

HanoiWorld : A Joint Embedding Predictive Architecture Based World Model for Autonomous Vehicle Controller

Tran Tien Dat^{1,2}, Nguyen Hai An^{1,2}, Nguyen Khanh Viet Dung^{1,2}, Nguyen Duy Duc^{1,2}

*¹Faculty of Mathematics and Informatics, Hanoi University of Science and Technology,
Vietnam*

²Troy University, United States

*{Dat.TT228006, An.NH227993}@sis.hust.edu.vn
{Dung.NKV207947, Duc.ND218129}@sis.hust.edu.vn
dtran220296@troy.edu*

Abstract

Current attempts of Reinforcement Learning for Autonomous Controller are data-demanding while the result are under-performed, unstable and unable to grapple and anchoring on the concept of safety, and over-concentrate on noise feature dues to the nature of pixel reconstruction. While current Self-Supervised Learning approaches that learning on high-dimensional representation by leveraging the Joint Embedding Predictive Architecture (JEPA) is interesting and effective alternative, as the idea is mimicking the natural of human’s brain in acquiring new skill using imagination and minimal sample of observations. This study introduces HanoiWorld, a JEPA-based world model that using recurrent neural network (RNN) for making longterm horizontal planning with effective inference time. Experiments conducted on Highway-Env package with difference enviroment showcase the effective capability of making driving plan while safety-awareness with considerable collision rate in comparison with SOTA baselines.

1 Introduction

Since the first experiment on an Autonomous Vehicle (AV) was conducted in 1986 at Carnegie Mellon University [24], the research domain of developing self-driving vehicles has made significant progress, both in technical and practical applications in the real world. The vehicles are expected to operating safely while handling the challenges of uncertainty, partial observability, and multi-agent enviroment (interaction between the ego-vehicle and the surroundings vehicles, and obstacle as pedestrians, etc.) [33, 67]. However, as Dulac-Arnold, Mankowitz, and Hester [15] suggest that these challenges does limit the feasibility on deploying and experimenting on such reinforced-learning based

controller, and prior work only based on naive transferring paradigms from the physical-environment which lead to training instability due to noise, and data-fragmentation [12, 15, 34].

From the technical perspective, the decision of AV have traditionally relied on simulator for experience rollout, and the planning algorithms with reinforcement learning with the assumption of data-abundance as Hessel et al. [29] and Hoel et al. [30], whereas, the classical attempts - such as Monte Carlo Tree Search (MCTS) or belief-space planning under partially observable Markov decision processes (POMDPs) require massive computational overhead within the simulator for experiment rolling-out without doing the policy training Han et al. [27], Huang et al. [32], and Morimura et al. [45], while these method does offering for resolving the uncertainty, but the scalability is limited [44]. Additionally such rolling-out strategy tend to amplifying the error over-the-longtime horizon, and lead to the raise of the model inaccuracy Robine, Uelwer, and Harmeling [46] and Xiao et al. [61]. Additionally, the observation-level prediction tends to prioritize the visual, and kinematic fidelity as the reconstruction challenge, which may not truly encapsulated the decision-relevant manifold, and leading to inefficiencies in object control [1, 36, 48].

Inspiration from human’s capability of acquiring new skill as driving by leveraging the capability of imaginary on the plausible future scenario based on the current interaction with the environment - the affordance based theory, and human memory Gibson [21], Ha and Schmidhuber [23], and LeCun [36], which can be formalized into the model-designing implementation as using representation using the self-supervised learning paradigm for learning the environment dynamicity. Joint-Embedding Predictive Architectures (JEPA) propose learning latent spaces by directly predicting future representations, without reconstructing raw observations, but only structure alignment between encoder and enforcing the information variability, but not stochasticity on noise for preventing the embedding collapse as Bardes, Ponce, and LeCun [5]. State-of-the-art model as V-JEPA-2 extends the idea to large-scale video data for learning the action dynamicity in the yielded representation from the passive-observation, which can be used later for producing training signal for lightweight action-conditioned controller [1]. In parallel, the recurrent state-space models (RSSMs) have been shown to provide an effective mechanism for maintaining compact latent memories that approximate Markovian dynamics under partial observability with the capability of long-term planning using minimal amount of representation [25, 26]. The attempts of using encoder with long-term RSSM planning model show-case the efficiency and overhead-minimalizing with increasing in training utility than MCTS based approach, while still yield out sensory grounded action from agents.

Based on these aforementioned works, this result argue that *world-model designing can be potential benefit from the high-quality self-supervised learning embedding from pre-*

trained encoder as V-JEPA 2 and combine with the usage of long-term planner which can reduce and minimize the cost of inference while remaining accuracy, and tunable model driving quality.

The contribution of this studies include 4 keys essential contributions as follow:

- A unified perspective on world-model design for autonomous vehicles that emphasizes predictive, representation-level modeling over observation-level simulation which is called HanoiWorld.
- Suggesting a JEPA-based encoding strategy, inspired by V-JEPA-2 fine-tuning, for learning decision-relevant latent representations from large-scale video data.
- The integration of an RSSM-based latent memory to support approximate Markovian state transitions under partial observability.
- A demonstration that a simple MLP-based actor–critic controller can be trained effectively within the learned latent world model, avoiding expensive planning algorithms and complex policy architectures.

The rest of the paper shall be showcased as follows; Section 2 focusing on the related works and provide the whole conceptual and theoretical foundation on the challenges and related solution; Section 3 discusses the suggested proposed world-model designing; Section 4 will attempts to provide the experiment description, the usecase and result discussion. Finally, the report shall be conclude on Section 5.

Additionally, we will release the experimental codebase to facilitate reproducibility at HANOI–WORLD codebase.

2 Related Works

World Model have been proposed as the novel-approach as the novel solution for training the reinforcement learning based controller dues to the unreliability, excessiveness, and inefficent of the model that based on real-world interaction as Dulac-Arnold, Mankowitz, and Hester [15] suggested, while the vehicle constantly work under occlusion with unlimited knowledge on the world trigger the problems of inevitable uncertainty. These challenges co-align and trigger the need of reframe the training approach for the reinforcement-learning controller using the compact while semantic vivid representation for attaining the scalability and reliability that inspired on human’s biological mechanism on learning using affordance and imaginary [21, 36].

2.1 From Model-Based Reinforcement Learning to Driving World Models

The conceptual roots of driving world models lie in model-based reinforcement learning (MBRL), where an agent learns a dynamics model and then plans by simulating futures. Progress in MBRL was enabled by shared tooling and standardized benchmarks. OpenAI Gym [8] made it easy to compare algorithms across tasks via a uniform API, while the DeepMind Control Suite [52] provided a curated set of continuous-control environments that encouraged rigorous evaluation of learning and control. These infrastructures fostered iterative improvements in learning dynamics models, representation learning, and planning.

Modern latent-dynamics agents exemplify the world model as an imagination engine viewpoint. DreamerV3 [25] demonstrates strong and stable performance across a wide range of environments by learning a compact latent state and optimizing behavior via imagined rollouts. Although most Dreamer-style results are reported outside real driving, the core design principles—predictive latent state, stochastic dynamics, and planning or policy improvement through imagined trajectories—strongly influence driving world model designs. Complementary perspectives on embodied agent design and generalization in RL emphasize that robustness and scalable training protocols matter as much as raw model capacity [28]. In autonomous driving, these principles interact with additional constraints: safety, distribution shift, long-horizon decision-making, and multi-agent interactions.

A growing line of work argues that the value of a world model should be judged by downstream utility (e.g., improved planning or safer decisions) rather than by generative fidelity alone. Planning-centric views highlight that an agent can exploit imperfections in a learned model, producing good rollouts that do not correspond to the real world. Analyses of embodied world models stress safety as a first-class concern and call for evaluation protocols that expose failure modes, especially those that emerge only in closed-loop control [4]. These concerns become acute in autonomous driving, where rare events dominate risk and a small modeling error can cascade into catastrophic outcomes.

2.2 Self-Supervised and Predictive Representation Learning for Driving

Autonomous driving provides abundant unlabeled sensor streams but comparatively limited dense annotations, motivating self-supervised learning (SSL) as a foundation for world models. In vision, SSL matured from contrastive and clustering-based approaches to predictive and distillation-based schemes. DINO [10] showed that self-distillation without labels can learn semantically meaningful features, and such ideas have inspired

driving-specific pretraining efforts that seek transferable representations across time, viewpoint, and weather.

However, driving data introduces distinct pitfalls for generic SSL. Contrastive learning requires defining positive pairs that represent the same underlying content under augmentations; in driving scenes with many objects and rapid ego-motion, naive augmentations can destroy correspondence and lead to negative transfer. Generative SSL that reconstructs masked inputs can be expensive for 3D data and may force the model to predict arbitrary surface details rather than planning-relevant semantics. Several recent works therefore advocate embedding-level prediction and variance regularization as alternatives to either contrastive pairs or explicit reconstruction [41, 42, 70].

A particularly influential conceptual framework is the Joint Embedding Predictive Architecture (JEPA) viewpoint, which proposes learning representations by predicting the embeddings of unknown parts of the input given the known parts, rather than reconstructing pixels or using negative pairs [36]. Driving provides a compelling application: masked regions in LiDAR or camera space may correspond to multiple plausible surfaces, yet the semantics (e.g., rear of a car, free space behind a truck) can remain stable in an embedding space. JEPA-style methods can therefore better align with the uncertainty intrinsic to partial observability. For example, JEPA-based LiDAR pretraining predicts BEV embeddings for masked regions and uses explicit variance regularization to prevent representation collapse, yielding consistent gains in downstream 3D detection while reducing pretraining compute relative to dense reconstruction [70].

World models also benefit from discrete or structured latent spaces that stabilize learning and improve sample efficiency. Vector-quantized representations provide one path, but codebook collapse can limit capacity. Online codebook learning strategies such as clustering-based VQ updates aim to keep all codevectors active, improving utilization and reconstruction/generation quality [69]. In driving, discretized latents may improve controllability, support efficient rollouts, and provide a bridge between geometric and semantic factors.

2.3 Spatial World States: BEV, Occupancy, and Geometric Abstractions

Many driving world models adopt spatially grounded world states rather than purely abstract latent vectors. BEV representations offer a convenient coordinate frame that aligns with planning: it naturally represents lanes, drivable space, and other agents, and it facilitates sensor fusion. BEV representations also reduce the burden of viewpoint variation, enabling models to focus on dynamics rather than perspective transformations. Consequently, BEV features are widely used as intermediate states for both perception and prediction, and they serve as a natural substrate for world modeling [38, 39].

Occupancy-based representations extend BEV by modeling 3D free space and occlusion, which are critical for safety. A world model that predicts occupancy can support collision checking, visibility reasoning, and planning under uncertainty. Recent LiDAR-oriented world models stress that camera-only generation may produce visually plausible but geometrically inconsistent futures, whereas occupancy prediction can enforce physical constraints and preserve 3D structure [6, 37]. These works highlight the importance of representing not just objects but also empty space, since the absence of obstacles is as planning-relevant as their presence.

Geometric abstractions are also closely tied to mapping and scene priors. High-definition maps encode lane topology, boundaries, and crosswalks, and several world modeling pipelines treat maps as part of the world state, either as conditioning signals for generation or as latent factors to be predicted. Methods that jointly reason about agent trajectories and map structure aim to ensure that generated futures obey road geometry and traffic rules [58, 68]. In addition, surveys of physical world models emphasize that effective world models should capture not only statistical regularities but also physically grounded structure and causal relations, particularly when extrapolating beyond the training distribution [12].

2.4 Transformers, Attention, and Interaction-Centric Modeling

Transformers and attention mechanisms have become central to autonomous driving because they support long-range dependencies and flexible fusion across heterogeneous inputs. In perception and prediction, attention helps focus computation on the most relevant actors and regions of the scene, and it provides a natural way to model interactions among agents. Transformer-based architectures are therefore widely used in modern driving world models, especially when combining multi-view images, point clouds, and map features [56].

Interaction-centric modeling is essential because other agents react to each other and to the ego vehicle. World models that treat agents as independent can systematically fail in dense traffic, merges, intersections, and other interactive contexts. Recent work emphasizes representations that capture agent-agent coupling, intent, and right-of-way, often using attention or graph-style message passing to represent joint futures [16, 40]. These ideas are consistent with broader trends in embodied learning, where the world model must represent not only passive dynamics but also the consequences of actions and the strategic responses of others.

Transformers are also influential in self-supervised pretraining and in building scalable foundation-style models that transfer across tasks. Papers exploring large-scale training regimes for world models argue that representation, prediction, and planning can be co-trained when the model is sufficiently expressive and the training data is diverse [60,

62]. This perspective connects to the wider discussion of how to unify perception and control under a single learned model, and it motivates architectures that can condition on both sensory history and action sequences.

2.5 Multi-Agent and V2X World Models: Cooperative Perception to Cooperative Prediction

Autonomous driving is inherently multi-agent, and connected autonomy introduces additional channels of information via V2X communication. Cooperative perception is an early instance of distributed world modeling, where multiple vehicles or infrastructure sensors share data or features to reduce occlusion and extend sensing range. Public benchmarks for cooperative perception have enabled systematic evaluation of fusion strategies, including early, late, and intermediate fusion [63]. Intermediate feature sharing is often favored as a balance between accuracy and bandwidth, and it has motivated learned fusion modules that can tolerate localization noise and intermittent communication.

Transformer-based fusion architectures extend cooperative perception by enabling attention over agents and multi-scale features. V2X-focused transformers incorporate mechanisms to handle pose uncertainty, varying sensor modalities, and temporal misalignment due to communication delay [64]. These settings highlight a key difference between single-agent and multi-agent world models: the state is not merely what the ego sees, but a distributed set of partial observations that must be reconciled into a coherent representation.

Real-world datasets are crucial for validating these ideas because simulation can underrepresent sensor artifacts and the true distribution of road interactions. V2V4Real provides real-world multi-vehicle data with LiDAR, RGB, 3D bounding boxes, and HD maps, designed explicitly for cooperative perception tasks such as cooperative detection, tracking, and sim-to-real adaptation [65]. Such datasets suggest that future world models for autonomous driving should be designed from the start to support multi-agent fusion and prediction, rather than retrofitting single-agent models to cooperative settings.

Cooperative world models must go beyond current-state fusion to future-state prediction under shared information. This entails modeling how distributed observations evolve, how agents may act in response to each other, and how communication delays affect belief updates. Works that study cooperative forecasting and interaction in distributed settings argue for robust, uncertainty-aware fusion and for models that degrade gracefully when messages are delayed or missing [53, 54]. This line of research also connects to simulators and benchmark environments that can test communication-aware policies under controlled conditions [3, 7].

2.6 Safety-Critical Modeling: Accident Prediction, Risk Anticipation, and Vulnerable Road Users

A primary promise of world models in driving is improved anticipation of rare and safety-critical events. Traditional trajectory prediction benchmarks emphasize average errors on non-critical behavior, which may not correlate with accident risk. Accident prediction benchmarks therefore play an important role in driving world model evaluation. DeepAccident introduces motion and accident prediction in V2X contexts, targeting the challenging setting where the model must predict not only where agents will move, but whether and when a collision will occur and which participants will be involved [57]. Such benchmarks force world models to represent risk factors that may be subtle or long-range, such as occluded cross traffic or rapidly changing gaps.

Risk anticipation also motivates explicit modeling of occlusion and free space, which are critical to collision avoidance. Occupancy-based world models and LiDAR-centric methods emphasize that representing unknown regions and visibility is crucial: predicting an agent behind an occluder is a fundamentally different problem from predicting a visible agent. Recent risk-oriented world modeling studies therefore combine geometric priors with predictive uncertainty, aiming to represent multiple plausible futures rather than a single deterministic trajectory [6, 22].

Safety also depends on modeling vulnerable road users (VRUs) such as pedestrians and cyclists, especially in dense, mixed traffic. High-resolution trajectory datasets that include VRUs provide essential supervision for interaction-aware world models. On-SiteVRU offers high-density trajectories across complex urban scenarios with fine temporal resolution and contextual information, enabling research on VRU behavior modeling, interaction risk, and safety evaluation [66]. These datasets complement vehicle-centric benchmarks and highlight that world models must reason about heterogeneous participants with different kinematics, goals, and social norms.

Finally, recent analyses argue that safety evaluation for world models should be multi-dimensional: generative realism is insufficient, and closed-loop planning tests can expose model exploitation or compounding errors. The safety challenge perspective calls for stress tests, counterfactual evaluation, and metrics that quantify whether planning with the model improves safety outcomes under distribution shift [4]. This theme strongly motivates research into world models that are not only expressive but also calibrated, robust, and auditable.

2.7 Generative Simulation: Diffusion, Video World Models, and Language-Conditioned Scenario Generation

Generative world models have expanded rapidly, driven by diffusion models and advances in controllable video generation. In autonomous driving, video generation is appealing because it can synthesize diverse scenes, including rare events, that are hard to capture in real data. Several works frame driving world modeling as conditional generation of future observations given current context and candidate actions, often incorporating maps, bounding boxes, or trajectories as conditioning signals [13, 18]. Surveys and methodological papers in this area emphasize that diffusion-based dynamics modeling can produce high-fidelity samples and can incorporate structured priors, but they also note challenges in temporal consistency and action controllability [4, 12].

DriveDreamer-2 demonstrates a particularly influential direction: integrating language models with world models to enable user-driven simulation. In DriveDreamer-2, an LLM converts user prompts into agent trajectories, a diffusion-based component generates an HD map consistent with those trajectories, and a unified multi-view video generator synthesizes multi-camera driving videos [68]. This pipeline aims to generate long-tail scenarios (e.g., abrupt cut-ins) in a user-friendly way and to improve downstream perception training. Related efforts such as DriveDreamer variants emphasize improving cross-view coherence and temporal stability and highlight how structured intermediate representations (trajectories, maps) can improve controllability [58].

Generative modeling is also used for counterfactual evaluation and data augmentation in planning-centric contexts. Dreamer-style rollouts provide imagined futures in latent space, while diffusion or video models provide rich sensory predictions. Integrative systems such as CarDreamer seek to couple world modeling with downstream decision-making and to use learned imagination as a training signal for driving policies [20]. Other large-scale training efforts explore how to train world models that support both prediction and planning, often combining self-supervised objectives with policy learning [1, 60]. Across these works, a recurring open problem is aligning generative fidelity with decision utility: a model can produce realistic-looking videos while still being unreliable for safety-critical planning.

2.8 Evaluation and Benchmarking: Utility, Generalization, and the Reality Gap

A persistent difficulty is that there is no single metric that fully captures the quality of a driving world model. Generative metrics as FID - Fr chet Inception Distance, or FVD - Fr chet Video Distance, which can be expressed as the Eq. (1) and Eq. (2). These metrics measure visual realism, however as may ignore physical plausibility or planning

relevance [4].

$$\text{FID}(\mathcal{X}_r, \mathcal{X}_g) = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|_2^2 + \text{Tr} \left(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2 (\boldsymbol{\Sigma}_r \boldsymbol{\Sigma}_g)^{\frac{1}{2}} \right) \quad (1)$$

$$\text{FVD}(\mathcal{V}_r, \mathcal{V}_g) = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|_2^2 + \text{Tr} \left(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2 (\boldsymbol{\Sigma}_r \boldsymbol{\Sigma}_g)^{\frac{1}{2}} \right) \quad (2)$$

Given that:

- \mathcal{X}_r and \mathcal{X}_g denote the sets of real and generated image samples, respectively.
- \mathcal{V}_r and \mathcal{V}_g denote the sets of real and generated video samples, respectively.
- $\boldsymbol{\mu}_r \in R^d$ and $\boldsymbol{\mu}_g \in R^d$ represent the empirical mean vectors of deep feature embeddings extracted from real and generated samples.
- $\boldsymbol{\Sigma}_r \in R^{d \times d}$ and $\boldsymbol{\Sigma}_g \in R^{d \times d}$ denote the empirical covariance matrices of the corresponding feature embeddings.
- $\|\cdot\|_2$ denotes the Euclidean (ℓ_2) norm.
- $\text{Tr}(\cdot)$ denotes the trace operator of a square matrix.
- $(\boldsymbol{\Sigma}_r \boldsymbol{\Sigma}_g)^{\frac{1}{2}}$ denotes the matrix square root of the product of the two covariance matrices.
- d denotes the dimensionality of the feature embedding space.

Trajectory metrics based as ADE - Average Distance Error, and FDE - Final Distance Error measure average error but can miss long-tail risk as Baraldi et al. [4] and Caesar [9] suggests. In addition, planning-centric evaluation measures downstream driving performance, but it requires closed-loop testing and can be confounded by simulator limitations. Recent works therefore argue for multi-axis evaluation that includes: (i) predictive accuracy and calibration, (ii) robustness under distribution shift, (iii) usefulness for downstream tasks such as detection, tracking, or planning, and (iv) safety-critical stress tests [4, 9].

Datasets and benchmarks shape progress by determining what is measurable. Waymo Open provides scale and diversity for training and evaluating models that must handle real sensor noise and complex urban scenarios [50]. Cooperative datasets such as V2V4Real add the challenges of multi-agent fusion, localization error, and communication constraints [65]. Accident-focused benchmarks like DeepAccident stress rare-event anticipation and interaction under occlusion [57]. VRU-focused datasets like OnSiteVRU emphasize mixed traffic and fine-grained interactions [66]. Together, these resources suggest that general driving world models must learn from diverse data sources and must be evaluated across diverse tasks.

The reality gap between simulation and the real world remains a central concern. Simulation allows scalable closed-loop testing, but it may underrepresent rare behaviors or sensor artifacts. Several works therefore emphasize sim-to-real adaptation, hybrid training (real + synthetic), and evaluation protocols that detect overfitting to simulator biases [39, 65]. Framework discussions and surveys highlight the need for standardized reporting and reproducibility across toolchains, since seemingly small implementation choices can dominate conclusions [12, 34].

2.9 Additional Consideration: Surveys, Physics, and Platform Considerations

Several additional threads in the provided corpus help connect autonomous-driving world models to broader scientific and engineering questions. First, surveys and position pieces emphasize that “world modeling” is an umbrella term spanning representations, learning objectives, and downstream uses, and they argue that progress depends on principled choices about what is modeled explicitly versus implicitly as [14, 36]. Recent survey-style works discuss how physical knowledge (dynamics, constraints, and causal structure) can be embedded into learned representations to improve robustness and interpretability [12, 13]. These discussions are particularly relevant in driving because failures often arise from violations of basic physical plausibility (e.g., impossible motion, inconsistent occlusion) or from spurious correlations in training data - and ones solution suggested by Kong et al. [35] and Sobal et al. [48] where integrate primitive and deterministic physical model of the world as microscopic approximation is considerable for learning the physical-affordance alignment as the purpose that world-model that is aim-at [21].

Furthermore, works of Audinys [3], Bouzaiene [7], Brockman et al. [8], and Tassa [52] strongly argue that evaluation in embodied domains is inseparable from the environment and platform used for training and testing. Practical benchmarking choices sensor suites, map availability, traffic participant diversity, and even simulator fidelity affect what a world model learns and how it generalizes. Platform- and tooling-oriented perspectives highlight that reproducible research requires careful specification of environments, data pipelines, and evaluation procedures. In driving, this connects to the well-known tension between rich simulation for closed-loop testing and real-world data for realism; which does suggest the novel hybrid pipelines that use simulation to explore long-tail scenarios and real data to anchor the model to real sensor statistics, which attempts in mimicing human’s capability of acknowledging the physical-affordance [21].

Finally, works as [1, 14, 36] in the World-Modeling showed that, difference modalities can lead to effective representations for downstream and continuous task training, however, such internal representation shall be robust under environment stochasticity

and unpredictability, which can be resolved by sequential-based reasoning model with latent-overshooting as Hafner et al. [26].

2.10 Design Implications

Across these literatures, several design lessons emerge, which does become our result theoretical foundation; First, predictive representation learning that operates in embedding space (e.g., JEPA-style objectives) is increasingly favored over pixel-level reconstruction in driving, because it better matches partial observability and the capability of learning the latent-representation across multi-modal futures [36, 70]. Second, spatially grounded states such as BEV and occupancy are practical and planning-aligned world states, particularly when fused with map priors and explicit modeling of free space [6, 37, 38]. Third, multi-agent interaction and V2X communication are becoming core requirements: a driving world model must reconcile distributed observations and predict joint futures under delay and uncertainty [53, 64, 65]. Finally, safety demands benchmarks that target rare events, accident prediction, shall be balance with efficiency and performance based metrics as reward is a challenges need to be address in later studies [4, 57, 66].

3 Proposed World Model and System Architecture

This section shall focus in suggestion in world model contruction, the whole agent interaction flow, and training algorithm that shall be leveraged

3.1 Notation

Our problem can be formalizing as the deterministic Markov Decision Process (MDP) with the usage of internal continuous stochastic states and derterministic state, and yeilding the deterministic continuous action in the finite-horizontal enviroment as [25, 49] suggested. The MDP can be defined as the tuples in Eq. (3)

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma), \quad (3)$$

where \mathcal{S} denotes the (possibly continuous) state space, \mathcal{A} denotes the (possibly continuous) action space, $\mathcal{P}(s_{t+1} | s_t, a_t)$ represents the state transition dynamics which including the stochastic states s_t capturing the enviroment randomness pattern within the transition, and determistic states h_t yeild from the recurrent-planner for historical context, $r : \mathcal{S} \times \mathcal{A} \rightarrow R$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor - and for the setting we are adopt directly the setting of DreamerV3 by setting $\gamma = 0.997$.

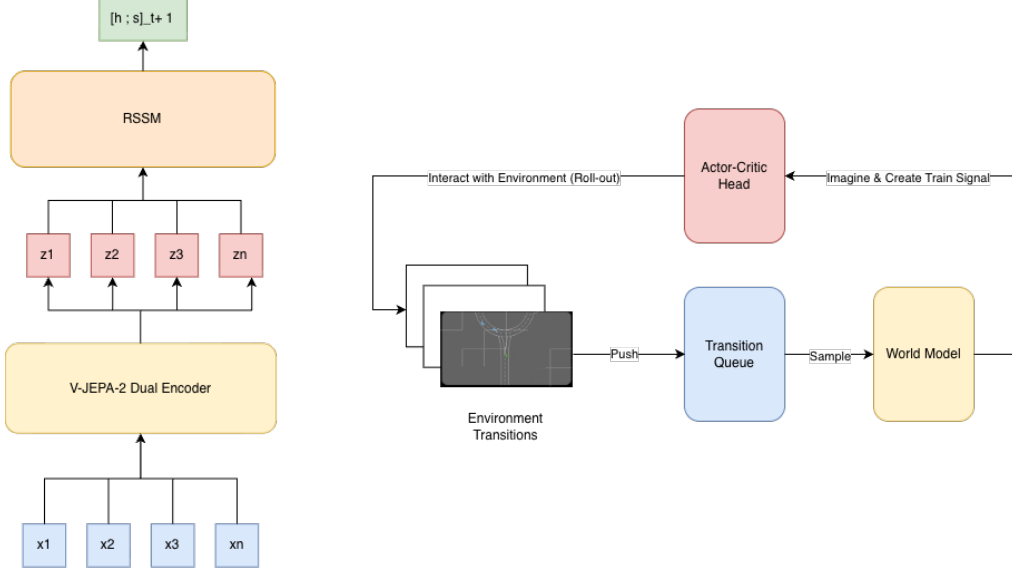


Figure 1: The proposed HanoiWorld World Model (left) include an visual encoder based on V-JEPA-2 checkpoint proposed by [2] and the RSSM-backbone suggested by [26] for making long-term planning on the next possible transition of enviroment - the green block. The overall system arrchitueural of the enviroment on the (right) which been design aiming for both effective rolling-out while model training from the enviroment by creating such feedback loop on the agent interaction with data-scarcity

The action space we attempts including the ego’s vehicle acceleration, and steering angle (in radian) as the response for making the interaction to the simmulation.

3.2 Overall System Architecture

The proposed overall system archtiecture include modules as enviroment interface, which build based on the HighwayEnv package provided by [17] for agent interaction, we additionally including the Transtion Queues with limited in size as the experience collector for the World Model can be sampling randomly certain episode from the queue that have with satisfied sequence length T , the World Model shall provide the imaginary embedding as the train-signal toward the Actor-Critic controler head for yielding the action for reacting toward the enviroment and create an close interacting cycle until the agent reach the terminatin states, the architecture of the whole HanoiWorld is presented on the right side of Figure 1.

The transition queue shall storing the list of episode (eg. reference of valid episode) that have the length of transition satisfy certain threadshold - T , and the sequence shall contain the Bird-Eye-View RGB images and the metadata on the agent’s interaction as the agent’s action, ego’s actions, and the description on the reward been yield across the simulation for those sequence.

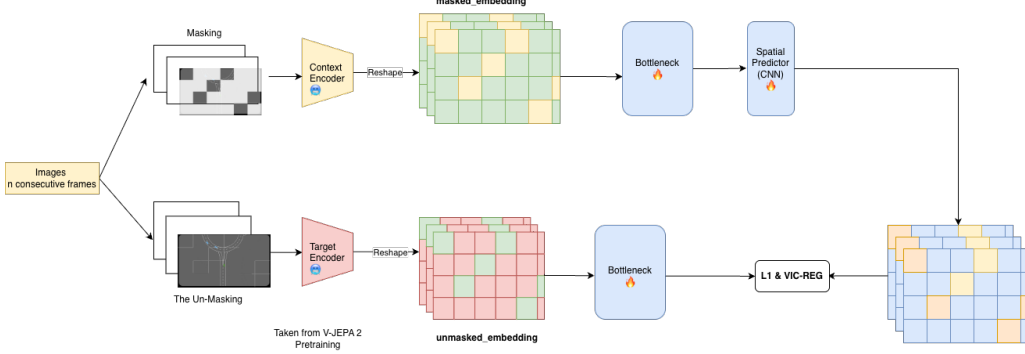


Figure 2: Overview of the proposed HanoiWorld encoder architecture. A pretrained and frozen V-JEPA 2 encoder [2] is used as a high-quality representation backbone to improve training efficiency and embedding robustness under limited-data settings. A downstream bottleneck Multi-Layer Perceptron (MLP) is trained to project the high-dimensional representations into a compact and task-compatible latent space of size 1024×128 . In parallel, the student encoder branch incorporates an additional 2D convolutional neural network (CNN) module to predict spatial representations. Both branches are jointly optimized using an ℓ_1 alignment loss as Eq. (5) and the VICReg regularization objective [5].

3.3 HanoiWorld - The World Model

V-JEPA 2 Based Encoder

The main innovation of the HanoiWorld is the inclusion of the strong pretrained image encoder based on self-supervised learning manner proposed by Assran et al. [2]. Specifically, the V-JEPA 2 aim at learning the essential knowledge on enviroment interaction as the motion, object movement , without focusing on the stochasitic and noisy pixel detail as reconstruction attempts as DreamerV3-based encoders Hafner [25] and Hu et al. [31]. The V-JEPA 2 encoder shall be trained to predict the masked version of the corresponding input follow self-supervising manner for more 1 million hours of video, and be stablized using Exponential Moving Average (EMA) in the Student-Teacher scheme for preventing the embedding colaspe while maintaining essential feature are strucutrally preserved Caron et al. [11].

EMA Teacher Encoder. Let θ denote the parameters of the student encoder E_θ and $\bar{\theta}$ denote the parameters of the teacher encoder. The teacher parameters are updated using an exponential moving average:

$$\bar{\theta} \leftarrow \tau \bar{\theta} + (1 - \tau) \theta, \quad (4)$$

where $\tau \in [0, 1)$ is the momentum coefficient. Gradients are stopped through the teacher encoder to prevent representation collapse, and in our experiment, we set $\tau = 0.996$.

Masked Representation Prediction. Given a video y and a masked view x obtained by removing a subset of spatio-temporal patches, the encoder E_θ extracts representations from the visible tokens. A predictor P_ϕ is trained to predict the representations of the masked tokens.

The main training objective is defined as the L1-Loss as the Eq. (5):

$$\mathcal{L}_{\text{align}} = \|P_\phi(\Delta_y, E_\theta(x)) - \text{sg}(E_{\bar{\theta}}(y))\|_1, \quad (5)$$

where Δ_y denotes learnable mask tokens indicating the locations of the masked patches, and $\text{sg}(\cdot)$ denotes the stop-gradient operator. The loss is applied only to the masked patches, and we leverage the patch-mask with random-masking as Mo and Tong [43] suggest for the finetuning process.

The leverage of the dual-branch self-supervised training with masking pattern does occur in work of Zhu et al. [71], in which suggest that by let the encoder have to force to guess the masked patch from the sampel - which does creating challenge for preventing embedding colaspe while occupancy semenatic can be learn and predicted without the need of full geometrical structure on the enviroment.

RSSM-Based Reasoning Model

The RSSM (Recurrent State-Space Model) our design is follow the codebase provided with the DreamerV3 architecture by [25, 26], the purpose of RSSM considered as the internal-memory module that storing the deterministic memory for long-term historical semantics, and the stochastic latent values as the uncertainty prediction based on the enviroment encoded-signals for capturing the world’s evolution, and agent’s expected prior conditioned by the world’s state transformation. The generative process on yeilding the World Dynamics can be formulated as the follow Eq. (6)

$$\begin{aligned} h_t &= f_\theta(h_{t-1}, z_{t-1}, a_{t-1}), \\ z_t &\sim p_\theta(s_t \mid h_t), \\ z_t &\sim p_\theta(z_t \mid h_t, z_t), \\ r_t &\sim p_\theta(r_t \mid h_t, z_t). \end{aligned} \quad (6)$$

Given that:

- f_θ denotes the recurrent dynamics function parameterized by θ , implemented as a recurrent neural network.
- h_t represents the deterministic latent state at time step t .
- a_t denotes the action executed by the agent at time step t .

- z_t corresponds to the observation at time step t , which in our setting is the V-JEPA-2 encoded embedding for the corresponding time step.
- r_t denotes the predicted reward yielded by the agent at time step t .
- All conditional distributions $p_\theta(\cdot)$ are parameterized by multi-layer perceptron (MLP) decoder networks.

Additionally, as the HanoiWorld migrating the RSSM-codebased from the DreamerV3, the model does additional including a continue-predictor for guessing whether the episode shall be continue given the encoder embedding o_t , and historical latent images z_t [25] - showcased with Eq. (7), given the continuation signal is a Bernouli random variable for simplified assumption.

$$c_t \sim p_\theta(c_t \mid h_t, z_t), \quad (7)$$

where $c_t \in \{0, 1\}$ indicates whether the episode continues at time step t .

The continue-predictor training as the logistic-regression model using the binary-cross-entropy as the original DreamerV3 suggestion with the Eq. (8)

$$\mathcal{L}_{\text{cont}} = -[c_t \log \hat{c}_t + (1 - c_t) \log(1 - \hat{c}_t)]. \quad (8)$$

- $c_t \in \{0, 1\}$ denotes the ground-truth binary continuation label at time step t .
- $\hat{c}_t \in (0, 1)$ denotes the predicted continuation probability produced by the model.

3.4 The Actor-Critic Training Head

The HanoiWorld Actor-Critic framework is directly derived from the reward- and value-based learning paradigm of DreamerV3. It shares core conceptual foundations with the classical theory proposed by Sutton [51], which emphasizes a separation between an *Actor* network (policy network) that reacts by selecting actions and a *Critic* network that evaluates the agent’s performance through a value function in order to support more effective planning. However, unlike the classical Actor-Critic formulation in Sutton [51], which operates on actual environment states, the DreamerV3 Actor-Critic operates entirely in a learned latent environment. Specifically, both the policy network π_θ and the value network v_ψ are defined over the latent state ℓ_t produced by the RSSM, as formalized in Eq. (9).

$$\ell_t = (h_t, z_t), \quad a_t \sim \pi_\theta(a_t \mid \ell_t), \quad v_\psi(R_t \mid \ell_t). \quad (9)$$

Given that:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}. \quad (10)$$

- $\ell_t = (h_t, z_t)$ denotes the latent state at time step t , consisting of the deterministic and stochastic components of the RSSM and V-JEPA-2 encoder respectively.
- R_t denotes the return at time step t , defined as the discounted sum of future rewards over an episode, with discount factor $\gamma = 0.997$.

3.5 Training Procedure an Objective

HanoiWorld training objective including the Spatial Predictor and Predictor training in the Encoder module, and training the RSSM-dynamics based model, with the Actor-Critic training follow. Given that HanoiWorld training tactics only mimicking the training procedure of DreamerV3 on the Dynamics, and Actor-Critic network, not the reconstructed-based training for encoder as [25].

Encoder Training

For the Bottleneck and Spatial Predictor training, beside using the L1-loss as Eq. (5), we consider the usage of Variance Correlation and Covariance Regularization suggested by Bardes, Ponce, and LeCun [5] - check Eq. (11) and Eq. (12)

$$\mathcal{L}_{\text{var}} = \frac{1}{D} \sum_{d=1}^D \max\left(0, 1 - \sqrt{\text{Var}(\tilde{z}_{:,d}) + \varepsilon}\right), \quad (11)$$

$$\mathcal{L}_{\text{cov}} = \frac{1}{D} \sum_{i \neq j} (\text{Cov}(\tilde{z})_{ij})^2, \quad (12)$$

$$\text{Cov}(\tilde{z}) = \frac{1}{BN - 1} (\tilde{\mathbf{z}} - \bar{\mathbf{z}})^\top (\tilde{\mathbf{z}} - \bar{\mathbf{z}}), \quad (13)$$

Given that:

- D denotes the embedding dimensionality ($D=128$), and $\tilde{\mathbf{z}} \in R^{(BN) \times D}$ represents the flattened embeddings obtained by reshaping the batch and token dimensions.
- $\tilde{z}_{:,d}$ refers to the d -th embedding dimension across all BN samples, and ε is a small constant added for numerical stability.
- The variance regularization in Eq. (11) enforces a minimum standard deviation for each embedding dimension, preventing representational collapse.
- ϵ standfor the small constant for numerical stability as $\epsilon = 1\text{e-}4$
- The covariance matrix $\text{Cov}(\tilde{z})$ in Eq. (13) is computed after centering the embeddings by their mean $\bar{\mathbf{z}} = \frac{1}{BN} \sum_{i=1}^{BN} \tilde{\mathbf{z}}_i$.

- The covariance regularization in Eq. (12) penalizes the squared off-diagonal entries of the covariance matrix, encouraging decorrelation between different embedding dimensions.

The Bottleneck are train that follow the weighted sum loss function as follow [2, 5, 71], and with the weights are setup toward (α, β, γ) to (1.0, 1.0, 0.1) for stability and prevent collapse, and lossely controlling the covariance between teacher’s bottleneck and student’s predictor.

$$\mathcal{L}_{\text{encoder}} = \alpha \mathcal{L}_{\text{align}} + \beta \mathcal{L}_{\text{var}} + \gamma \mathcal{L}_{\text{cov}}. \quad (14)$$

We does finetune our encoder on the 2D Bird-Eye-View dataset which have been pre-processed from the Nuscene dataset proposed by Caesar [9] by rendering from the LiDAR-cloud based with additional metadata on the obstacle and enviroment movement toward the RGB based images that V-JEPA 2 encoder can working with.

RSSM-dynamic model training

For the RSSM module training, we leverage the usage of predictor loss which is the inverse summation on the log probability of the probability in guessing the actual observation on the enviroment’s transition, the reward function, and continue flag as Eq (15); the dynamicity loss by minimalizing the KL-diveregnce between actual posterior distribution encoder yeild out from actual observation stably, with prior that world model imagine from prior-latent memory as Eq (16); and the representation loss for anchoring the encoder without drifting away from latent world model expectation in Eq (17). This loss-setup follows the work of [25] suggested and using the identical weight-set as DreamerV3 been configured before.

$$\mathcal{L}_{\text{pred}}(\phi) = -\log p_{\phi}(x_t | z_t, h_t) - \log p_{\phi}(r_t | z_t, h_t) - \log p_{\phi}(c_t | z_t, h_t), \quad (15)$$

$$\mathcal{L}_{\text{dyn}}(\phi) = \max\left(1, \text{KL}\left(\text{sg}[q_{\phi}(z_t | h_t, x_t)] \parallel p_{\phi}(z_t | h_t)\right)\right), \quad (16)$$

$$\mathcal{L}_{\text{rep}}(\phi) = \max\left(1, \text{KL}\left(q_{\phi}(z_t | h_t, x_t) \parallel \text{sg}[p_{\phi}(z_t | h_t)]\right)\right). \quad (17)$$

REINFORCE-based Actor-Critic learning

The algorithm HanoiWorld Actor-Critic component are based on the Reinforce algorithm suggest by Williams [59], which suggest the agent (specific the policy network of the Actor) shall optimize and favor the action that yeilded best return. However, the algorithm we used follow the implementation of DreamerV3 from [25], which suggest both actor and critic with imagined roll out from latent RSSMs, with the critic layer

yield the prediction on the possible cumulative reward based on the latent feature on the future, while the actor’s policy network are optimized based on the advantage signal, and the entropy regularization for trigger agent exploration on potential action with imaginative near-future - and the entropy are being scale by a fixed constant. The actor critics algorithm training is explicit describe on the Algorithm 1.

Algorithm 1: Actor–Critic Training via Imagined Rollouts

Input: World model parameters ϕ ;
 Actor parameters θ ;
 Critic parameters ψ ;
 Imagined horizon H ;
 Discount factor γ ;
 λ -return parameter λ ;
 Entropy coefficient $\beta = 3\text{e-}4$
 Sample posterior latent state (h_0, z_0) from real experience;
 Initialize imagined trajectory buffers;
for $t = 0$ **to** $H - 1$ **do**
 Compute feature $f_t \leftarrow f_\phi(h_t, z_t)$;
 Sample action $a_t \sim \pi_\theta(\cdot \mid f_t)$;
 Predict next latent state $(h_{t+1}, z_{t+1}) \sim p_\phi(\cdot \mid h_t, z_t, a_t)$;

4 Experimental Result

This section of the paper shall focusing on the experiment that been used for evaluating the performance of the HanoiWorld in comparison with difference baseline approaches

4.1 Experiment Description

For evaluating the HanoiWorld performance in comparison toward across difference alternatives, experiments shall be conducted within the environment from the HighwayEnv package provided by Foundation [17], these include *Highway*, *Roundabout*, *Merge*, which shall be discussed in detail in subsection 4.2

Comparative World Model design that we chosen in designing are including the DreamerV3 proposed by Hafner [25], the VQ-VAE + ConvLSTM based suggested by Robine, Uelwer, and Harmeling [46], and the HanoiWorld proposed agents. The baseline model described as below:

- The VQ-VAE encoder based with ConvLSTM planner suggested by Robine, Uelwer, and Harmeling [46] is a small and discrete latent world model which aiming for efficient inference and planning on the latent space, as the inheritance on the idea of Ha and Schmidhuber [23]. The VQ-VAE based solution is design on 100 thousands

Algorithm 1: Algorithm 1 (continued)

```

for  $t = 0$  to  $H - 1$  do
  | Predict reward  $\hat{r}_{t+1} \leftarrow r_\phi(h_{t+1}, z_{t+1})$ ;
  | Predict continuation  $\hat{c}_{t+1} \leftarrow c_\phi(h_{t+1}, z_{t+1})$ ;
  Compute value predictions  $V_t \leftarrow V_\psi(f_t)$  for all  $t \in \{0, \dots, H\}$ ;
  Compute  $\lambda$ -return targets;
  Set effective discount  $\gamma_t \leftarrow \gamma \cdot \hat{c}_t$ ;
  Set bootstrap target  $G_H^\lambda \leftarrow V_H$ ;
  for  $t = H - 1$  to  $0$  do
    |  $G_t^\lambda \leftarrow \hat{r}_{t+1} + \gamma_{t+1}((1 - \lambda)V_{t+1} + \lambda G_{t+1}^\lambda)$ ;
  Critic update;
  Minimize negative log-likelihood of  $\lambda$ -returns;;
   $\mathcal{L}_{\text{value}}(\psi) \leftarrow -E_t[\log p_\psi(G_t^\lambda | f_t)]$ ;
  Actor update;
  Compute advantage (baseline):  $A_t \leftarrow \text{sg}(G_t^\lambda - V_t)$ ;
  Minimize actor loss;;
   $\mathcal{L}_{\text{actor}}(\theta) \leftarrow -E_t[\log \pi_\theta(a_t | f_t) A_t + \beta \mathcal{H}(\pi_\theta(\cdot | f_t))]$ ;
  Update critic parameters  $\psi$ ;
  Update actor parameters  $\theta$ ;

```

episode on the game of Atari by using the Proximal Policy Optimization (PPO) algorithm proposed by [47] for latent space planning on discrete imagined rollouts.

- The DreamerV3 by Hafner [25] currently is reached State-Of-The-Art level performance with the capability of long-term horizontal planning by the RSSMs networks, which does inspired the design of HanoiWorld by attempts on learning the robust, and generalizing model for environment agnostic, which use to predict the plausible next latent state, reward, and continuation, the model is leveraged for yielding the training signal to multi-layer-perceptron layers on reward, and values prediction based on imagined rollout, with Actor follow the REINFORCE gradient update, or hybridizing update learning signal from RSSM for robust and stable learning

For the evaluation metric, the selected metric 2 metric - the average reward (e.g score), and the collision rate as the following:

- *The average reward/score* is the metric that Ha and Schmidhuber [23] used for evaluating the efficiency of the agents on planing efficiency - which is the average of the reward signal across steps over episodes - which showcased with the Eq. (18)
- *The Collision Rate* is the metric suggested by Gao et al. [19] which can be estimated by the proportion of episode that ego vehicles occur the collision, which expressed with Eq. (19)

$$\bar{R} = \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{T_i} r_{i,t}, \quad (18)$$

Given that:

- \bar{R} denotes the average driving reward over all evaluation episodes.
- N is the total number of evaluation episodes.
- $i \in \{1, \dots, N\}$ indexes the episode.
- t indexes the time steps within an episode.
- $r_{i,t}$ is the instantaneous reward received at time step t in episode i .

Interpretation: Higher values of \bar{R} indicate better driving behavior, meaning smoother control (less wobbling) and safer trajectories (fewer crashes).

$$\text{Collision Rate} = \frac{\# \text{ collision episodes}}{\text{total episodes that being evaluated}} \quad (19)$$

4.2 Case Studies

The experiment had been conducted on the interactive environment for examining the model's adaptability and generalizabilities across different driving situations as the different road-topology, lane merging scenarios, and navigation in the roundabout.

The information and tabular description of each scenario are provided within the table 1.

Aspect	Highway	Merge	Roundabout
Task	highway_highway	highway_merge	highway_roundabout
Observation	RGB image (64×64)	RGB image (64×64)	RGB image (64×64)
Action space	Continuous	Discrete	Discrete
Episode time limit	200	800	800
Traffic setting	Dense straight highway	On-ramp merging traffic	Circular multi-agent traffic
Vehicle count	50	Environment-controlled	Environment-controlled
Vehicle density	1.5	Implicit (spawn-based)	Implicit (spawn-based)
Speed target (m/s)	23–27	20–28	8–15
Collision penalty	–5.0	–1.0	–1.0
Safe-distance shaping	Strong penalty (tailgating)	Reward + penalty	Reward + penalty (stronger)
Lane-change shaping	Smart lane change rewarded	Lane changes penalized	Slight penalty
Progress shaping	Moderate	Strong	Moderate
Heading alignment reward	Yes	Yes	Yes (strong)
Survival reward	–	Low	Moderate
Success reward	1.0	0.8	1.0
Reward shaping weight	0.85	0.8	0.8
Imagination gradient	REINFORCE	REINFORCE	REINFORCE
Primary behavioral objective	Stable high-speed driving	Safe and anticipatory merging	Social negotiation and yielding

Table 1: Comparison of Highway-based autonomous driving environments and their core configuration and reward-shaping characteristics.

Task description on each enviroment

The highway-env (e.g highway-v0) - Figure 3 is designed to showcase the model's capability of controlling the vehicle's movement on multi-lane scenarios where ego vehicle shall interact and make lane changing decision in corresponsse to difference agent's (surrounding vehicles can accelerate or suddently de-accelerate, and change-lanes). The enviroment attempts to test the model's capability on maintain stablizing, and safe-driving speed, and the flexibility by making lane switch as the correspondence toward surrounding agents [17].

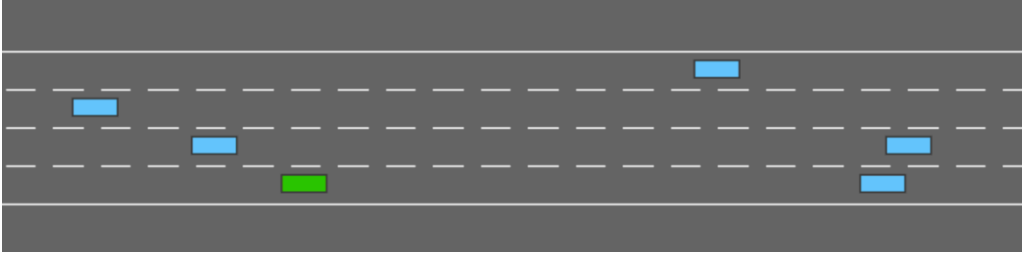


Figure 3: The highway-v0 Enviroment - a snapshot taken from [17]

The merge-env (e.g merge-v0) - Figure 4, within the simulation, the agent shall have to make the attempts on switching from 1 road-branch toward difference for addressing the lane-merging challenge, which is challenging dues to the speed controlling, and detail planning of the agent in preventing the collision during the merging process [17]. The goal of the scenario is the model attempts understand and predicted the quick physical transition on difference agents for making appropriate merging-decision.

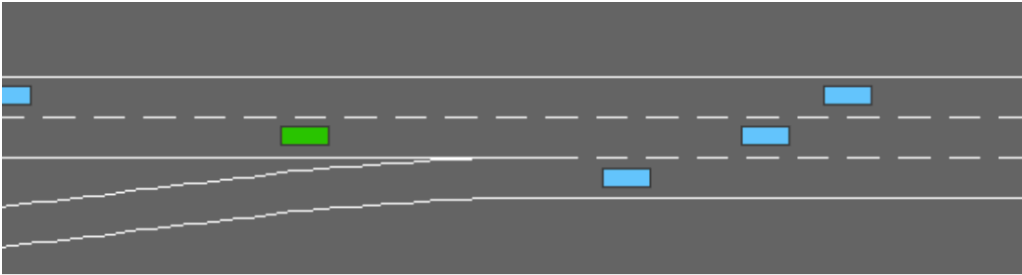


Figure 4: The roundabout-v0 Enviroment - a snapshot taken from [17]

The roundabout (roundabout-v0) as Figure 5 is the complex and realistic version of the merge-v0 where the agents have more than 1 suggestion in navigation, and the density of the navigation across each dirrection is stochastic as the prior usecases. The purpose of this situation is considered as the roburstness checking on the performance with more closer toward realistic driving enviroment, where the ego's shall make the planning on when to navigate, stop to entering the roundabout without trggering an accident [17].

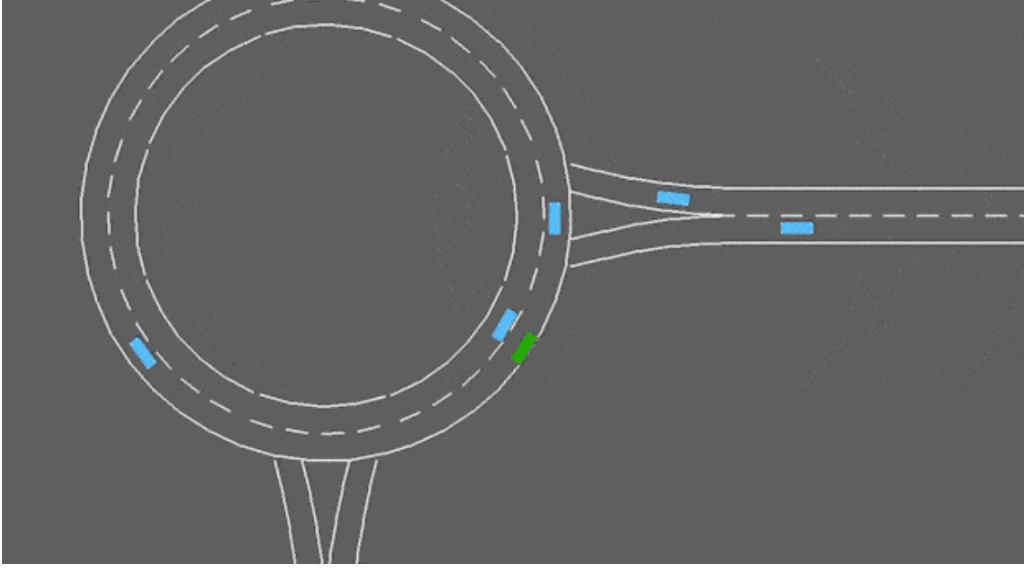


Figure 5: The roundabout-v0 Enviroment - a snapshot taken from [17]

4.3 Experimental Procedure

For the baseline - both the DreamerV3, and VQ-VAE based model shall be trained within 100 thousand steps including the rolling-out for training based on the sample; while HanoiWorld dues to the larger model-structure but the RSSM-planning module are still effective in the inference - it shall be trained under 5000 steps with the prior-pretrained V-JEPA 2 with the bottleneck and spatial embedding predictor. All model world-model (except for the HanoiWorld encoder) shall be trained primitively within the Highway-env simmulation.

On the evaluation phase of the experiment, difference model shall be evaluating and inferred within 100 vary length episode on 3 difference enviroments, and returning the average reward/score and the overall collision rate within 100 episodes, and the result shall be record both the mean-values, and the standard deviation of the measurement across episodes.

4.4 Evaluation

The experiment result shall be showcased with the table 2 for collision rate, and table 3 for average reward scoring

Table 2: Collision rate over 100 evaluation episodes (mean \pm std).

ENV	highway-v0	merge-v0	roundabout-v0
DreamerV3	0.550 ± 0.497	0.030 ± 0.171	0.500 ± 0.500
VQ-VAE	1.000 ± 0.000	0.290 ± 0.454	0.570 ± 0.495
HanoiWorld	0.200 ± 0.400	0.970 ± 0.170	0.340 ± 0.473

Table 3: Average episode reward over 100 evaluation episodes (mean \pm std).

ENV	highway-v0	merge-v0	roundabout-v0
DreamerV3	51.065 \pm 52.700	41.973 \pm 2.896	9.423 \pm 6.171
VQ-VAE	3.121 \pm 12.047	30.114 \pm 11.664	3.826 \pm 4.643
HanoiWorld	13.163 \pm 23.277	13.480 \pm 5.703	9.818 \pm 6.252

From the perspective of the collision rate, HanoiWorld demonstrates superior performance compared to the corresponding baselines on the Highway and Roundabout environments, with average collision rates of 0.200 and 0.340, respectively. This outperforms DreamerV3, which exhibits mid-tier collision performance, while VQ-VAE performs worst among the evaluated methods. However, HanoiWorld reveals a notable weakness in the lane-merging scenario, where it exhibits the highest expected collision rate - while DreamerV3 show it's stronger capability of navigation in this situation with the rate of 0.030. From the model's stability - which is expressed through the collision rate stability, our HanoiWorld does show statistical stability over domain-wise with lowest standard deviation, however these does show the signal that HanoiWorld can under poor navigation performance in merge-v0 scenario where the collision rate near to 1.

Extend toward the average episode reward, the HanoiWorld only showcased it's performance competitively with SOTA baselines of DreamerV3 under the roundabout-v0 environments with certain raise on the average reward 9.818 with 9.423 respectively, whereas, the DreamerV3 still show the effectively planning efficiency with the highest planning performance, with the highway-v0 show significant stable planning then merge-v0 scenarios, while the HanoiWorld performance under these 2 task is in the middle on the highway case; and poor planning in the merge-v0 simulation even with smaller parameter model as VQ-VAE baselines.

4.5 Results Discussion

As the result been showcased, the HanoiWorld can effectively planning on the both simple (highway-v0) and complex environment (roundabout-v0) with the competitive performance, while it's showcased such poor generalization with morderling challenging as merging - where the frequencies of collision is substantially high, while the performance on reward yielding as ones could observe is consistence across environment, which proof the HanoiWorld can generalizing the planning tatic across difference environment. However, on the merge-v0 we assume the model have the signal of reward hacking by favoring the action of collision in yeilding consistence reward across episode.

In additionally, as we observe, in the case where model need to showcased the flexibility in road navigation as roundabout-v0 in lane selection, the model currently favor the safe-lane selection, which does show that our entropy based regularization for making decision is underperformed and need to be carefully studies with abalation.

Even though HanoiWorld showcases competitive and worth-considering performance in the comparison with state-of-the-art baselines as DreamerV3, there are certain aspect that this studies can be addressed within the nearby future, as the more global-contextual introduction on the enviroment for gating on the affordance relationship between the enviroment and the agent’s perspective by using text-conditioned language encoder as Hu et al. [31], or occupancies encoder on the whole enviroment graph as the Vasudevan et al. [55] and Zhao [68] for generalzing behavior; additionally the studies does not inspect the imaginary rolling out as the work for Ha and Schmidhuber [23] – which can consider for later studies on checking the performance of the ego’s with difference imaginary tempera-ture for hardness controlling as the migration for reward-regularization within the latent domain.

5 Conclusion

This study showcased HanoiWorld, a JEPA-based worldmodel for simmulation generation in training autonomous vehicle controller in reinforcement learning, as the result have proved HanoiWorld is capable in making effective planning strategy within the latent domain as the SOTA baseline, and lighter competitor, make the attemptss seem compelling than the pixel-reconstruction based approach dues to computational efficency. Additionally, our model does show that in some certain scenarios HanoiWorld reach lower collision rate, which can hypothesize the model do learn the concept of safetyiness both across enviroment, however the planning mechanism of the model are under-performing in comparing with SOTA baseline as DreamerV3.

Remarkably, the current study does not attempt in the integration of more global contextual condition using multi-modality inputs as the language, global graph; in addition with the latent-probing for actual evaluating – understanding the planing mechanism of the WorldModel under difference enviromental hardness; suggesting the potential of further dirrection that later study can be inspired and continues using HanoiWorld’s result.

References

- [1] Mahmoud Assran et al. “Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture”. In: *arXiv preprint arXiv:2301.08243* (2025).
- [2] Mido Assran et al. *V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning*. arXiv.org, 2025.
- [3] Lukas Audinys et al. “Benchmarking World Models for Autonomous Systems”. In: *arXiv preprint* (2025).
- [4] Lorenzo Baraldi et al. “The Safety Challenge of World Models for Embodied AI Agents: A Review”. In: (2025). (Visited on 12/31/2025).
- [5] Adrien Bardes, Jean Ponce, and Yann LeCun. “VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning”. In: *arXiv:2105.04906 [cs]* (Jan. 2022).
- [6] Julian Bogdoll et al. “Occupancy-Based World Modeling for Safety-Critical Autonomous Driving”. In: *arXiv preprint* (2025).
- [7] Khaled Bouzaiene. “Simulation Platforms for Autonomous Driving Research”. In: *arXiv preprint* (2025).
- [8] Greg Brockman et al. “OpenAI Gym”. In: *arXiv preprint arXiv:1606.01540* (2016).
- [9] Holger Caesar et al. “nuScenes: A Multimodal Dataset for Autonomous Driving”. In: *CVPR* (2021).
- [10] Mathilde Caron et al. “Emerging Properties in Self-Supervised Vision Transformers”. In: *ICCV* (2021).
- [11] Mathilde Caron et al. “Emerging Properties in Self-Supervised Vision Transformers”. In: *arXiv:2104.14294 [cs]* (May 2021).
- [12] Rui Chen et al. “Physical World Modeling for Autonomous Agents”. In: *arXiv preprint* (2025).
- [13] Yifan Chu et al. “World Models: A Survey”. In: *arXiv preprint* (2025).
- [14] Alexandre Dawid and Yann LeCun. “Energy-Based World Models”. In: *arXiv preprint* (2024).
- [15] Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. “Challenges of Real-World Reinforcement Learning”. In: *arXiv:1904.12901 [cs, stat]* (Apr. 2019).
- [16] Qian Feng et al. “Multi-Agent Intent Modeling for Autonomous Driving”. In: *arXiv preprint* (2025).
- [17] Farama Foundation. *Gymnasium Documentation*. gymnasium.farama.org, 2025.
- [18] Yiheng Fu et al. “Driving Diffusion Models”. In: *arXiv preprint arXiv:2405.02345* (2024).

- [19] Dechen Gao et al. *CarDreamer: Open-Source Learning Platform for World Model based Autonomous Driving*. arXiv.org, 2024. (Visited on 10/24/2025).
- [20] Rui Gao et al. “CarDreamer: Learning World Models for Autonomous Driving”. In: *arXiv preprint arXiv:2406.10101* (2024).
- [21] James J. Gibson. *THE ECOLOGICAL APPROACH TO VISUAL PERCEPTION*. 1979.
- [22] Yifan Guan et al. “Large-Scale Training of Driving World Models”. In: *arXiv preprint* (2025).
- [23] David Ha and Jürgen Schmidhuber. “World Models”. In: *World Models 1* (2018), e10. DOI: 10.5281/zenodo.1207631.
- [24] Shlomi Hachohen, Oded Medina, and Shraga Shoval. “Autonomous Driving: A Survey of Technological Gaps Using Google Scholar and Web of Science Trend Analysis”. In: *IEEE Transactions on Intelligent Transportation Systems* 23 (May 2022), pp. 21241–21258. DOI: 10.1109/tits.2022.3172442.
- [25] Danijar Hafner et al. “Mastering Diverse Domains through World Models”. In: *arXiv preprint arXiv:2301.04104* (2023).
- [26] Danijar Hafner et al. “Learning Latent Dynamics for Planning from Pixels”. In: *arXiv:1811.04551 [cs, stat]* (June 2019).
- [27] Ye Han et al. *A Value Based Parallel Update MCTS Method for Multi-Agent Cooperative Decision Making of Connected and Automated Vehicles*. arXiv.org, 2024. (Visited on 11/12/2025).
- [28] Nicklas Hansen et al. “Generalization in Model-Based Reinforcement Learning”. In: *NeurIPS* (2023).
- [29] Matteo Hessel et al. *Rainbow: Combining Improvements in Deep Reinforcement Learning*. arXiv.org, 2017.
- [30] Carl-Johan Hoel et al. “Combining Planning and Deep Reinforcement Learning in Tactical Decision Making for Autonomous Driving”. In: *IEEE Transactions on Intelligent Vehicles* (2019), pp. 1–1. DOI: 10.1109/tiv.2019.2955905. (Visited on 02/04/2020).
- [31] Anthony Hu et al. *GAIA-1: A Generative World Model for Autonomous Driving*. arXiv.org, Sept. 2023.
- [32] Zhiyu Huang et al. *Learning Online Belief Prediction for Efficient POMDP Planning in Autonomous Driving*. arXiv.org, 2024. (Visited on 11/18/2025).
- [33] Fan Jia et al. *ADriver-I: A General World Model for Autonomous Driving*. arXiv.org, Oct. 2023. (Visited on 11/02/2025).

- [34] Ming Jia et al. “Survey of World Models in Autonomous Driving”. In: *arXiv preprint* (2025).
- [35] Jason Kong et al. *Kinematic and dynamic vehicle models for autonomous driving control design*. IEEE Xplore, June 2015. DOI: 10.1109/IVS.2015.7225830. (Visited on 12/10/2022).
- [36] Yann LeCun. “A Path Towards Autonomous Machine Intelligence”. In: *arXiv preprint arXiv:2203.03620* (2022).
- [37] Wei Li et al. “LiDAR-Centric World Models for Autonomous Driving”. In: *arXiv preprint* (2025).
- [38] Yin hao Li et al. “BEVFormer: Learning Bird’s-Eye-View Representation from Multi-Camera Images”. In: *ECCV* (2022).
- [39] Zhi Li et al. “Occupancy Prediction for Autonomous Driving”. In: *CVPR* (2024).
- [40] Han Long et al. “Interaction-Aware World Models for Dense Traffic”. In: *arXiv preprint* (2025).
- [41] Chen Min et al. “Occupancy-MAE: Self-Supervised Learning for Autonomous Driving”. In: *CVPR* (2024).
- [42] Nguyen Minh et al. “Masked Autoencoders for Point Clouds”. In: *ECCV* (2022).
- [43] Shentong Mo and Shengbang Tong. *Connecting Joint-Embedding Predictive Architecture with Contrastive Self-supervised Learning*. Oct. 2024. (Visited on 11/06/2025).
- [44] Tomov S Momchil et al. *TreeIRL: Safe Urban Driving with Tree Search and Inverse Reinforcement Learning*. arXiv.org, 2025. (Visited on 11/12/2025).
- [45] Tetsuro Morimura et al. *Policy Gradient Algorithms with Monte Carlo Tree Learning for Non-Markov Decision Processes*. 2024. (Visited on 11/18/2025).
- [46] Jan Robine, Tobias Uelwer, and Stefan Harmeling. *Smaller World Models for Reinforcement Learning*. arXiv.org, 2020. (Visited on 11/23/2025).
- [47] John Schulman et al. *Proximal Policy Optimization Algorithms*. arXiv.org, Aug. 2017.
- [48] Vlad Sobal et al. *Separating the World and Ego Models for Self-Driving*. arXiv.org, 2022. (Visited on 10/30/2025).
- [49] Hao Sun et al. *Supervised Q-Learning can be a Strong Baseline for Continuous Control*. Openreview.net, 2022. (Visited on 11/12/2025).
- [50] Pei Sun et al. “Scalability in Perception for Autonomous Driving: Waymo Open Dataset”. In: *CVPR* (2020).

- [51] Richard Sutton. “Dyna, an integrated architecture for learning, planning, and reacting”. In: *SIGART newsletter* 2 (July 1991), pp. 160–163. DOI: 10.1145/122344.122377.
- [52] Yuval Tassa et al. “DeepMind Control Suite”. In: *arXiv preprint arXiv:1801.00690* (2018).
- [53] Xiaoyu Tu et al. “Delay-Aware Cooperative Perception for Autonomous Driving”. In: *arXiv preprint* (2025).
- [54] Xiaoyu Tu et al. “Robust Cooperative World Models under Communication Constraints”. In: *arXiv preprint* (2025).
- [55] Arun Balajee Vasudevan et al. *Planning with Adaptive World Models for Autonomous Driving*. arXiv.org, 2024. (Visited on 10/24/2025).
- [56] Siva Vasudevan et al. “Transformer-Based Planning for Autonomous Driving”. In: *CVPR* (2024).
- [57] Shaoshuai Wang et al. “DeepAccident: A Motion and Accident Prediction Benchmark for V2X Autonomous Driving”. In: *NeurIPS* (2023).
- [58] Xiaofeng Wang et al. “DriveDreamer: Towards Realistic World Models for Autonomous Driving”. In: *arXiv preprint arXiv:2310.05008* (2023).
- [59] Ronald J. Williams. “Simple statistical gradient-following algorithms for connectionist reinforcement learning”. In: *Machine Learning* 8 (May 1992), pp. 229–256. DOI: 10.1007/bf00992696.
- [60] Yifan Wu et al. “Scaling World Models for Autonomous Driving”. In: *arXiv preprint* (2025).
- [61] Chenjun Xiao et al. *Learning to Combat Compounding-Error in Model-Based Reinforcement Learning*. arXiv.org, 2019. (Visited on 12/29/2025).
- [62] Yiming Xie et al. “Unified World Modeling and Planning for Autonomous Driving”. In: *arXiv preprint* (2025).
- [63] Runsheng Xu et al. “OPV2V: An Open Benchmark Dataset and Fusion Pipeline for Perception with Vehicle-to-Vehicle Communication”. In: *ICRA* (2021).
- [64] Runsheng Xu et al. “V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer”. In: *ECCV* (2022).
- [65] Runsheng Xu et al. “V2V4Real: A Real-World Large-Scale Dataset for Vehicle-to-Vehicle Cooperative Perception”. In: *CVPR* (2023).
- [66] Zhangcun Yan et al. “OnSiteVRU: A High-Resolution Trajectory Dataset for Vulnerable Road Users”. In: *arXiv preprint arXiv:2503.23365* (2025).

- [67] Yumeng Zhang et al. *BEVWorld: A Multimodal World Simulator for Autonomous Driving via Scene-Level BEV Latents*. arXiv.org, June 2024. (Visited on 11/02/2025).
- [68] Guosheng Zhao et al. “DriveDreamer-2: LLM-Enhanced World Models for Diverse Driving Video Generation”. In: *arXiv preprint arXiv:2403.06845* (2024).
- [69] Chuanxia Zheng and Andrea Vedaldi. “Online Clustered Codebook”. In: *arXiv preprint arXiv:2307.15139* (2023).
- [70] Haoran Zhu et al. “Self-Supervised Representation Learning with Joint Embedding Predictive Architecture for Automotive LiDAR Object Detection”. In: *arXiv preprint arXiv:2501.04969* (2025).
- [71] Haoran Zhu et al. *Self-Supervised Representation Learning with Joint Embedding Predictive Architecture for Automotive LiDAR Object Detection*. arXiv.org, 2025. (Visited on 11/02/2025).

A Additional Results

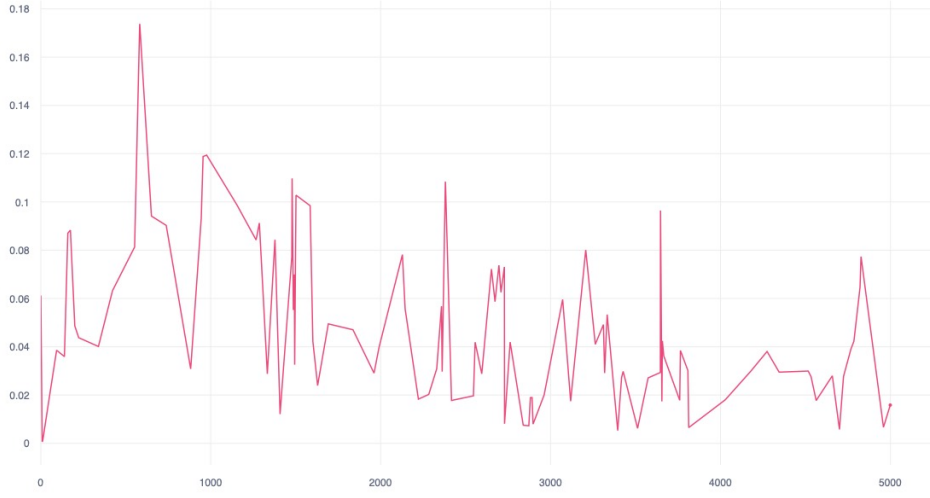


Figure 6: Merge environment: continuation loss during RSSM training on step level.

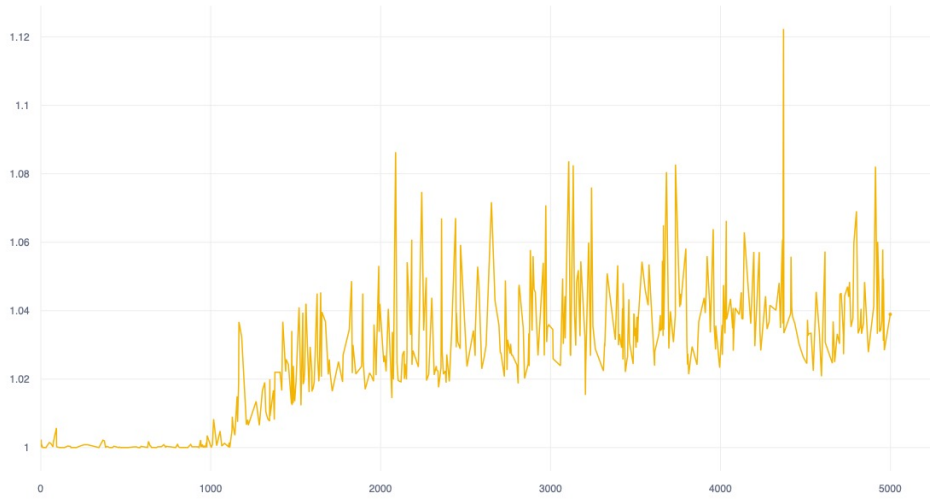


Figure 7: Merge environment: dynamics loss during RSSM training on step level.

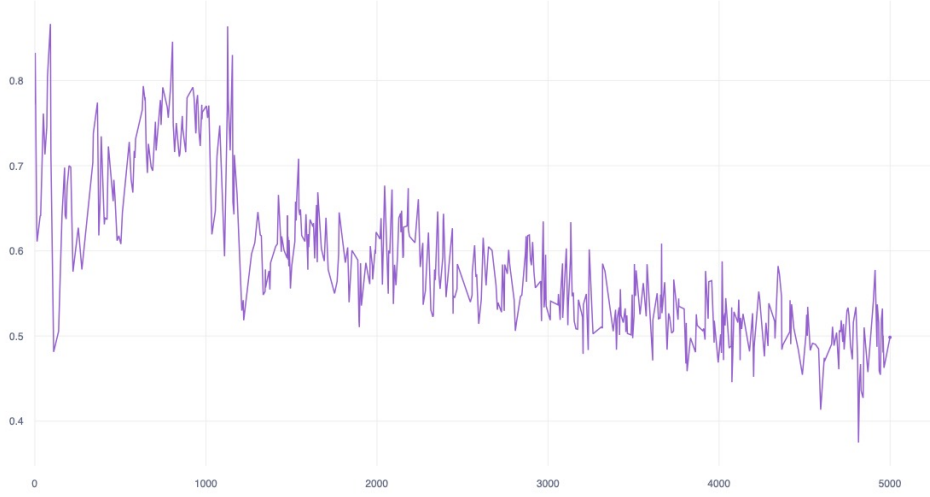


Figure 8: Merge environment: KL regularization loss during RSSM training on step level.

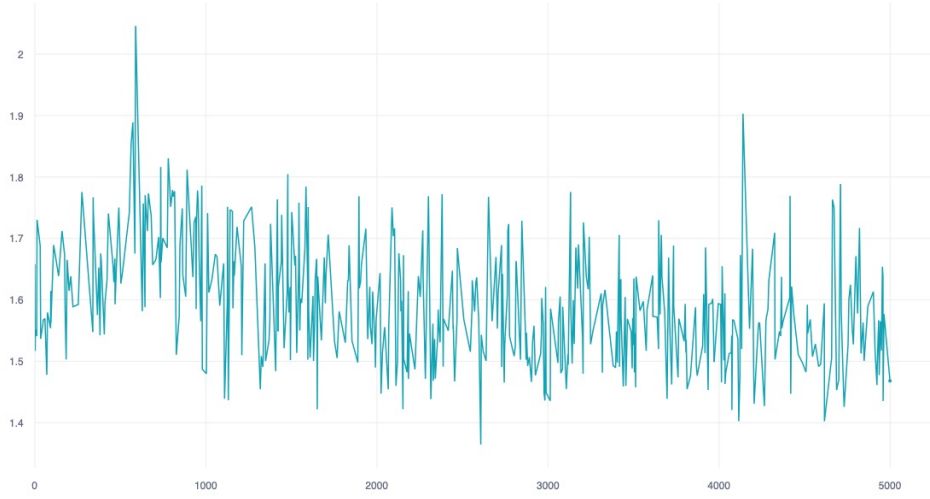


Figure 9: Merge environment: HanoiWorld model loss during RSSM training on step level.

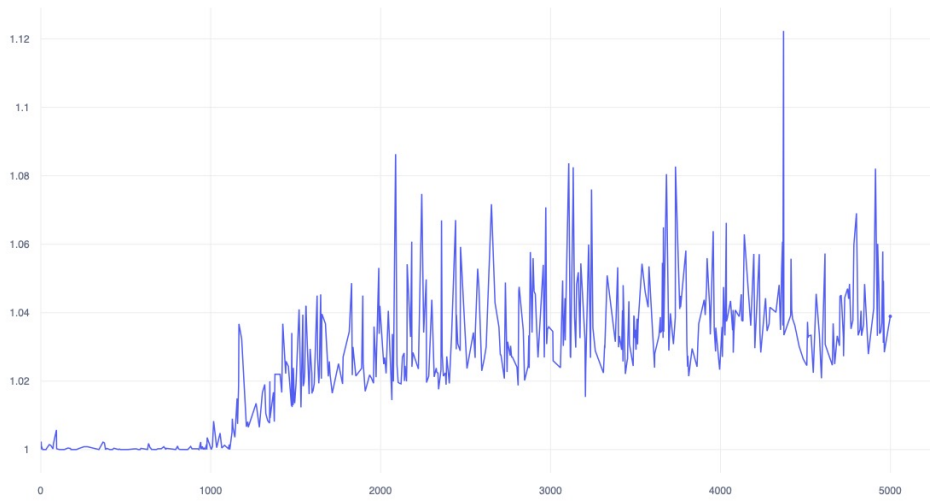


Figure 10: Merge environment: HanoiWorld representation loss during RSSM training on step level.

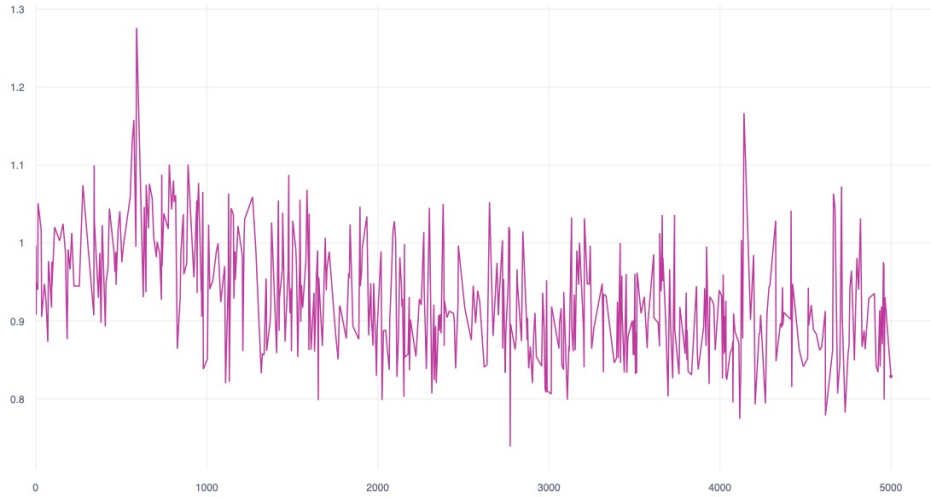


Figure 11: Merge environment: HanoiWorld reward loss predictor during RSSM training on step level.

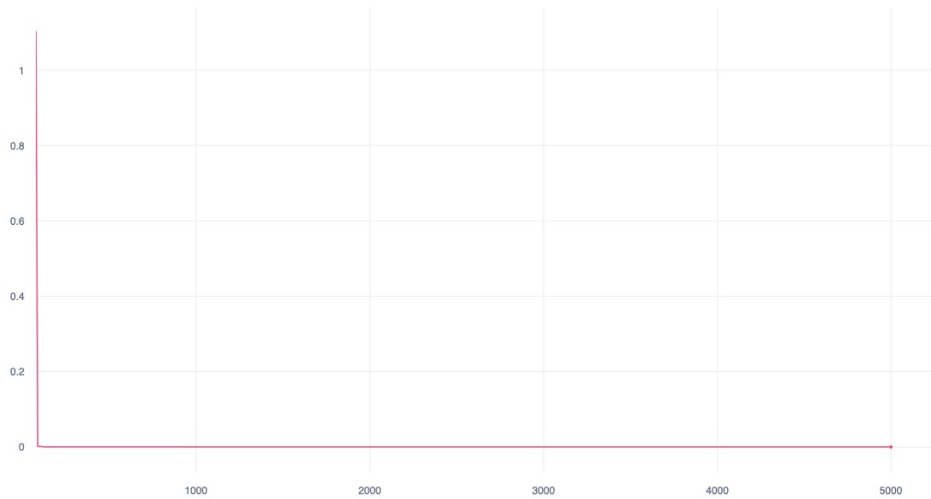


Figure 12: Highway environment: continuation loss during RSSM training.

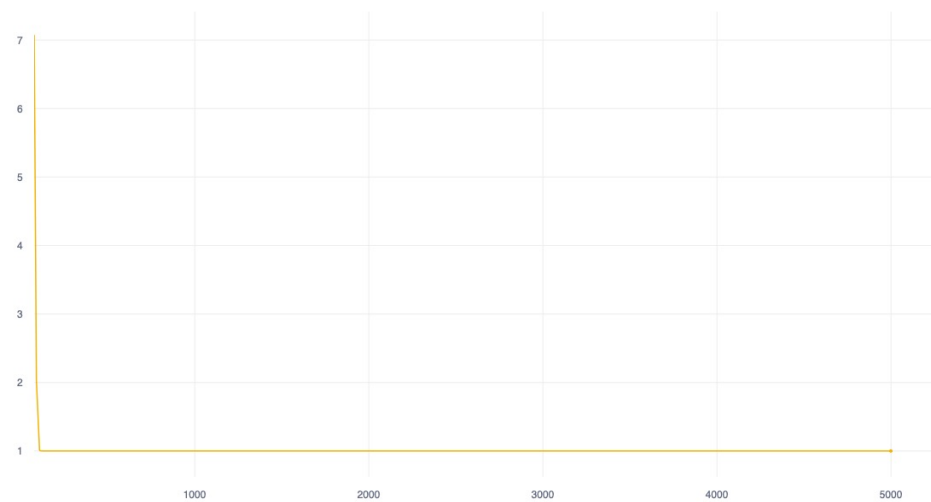


Figure 13: Highway environment: dynamics loss during RSSM training.

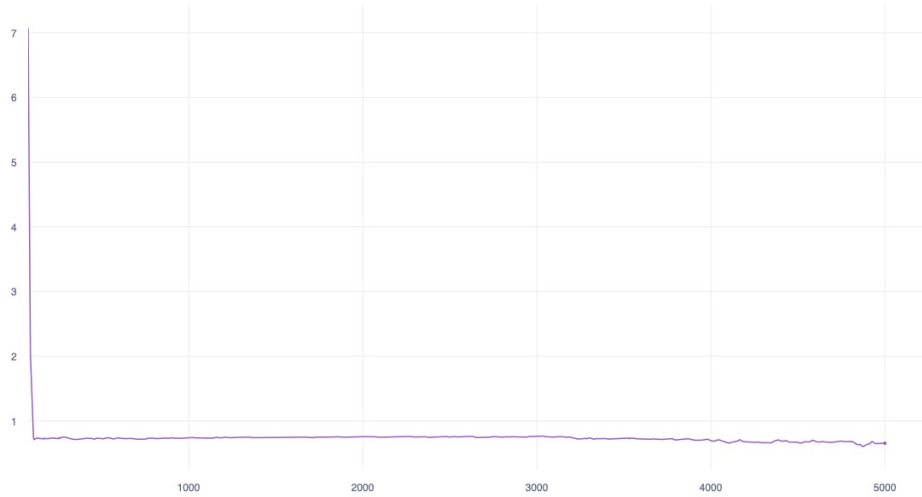


Figure 14: Highway environment: KL regularization loss during RSSM training.

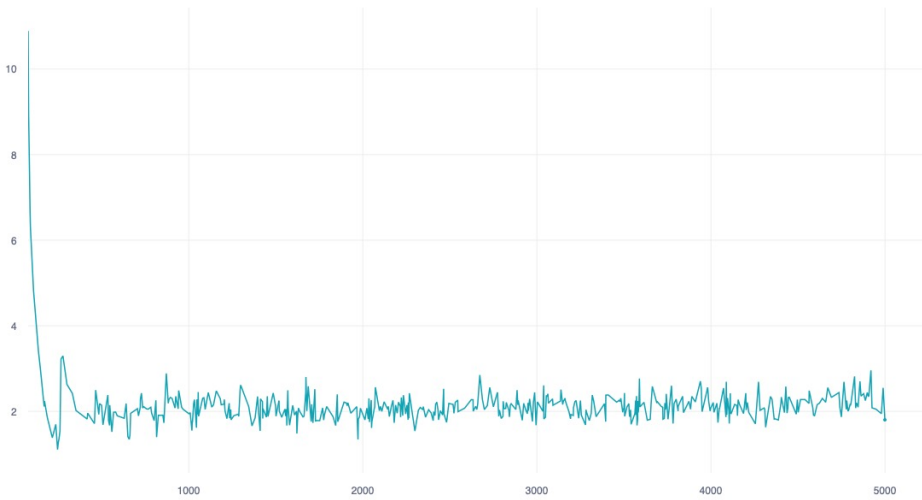


Figure 15: Highway environment: RSSM model loss during RSSM training.

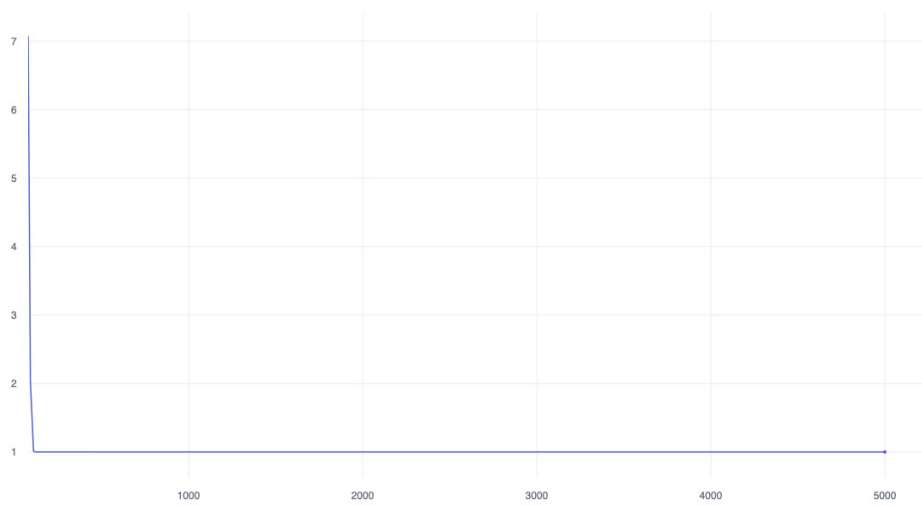


Figure 16: Highway environment: RSSM represent loss during RSSM training.

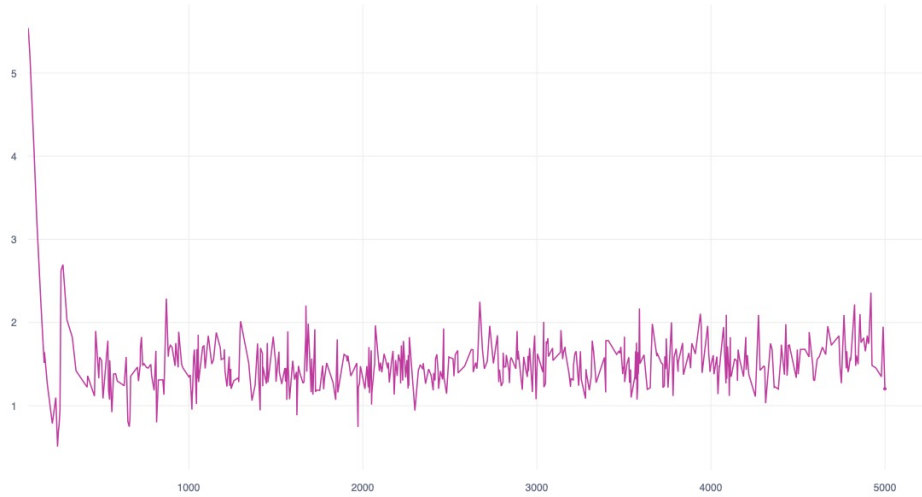


Figure 17: Highway environment: RSSM reward predictor loss during RSSM training.

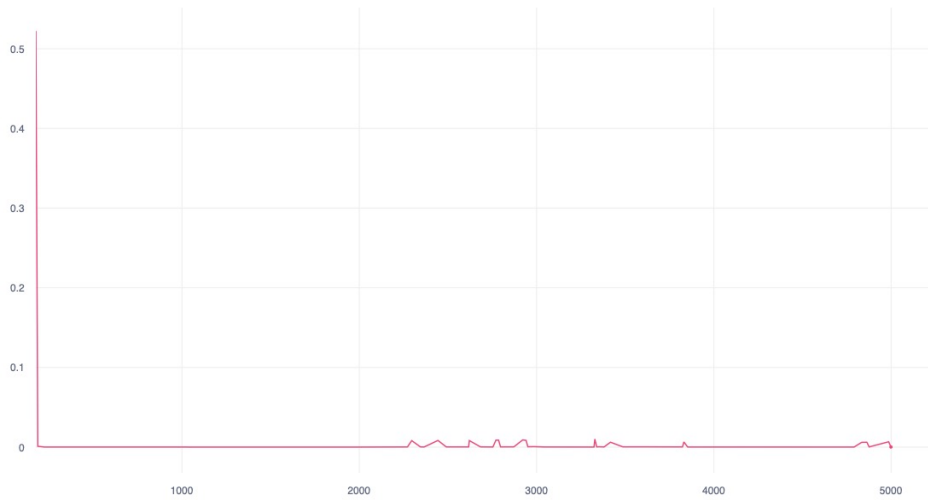


Figure 18: Roundabout environment: continuation loss during RSSM training.

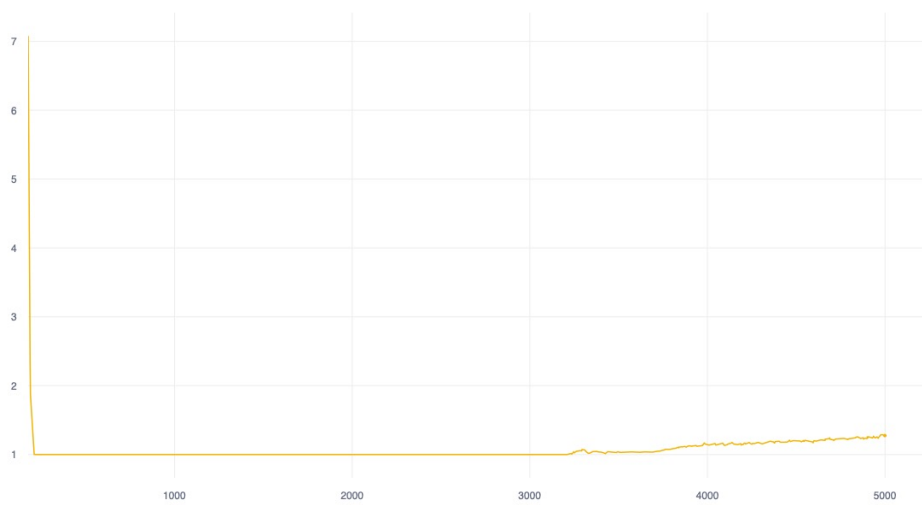


Figure 19: Roundabout environment: dynamics loss during RSSM training.

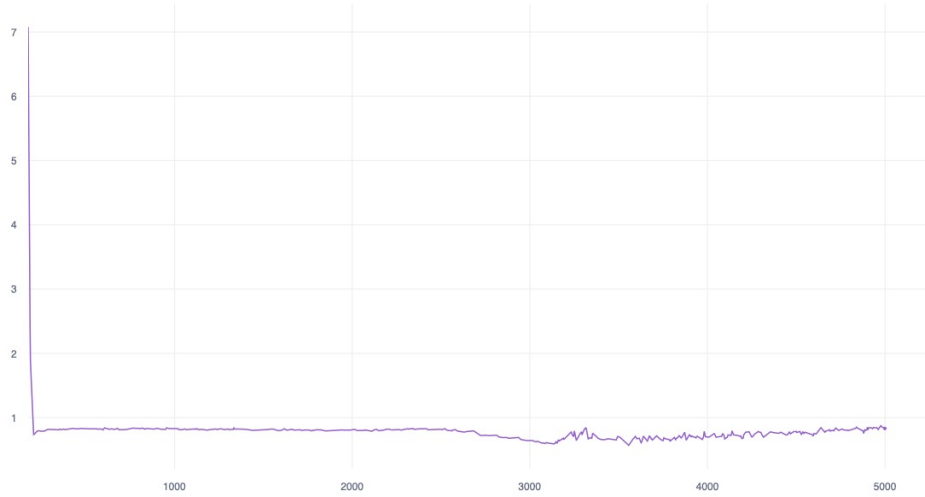


Figure 20: Roundabout environment: KL regularization loss during RSSM training.

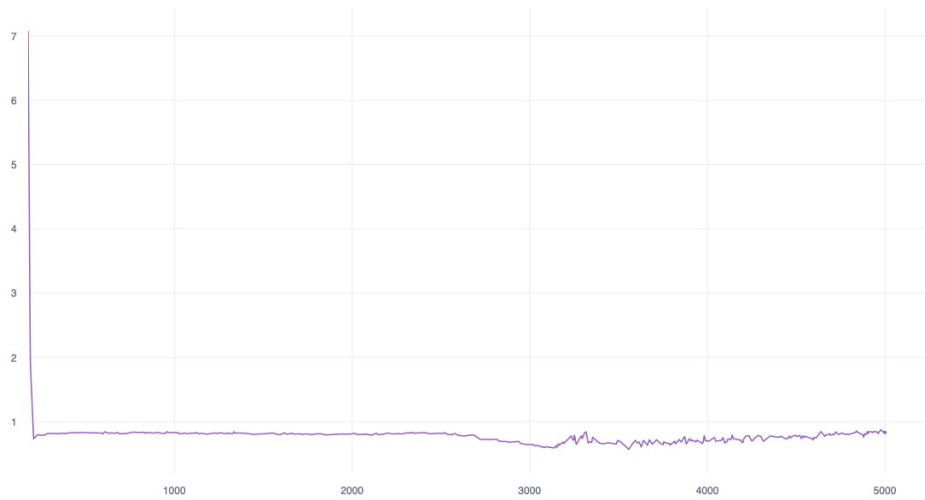


Figure 21: Roundabout environment: RSSM model loss regularization loss during RSSM training.

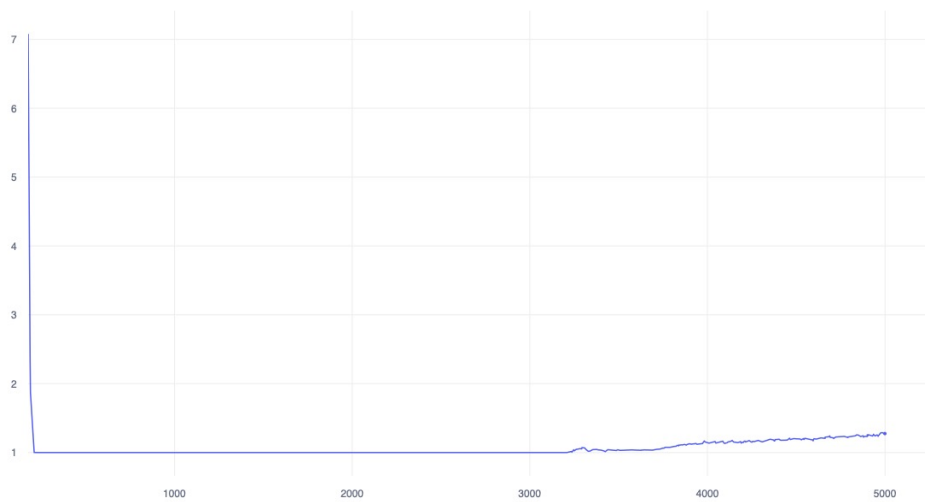


Figure 22: Roundabout environment: RSSM representation loss regularization loss during RSSM training.

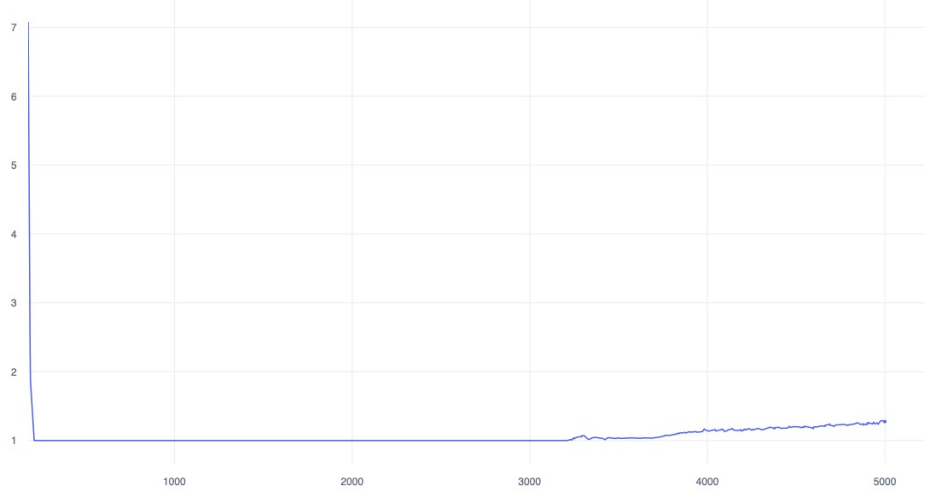


Figure 23: Roundabout environment: RSSM reward predictor loss during RSSM training.

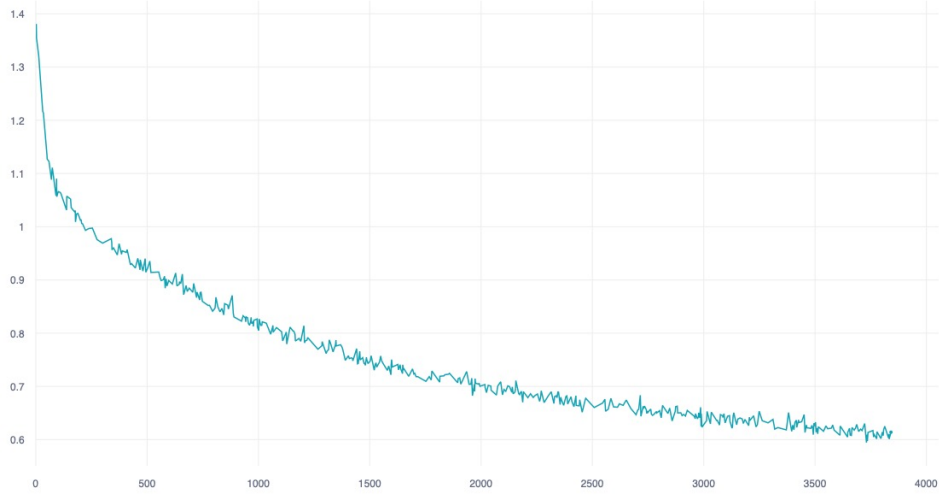


Figure 24: L1-Loss between student's spatial predictor & teacher bottleneck layer on V-JEPA 2 encoder.

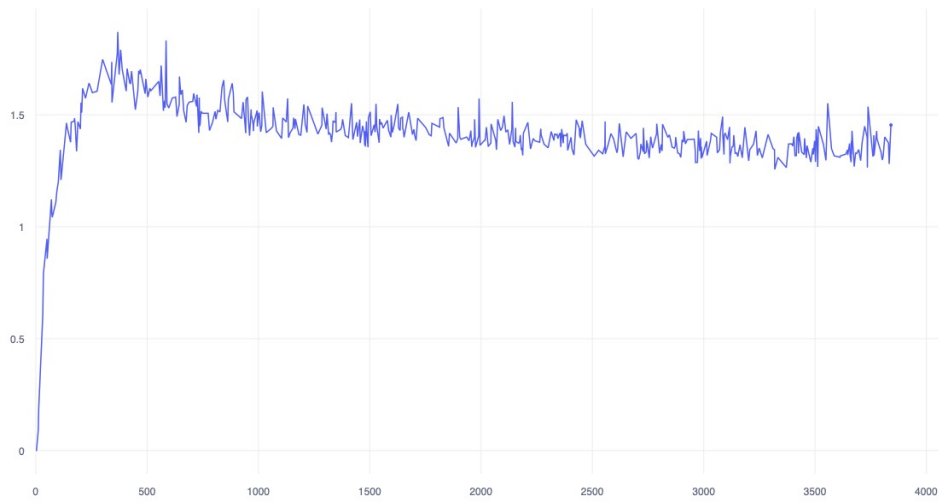


Figure 25: Covariance regularization between student's spatial predictor & teacher bottleneck layer on V-JEPA 2 encoder.

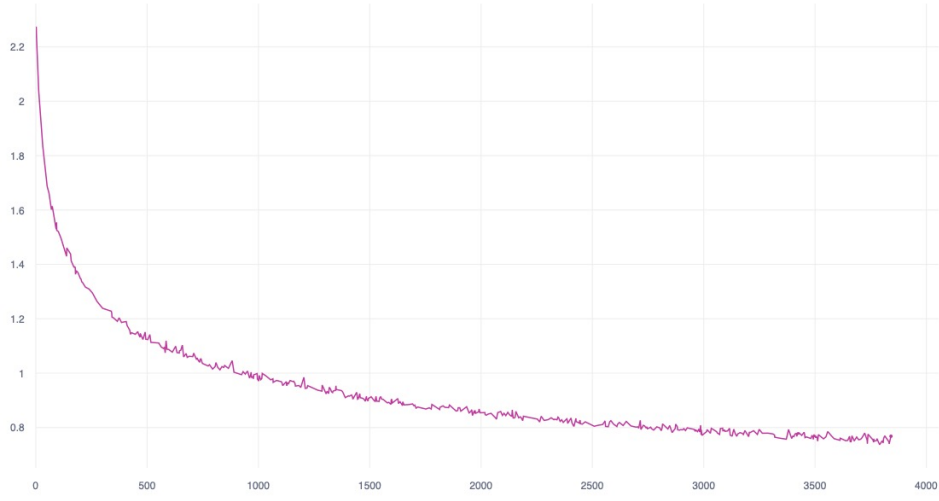


Figure 26: Encoder total loss on embedding predicting between student's spatial predictor & teacher bottleneck layer on V-JEPA 2 encoder.

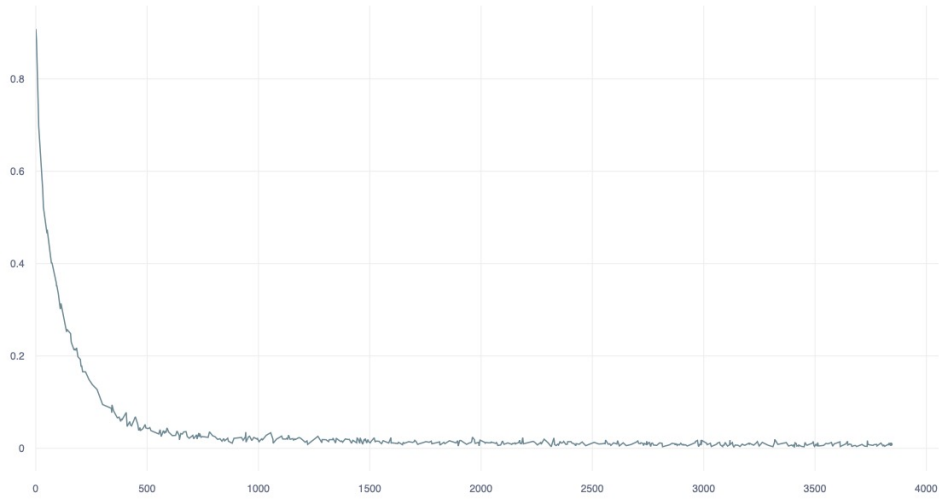


Figure 27: Encoder variance regularization on embedding predicting between student's spatial predictor & teacher bottleneck layer on V-JEPA 2 encoder.