{9}$$

HarmonyDream (Ma et al., 2024) demonstrated that observation modeling tends to dominate this objective due to visual inputs' high dimensionality compared to scalar rewards. Their work introduced a variational formulation:

$$\mathcal{L}(\theta, w_o, w_r, w_d) = \sum_{i \in \{o,r,d\}} \mathcal{H}(\mathcal{L}_i(\theta), \frac{1}{w_i}) = \sum_{i \in \{o,r,d\}} w_i \mathcal{L}_i(\theta) + \log(\frac{1}{w_i}), \tag{10}$$

where $\mathcal{H}(\mathcal{L}_i(\theta), w_i) = w_i \mathcal{L}_i(\theta) + \log(1/w_i)$ dynamically balances the losses by maintaining $\mathbb{E}[w^* \cdot \mathcal{L}] = 1$. This harmonization technique can substantially enhance sample efficiency and performance. Our work extends these insights through the integration of dynamic task balancing mechanisms into our EDELINE world model architecture.

B. Actor-Critic Learning Objectives

We follow DIAMOND (Alonso et al., 2024) in the design of our agent behavior learning. Let o_t, r_t, and d_t denote the observations, rewards, and boolean episode terminations predicted by our world model. We denote H as the imagination horizon, V_θ as the value network, π_θ as the policy network, and a_t as the actions taken by the policy within the world model.

For value network training, we use λ-returns to balance bias and variance in the regression target. Given an imagined trajectory of length H, we define the λ-return recursively:

$$\Lambda_t = \begin{cases} r_t + \gamma(1-d_t)[(1-\lambda)V_\theta(o_{t+1}) + \lambda\Lambda_{t+1}] & \text{if } t < H \\ V_\theta(o_H) & \text{if } t = H. \end{cases} \tag{11}$$

The value network V_θ is trained to minimize $\mathcal{L}_V(\theta)$, the expected squared difference with λ-returns over imagined trajectories:

$$\mathcal{L}_V(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{H-1}(V_\theta(\mathbf{x}_t) - \text{sg}(\Lambda_t))^2\right], \tag{12}$$

where $\text{sg}(\cdot)$ denotes the gradient stopping operation, following standard practice (Hafner et al., 2024; Micheli et al., 2023).

For policy training, we leverage the ability to generate large amounts of on-policy trajectories in imagination using a REINFORCE objective (Sutton & Barto, 2018). The policy is trained to minimize:

$$\mathcal{L}_\pi(\theta) = -\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{H-1}\log(\pi_\theta(a_t|o_{\leq t}))\text{sg}(\Lambda_t - V_\theta(o_t)) + \eta\mathcal{H}(\pi_\theta(a_t|o_{\leq t}))\right], \tag{13}$$

where $V_\theta(o_t)$ serves as a baseline to reduce gradient variance, and the entropy term $\mathcal{H}$ with weight η encourages sufficient exploration.