HanoiWorld：基于联合嵌入预测架构的自动驾驶车辆控制器世界模型

Tran Tien Dat[1,2], Nguyen Hai An[1,2], Nguyen Khanh Viet Dung[1,2], Nguyen Duy Duc[1,2]

[1]数学与信息学学院，河内科技大学，

越南

[2]特洛伊大学，美国

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

arXiv:2601.01577v1 [cs.RO] 4 Jan 2026

摘要

当前将强化学习用于自主控制器的尝试需要大量数据，但效果不佳、表现不稳定，且无法把握并扎根于安全这一概念；此外，由于像素重建的特性，这些方法过度关注噪声特征。与此同时，当前基于自监督学习的方法利用联合嵌入预测架构（JEPA）学习高维表征，是一种颇具吸引力且有效的替代方案，因为这一思路模仿了人脑获取新技能的方式：通过想象和少量观测进行学习。本研究提出 Hanoi-World，这是一种基于 JEPA 的世界模型，采用循环神经网络（RNN）进行长期水平规划，并具有较高的推理效率。在不同环境下使用 Highway-Env 软件包开展的实验表明，与最先进的基线方法相比，该模型具备有效制定自动驾驶计划的能力，同时具有安全意识，且碰撞率颇具竞争力。

1 引言

自1986年卡内基梅隆大学开展首个自动驾驶车辆（AV）实验以来，开发自动驾驶车辆的研究领域在技术和真实世界的实际应用方面都取得了显著进展。人们期望车辆能够安全运行，同时应对不确定性、部分可观测性和多智能体环境带来的挑战（例如自车与周围车辆以及行人等障碍物之间的交互） [2, 3]。然而，正如 [4]所指出的，这些挑战确实限制了部署和测试此类基于强化学习的控制器的可行性，而以往的工作仅基于从物理环境进行简单迁移的范式

，这会因噪声和数据碎片化而导致训练不稳定性 [4–6]。

从技术角度来看，自动驾驶车辆的决策传统上依赖模拟器进行经验展开，并采用强化学习规划算法，且假设数据充足 [7, 8], ；而经典方法——例如蒙特卡洛树搜索（MCTS）或基于部分可观测马尔可夫决策过程（POMDP）的信念空间规划——需要在模拟器中投入大量计算开销进行实验滚动展开，且不进行策略训练 [9–11], 。尽管这些方法确实能够解决不确定性问题，但其可扩展性有限 [12]。此外，这种滚动展开策略往往会在较长时间范围内放大误差，导致模型不准确性上升 [13, 14]。此外，观测层面的预测往往优先考虑视觉和运动学保真度，将其作为重建难题，但这可能无法真正捕捉决策相关流形，进而导致物体控制效率低下 [15–17]。

灵感源于人类通过自动驾驶获取新技能的能力：人类能够基于当前与环境的互动，想象可能的未来场景——这体现了可供性理论以及人类记忆 [17–19],。这一能力可以形式化为模型设计与实现：采用基于表征的自监督学习范式，学习环境动力学。联合嵌入预测架构（JEPA）提出通过直接预测未来表征来学习潜在空间，而无需重构原始观测；其仅在编码器之间进行结构对齐，并强化信息可变性，而非噪声的随机性，以防止嵌入坍塌，如 [20]所示。最先进模型 V-JEPA-2 将这一思想扩展到大规模视频数据，从被动观测中学习动作动力学，并将其编码到所得表征中，之后可用于为轻量级动作条件控制器生成训练信号 [16]。与此同时，循环状态空间模型（RSSM）已被证明能够在部分可观测条件下，有效维护近似马尔可夫动态的紧凑潜在记忆，并以最少的表征实现长期规划 [21, 22]。采用编码器与长期 RSSM 规划模型的尝试表明，相较于基于 MCTS 的方法，该方案在提高训练效用的同时具有效率高、额外开销低的优势，并且仍能从交通参与者中产出以感知为基础的动作。

基于上述研究成果，本研究认为，世界模型设计有望受益于预训练编码器（如 *V-JEPA 2*）所生成的高质量自监督学习嵌入，并结合长期规划器的使用，在保持准确性的同时降低并最小化推理成本，以及实现可调节的模型驾驶质量。

T本研究的贡献包括以下 4 项重要贡献：

- 提出一种面向自动驾驶汽车的统一世界模型设计视角，强调在表征层面进行预测建模，而非在观测层面进行仿真，并将其命名为 HanoiWorld。

- 提出一种受 V-JEPA-2 微调启发的 JEPA 编码策略，用于从大规模视频数据中学习与决策相关的潜在表征。
- 集成基于 RSSM 的潜在记忆，以在部分可观测性下支持近似马尔可夫状态转移。

- 证明简单的基于多层感知机的行动者—评论家控制器能够在学习得到的潜在世界模型中有效训练，无需采用计算开销高昂的规划算法和复杂的策略架构。

本文其余部分安排如下：第2节聚焦于相关工作，并为相关挑战及其解决方案提供完整的概念和理论基础；第3节讨论所提出的世界模型设计；第4节将介绍实验描述、用例及结果讨论。最后，第5节对全文进行总结。

此外，我们将在 HANOI–WORLD代码库中发布实验代码库，以促进研究结果的可复现性。

2 相关工作

世界模型被提出，作为训练基于强化学习的控制器的一种新颖方法，以解决基于真实世界交互的模型所存在的不可靠、过度复杂和低效等问题，正如 [4] 所指出的那样；与此同时，车辆在持续遭遇遮挡、对周围世界了解有限的情况下运行，也引发了不可避免的不确定性问题。这些挑战相互交织，促使人们重新思考强化学习控制器的训练方法，转而采用紧凑而语义丰富的表征，以实现可扩展性和可靠性。这一思路受到人类生物学习机制的启发，即利用可供性和想象 [17, 19]进行学习。

2.1 从基于模型的强化学习到驾驶世界模型

驾驶世界模型的概念根源在于基于模型的强化学习（MBRL）：智能体学习动力学模型，然后通过模拟未来状态进行规划。MBRL 的发展得益于共享工具和标准化基准测试。OpenAI Gym [23]通过统一的

API，使跨任务比较算法变得更加容易；而 DeepMind 控制套件 [24] 则提供了一组精心整理的连续控制环境，促进了对学习和控制的严格评估。这些基础设施推动了动力学模型学习、表征学习和规划的迭代改进。

现代潜在动力学智能体体现了“将世界模型视为想象引擎”的观点。DreamerV3 [22] 通过学习紧凑的潜在状态，并借助想象展开轨迹优化行为，在各种环境中展现出强劲且稳定的表现。尽管大多数 Dreamer 风格的结果都来自真实驾驶之外的场景，但其核心设计原则——预测性潜在状态、随机动力学，以及通过想象轨迹进行规划或策略改进——对自动驾驶世界模型的设计产生了深远影响。关于具身智能体设计和强化学习泛化的互补观点强调，鲁棒性和可扩展训练方案与模型的原始容量同样重要 [25]。在自动驾驶中，这些原则还受到其他约束的影响：安全、分布偏移、长时程决策和多智能体交互。

越来越多的研究认为，评估世界模型的价值，应看其下游效用（例如改进规划或提升决策安全性），而不能只看生成保真度。规划中心视角指出，智能体可以利用学习得到的模型中的不完善之处，生成“良好”的展开轨迹，但这些轨迹并不对应真实世界。对具身世界模型的分析强调，安全应被视为首要考量，并呼吁制定能够揭示失效模式的评估方案，尤其要关注那些仅在闭环控制中才会出现的失效模式 [26]。在自动驾驶中，这些问题尤为突出，因为罕见事件主导着风险，而微小的建模误差也可能逐步演变为灾难性后果。

2.2 面向自动驾驶的自监督与预测式表征学习

自动驾驶能够提供大量无标签传感器流，但密集标注相对有限，因此自监督学习（SSL）成为构建世界模型的基础。在视觉领域，自监督学习从对比式和基于聚类的方法逐渐发展到预测式和基于蒸馏的方案。DINO [27] 表明，无需标签的自蒸馏也能学习到语义上有意义的特征；这一思路启发了驾驶专用预训练研究，旨在学习能够跨时间、视角和天气迁移的表征。

然而，驾驶数据给通用自监督学习带来了独特的挑战。对比学习需要定义能够在数据增强下代表相同底层内容的“正样本对”；在物体众多且自车运动迅速的驾驶场景中，简单的数据增强可能会破坏对应关系，导致负迁移。重建掩码输入的生成式自监督学习在处理三维数据时成本可能很高，还可能迫使模型

预测任意表面细节，而非规划相关语义。因此，近期有多项研究主张采用嵌入级预测和方差正则化，作为对比样本对或显式重建的替代方案 [28–30]。

一个极具影响力的概念框架是联合嵌入预测架构（JEPA）视角。该视角主张：根据输入中已知的部分，预测未知部分的嵌入，从而学习表征，而不是重建像素或使用负样本对 [17]。自动驾驶提供了一个很有说服力的应用场景：激光雷达或相机空间中的掩码区域可能对应多个合理的表面，但其语义（例如“汽车后方”“卡车后方的自由空间”）在嵌入空间中仍可能保持稳定。因此，JEPA 风格的方法能够更好地适应部分可观测性所固有的不确定性。例如，基于 JEPA 的激光雷达预训练会预测掩码区域的鸟瞰视图嵌入，并使用显式的方差正则化来防止表征坍塌；与稠密重建相比，这种方法在减少预训练计算量的同时，还能持续提升下游三维目标检测性能 [30]。

世界模型也受益于离散或结构化的潜在空间，这类空间能够稳定学习过程并提高样本效率。向量量化表征提供了一种实现路径，但码本坍塌可能限制容量。在线码本学习策略（例如基于聚类的向量量化更新）旨在保持所有码向量处于活跃状态，从而提高利用率以及重建/生成质量 [31]。在自动驾驶中，离散化潜变量可以改善可控性、支持高效的展开轨迹，并在几何因素与语义因素之间建立桥梁。

2.3 空间世界状态：鸟瞰视图、占用与几何抽象

许多驾驶世界模型采用空间锚定的世界状态，而非纯粹抽象的潜向量。鸟瞰视角表征提供了一种便捷的坐标系，与规划相一致：它能够自然地表示车道、可行驶空间和其他交通参与者，并有助于传感器融合。鸟瞰视角表征还减轻了视角变化带来的负担，使模型能够专注于动力学，而非透视变换。因此，鸟瞰视角特征被广泛用作感知和预测的中间状态，并成为世界建模的天然基础 [32, 33]。

基于占用的表征通过对三维自由空间和遮挡进行建模，扩展了鸟瞰视图，这对安全至关重要。能够预测占用情况的世界模型可支持碰撞检测、可见性推理和不确定性下的规划。近期面向激光雷达的世界模型研究强调，仅基于相机的生成可能产生视觉上合理、但几何上不一致的未来状态；而占用预测则可以施加物理约束并保留三维结构 [34, 35]。这些研究凸显了不仅要表征物体，也要表征空旷空间的重要性，因为不存在障碍物与存在障碍物同样具有

与规划相关的意义。

几何抽象也与地图构建和场景先验密切相关。高精地图编码了车道拓扑、边界和人行横道；在若干世界建模流程中，地图被视为世界状态的一部分，可以作为生成的条件信号，也可以作为待预测的潜在因子。联合推理智能体轨迹与地图结构的方法旨在确保生成的未来符合道路几何结构和交通规则 [36, 37]。此外，关于“物理世界模型”的综述强调，有效的世界模型不仅应捕捉统计规律，还应捕捉以物理为基础的结构和因果关系，尤其是在超出训练分布进行外推时 [6]。

2.4 Transformer、注意力机制与以交互为中心的建模

Transformer和注意力机制已成为自动驾驶的核心技术，因为它们能够支持长程依赖，并灵活融合异构输入。在感知和预测中，注意力机制有助于将计算集中于场景中最相关的交通参与者和区域，也为建模交通参与者之间的交互提供了自然的方法。因此，基于Transformer的架构广泛应用于现代驾驶世界模型，尤其适用于结合多视角图像、点云和地图特征的场景 [38]。

以交互为中心的建模至关重要，因为其他交通参与者会彼此互动，也会对自车做出反应。将交通参与者视为彼此独立的世界模型，在密集交通、汇入路段、交叉路口及其他存在交互的情境中，可能会系统性地失效。近期研究强调，表征应能捕捉智能体间耦合、意图和路权，通常借助注意力机制或图式消息传递来表示联合未来 [39, 40]。这些理念与具身学习的整体趋势一致：世界模型不仅要表征被动动力学，还必须呈现动作的后果以及其他参与者的策略性响应。

Transformer 在自监督预训练以及构建可跨任务迁移的、具有可扩展性的“基础模型”方面也发挥着重要作用。探索世界模型大规模训练方案的论文认为，当模型具有足够强的表达能力且训练数据多样时，表征、预测和规划可以进行协同训练[41, 42]。这一观点与如何在单个学习得到的模型中统一感知与控制这一更广泛的讨论相联系，并促使人们设计能够同时基于感官历史和动作序列进行条件化的架构。

2.5 多智能体与车联网世界模型：从协同感知到协同预测

自动驾驶本质上是多智能体的，而互联自动驾驶则通过车联网通信引入了额外的信息渠道。协同感知是“分布式世界建模”的早期实例之一，在这一过程中，多辆车辆或基础设施传感器共享数据或特征，以减少遮挡并扩展感知范围。用于协同感知的公开基准推动了对融合策略的系统性评估，包括早期融合、后期融合和中间融合 [43]。中间特征共享通常被视为准确性与带宽之间的平衡方案，并推动了能够容忍定位噪声和间歇性通信的学习式融合模块的发展。

基于Transformer的融合架构通过支持对交通参与者和多尺度特征进行注意力计算，拓展了协同感知。面向车联网的Transformer引入了相应机制，以处理位姿不确定性、传感器模态差异，以及通信延迟导致的时间错位 [44]。这些场景凸显了单智能体与多智能体世界模型之间的关键区别：“状态”并非仅仅是自车所看到的内容，而是一组分布式的局部观测，必须将其整合为连贯的表征。

真实世界数据集对于验证这些理念至关重要，因为仿真可能无法充分呈现传感器伪影和道路交互的真实分布。V2V4Real提供了真实世界的多车数据，包含激光雷达、RGB、三维边界框和高精地图，专为协同检测、跟踪和仿真到现实迁移等协同感知任务而设计 [45]。此类数据集表明，未来的自动驾驶世界模型应从一开始就面向多智能体融合与预测进行设计，而非将单智能体模型改造后用于协同场景。

协同世界模型必须超越共享信息下的“当前状态融合”，实现“未来状态预测”。这需要对分布式观测如何演变、交通参与者可能如何相互响应，以及通信延迟如何影响信念更新进行建模。研究分布式场景下协同预测与交互的工作指出，应采用稳健且具备不确定性感知能力的融合方法，并构建能够在消息延迟或丢失时平稳退化的模型 [46, 47]。这项研究还与模拟器和基准测试环境相关，可在受控条件下测试通信感知策略 [48, 49]。

2.6 安全关键建模：事故预测、风险预判与弱势道路使用者

驾驶世界模型的一个主要价值在于更好地预判罕见且安全关键事件。传统的轨迹预测基准测试强调非关键行为中的平均

误差，而这类误差可能与事故风险并不相关。因此，事故预测基准在驾驶世界模型评估中发挥着重要作用。DeepAccident 在车联网环境中引入了运动和事故预测，针对一种具有挑战性的场景：模型不仅必须预测交通参与者将向何处移动，还要预测是否会发生碰撞、何时发生碰撞以及哪些参与者将受到影响 [50]。此类基准测试迫使世界模型对可能细微或具有长程影响的风险因素进行表征，例如被遮挡的交叉车流或快速变化的间距。

风险预判也促使人们对遮挡和自由空间进行显式建模，而这对碰撞规避至关重要。基于占用的世界模型和以激光雷达为中心的方法强调，对“未知”区域和可见性进行表征至关重要：预测遮挡物后方的智能体，与预测可见智能体是本质上不同的问题。因此，近期面向风险的世界建模研究将几何先验与预测不确定性相结合，旨在表征多种合理的未来情景，而非单一的确定性轨迹 [34, 51]。

安全也取决于对弱势道路使用者（VRU）进行建模，例如行人和骑行者，尤其是在密集的混合交通中。包含弱势道路使用者的高分辨率轨迹数据集为交互感知世界模型提供了必要的监督信号。On-SiteVRU 提供复杂城市场景中的高密度轨迹数据，具有精细的时间分辨率和情境信息，支持对弱势道路使用者行为建模、交互风险和安全评估的研究 [52]。这些数据集补充了以车辆为中心的基准测试，并凸显出世界模型必须能够推理具有不同运动学特征、目标和社会规范的异质参与者。

最后，近期分析指出，世界模型的安全评估应当是多维度的：生成真实性还不够，闭环规划测试可以揭示模型利用或累积误差。“安全挑战”视角呼吁开展压力测试、反事实评估，并制定指标来量化在分布偏移下，使用该模型进行规划是否能改善安全结果 [26]。这一主题有力地推动了相关研究：世界模型不仅要有较强的表达能力，还必须经过校准、具备鲁棒性且可审计。

2.7 生成式仿真：扩散、视频世界模型与语言条件场景生成

生成式世界模型在扩散模型和可控视频生成技术进步的推动下迅速发展。在自动驾驶领域，视频生成颇具吸引力，因为它能够合成各种场景，包括难以在真实数据中捕捉到的罕见事件。一些研究将驾驶世界建模视为给定当前上下文和候选动作条件下对未来观测进行条件生成的过程，通常还会将地图、边界框或轨迹作为条件信号 [53, 54]。综述和

该领域的方法论文指出，基于扩散的动态建模能够生成高保真样本，并可融入结构化先验，但也提到了时间一致性和动作可控性方面的挑战 [6, 26]。

DriveDreamer-2 展示了一条颇具影响力的方向：将语言模型与世界模型相结合，从而实现用户驱动的仿真。在 DriveDreamer-2 中，大语言模型将用户提示转换为智能体轨迹，基于扩散的组件生成与这些轨迹一致的高清地图，统一的多视角视频生成器则合成多摄像头驾驶视频 [36]。这条流程旨在以用户友好的方式生成长尾场景（例如突然加塞），并改善下游感知训练。DriveDreamer变体等相关工作着重提升跨视角一致性和时间稳定性，并表明结构化的中间表示（轨迹、地图）有助于提高可控性 [37]。

生成式建模也用于以规划为中心的场景中的反事实评估和数据增强。“Dreamer”风格的展开轨迹在潜在空间中提供想象出的未来，而扩散模型或视频模型则提供丰富的感官预测。CarDreamer 等集成系统旨在将世界建模与下游决策相结合，并将学习得到的想象用作驾驶策略的训练信号 [55]。其他大规模训练工作则探索如何训练同时支持预测与规划的世界模型，通常会结合自监督目标与策略学习 [16, 41]。纵观这些工作，一个反复出现的开放问题是如何让生成保真度与决策效用相协调：模型可能生成看起来逼真的视频，却仍不适用于安全关键规划。

2.8 评估与基准测试：实用性、泛化与现实差距

一个持续存在的难题是，没有单一指标能够全面反映驾驶世界模型的质量。生成指标，如 FID——弗雷歇初始距离，或 FVD——弗雷歇视频距离，可分别由公式（1）和公式（2）表示。这些指标衡量视觉真实感，但可能忽略物理合理性或规划相关性 [26]。

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

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

鉴于：

- $\mathcal{X}_r$ 和$\mathcal{X}_g$ 分别表示真实图像样本和生成图像样本的集合。

- $\mathcal{V}_r$ 和$\mathcal{V}_g$ 分别表示真实视频样本和生成视频样本的集合。
- $\boldsymbol{\mu}_r \in \mathbb{R}^d$ 和 $\boldsymbol{\mu}_g \in \mathbb{R}^d$ 表示从真实样本和生成样本中提取的深度特征嵌入的经验均值向量。
- $\boldsymbol{\Sigma}_r \in \mathbb{R}^{d\times d}$ 和 $\boldsymbol{\Sigma}_g \in \mathbb{R}^{d\times d}$ 表示相应特征嵌入的经验协方差矩阵。
- $\|\cdot\|_2$ 表示欧几里得（ℓ_2）范数。
- $\mathrm{Tr}(\cdot)$ 表示方阵的迹算子。
- $(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g)^{\frac{1}{2}}$-表示两个协方差矩阵乘积的矩阵平方根。
- d 表示特征嵌入空间的维数。

基于平均距离误差（ADE）和最终距离误差（FDE）的轨迹指标衡量的是平均误差，但可能遗漏长尾风险，正如 [26, 56] 所指出的。此外，以规划为中心的评估衡量下游驾驶性能，但需要闭环测试，并且可能受到模拟器局限性的干扰。因此，近期研究主张进行多维度评估，包括：（i）预测准确性与校准，（ii）分布偏移下的鲁棒性，（iii）对检测、跟踪或规划等下游任务的有用性，以及（iv）安全关键压力测试 [26, 56]。

数据集和基准测试通过决定何为可测量之物来塑造研究进展。Waymo Open 为训练和评估必须处理真实传感器噪声与复杂城市场景的模型提供了规模与多样性 [57]。V2V4Real 等协作数据集增加了多智能体融合、定位误差和通信约束方面的挑战 [45]。DeepAccident 等事故聚焦型基准强调遮挡条件下的罕见事件预判与交互 [50]。OnSiteVRU 等弱势道路使用者聚焦型数据集则突出混合交通与细粒度交互 [52]。这些资源共同表明，“通用”驾驶世界模型必须从多样化数据源中学习，并必须在多样化任务上进行评估。

仿真与真实世界之间的现实差距仍是一个核心问题。仿真支持可扩展的闭环测试，但可能无法充分呈现罕见行为或传感器伪影。因此，若干研究强调仿真到现实迁移、混合训练（真实 + 合成），以及能够检测对模拟器偏差过拟合的评估协议 [33, 45]。框架讨论和综述指出，需要在不同工具链之间进行标准化报告并确保可复现性，因为看似微小的实现选择就可能主导结论 [5, 6]。

2.9 其他考量：综述、物理与平台相关考量

所提供语料中的若干其他研究脉络，有助于将自动驾驶世界模型与更广泛的科学和工程问题联系起来。首先，综述和观点文章强调，“世界建模”是一个涵盖表征、学习目标和下游用途的总称，并指出进展取决于对哪些内容应显式建模、哪些应隐式建模作出有原则的选择 [17, 58]。近期的综述类研究探讨了如何将物理知识（动力学、约束和因果结构）嵌入学习表征，以提升鲁棒性和可解释性 [6, 54]。这些讨论与自动驾驶尤为相关，因为失败往往源于违反基本的物理合理性（例如，不可能运动、不一致遮挡），或源于训练数据中的伪相关——而 [15, 59] 提出的一种解决方案是，将原始且确定性的世界物理模型作为微观近似加以整合，以学习物理与可供性之间的对齐；这正是该世界模型所追求的目标 [19]。

此外， [23, 24, 48, 49] 的研究强烈指出，在具身领域中，评估与训练和测试所用的环境及平台密不可分。实际基准测试的选择——传感器套件、地图可用性、交通参与者多样性，乃至模拟器保真度——都会影响世界模型的学习内容及其泛化方式。以平台和工具为导向的观点强调，可复现研究需要仔细说明环境、数据管线和评估流程。在自动驾驶领域，这与一个广为人知的矛盾有关：闭环测试需要丰富的仿真，而真实性则依赖真实世界数据；这确实促使人们探索新型混合管线，利用仿真探索长尾场景，并以真实数据为模型提供真实传感器统计特征的基准，从而尝试模仿人类感知物理可供性的能力 [19]。

最后，正如 [16, 17, 58] 在世界建模领域的研究所示，不同模态能够为下游任务和连续任务训练生成有效的表征；然而，这种内部表征应能抵御环境的随机性和不可预测性，而基于序列的推理模型可以通过潜在空间过度预测来解决这一问题 [21]。

2.10 设计启示

综观这些文献，可以总结出几条设计经验，它们构成了本文的理论基础。首先，在自动驾驶领域，越来越倾向于采用在嵌入空间中运行的预测性表征学习（例如 JEPA 式目标），而非像素级重建，因为前者更契合部分可观测性，也更有利于学习跨多模态未来的潜在表征能力

。其次，鸟瞰视图和占用等具有空间基础的状态，是实用且符合规划需求的世界状态，尤其是在融合地图先验并显式建模自由空间时[32, 34, 35]。第三，多智能体交互和车联网通信正成为核心要求：驾驶世界模型必须协调分布式观测，并在延迟和不确定性条件下预测联合未来 [44–46]。最后，安全要求基准测试能够针对罕见事件和事故预测；同时还应兼顾效率和基于性能的指标，因为奖励仍是后续研究需要解决的一项挑战 [26, 50, 52]。

3 拟议的世界模型与系统架构

本节将重点介绍世界模型的构建方案、完整的智能体交互流程，以及将采用的训练算法。

3.1 符号说明

我们的问题可形式化为一个确定性马尔可夫决策过程（MDP），其中使用内部连续随机状态和确定性状态，并在有限时域环境中生成确定性连续动作，如 [22, 60] 所示。该 MDP 可定义为式（3）中的元组。

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

其中，$\mathcal{S}$ 表示（可能为连续的）状态空间，$\mathcal{A}$表示（可能为连续的）动作空间，$\mathcal{P}(s_{t+1} \mid s_t, a_t)$ 表示状态转移动态，其中包含随机状态 s_t ，用于捕捉转移过程中的环境随机性模式；确定性状态 h_t 则由循环规划器根据历史上下文生成； $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 为奖励函数， $\gamma \in [0, 1)$ 为折扣因子。对于所采用的设置，我们直接沿用 DreamerV3 的设置，将 $\gamma = 0.997$ 设为折扣因子。

我们尝试使用自车加速度和转向角（单位为弧度）作为动作空间，以便与仿真环境进行交互。

3.2 整体系统架构

所提出的整体系统架构包含多个模块，例如基于 [61] 提供的 HighwayEnv 软件包构建的环境接口，用于智能体交互。我们还加入了容量有限的转移队列，作为世界模型的经验收集器；世界模型可以从队列中随机采样满足序列长度 T 要求的回合，并提供想象的

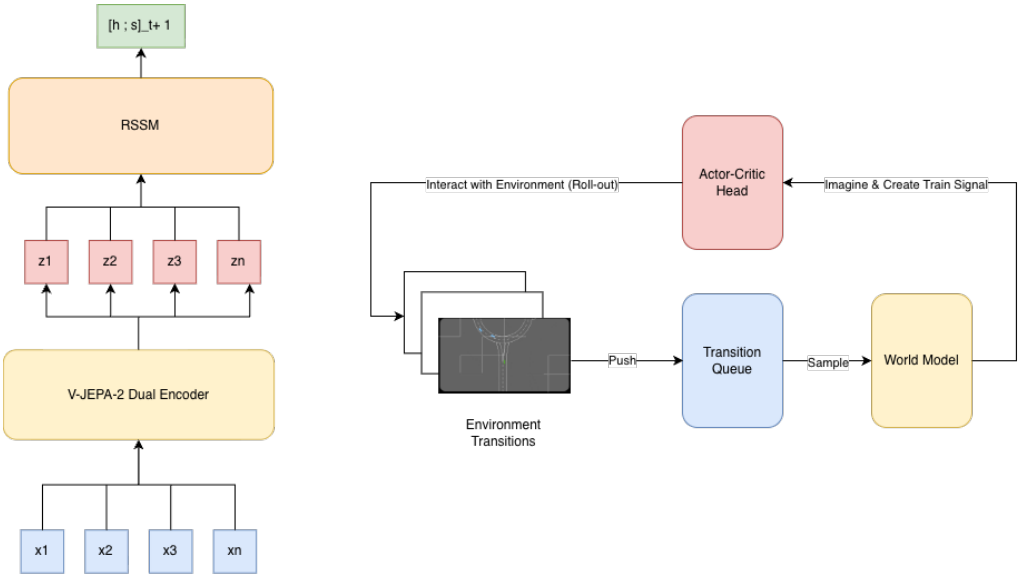

图 1： 所提出的 HanoiWorld 世界模型（左）包括一个基于 [62] 提出的 V-JEPA-2 检查点的视觉编码器，以及 [21] 提出的 RSSM 主干网络，用于对环境中的下一种可能转移（绿色模块）进行长期规划。右侧展示了环境的整体系统架构，其设计旨在通过创建智能体交互反馈回路，在数据稀缺的情况下实现有效的滚动展开和模型训练。

嵌入作为训练信号，传递给演员-评论家控制器头，以生成用于响应环境的动作，并形成一个紧密的交互循环，直到智能体到达终止状态。整个 HanoiWorld 的架构如图 1 右侧所示。

转移队列应存储回合列表（例如，有效回合的引用），这些回合的转移长度满足特定阈值—— T；该序列应包含鸟瞰 RGB 图像，以及智能体交互的元数据，例如智能体的动作、自车的动作，以及在仿真过程中为这些序列产生的奖励描述。

3.3 HanoiWorld——世界模型

基于 V-JEPA 2 的编码器

HanoiWorld 的主要创新在于纳入了 [62] 提出的、采用自监督学习方式训练的强大预训练图像编码器。具体而言，V-JEPA 2 旨在学习与环境交互相关的基本知识，例如运动和物体移动，而不像 DreamerV3 系列编码器 [22, 63]那样通过重建尝试关注随机且嘈杂的像素细节。V-JEPA 2 编码器将通过自监督方式训练，根据对应输入预测其掩码版本，并使用超过 100 万小时的视频数据进行训练；同时通过指数

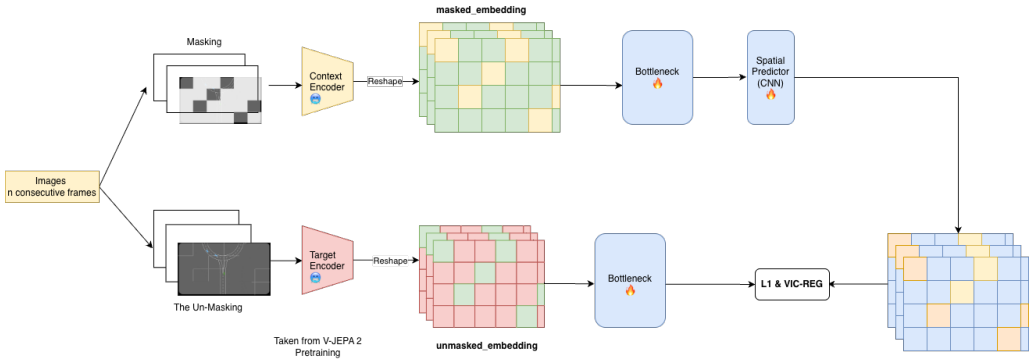

图 2： 所提出的 HanoiWorld 编码器架构概览。使用预训练且冻结的 V-JEPA 2 编码器 [62] 作为高质量表征骨干网络，以提升有限数据场景下的训练效率和嵌入鲁棒性。训练下游瓶颈多层感知机（MLP），将高维表征投影到大小为 1024×128的紧凑且适用于任务的潜在空间中。与此同时，学生编码器分支引入额外的二维卷积神经网络（CNN）模块来预测空间表征。两个分支通过式（5）中的 ℓ_1 对齐损失和 VICReg 正则化目标 [20]进行联合优化。

移动平均（指数移动平均，EMA）学生-教师方案可防止嵌入坍塌，同时在结构上保留基本特征 [64]。

EMA 教师编码器。 令 θ 表示学生编码器 E_θ 的参数，并令 $\bar{\theta}$ 表示教师编码器的参数。教师参数通过指数移动平均进行更新：

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

其中 $\tau \in [0, 1)$ 为动量系数。为防止表征坍塌，梯度不会通过教师编码器反向传播；在我们的实验中，设置 $\tau = 0.996$。

掩码表征预测。 给定一个视频 y 和一个通过移除部分时空图块得到的掩码视图 x，编码器 E_θ 从可见令牌中提取表征。训练预测器 P_ϕ 来预测掩码令牌的表征。

主要训练目标定义为 L1 损失，如式（5）所示：

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

其中， Δ_y 表示用于指示掩码图块位置的可学习掩码令牌，sg(·)表示停止梯度算子。损失仅应用于掩码图块，并且在微调过程中，我们采用随机掩码的图块掩码，如 [65] 所示。

The 利用采用掩码模式的双分支自监督训练 s

这出现在 [66], 的研究中。该研究提出，让编码器必须猜测样本中的掩码区域，从而制造挑战，防止嵌入塌陷；与此同时，无需获取环境的完整几何结构，也能学习并预测占用语义。

基于 RSSM 的推理模型

我们设计的 RSSM（循环状态空间模型）遵循 [21, 22], 提供的 DreamerV3 架构代码库。RSSM 被视为内部记忆模块，用于存储承载长期历史语义的确定性记忆；随机潜变量则基于环境编码信号进行不确定性预测，以捕捉世界演化，并由世界状态转移条件化智能体的预期先验。生成世界动力学的过程可由下方的公式（6）表示。

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

其中:

- f_θ 表示由 θ参数化、以循环神经网络实现的循环动力学函数。
- h_t 表示时间步 t的确定性潜状态。
- a_t 表示智能体在时间步 t执行的动作。
- z_t 对应于时间步 t的观测，在我们的设定中，它是相应时间步的 V- JEPA-2 编码嵌入。
- r_t 表示智能体在时间步 t获得的预测奖励。
- 所有条件分布 $p_\theta(\cdot)$ 均由多层感知机（MLP ）解码器网络。

此外，由于 HanoiWorld 从 Dream-erV3 迁移了基于 RSSM 代码的架构，该模型还加入了继续预测器，根据编码器嵌入 o_t 和历史潜在图像 z_t[22]来判断回合是否会继续——如式（7）所示。为简化假设，延续信号被视为伯努利随机变量。

$$c_t \sim p_\theta(c_t \mid h_t, z_t), \tag{7}$$

其中 $c_t \in \{0, 1\}$ 表示回合在时间步 t 是否继续。

按照 DreamerV3 的原始建议，使用式（8）中的二元交叉熵，将继续预测器作为逻辑回归模型进行训练

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

- $c_t \in \{0, 1\}$ 表示时间步的真实二元延续标签 t.
- $\hat{c}_t \in (0, 1)$ 表示模型生成的预测延续概率 .

3.4 演员-评论家训练头

HanoiWorld 演员-评论家框架直接源自 DreamerV3 基于奖励和价值的学习范式。它与 [67], 提出的经典理论共享核心概念基础；该理论强调将行动者 网络（策略网络）与评论家 网络分离：前者通过选择动作作出响应，后者则通过价值函数评估智能体的表现，从而支持更有效的规划。然而，与在 [67], 实际环境状态上运行的经典演员-评论家形式不同，DreamerV3 演员-评论家完全在学习得到的潜在环境中运行。具体而言，策略网络 π_θ 和价值网络 v_ψ 都定义在由 RSSM 生成的潜在状态 ℓ_t 上，如式（9）所示。

$$\ell_t = (h_t, z_t), \qquad a_t \sim \pi_\theta(a_t \mid \ell_t), \qquad v_\psi(R_t \mid \ell_t). \tag{9}$$

已知：

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

- $\ell_t = (h_t, z_t)$ 表示时间步 t 的潜在状态，分别由 RSSM 的确定性分量和 V-JEPA-2 编码器的随机性分量组成。
- R_t 表示时间步 t 的回报，定义为一个回合内未来奖励的折扣总和，折扣因子为 $\gamma = 0.997$。

3.5 训练流程与目标

HanoiWorld 的训练目标包括在编码器模块中训练空间预测器和预测器、训练基于RSSM动力学的模型，随后进行演员-评论家训练。鉴于 HanoiWorld 的训练策略仅模仿 DreamerV3 在动力学和演员-评论家网络上的训练流程，并未像 [22] 那样采用基于重建的编码器训练。

编码器训练

对于瓶颈和空间预测器的训练，除了使用式（5）中的 L1 损失外，我们还考虑采用 [20] 提出的方差相关性和协方差正则化——参见式（11）和式（12）。

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

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

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

已知：

- D 表示嵌入维度（D=128），而 $\tilde{\mathbf{z}} \in \mathbb{R}^{(BN)\times D}$ 表示通过重塑批次和词元维度得到的展平嵌入。
- $\tilde{z}_{:\,d}$ 表示所有 BN 个样本中第 d个嵌入维度，而 ε 是一个用于数值稳定性的小常数。
- 式（11）中的方差正则化为每个嵌入维度设置了最小标准差，从而防止表征坍塌。

- ϵ 表示用于数值稳定性的小常数，如 $\epsilon = 1\text{e-}4$
- 公式（13）中的协方差矩阵 Cov($\tilde{z}$) 是在用均值 $\bar{\mathbf{z}} = \frac{1}{BN}\sum_{i=1}^{BN}\tilde{\mathbf{z}}_i$对嵌入进行中心化后计算得到的。 —

公式（12）中的协方差正则化会惩罚协方差矩阵非对角线元素的平方，从而鼓励不同嵌入维度之间去相关。

瓶颈的训练采用以下加权和损失函数 [20,62, 66], ，权重设置为 (α, β, γ) 对应 (1.0, 1.0, 0.1)，以确保稳定性并防止坍塌，同时松散地控制教师的瓶颈与学生的预测器之间的协方差。

$$\mathcal{L}_{\text{encoder}} = \alpha\,\mathcal{L}_{\text{align}} + \beta\,\mathcal{L}_{\text{var}} + \gamma\,\mathcal{L}_{\text{cov}}. \tag{14}$$

我们在二维鸟瞰视图数据集上对编码器进行微调。该数据集由 [56] 提出的 NuScenes 数据集预处理而来：将基于激光雷达点云的数据渲染为 RGB 图像，并添加有关障碍物和环境运动的元数据，以便 V-JEPA 2 编码器能够处理这些图像。

RSSM 动态模型训练

对于 RSSM 模块的训练，我们采用预测器损失，即对环境转移中的实际观测、奖励函数和继续标志的预测概率取对数后求和，再取其相反数，如公式（15）所示；采用动态性损失，通过最小化由实际观测稳定地产生的编码器后验分布与世界模型根据先验潜在记忆所想象出的先验分布之间的 KL 散度来计算，如公式（16）所示；并采用表征损失，以锚定编码器，避免其偏离潜在世界模型的预期，如公式（17）所示。该损失设置遵循 [22]提出的方法，并使用与 DreamerV3 此前配置相同的权重集。

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

$$\mathcal{L}_{\text{dyn}}(\phi) = \max\Big(1,\ \mathrm{KL}\Big(\mathrm{sg}[q_\phi(z_t \mid h_t, x_t)]\ \Big\|\ p_\phi(z_t \mid h_t)\Big)\Big), \tag{16}$$

$$\mathcal{L}_{\text{rep}}(\phi) = \max\Big(1,\ \mathrm{KL}\Big(q_\phi(z_t \mid h_t, x_t)\ \Big\|\ \mathrm{sg}[p_\phi(z_t \mid h_t)]\Big)\Big). \tag{17}$$

基于 REINFORCE 的演员-评论家学习

HanoiWorld 算法中的演员-评论家组件基于 [68], 提出的 REINFORCE 算法，该算法建议智能体（具体而言是演员的策略网络）优化并偏好能够带来最佳回报的动作。然而，我们采用的算法遵循 [22], 中的 DreamerV3 实现。该实现建议演员和评论家均基于潜在 RSSM 进行想象式展开，其中评论家层根据未来的潜在特征预测可能的累积奖励，而演员的策略网络则基于优势信号进行优化，并通过熵正则化促使智能体探索想象的近期未来中的潜在动作；熵则按固定常数缩放。演员-评论家算法的训练过程在算法 1 中有明确说明。

4 实验结果

本节将重点介绍用于评估 HanoiWorld 性能的实验，并将其与不同的基线方法进行比较。

4.1 实验描述

为评估 HanoiWorld 相较于不同方案的性能，将在 HighwayEnv 环境中开展实验。

算法 1： 通过想象展开进行演员-评论家训练

输入： 世界模型参数 ϕ;
演员参数 θ;
评论家参数 ψ;
想象时域 H;
折扣因子 γ;
λ-回报参数 λ;
熵系数 $\beta = 3\text{e-}4$
从真实经验中采样后验潜状态 $(h_0,\ z_0)$;
初始化想象轨迹缓冲区;
对于 $t = 0$ **至** $H-1$ **执行**

- 计算特征 $f_t \leftarrow f_\phi(h_t,\ z_t)$;
- 采样动作 $a_t \sim \pi_\theta(\cdot \mid f_t)$;
- 预测下一个潜在状态 $(h_{t+1},\ z_{t+1}) \sim p_\phi(\cdot \mid h_t,\ z_t,\ a_t)$;

算法 1： 算法 1（续）

对于 $t = 0$ **到** $H-1$ **执行**

- 预测奖励 $\hat{r}_{t+1} \leftarrow r_\phi(h_{t+1},\ z_{t+1})$;
- 预测延续 $\hat{c}_{t+1} \leftarrow c_\phi(h_{t+1}, z_{t+1})$;

计算所有 $t \in \{0, \ldots, H\}$ 的价值预测 $V_t \leftarrow V_\psi(f_t)$;

计算 λ-回报目标;
设置有效折扣 $\gamma_t \leftarrow \gamma \cdot \hat{c}_t$;
设置自举目标 $G_H^\lambda \leftarrow V_H$;
对于 $t = H-1, H-2, \ldots, 0$ **do**

- $G_t^\lambda \leftarrow \hat{r}_{t+1} + \gamma_{t+1}\Big((1-\lambda)\, V_{t+1} + \lambda\, G_{t+1}^\lambda\Big)$;

评论家更新;
最小化 λ-回报的负对数似然：;
$\mathcal{L}_{\text{value}}(\psi) \leftarrow -\mathbb{E}_t\Big[\text{日志 } p_\psi(G_t^\lambda \mid f_t)\Big]$;

行动者更新;
计算优势（基线）： $A_t \leftarrow \text{sg}(G_t^\lambda - V_t)$;
最小化行动者损失：;
$\mathcal{L}_{\text{actor}}(\theta) \leftarrow -\mathbb{E}_t[\text{日志 } \pi_\theta(a_t \mid f_t)\ A_t + \beta\, \mathcal{H}(\pi_\theta(\cdot \mid f_t))]$;
更新评论家参数 ψ;
更新演员参数 θ;

由 [61], 提供的套件包括 高速公路、环岛、并道，这些内容将在第 4.2 小节中详细讨论。

我们在设计中选择的对比世界模型包括 [22], 提出的 DreamerV3、 + 提出的基于 VQ-VAE 和卷积长短期记忆网络的模型，以及提出 HanoiWorld 的智能体 [13], 。基线模型介绍如下：

- 基于 VQ-VAE 编码器和卷积长短期记忆网络规划器的方案由 [13] 提出，是一种小型离散潜在世界模型，旨在实现潜在空间中的高效推理与规划，其思想继承自 [18]。该 VQ-VAE 方案使用近端策略优化（PPO）算法，在 Atari 游戏的 10 万个回合上进行设计，由 [69] 提出，用于在离散的想象展开轨迹中进行潜在空间规划。

- 目前， [22] 提出的 DreamerV3 已达到最先进水平，并凭借 RSSM 网络实现了长时域规划能力。这启发了 HanoiWorld 的设计，旨在学习鲁棒且具有泛化能力的环境无关模型，用于预测合理的下一个潜在状态、奖励和延续。该模型通过想象展开为多层感知机层提供训练信号，以预测奖励和价值；行动者则采用 REINFORCE 梯度更新，或结合来自 RSSM 的学习信号进行混合更新，从而实现鲁棒、稳定的学习。

对于评估指标，选定的两个指标——平均奖励（例如得分）和碰撞率如下：

- 平均奖励/得分是 [18]用于评估交通参与者规划效率的指标，即跨回合各步骤奖励信号的平均值，如公式（18）所示

- 碰撞率是 [70] 建议的指标，可通过自车发生碰撞的回合数占总回合数的比例进行估算，如公式（19）所示

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

鉴于：

- $\bar{R}$ 表示所有评估回合中的平均自动驾驶奖励。
- N 表示评估回合的总数。
- $i \in \{1, \ldots, N\}$ 索引回合。
- t表示一个回合中的时间步。

- $r_{i,\ t}$ 表示在回合 i的时间步 t获得的即时奖励。

解释： R 的值越高，表示自动驾驶表现越好，即控制更加平稳（摆动更少），轨迹更加安全（碰撞更少）。

$$\text{Collision Rate} = \frac{\#\,\text{collision episodes}}{\text{total episodes that being evaluated}} \tag{19}$$

4.2 案例研究

该实验在交互式环境中进行，用于检验模型在不同驾驶情境下的适应性和泛化能力，包括不同的道路拓扑结构、车道合流场景以及环岛导航。

每个场景的信息和表格描述见表1。

每个环境的任务描述

Highway-Env（例如 highway-v0）——图3旨在展示模型在多车道场景中控制车辆运动的能力。在该场景中，自车需要与交通参与者互动，并针对不同智能体的行为做出车道切换决策（周边车辆可能加速、突然减速或变更车道）。该环境旨在测试模型保持稳定且安全的行驶速度的能力，以及通过根据周边交通参与者的情况进行车道切换来体现的灵活性 [61]。

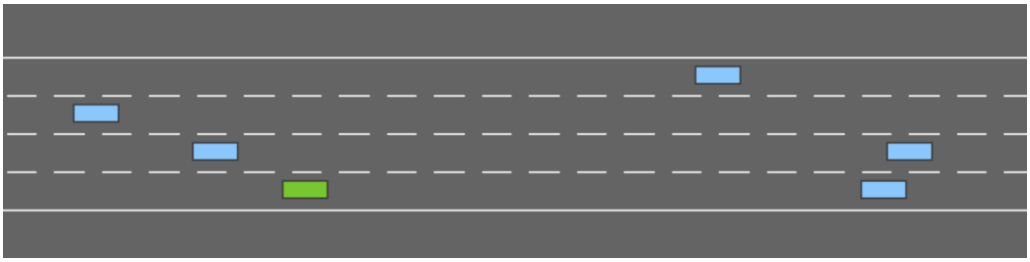

图 3： 高速公路环境 highway-v0——取自 [61]的一个快照

并道环境（例如 merge-v0）——图 4。在仿真中，智能体必须尝试从一条道路分支切换到另一条道路分支，以应对车道合流挑战。由于速度控制以及智能体在合流过程中避免碰撞所需的详细规划，这项任务颇具挑战性 [61]。该场景的目标是让模型尝试理解并预测不同交通参与者的快速物理状态转换，从而做出恰当的并道决策。

图 5 所示的环岛（roundabout-v0）是 merge-v0 的复杂且逼真的版本，交通参与者在导航时有多个选择，而且密度

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.

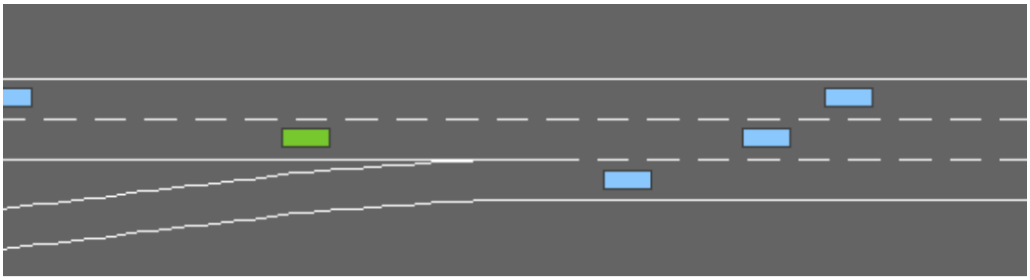

图 4： roundabout-v0 环境——取自 [61]的快照

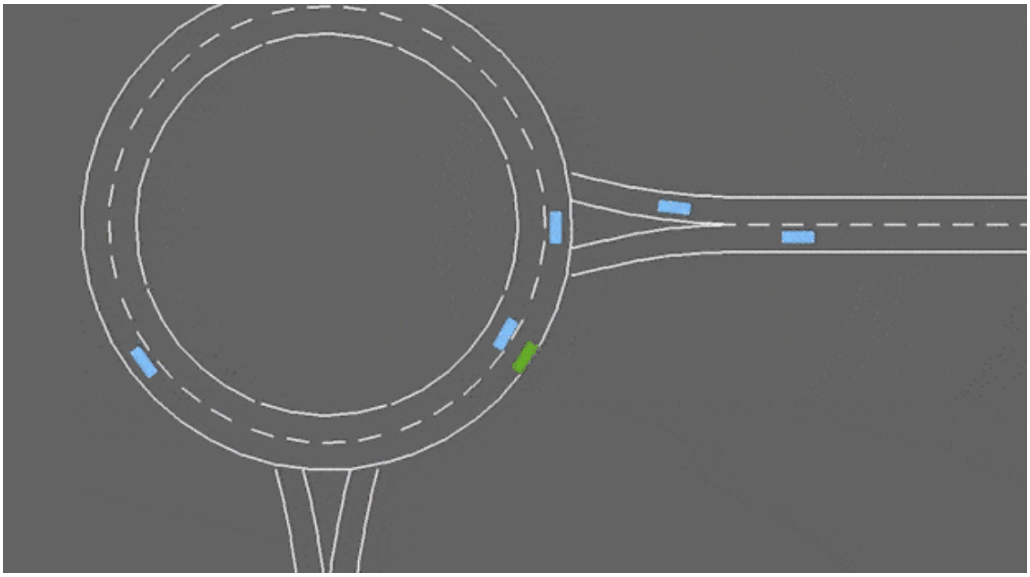

图 5： roundabout-v0 环境——取自 [61]的快照

各方向的导航过程具有随机性，与先前的用例相同。该场景旨在检验模型在更接近真实驾驶环境下的鲁棒性；在这种环境中，自车应规划何时进入环岛，以及何时停车，以避免引发事故 [61]。

4.3 实验流程

对于基线模型，DreamerV3 和基于 VQ-VAE 的模型都应训练 10 万步，并根据样本进行训练期间的滚动展开；而 HanoiWorld 尽管模型结构更大，其 RSSM 规划模块在推理时仍然有效，因此应使用带有瓶颈和空间嵌入预测器的预训练 V-JEPA 2 先验，训练 5000 步。除 HanoiWorld 编码器外，所有世界模型均应直接在 High way-Env 仿真环境中训练。

在实验评估阶段，将在 3 种不同的环境中对不同模型进行评估和推理，每个环境包含 100 个长度各异的回合，并返回平均奖励/得分以及这 100 个回合的总体碰撞率，结果

应记录各回合测量结果的均值和标准差。

4.4 评估

实验结果将通过表2中的碰撞率和表3中的平均奖励评分进行展示

表2： 100个评估回合中的碰撞率（均值 ±标准差）。

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

表 3： 100 个评估回合的平均回合奖励（均值 ±标准差）。

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

从碰撞率来看，HanoiWorld 在高速公路和环岛环境中的表现优于相应的基线，平均碰撞率分别为 0.200 和 0.340。这一表现优于 DreamerV3，后者的碰撞表现处于中等水平，而 VQ-VAE 在所评估的方法中表现最差。然而，HanoiWorld 在车道合并场景中暴露出明显弱点，其预期碰撞率最高；而 DreamerV3 在这种情况下展现出更强的导航能力，碰撞率为 0.030。从模型稳定性来看——此处以碰撞率稳定性衡量——HanoiWorld 在不同领域中确实表现出统计稳定性，其标准差最低；但与此同时，这也表明 HanoiWorld 在 merge-v0 场景中的导航性能可能较差，碰撞率接近 1。

从平均回合奖励来看，HanoiWorld 仅在 roundabout-v0 环境中展现出与 DreamerV3 最先进基线方法相当的竞争力，平均奖励分别有所提升，达到 9。818 和 9。423。相比之下，DreamerV3 仍展现出高效的规划能力，规划表现最佳；在 highway-v0 场景中，其规划稳定性显著优于 merge-v0 场景。HanoiWorld 在这两项任务中的表现则处于中游：在 highway-v0 场景中表现一般，而在 merge-v0 仿真中的规划效果较差，甚至不及参数规模更小的 VQ-VAE 基线方法。

4.5 结果讨论

结果表明，HanoiWorld 能够在简单环境（highway-v0）和复杂环境（roundabout-v0）中有效进行规划，且表现具有竞争力；但在难度更高的并道场景中，其泛化能力较差，碰撞频率显著偏高。与此同时，可以观察到，其奖励表现跨环境保持一致，这证明 HanoiWorld 能够将规划策略泛化到不同环境中。不过，在 merge-v0 中，我们认为该模型可能存在奖励黑客行为的迹象：它偏好选择会导致碰撞的动作，以便在各回合中持续获得奖励。

此外，正如我们所观察到的，当模型需要在车道选择任务中（如 roundabout-v0）展现道路导航的灵活性时，目前模型更倾向于安全车道选择。这表明，我们用于决策的基于熵的正则化效果不佳，需要通过消融实验进行仔细研究。

尽管与 DreamerV3 等最先进的基线方法相比，HanoiWorld 展现出了有竞争力且值得关注的性能，但这项研究仍有一些方面可在近期进一步探讨。例如，可以通过文本条件语言编码器（如[63], ）或整个环境图上的占用编码器（如 [36, 71] ），更全面地引入环境的全局上下文，以刻画环境与智能体视角之间的可供性关系，从而实现行为泛化。此外，本研究并未考察想象展开这一工作方向（ [18] ），后续研究可据此考察自车在不同想象温度下的性能，以控制难度，并探索在潜在域中迁移奖励正则化的方法。

5 结论

本研究展示了 HanoiWorld，这是一种基于 JEPA 的世界模型，可用于生成模拟环境，以便在强化学习中训练自动驾驶车辆控制器。结果证明，HanoiWorld 能够在潜在域中制定有效的规划策略，达到最先进基线方法的水平；同时，它也是一种更轻量的竞争方案。相较于基于像素重建的方法，HanoiWorld 在计算效率方面更具优势，因此这一尝试颇具吸引力。此外，我们的模型确实表明，在某些特定场景中，HanoiWorld 的碰撞率较低，这使我们可以推测，该模型能够学习到跨不同环境的安全概念。然而，与 DreamerV3 等最先进基线方法相比，该模型的规划机制表现仍不够理想。

值得注意的是，当前研究并未尝试利用语言、全局图等多模态输入来整合更多全局上下文条件，也未通过潜在探测对不同环境难度下的世界模型进行实际评估，以理解其规划机制。

这表明后续研究可以受到启发，沿着这一潜在方向继续推进，并借鉴 HanoiWorld 的成果。

参考文献

[1] Shlomi Hacohen、Oded Medina 和 Shraga Shoval。自动驾驶：基于 Google Scholar 和 Web of Science 趋势分析的技术差距综述。*IEEE*智能交通系统学报，23:21241–21258，2022年05月。[2] Fan Jia、Weixin Mao、Yingfei Liu、Yucheng Zhao、Yuqing Wen、Chi Zhang、Xiangyu Zhang 和 Tiancai Wang。Adriver-i：面向自动驾驶的通用世界模型，2023年10月。[3] Yumeng Zhang、Shi Gong、Kaixin Xiong、Xiaoqing Ye、Xiaofan Li、Xiao Tan、Fan Wang、Jizhou Huang、Hua Wu 和 Haifeng Wang。Bevworld：通过场景级鸟瞰视图潜变量实现的自动驾驶多模态世界模拟器，2024年06月。[4] Gabriel Dulac-Arnold、Daniel Mankowitz 和 Todd Hester。现实世界强化学习面临的挑战。*arXiv:1904.12901 [cs, stat]*, 04 2019.[5] Ming Jia 等。自动驾驶世界模型综述。*arXiv*预印本，2025年。[6] Rui Chen 等。面向自主智能体的物理世界建模。*arXiv*预印本，2025年。

[7] Matteo Hessel、Joseph Modayil、van Hasselt、Tom Schaul、Georg Ostrovski、Will Dabney、Dan Horgan、Bilal Piot、Mohammad Azar 和 David Silver。《Rainbow：结合深度强化学习中的改进》，2017 年。

[8] Carl-Johan Hoel、Katherine Driggs-Campbell、Krister Wolff、Leo Laine 和 Mykel Kochenderfer。将规划与深度强化学习相结合，用于自动驾驶中的战术决策。*IEEE*智能车辆学报，第 1–1 页，2019年。[9] Ye Han、Lijun Zhang、Dejian Meng、Zhuang Zhang、Xingyu Hu 和 Songyu Weng。面向联网自动驾驶车辆多智能体协同决策的基于价值的并行更新 MCTS 方法，2024年。[10] Tetsuro Morimura、Kazuhiro Ota、Kenshi Abe 和 Peinan Zhang。针对非马尔可夫决策过程的结合蒙特卡罗树学习的策略梯度算法，2024年。[11] Zhiyu Huang、Chen Tang、Chen Lv、Masayoshi Tomizuka 和 Wei Zhan。学习在线信念预测，以实现自动驾驶中的高效 POMDP规划，2024年。[12] Tomov S Momchil、Sang Uk Lee、Hansford Hendrago、Jinwook Huh、Teawon Han、Forbes Howington、da Silva、Gianmarco Bernasconi、Marc Heim、Samuel Findler、Xi- aonan Ji、Alexander Boule、Michael Napoli、Kuo Chen、Jesse Miller、Boaz Floor，以及

Yunqing Hu。Treeirl：利用树搜索和逆强化学习实现安全城市驾驶，2025年。[13] Jan Robine、Tobias Uelwer 和 Stefan Harmeling。用于强化学习的更小型世界模型，2020年。[14] Chenjun Xiao、Yifan Wu、Chen Ma、Dale Schuurmans 和 Martin M¨uller。学习应对基于模型的强化学习中的误差累积，2019 年。[15] Vlad Sobal、Alfredo Canziani、Nicolas Carion、Kyunghyun Cho 和 Yann LeCun。为自动驾驶区分世界模型与自我模型，2022年。[16] Mahmoud Assran、Adrien Bardes、Jean Ponce 和 Yann LeCun。利用联合嵌入预测架构从图像中进行自监督学习。*arXiv* 预印本 *arXiv.2301.08243*，2025年。[17] Yann LeCun。迈向自主机器智能的道路。*arXiv* 预印本 *arXiv:2203.03620*，2022年。[18] David Ha 和 J¨urgen Schmidhuber。世界模型。世界模型，1:e10，2018年。[19] James J. Gibson。视觉感知的生态学方法，1979年。[20]Adrien Bardes、Jean Ponce 和 Yann LeCun。VICReg：面向自监督学习的方差-不变性-协方差正则化。*arXiv:2105.04906 [cs]*, 01 2022.[21] Danijar Hafner、Timothy Lillicrap、Ian Fischer、Ruben Villegas、David Ha、Honglak Lee 和 James Davidson。学习潜在动力学，以实现基于像素的规划。*arXiv:1811.04551 [cs, stat]*, 06 2019.[22] Danijar Hafner 等。通过世界模型掌握多样化领域。*arXiv* 预印本 *arXiv:2301.04104*，2023年。[23] Greg Brockman、Vicki Cheung、Ludwig Pettersson、Jonas Schneider、John Schulman、Jie Tang 和 Wojciech Zaremba。OpenAI Gym。*arXiv* 预印本 *arXiv:1606.01540*，2016年。[24] Yuval Tassa 等。DeepMind 控制套件。*arXiv* 预印本 *arXiv:1801.00690*，2018年。[25]Nicklas Hansen 等。基于模型的强化学习中的泛化。*NeurIPS*，2023年。[26] Lorenzo Baraldi、Zifan Zeng、Chongzhe Zhang、Aradhana Nayak、Hongbo Zhu、Feng Liu、Qunli Zhang、Peng Wang、Shiming Liu、Zheng Hu、Angelo Cangelosi 和 Lorenzo Baraldi。具身人工智能智能体世界模型的安全挑战：综述，2025年。

[27] Mathilde Caron 等. 自监督视觉 Transformer 中的新兴属性。国际计算机视觉大会，2021年。[28] Nguyen Minh 等. 点云掩码自编码器。*ECCV*，2022年。[29] Chen Min 等. Occupancy-mae：面向自动驾驶的自监督学习。*CVPR*，2024年。[30] Haoran Zhu 等. 采用联合嵌入预测架构进行自监督表征学习，用于汽车激光雷达目标检测。*arXiv* 预印本 *arXiv:2501.04969*，2025年。[31] Chuanxia Zheng 和 Andrea Vedaldi. 在线聚类码本。*arXiv* 预印本 *arXiv:2307.15139*，2023年。[32] Yinhao Li 等. Bevformer：从多相机图像中学习鸟瞰视角表示。*ECCV*，2022年。[33] Zhi Li 等. 自动驾驶占用预测。*CVPR*，2024年。[34] Julian Bogdoll 等. 面向安全关键型自动驾驶的基于占用的世界建模。*arXiv* 预印本，2025年。[35] Wei Li 等. 面向自动驾驶的激光雷达中心世界模型。*arXiv* 预印本，2025年。[36] Guosheng Zhao 等. Drivedreamer-2：大语言模型增强的世界模型，用于生成多样化驾驶视频。*arXiv* 预印本 *arXiv:2403.06845*，2024年。[37] Xiaofeng Wang、Guosheng Zhao、Zheng Zhu、Xinze Chen、Guan Huang、Xiaoyi Bao 和 Xingang Wang. Drivedreamer：迈向逼真的自动驾驶世界模型。*arXiv* 预印本 *arXiv:2310.05008*，2023年。[38] Siva Vasudevan 等. 面向自动驾驶的基于变换器的规划。*CVPR*，2024年。[39] Han Long 等. 面向密集交通的交互感知世界模型。*arXiv* 预印本，2025年。[40] Qian Feng 等. 面向自动驾驶的多智能体意图建模。*arXiv* 预印本，2025年。[41] Yifan Wu 等. 扩展自动驾驶世界模型。*arXiv* 预印本，2025年。[42] Yiming Xie 等. 面向自动驾驶的统一世界建模与规划。*arXiv* 预印本，2025年。

[43] Runsheng Xu 等. Opv2v：一个用于车车通信感知的开放基准数据集与融合流程。*ICRA*，2021年。[44] Runsheng Xu 等. V2x-vit：采用视觉Transformer的车联网协同感知。*ECCV*，2022年。[45] Runsheng Xu 等. V2V4Real：一个用于车车协同感知的真实世界大规模数据集。*CVPR*，2023年。[46] Xiaoyu Tu 等. 面向自动驾驶的时延感知协同感知。*arXiv* 预印本，2025年。[47] Xiaoyu Tu 等. 通信约束下的鲁棒协同世界模型。*arXiv* 预印本，2025年。[48] Lukas Audinys 等. 自主系统世界模型基准测试。*arXiv* 预印本，2025年。[49] Khaled Bouzaiene. 自动驾驶研究仿真平台。*arXiv* 预印本，2025年。[50] Shaoshuai Wang 等. DeepAccident：面向车联网自动驾驶的运动与事故预测基准。*NeurIPS*，2023年。[51] Yifan Guan 等. 驾驶世界模型的大规模训练。*arXiv* 预印本，2025年。[52] Zhangcun Yan 等. OnSiteVRU：面向弱势道路使用者的高分辨率轨迹数据集。*arXiv* 预印本 *arXiv:2503.23365*，2025年。[53] Yiheng Fu 等. 驾驶扩散模型。*arXiv* 预印本 *arXiv:2405.02345*，2024年。[54] Yifan Chu 等. 世界模型综述。*arXiv* 预印本，2025年。[55] Rui Gao 等. CarDreamer：学习自动驾驶世界模型。*arXiv* 预印本 *arXiv:2406.10101*，2024年。[56] Holger Caesar 等. nuscenes：一个自动驾驶多模态数据集。*CVPR*，2021年。[57] Pei Sun 等. 自动驾驶感知的可扩展性：Waymo开放数据集。*CVPR*，2020年。[58] Alexandre Dawid 和 Yann LeCun. 基于能量的世界模型。*arXiv* 预印本，2024年。[59] Jason Kong、Mark Pfeiffer、Georg Schildbach 和 Francesco Borrelli. 用于自动驾驶控制设计的车辆运动学与动力学模型，2015年6月。

[60] Hao Sun、Ziping Xu、Meng Fang 和 Bolei Zhou。监督式 Q 学习可以成为连续控制的强大基线，2022年。

[61] Farama Foundation。Gymnasium 文档，2025年。

[62] Mido Assran、Adrien Bardes、David Fan、Quentin Garrido、Russell Howes、Mojtaba Komeili、Matthew Muckley、Ammar Rizvi、Claire Roberts、Koustuv Sinha、Artem Zholus、Sergio Arnaud、Abha Gejji、Ada Martin、Francois Robert Hogan、Daniel Dugas、Piotr Bojanowski、Vasil Khalidov、Patrick Labatut、Francisco Massa、Marc Szafraniec、Kapil Krishnakumar、Yong Li、Xiaodong Ma、Sarath Chandar、Franziska Meier、Yann LeCun、Michael Rabbat 和 Nicolas Ballas。V-JEPA 2：自监督视频模型助力理解、预测与规划，2025年。

[63] Anthony Hu、Lloyd Russell、Hudson Yeo、Zak Murez、George Fedoseev、Alex Kendall、Jamie Shotton 和 Gianluca Corrado。Gaia-1：用于自动驾驶的生成式世界模型，2023年9月。[64] Mathilde Caron、Hugo Touvron、Ishan Misra、Hervé Jégou、Julien Mairal、Piotr Bojanowski 和 Armand Joulin。自监督视觉 Transformer 中涌现的特性。*arXiv:2104.14294 [cs]*, 05 2021。[65] Shentong Mo 和 Shengbang Tong。将联合嵌入预测架构与对比式自监督学习相结合，2024年10月。[66] Haoran Zhu、Zhenyuan Dong、Kristi Topollai、Beiyao Sha 和 Anna Choromanska。利用联合嵌入预测架构进行汽车激光雷达目标检测的自监督表征学习，2025年。[67] Richard Sutton。Dyna：一种用于学习、规划和反应的集成架构。*SIGART* 新闻通讯，2：160–163，1991年7月。[68] Ronald J. Williams。用于联结主义强化学习的简单统计梯度跟随算法。*Machine Learning*，8：229–256，1992年5月。[69] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。近端策略优化算法，2017年8月。[70] Dechen Gao、Shuangyu Cai、Hanchu Zhou、Hang Wang、Iman Soltani 和 Junshan Zhang。CarDreamer：用于基于世界模型的自动驾驶的开源学习平台，2024年。[71] Arun Balajee Vasudevan、Neehar Peri、Jeff Schneider 和 Deva Ramanan。利用自适应世界模型进行自动驾驶规划，2024年。

A 补充结果

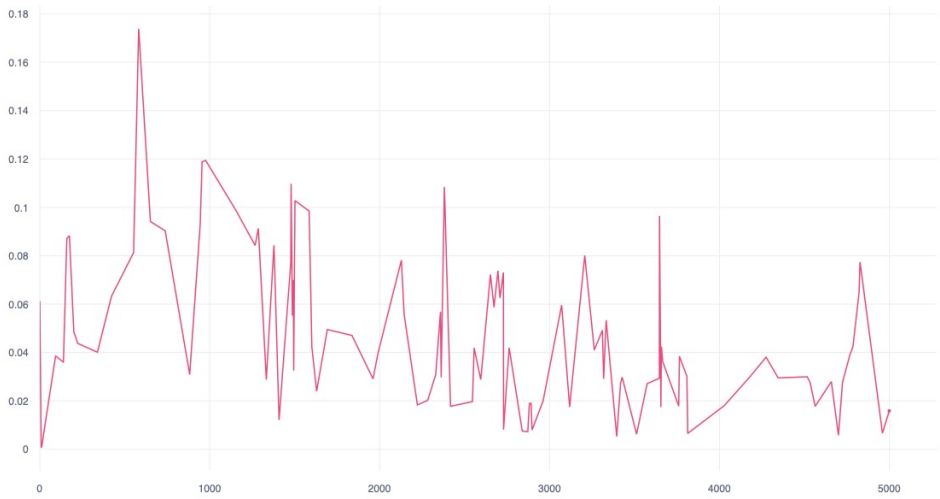

图6：合并环境：RSSM 的延续损失随训练步数的变化。

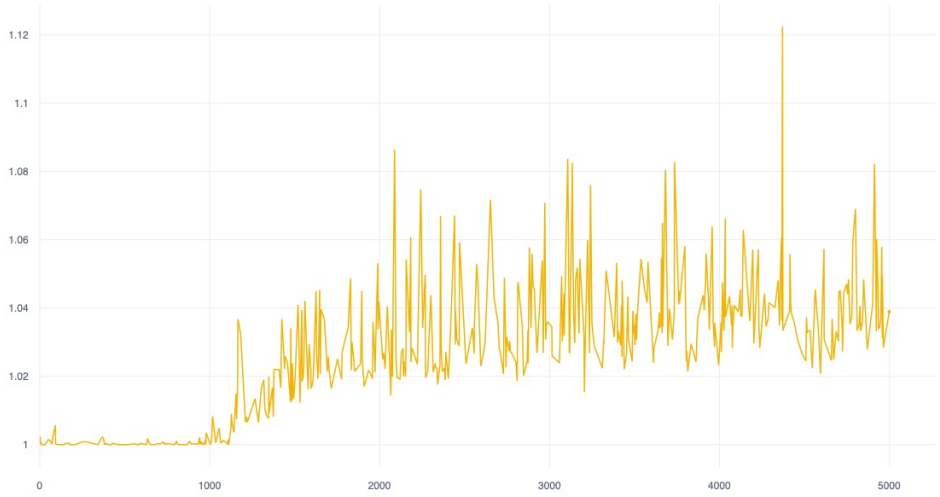

图 7：合并环境：RSSM 的动力学损失随训练步数的变化。

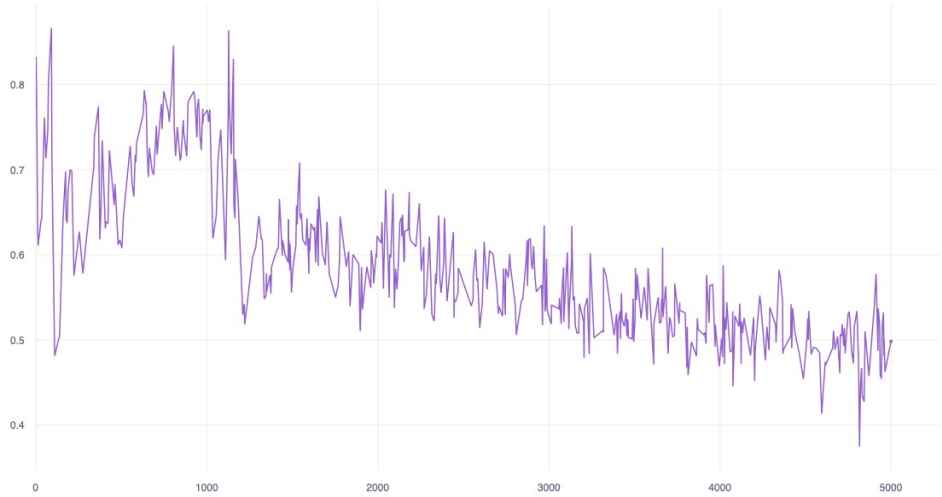

Figure 8: 并道环境：RSSM 的 KL 正则化损失随训练步数的变化 .

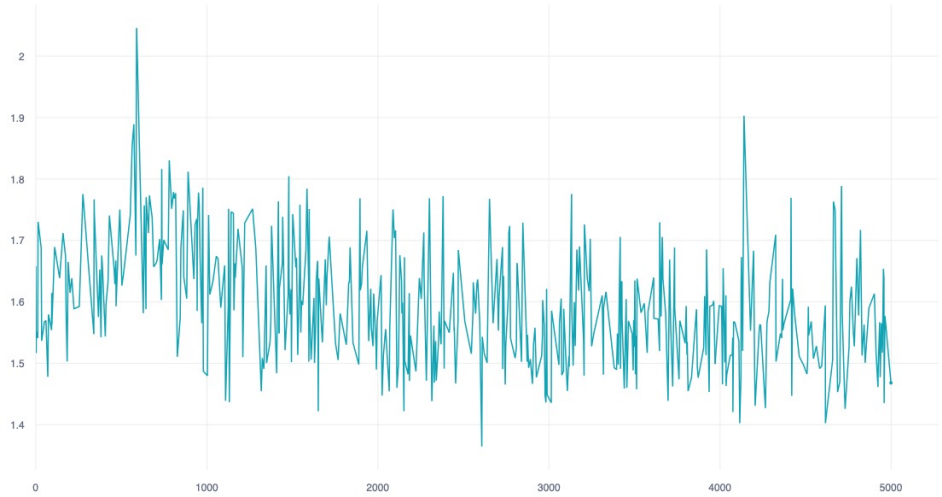

Figure 9: 并道环境：HanoiWorld 的 RSSM 模型损失随训练步的变化 s.

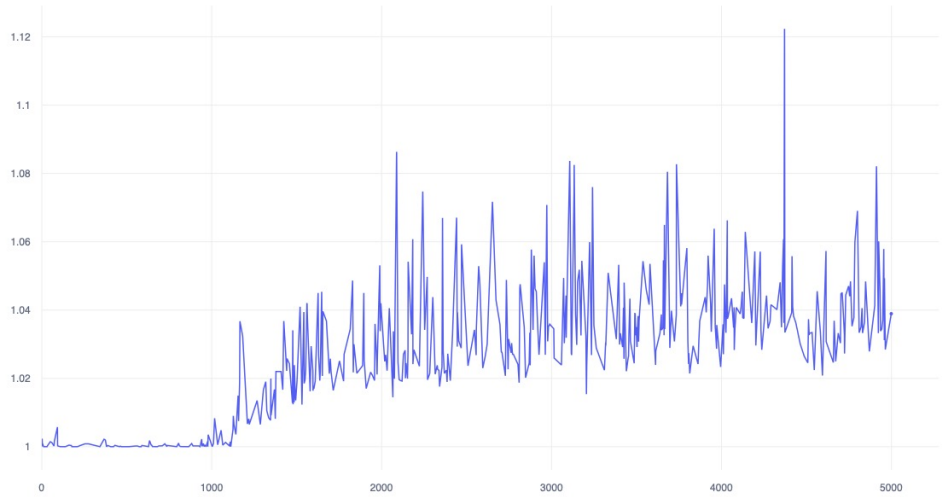

Figure 10: 并道环境：RSSM 的表征损失随训练步数的变化 .

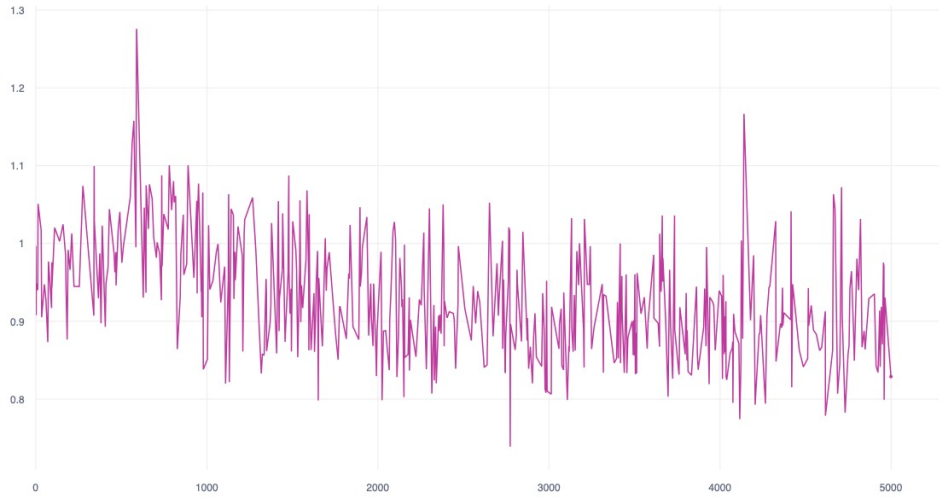

图 11： 合并环境：RSSM 的奖励预测损失随训练步数的变化。

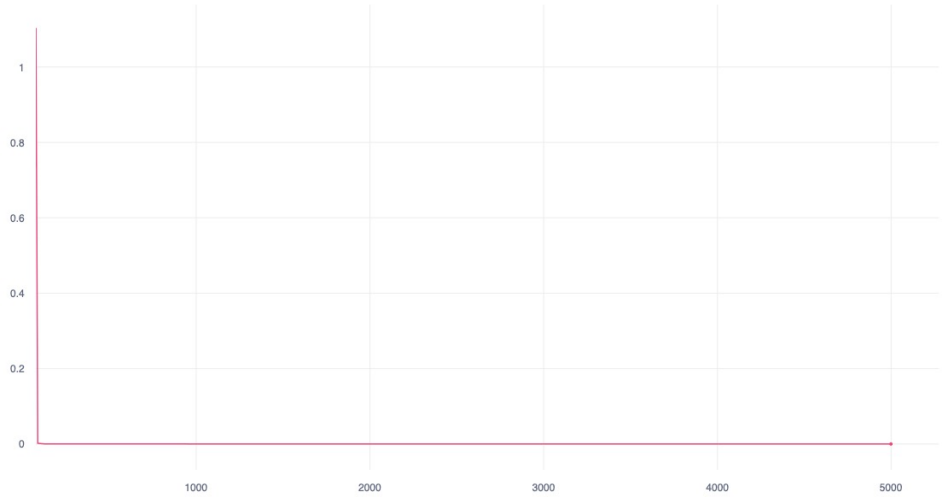

图 12： 高速公路环境：RSSM 的延续损失随训练步数的变化。

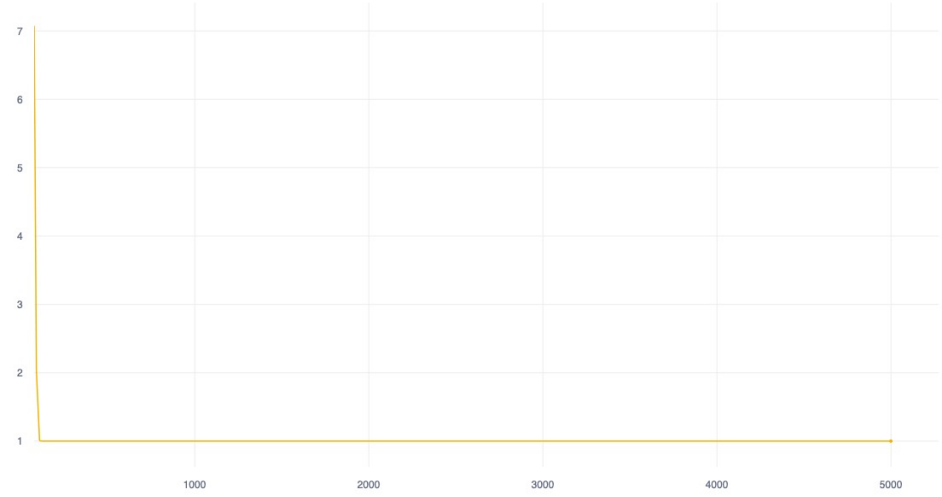

图 1　　3： 高速公路环境：RSSM 的动力学损失随训练步的变化　　s.

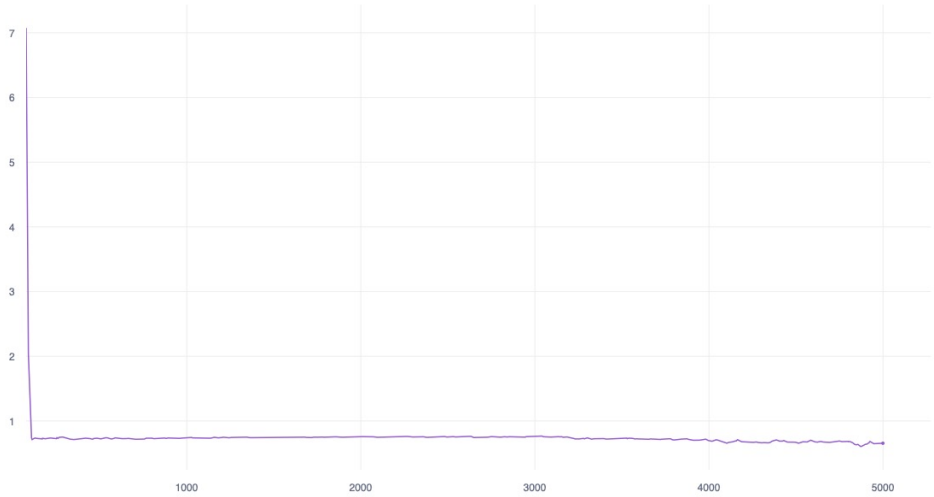

图 **14:** 高速公路环境：RSSM 在训练步数中的 KL 正则化损失 .

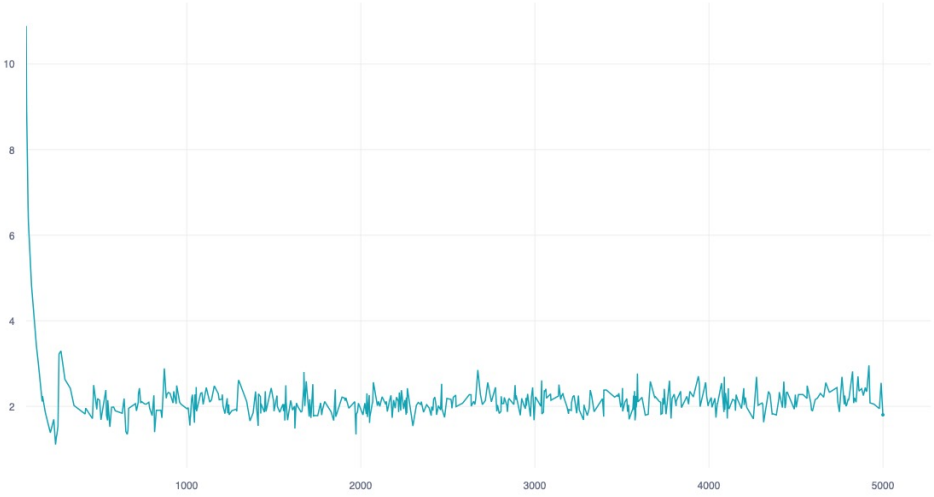

Figure 15: 高速公路环境：RSSM 模型总体损失随训练步数的变化。

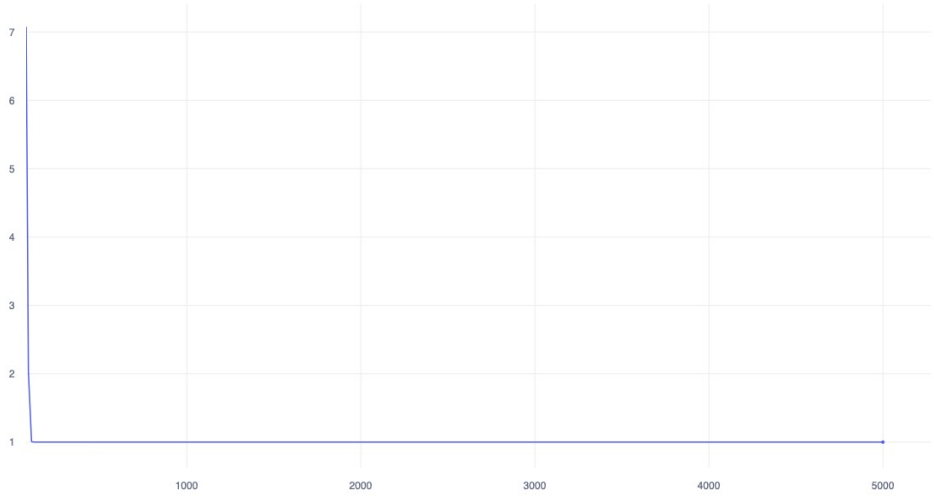

Figure 16: 高速公路环境：RSSM 在训练步中的表征损失 s.

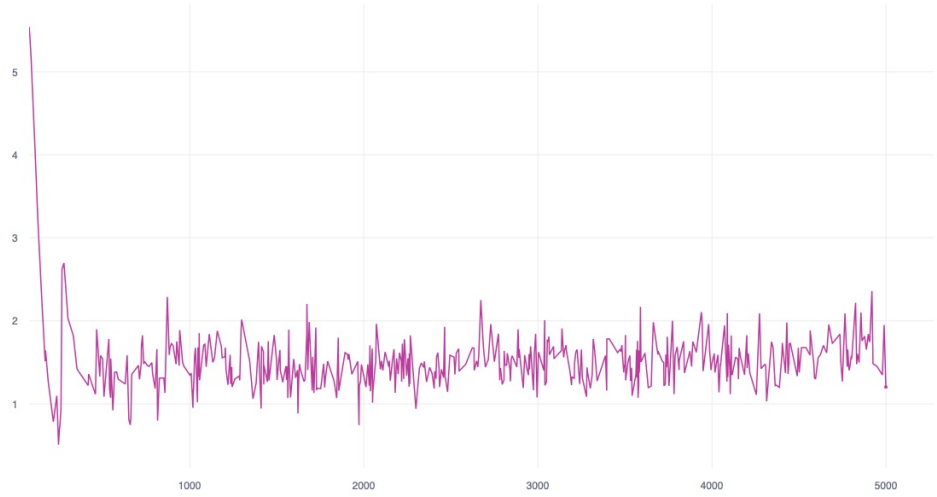

图 17：高速公路环境：RSSM 的奖励预测损失随训练步数的变化。

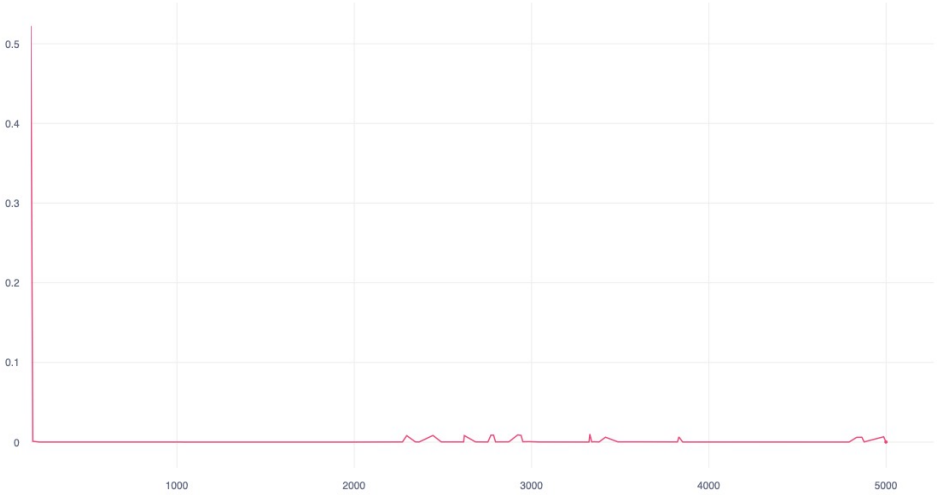

图 18：环岛环境：RSSM 的延续损失随训练步数的变化。

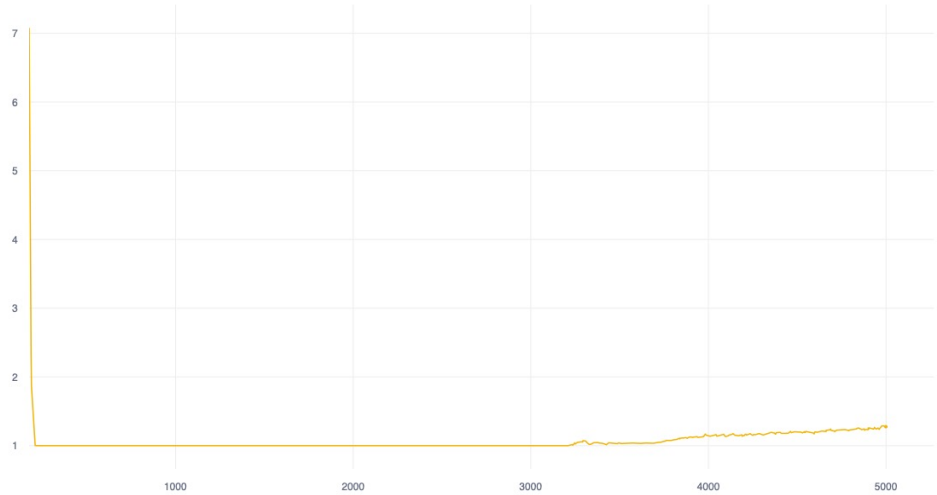

图 19：环岛环境：RSSM 的动力学损失随训练步数的变化。

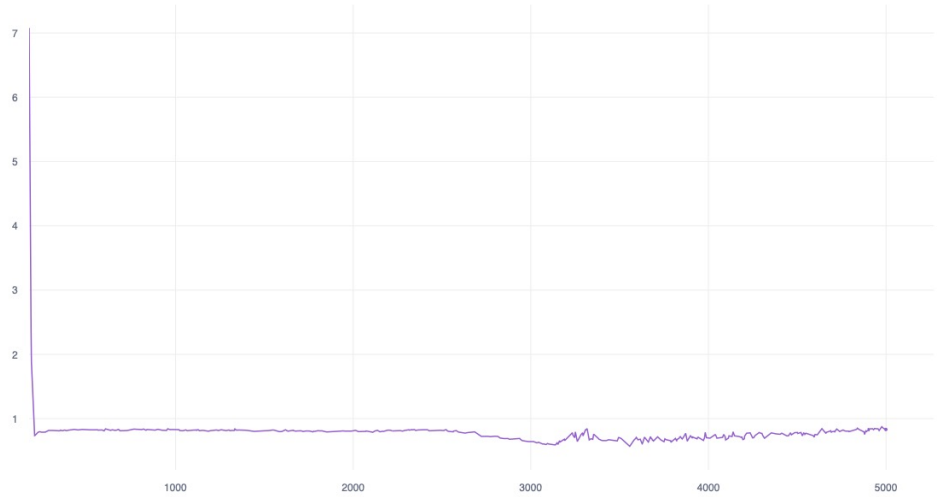

图 2　　0: 环岛环境：RSSM 的 KL 正则化损失随训练步数的变化　　s.

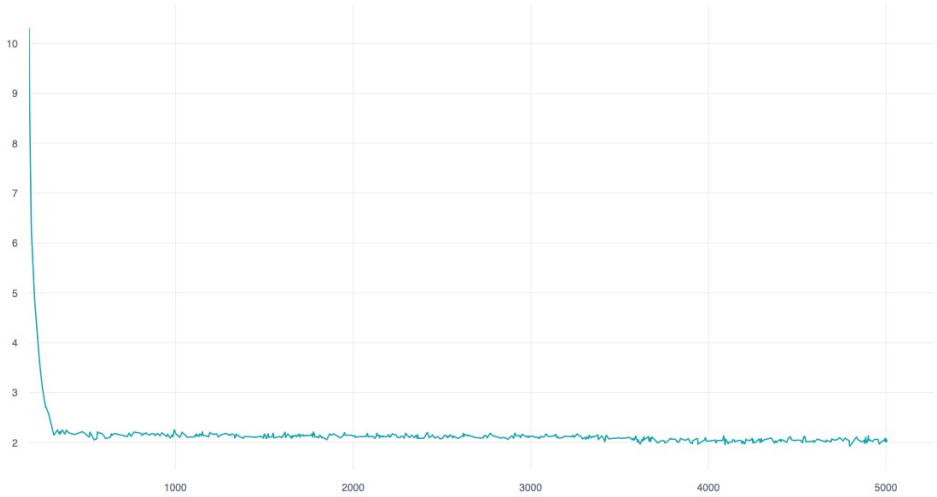

图 21: 环岛环境：RSSM 模型总体损失随训练步数的变化　　s.

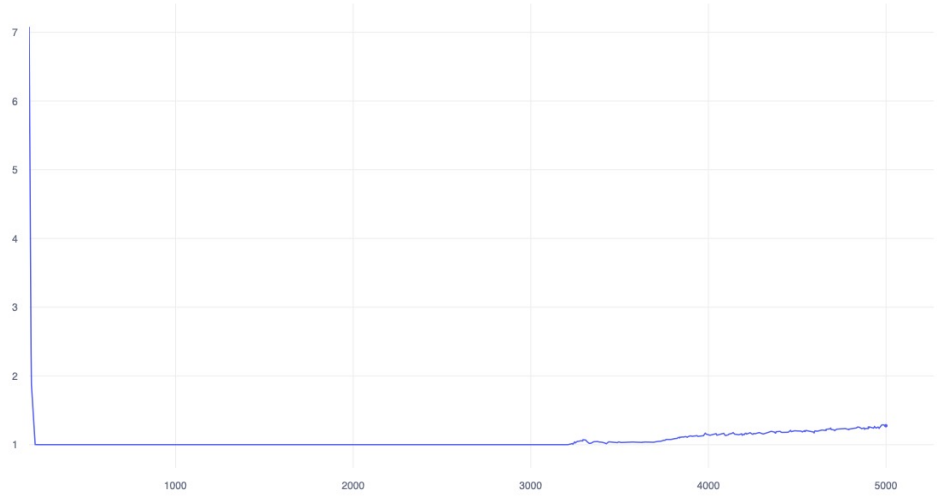

图 2　　2: 环岛环境：RSSM 的表征损失随训练步数的变化　　.

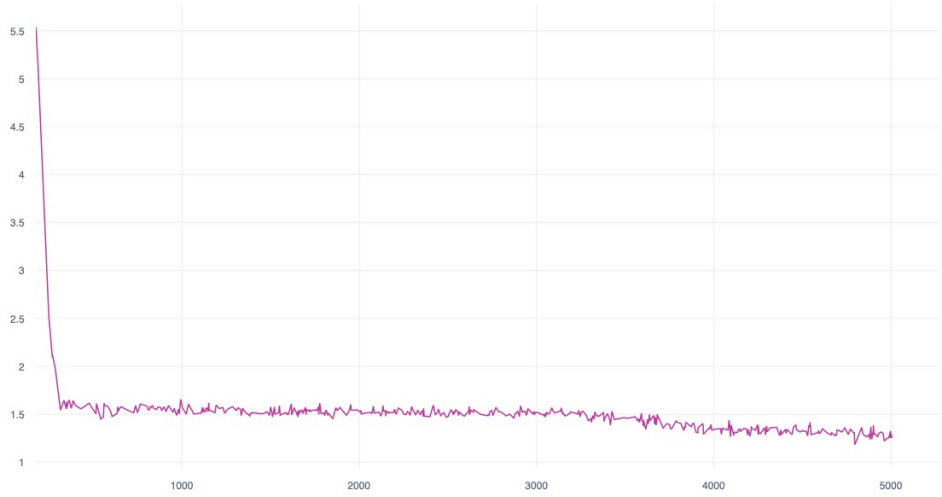

图 23： 环岛环境：RSSM 的奖励预测损失随训练步数的变化。

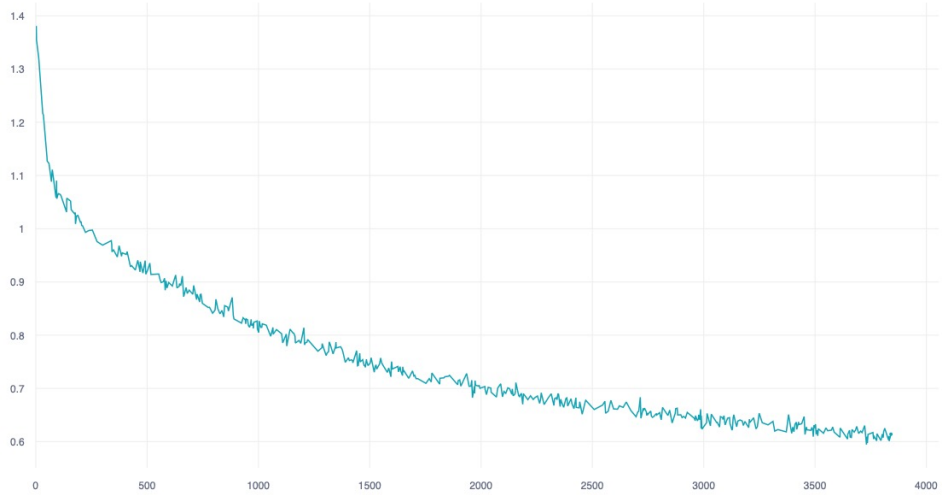

图 24： V-JEPA 2 编码器中学生空间预测器与教师瓶颈表征之间的 L1 对齐损失。

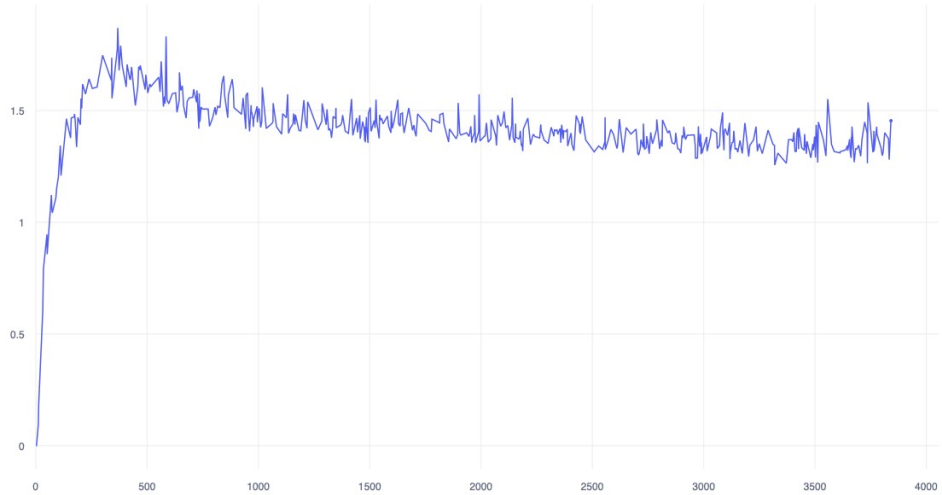

图 25： V-JEPA 2 编码器中学生空间预测器与教师瓶颈表征之间的协方差正则化损失。

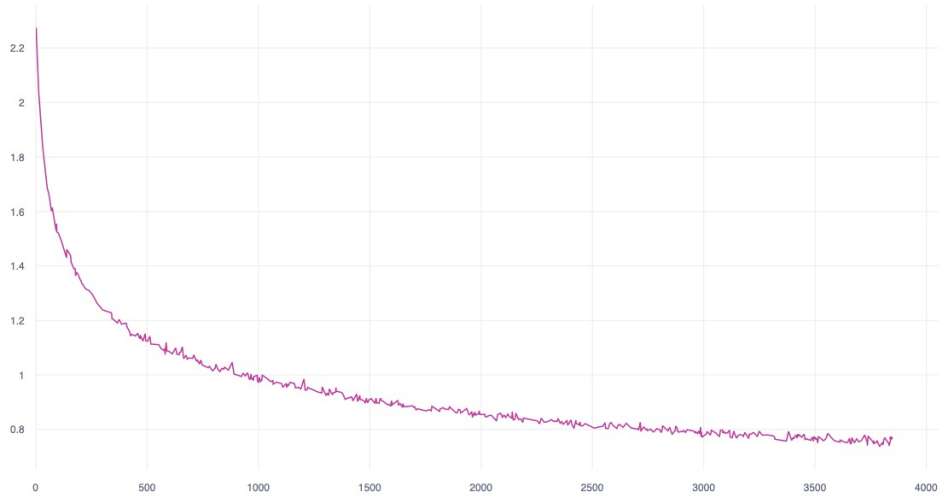

图 26： V-JEPA2 编码器中，学生空间预测器与教师瓶颈表征之间嵌入预测的编码器训练总损失。

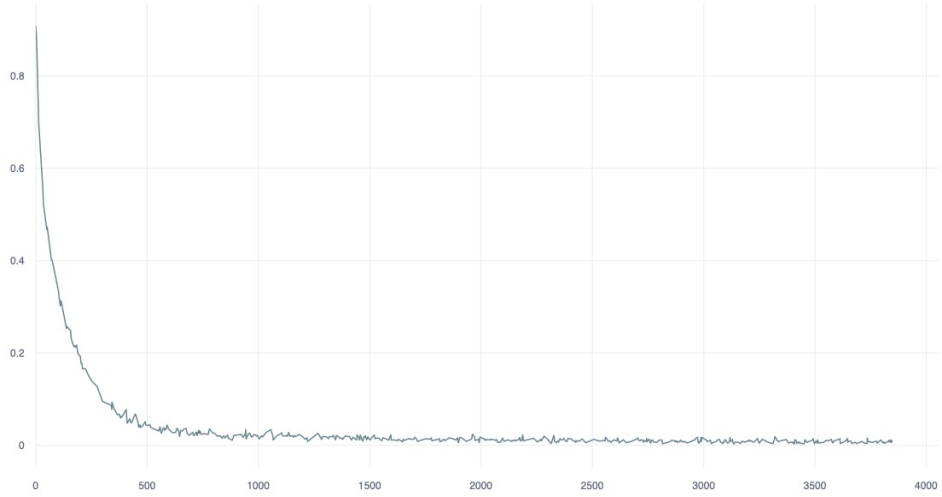

图 27： V-JEPA2 编码器中，学生空间预测器与教师瓶颈表征之间预测嵌入所应用的方差正则化损失。