Four Principles for Physically Interpretable World Models

Jordan Peper * JPEPER@UFL.EDU
Zhenjiang Mao * Z.MAO@UFL.EDU
Yuang Geng YUANG.GENG@UFL.EDU
Siyuan Pan PANSIYUAN@UFL.EDU
Ivan Ruchkin IRUCHKIN@ECE.UFL.EDU
Trustworthy Engineered Autonomy (TEA) Lab, University of Florida, Gainesville, FL, USA

Editors: G. Pappas, P. Ravikumar, S. A. Seshia

Abstract

As autonomous systems are increasingly deployed in open and uncertain settings, there is a growing need for trustworthy neuro-symbolic world models that can reliably predict future high-dimensional observations. The learned latent representations in world models lack direct mapping to meaningful physical quantities and dynamics, limiting their utility and interpretability in downstream planning, control, and safety verification. In this paper, we argue for a fundamental shift from *physically informed* to *physically interpretable* world models — and crystallize *four principles* that leverage symbolic knowledge to achieve these ends: (1) functionally organizing the latent space according to the physical intent, (2) learning aligned invariant and equivariant representations of the physical world, (3) integrating multiple forms and strengths of supervision into a unified training process, and (4) partitioning generative outputs to support scalability and verifiability. We experimentally demonstrate the value of each principle on two benchmarks. This paper opens several intriguing research directions to achieve and capitalize on full physical interpretability in learned world models.

Keywords: world models, representation learning, neuro-symbolic AI, trustworthy autonomy
Source code: https://github.com/Trustworthy-Engineered-Autonomy-Lab/piwm-principles

1. Introduction

Autonomous systems are increasingly deployed in open and uncertain environments (Saidi et al., 2022; Topcu et al., 2020) and use high-dimensional observations to perceive these environments in necessary detail. To achieve high performance, planning and control are often implemented with deep learning methods like reinforcement learning (RL) (Yang et al., 2022; Garg et al., 2019). Since RL training is sample-inefficient, it is impractical to perform in the real world — leading to controllers trained "in the imagination" of *world models* (Ha and Schmidhuber, 2018; Wu et al., 2022).

World models learn to approximate the physical world by predicting future observations based on current observations and actions. Popular neural world models compress observations into the latent space using an autoencoder, propagate these latent values forward in time based on learned temporal dependencies (Deng et al., 2023), and decode them into predicted observations. World models can be improved by injecting symbolic physical knowledge into their structure and training process. For example, Chen et al. (2022) automatically extracted physically meaningful variables from raw observations, yielding more stable long-horizon predictions than standard autoencoders. Brunton et al. (2016) similarly used sparse regression to recover governing equations of nonlinear dynamics from noisy data. Controllers also generalize better when expressed as neuro-symbolic predicates that combine vision-language models with predefined control primitives (Liang et al., 2024).

* First co-authors: equal contribution.

A major challenge of modern world models is their lack of *physical interpretability*. We define it as the degree to which a model's learned latent space corresponds meaningfully to the underlying physics: (a) how well latent embeddings map to physical variables, and (b) how closely latent dynamics emulate physical processes. Without sufficient physical interpretability, a world model offers limited utility in classical model-based autonomy and the design of physically grounded rewards for RL. We also cannot obtain physical guarantees from reachability analysis based on world models (Katz et al., 2022). The core reason for this uninterpretability is that deep learning thrives on distributed representations, in which each feature is partially encoded in multiple latent variables (Hinton, 1986). This challenge is further complicated by partial online observability of the physical state and the difficulty of precisely labeling the data (e.g., indicating which state is riskier in a video).

This paper calls for a paradigm shift from *physically informed* world models to *physically interpretable* ones. The former use symbolic physical knowledge to make learning more effective, efficient, and generalizable. The latter creates neuro-symbolic latent representations with explicit physical meaning, thus subsuming physically informed approaches. Physically meaningful representations bring in a plethora of desirable qualities such as reliability, verifiability, and debuggability.

By carefully analyzing the existing world model literature, this paper advances *four guiding principles* that underlie physical interpretability of learned world representations. Specifically, each principle asserts that **physically interpretable world models should:**

- **Principle 1:** . . . have a *functionally organized* latent space.
- **Principle 2:** . . . learn *aligned* invariant and equivariant representations of the physical world.
- **Principle 3:** . . . integrate *multiple forms and strengths* of supervision into training.
- **Principle 4:** . . . *partition* their generative outputs to support scalability and verifiability.

The next section identifies the interpretability gaps in the existing world models, while Section 3 details the four principles. In Section 4, we perform lightweight validation to demonstrate the value of these principles. Finally, Section 5 discusses the newly opened directions for future research.

2. World Models: State of the Art

Foundations of world models. Modern world models have led to state-of-the-art performance in autonomous planning and control while addressing the data-efficiency concerns of standard RL (Deng et al., 2023; Micheli et al., 2023; Robine et al., 2023). Early world models combined a variational autoencoder (VAE) with a recurrent neural network to predict latent dynamics (Ha and Schmidhuber, 2018). Later work refined both the encoder-decoder architecture and the surrogate dynamics: PlaNet introduced a recurrent state-space model (RSSM) for prediction (Hafner et al., 2019); Dreamer backpropagated gradients through imagined trajectories to improve latent prediction (Hafner et al., 2020); and DreamerV2 extended the RSSM to categorical latent variables (Hafner et al., 2022). More recent research combines autoregressive transformers with self-attention layers to capture detailed temporal dependencies (Robine et al., 2023), or diffusion models to mitigate compounding errors (Ding et al., 2024). World models have been used to optimize planning algorithms for autonomous vehicles in realistic environments: DriveDreamer (Wang et al., 2023b) generates realistic video trajectories from multi-modal inputs for policy optimization; DriveWorld (Min et al., 2024), OccWorld (Zheng et al., 2023), UniWorld (Min et al., 2023), and RenderWorld (Yan et al., 2024) forecast detailed 3D occupancy for motion planning.

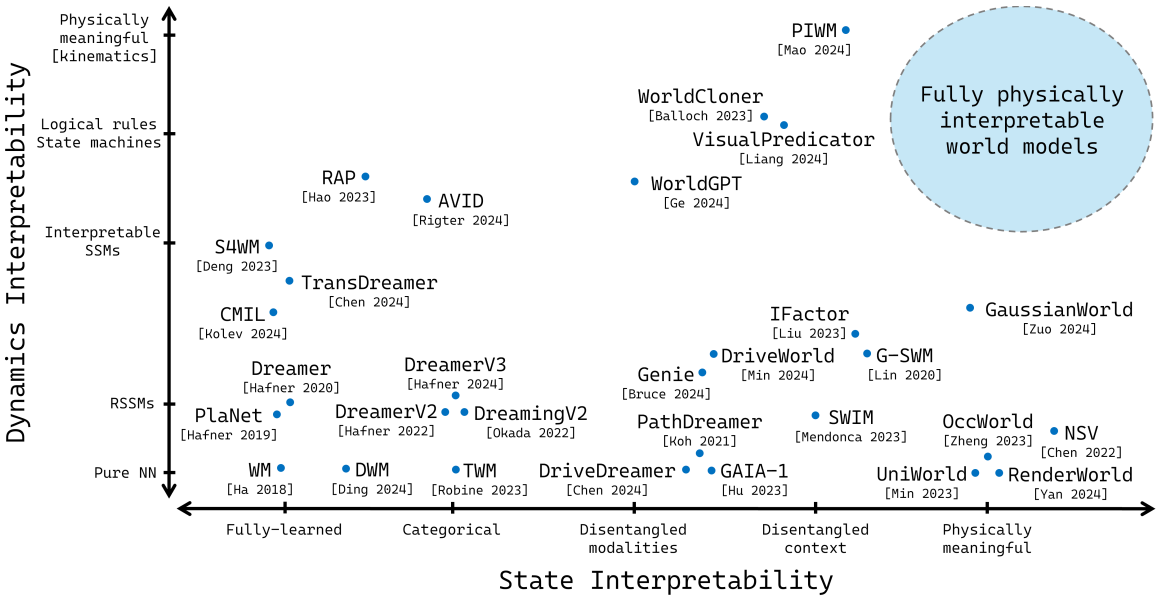

Figure 1: Existing world models by the interpretability of their state and dynamics.

Towards interpretable world models. The interpretability of world models remains a challenge. Relevant efforts toward disentangling latent variables (i.e., reducing their mutual dependency) include β-VAEs (Higgins et al., 2016) and causal VAEs (Yang et al., 2021). This disentanglement strategy is also employed in driving prediction frameworks like GNeVA (Lu et al., 2024a) and ISAP (Itkina and Kochenderfer, 2022). Under the umbrella of world models, G-SWM (Lin et al., 2020) investigated a principled modeling framework that inherits interpretable object and context latent separation from various spatial attention approaches (Kosiorek et al., 2018; Kossen et al., 2020; Jiang et al., 2020; Crawford and Pineau, 2020). Fremont et al. (2019) proposed `SCENIC` — a probabilistic language for generating physically constrained scenes with simulators. Such purely symbolic world models fall outside of our scope. More recent methods impose physical constraints for system identification Sridhar et al. (2023), motion prediction Tumu et al. (2023), and learnable ODE modeling Linial et al. (2021); Zhong and Meidani (2023); Mao et al. (2025). Incorporating partial knowledge of physics with weak supervision has also improved both the state and dynamics interpretability (Mao and Ruchkin, 2024). A recent Nature article leveraged the biological alignment of latent representations to predict microbiome community interactions and antibiotic resistance (Baig et al., 2023). *Neuro-symbolic world models* have also begun to emerge: VisualPredicator (Liang et al., 2024) learns a set of abstract states and high-level actions for strong out-of-distribution generalization, whereas WorldCloner (Balloch et al., 2023) learns symbolic rules to adapt the dynamics to open world novelty. Relevant neuro-symbolic research includes PhysORD (Zhao et al., 2024), which embeds physical laws into neural models, and work by Miao et al. (2025) transforming dashcam footage from a driving environment into a `SCENIC` script through a vision-language model.

Knowledge gap. We observe the lack of world models with full physical interpretability, as per Figure 1 (the underlying literature is listed in Table 2 in the Appendix). Some existing neuro-symbolic architectures scrape the threshold of physically interpretable dynamics, yet lack fluid state representations. On the other hand, multimodal transformer-based architectures preserve the physical context through 3D occupancy but predict with black-box mechanisms. Bridging this gap is key to transitioning from merely *physically informed* world models to fully *physically interpretable* ones.

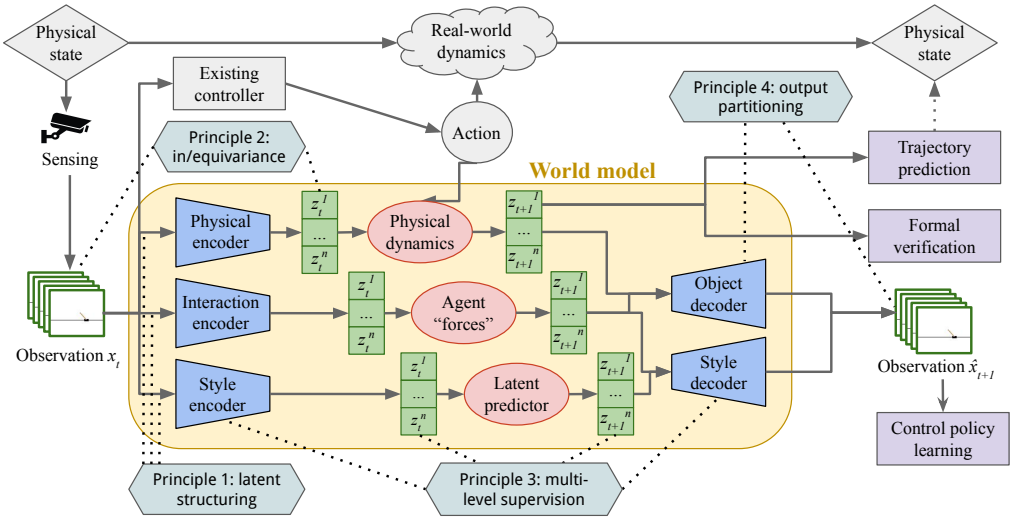

Figure 2: Overview of physically interpretable world models and four principles.

Our recent work highlights the need to address these open questions (Lu et al., 2024b) with predictive world models (Mao et al., 2024b) and their foundation-model variants (Mao et al., 2024a).

Benefits of Physical Interpretability. Aligning world models with fundamental physical principles (e.g., kinematics and conservation laws) has been shown to improve their out-of-distribution generalization and robustness (Mao and Ruchkin, 2024; Lin et al., 2020; Liang et al., 2024; Balloch et al., 2023; Greydanus et al., 2019). These principles prevent latching onto spurious correlations in training and constrain the models to traverse a physically meaningful manifold when extrapolating observations. Going further, physically interpretable representations would lead us to a *qualitatively new level* of safety and trustworthiness. It would make world models more transparent and debuggable by cross-checking them with real-world physics. It would also make generative components suitable for closed-loop verification of physical properties. Finally, physical representation would improve RL sample efficiency by shrinking the search space to physically feasible solutions.

3. Physical Interpretability Principles for World Models

This section puts forward *four guiding principles* for building physically interpretable world models. We begin by formally defining a world model:

Definition 1 (World Model) *A* world model *is a function* $f : X \to X$ *that maps an observation* $x_t \in X \subset \mathbb{R}^n$ *to* $x_{t+1} = f(x_t) = (\mathrm{dec} \circ \mathrm{dyn} \circ \mathrm{enc})(x_t)$*, where* t *is the discrete time index,* $\mathrm{enc} : X \to Z$ *maps the observation to a latent embedding* $z_t \in Z \subset \mathbb{R}^m$*,* dyn *is the latent dynamics propagating* z_t *to* z_{t+1}*, and* $\mathrm{dec} : Z \to X$ *maps embedding* z_{t+1} *to observation* x_{t+1}*.*

A world model learns its latent space by minimizing the gap between predicted and actual observations. However, due to being a black box, its latent space often lacks direct physical interpretability. To address this, we introduce a general concept of a physically interpretable world model:

Definition 2 (Physically Interpretable World Model) *A world model* f *is* physically interpretable *if: (i) there exists a latent-to-physical mapping* $v : \mathbb{R}^m \to \mathbb{R}^k$*, where* k *is the degrees of freedom*

of the physical environment, such that $z \xrightarrow{v} z_{phys}$*, where* $z_{phys} \in \mathbb{R}^k$ *is the minimal set of state variables required to fully describe the environment's true dynamics* $\mathrm{dyn}_{phys} : \mathbb{R}^k \to \mathbb{R}^k$*; and (ii) it holds that:* $v(\mathrm{dyn}(z)) = \mathrm{dyn}_{phys}(v(z))$ *for all* $z \in Z \subset \mathbb{R}^m$.

3.1. Principle 1: Functionally Organizing the Latent Space with Prior Knowledge

We propose *functionally organizing* a world model by modularizing the latent space and processing embeddings through separate branches, as seen in Figure 2. Each latent state is a vector z, which contains n distinct representations of a single observation, each dedicated to unique functionality. Let x represent the world model inputs (e.g., images), and $\mathrm{enc}_i(x) = z_i$ represent the encoder for a particular latent branch $f_i, i = 1..n$, as in Figure 2. Thus, the structured latent space becomes:

$$z = [\mathrm{enc}_1(x) \quad \mathrm{enc}_2(x) \quad \dots \quad \mathrm{enc}_n(x)]$$

Example. An autonomous driving engineer designs a world model with three branches: (1) absolute dynamics of the agent and environment itself, (2) relative dynamics between other agents, and (3) residual yet relevant features of the surroundings. Let L denote a loss function over $f(z)$ and x. The overall training loss should be proportional to the losses in each workflow branch:

$$\mathcal{L} \propto L_1(f_1(\mathrm{enc}_1(x)), x) + L_2(f_2(\mathrm{enc}_2(x)), x) + L_3(f_3(\mathrm{enc}_3(x)), x)$$

Recent work for the first branch aligned latent representations with physical properties (Mao and Ruchkin, 2024). For the second branch, earlier studies demonstrated that physical interactions between agents can be learned without supervision through graph neural networks (GNNs) (Kipf et al., 2018). In the context of world models, Lin et al. (2020) constructed a separate latent representation using a GNN to capture agent occlusions and interactions. Physics-informed neural networks (Raissi et al., 2019; Saemundsson et al., 2020) can also improve the physical interpretability of the world model's dynamics. For instance, Hamiltonian neural networks (Greydanus et al., 2019) learn and adhere to physical conservation laws, leading to impressive generalization. The third branch follows the typical strategy for creating uninterpretable world models and is considered a useful layer to the structured latent space (Lin et al., 2020). Latent space structuring has become increasingly prevalent to improve the performance of planning and control. For instance, a goal-based neural variational agent (GNeVA) uses separate polyline embeddings for the agent and the map, enabling interpretable generative motion prediction (Lu et al., 2024a). Similarly, an interpretable car trajectory prediction framework was proposed, integrating three distinct workflow branches: agent states, high-definition maps, and social context (Itkina and Kochenderfer, 2022).

> Principle 1: Physically interpretable world models should have a *functionally organized* latent space.

3.2. Principle 2: Exploiting Invariances and Equivariances in Input and Latent Spaces

Neural networks' impressive performance is due in part to their ability to learn rich *distributed representations* from training data. Rather than memorizing examples, these models construct hierarchical feature embeddings that capture data patterns and generalize to i.i.d. samples (Hinton, 1986). Nevertheless, training a model to internalize and imagine the world in a human-like manner

far from trivial (Ha and Schmidhuber, 2018). Encoding high-dimensional observations (e.g., images) through commonplace embedding methods (e.g., through autoencoders or encoder-only transformers) leaves the latent representation generally uninterpretable and task-agnostic. This raises concerns about whether spurious correlations distort the latent space or if it effectively encodes the details necessary for discriminating between features that should remain *functionally disentangled*.

Invariance and equivariance relations can help address uninterpretability in representation learning. These terms characterize how representations respond to observation-space transformations. If the representation of x shifts in an expected manner due to a transformation $g(x)$, then the representation model is said to be *equivariant* to that transformation. Likewise, if the representation does not shift under the transformation, then the model is said to be *invariant* to the transformation. For example, bisimulation metrics help learn latent obstacle representations invariant to changes in type, size, and brightness (Zhang et al., 2020). Pol et al. (2020) use contrastive loss to enforce action equivariance. Yet, integrating expert priors remains difficult; one method maps complex observation transformations to simpler latent ones via a symmetric embedding network (Park et al., 2022).

We categorize representations along two dimensions: (1) the nature of their transformation response (invariance versus equivariance) and (2) their degree of human alignment (aligned versus misaligned). A representation that is *aligned-invariant* remains unchanged when an observation undergoes a meaning-preserving transformation, while an *aligned-equivariant* representation transforms predictably when the observation's meaning is altered. In contrast, a representation is *misaligned-invariant* if it does not change under meaningful effects made to the observation (suggesting underfitting), and it is *misaligned-equivariant* if it changes in response to an observation transformation that should not affect the underlying meaning (suggesting a domain shift). Our training objective is to achieve invariance and equivariance alignment by ensuring that the post-transformation representations accurately reflect our human interpretation of the change.

Principle 2: Physically interpretable world models should learn *aligned invariant* and *aligned equivariant* representations of their environment.

Definition 3 (Equivariance) *Let $g_\theta : X \to X$ be an observation space transformation randomly parameterized by $\theta \sim \Theta$, where θ is a random variable drawn from Θ, and let $h_\Phi : Z \to Z$ be a latent space transformation randomly parameterized by $\phi \sim \Phi$, where ϕ is a random variable drawn from Φ. An encoder $\mathrm{enc} : X \to Z$ is an* equivariant *function if $\mathrm{enc}(g_\theta(x)) \overset{d}{=} h_\phi(\mathrm{enc}(x))$. Invariance is a special case of equivariance where $h_\phi(z) = z$.*

Following Definition 3, a simple loss function promotes aligned invariance and equivariance:

$$\mathcal{L}_{wm}(x) \propto \frac{1}{\dim(Z)} \mathbb{E}_{\Theta,\Phi} \left[\| \mathrm{enc}(g_\theta(x)) - h_\phi(\mathrm{enc}(x)) \|_2^2 \right],$$

where $\mathcal{L}_{wm}$ is the overall WM training loss, which is a function of the input observation x.

Example. Consider a vision-based autonomous car that hands over its neural-based controls to a simpler safety controller if a collision is predicted by its world model, consisting of a "physical" and "style" branch per Principle 1. Based on the prior knowledge, an engineer decides that scene brightness should not affect the physical latents; hence, the physics encodings should be *invariant* to changes in observation brightness. However, the style encodings should represent brightness in the resulting latent embedding and, thus, should be equivariant to the changes in brightness.

3.3. Principle 3: Multi-Level and Multi-Strength Supervision for Latent Representations

To bridge rich observations and physical meaning, world models must adapt to supervision signals of varying form and strength (Lee et al., 2013; Chen et al., 2020) — from exact state labels to trajectory-level constraints and weak self-supervision. They must also integrate these signals based on abstraction level (e.g., exact values, intervals, or missing data) to align representations with physical systems. Multi-level supervision tailors the loss functions and training process to the *level of abstraction* (e.g., full trajectories vs. specific state dimensions) and *strength* (e.g., exact labels vs. intervals). For instance, physical state labels allow for the direct alignment of latent representations with real-world quantities using supervised loss. When such labels are unavailable, temporal consistency and smoothness of trajectories can serve as implicit regularization techniques to constrain learned representations. Finally, self-supervision can leverage data-driven structures to discover meaningful latent representations in entirely unsupervised settings.

Principle 3: Physically interpretable world models should integrate multiple *forms* and *strengths* of supervision based on their availability and informativeness.

Supervised Learning: Strong supervision directly aligns specific latent dimensions with known physical states (e.g., positions, velocities), enabling fine-grained interpretability. In many cases, supervision signals are introduced directly into the embeddings to capture key features from labeled data (Zhuang et al., 2015). For example, in low-dimensional systems with position and velocity states $s = [p, v]$, additional latent dimensions ($z_{\text{extra}} \sim \mathcal{N}(0, 1)$) can improve reconstruction quality and stability (Chen et al., 2016; Alemi et al., 2018; Rezende et al., 2014).

Semi-Supervised Learning: When the labels are only available for some data, semi-supervised techniques can refine representations. Pseudo-labeling (e.g., Mean Teacher (Tarvainen and Valpola, 2017) and FixMatch (Sohn et al., 2020)) utilizes both labeled and unlabeled data to iteratively improve the latent space. In Motion2Vec (Tanwani et al., 2020), a small amount of labeled data is first used to initialize the embedding space; subsequently, RNNs predict pseudo-labels for unlabeled data, allowing the model to iteratively refine both the embedding and segmentation components.

Weak Supervision: Noisy or coarse labels, such as position constraints ($p \in [a, b]$), can be utilized via the trajectory smoothness loss: $\mathcal{L}_{\text{smooth}} = \sum_t \|p_t - 2p_{t+1} + p_{t+2}\|^2$. Temporal models like Kalman filters (Kalman, 1960) stabilize noisy trajectories in tasks such as autonomous driving. Interval signals as weak supervision can be directly incorporated into the loss (Mao and Ruchkin, 2024) or combined with contrastive learning to reinforce constraints (Sorokin and Gurevych, 2017).

Self-Supervised Learning: In the absence of labels, contrastive learning (Chen et al., 2020) aligns latent representations with task-specific similarity metrics (e.g., Euclidean distance or structural similarity). Contrastive world models (Poudel et al., 2022) explicitly employ representation learning losses to map similar states closer in the latent space. Plan2Explore (Sekar et al., 2020) generates self-supervised uncertainty-driven objectives to guide the representations.

Combining Supervision Levels: For a given dataset $\mathcal{D}$ with supervision signals (e.g., full trajectories, state variables, or interval constraints), the training objective should integrate matching losses to align the latent space with physical semantics. We advocate for using every available supervision. Given explicit state labels, use direct supervision. When only partial information is available, use weakly supervised constraints to refine the representations.

3.4. Principle 4: Output Space Partitioning for Verifiability

Ensuring the safety of vision-based autonomy is a critical and open challenge (Teeti et al., 2022; Geng et al., 2024; Fremont et al., 2019). The verification of such systems is difficult due to high-dimensional image inputs: classical techniques cannot handle this complexity, motivating new principles for pre-deployment safety guarantees (Althoff, 2015; Păsăreanu et al., 2023). One such attempt is to employ a generative image model to overapproximate observations from a given physical state and feed them into a state estimator or controller (Katz et al., 2022; Cai et al., 2024). Sadly, due to uninterpretable latent states, such "verification modulo generative models" does not provide guarantees for the physical world. Furthermore, decoder verification does not scale to large images.

To reduce decoder complexity, we propose to *partition the generated image* into physically meaningful parts. Specifically, a world model will contain multiple generators of output signals — each dedicated to its own object in the image. Each generator would be separately verifiable, and the results would be combined to provide world model-wide guarantees. This principle reduces each generator's size, making the verification of such world models tractable.

When applied to physically interpretable latent states, this principle can transfer verification guarantees to the physical world: the generators would represent the relationship between images and physical states, not uninterpretable latent ones. Specifications are written at the level of interpretable states (rather than images), after the information has been propagated through perception, control, and dynamics. This process grounds the verification in physical states rather than images.

> Principle 4: Interpretable world models should *partition generated observations* into segments from multiple simpler generators, enabling scalable verification.

Definition 4 (Partitioned World Model Generation) *A world model decoder* dec *translates a latent state* z *into a generated high-dimensional observation* $\hat{x}$*, expressed as* $\mathrm{dec}(z) = \hat{x}$*, by minimizing the reconstruction error between the original and reconstructed observations. Each image segment is produced by a separate decoder:* $\mathrm{dec}_1(z) = \hat{x}_1, \mathrm{dec}_2(z) = \hat{x}_2, ..., \mathrm{dec}_n(z) = \hat{x}_n$*. The combined generated image is represented as* $\hat{x} = \bigoplus_{i=1}^{n} \hat{x}_i$*, where* $\bigoplus$ *is a signal composition operation (e.g., overlaying image segments). The corresponding loss function* $\mathcal{L}_{gen}$ *is:*

$$\mathcal{L}_{gen} = ||x - \hat{x}||^2 + \lambda \sum_{i=1}^{N} ||x_i - \hat{x}_i||^2$$

The question of automatic partitioning of world model outputs can be answered by zero-shot approaches like the Segment Anything Model (SAM) (Kirillov et al., 2023). Recently, SAM was used to segment images to improve image and safety prediction (Mao et al., 2024a). A similar partitioning was used in the action space to scale up the verification of vision-based controllers via multiple low-dimensional approximations (Geng et al., 2024). Principle 4 propagates the physical meaning from different parts of the world model (established in Principle 1) to its generative outputs, effectively linking the high-dimensional observation with a lower-dimensional representation.

This principle has two remaining limitations. First, as the number of objects increases, partitioning becomes increasingly difficult, as in autonomous driving tasks with dynamic objects like cars and pedestrians. Additionally, the gap between a world model and the real world still needs to be formally quantified to obtain guarantees, which is an open problem for future research.

4. Experimental Validation

The objective of our experiments is to evaluate the impact of the four proposed principles on the interpretability of world model representations. We expect each principle to improve the prediction of future physical states compared to a baseline interpretable world model. The success is measured by the mean squared error (MSE) of state predictions over varying prediction horizons.

Two case studies are used to validate the principles: the *Lunar Lander* and *Cart Pole* environments from OpenAI Gym (Brockman et al., 2016). The state dimensions for the Lunar Lander and Cart Pole are 8 and 4, respectively, reflecting different levels of complexity in achieving interpretability. We utilize classical models, namely a Variational Autoencoder (VAE) for encoding observations and a Long Short-Term Memory (LSTM) network for temporal prediction. There are 64 latent dimensions in all experiments. In the baseline interpretable world model, only the first few dimensions are supervised with interpretable physical meanings, whereas the remaining dimensions are not supervised. Additional details are available in the Appendix and online repository.

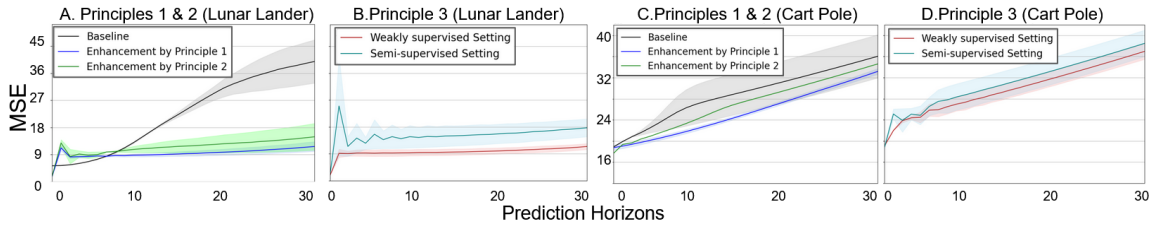

Figure 3: MSE of physical state prediction across different prediction horizons for Principles 1–3.

World model	Environment	Average MSE ↓	Average SSIM ↑	Model Size ↓
Baseline (monolithic)	Cart Pole	0.02856	0.997122	200,259
Partitioned 3-way	Cart Pole	0.05176	0.995614	**144,665**
Baseline (monolithic)	Lunar Lander	0.18801	0.8686	360,773
Partitioned 3-way	Lunar Lander	0.306	0.6289	**78,101**

Table 1: Model size reduction and reconstruction performance for validating Principle 4, $\lambda = 0.2$.

Principle 1: Here we split the encoder into the image part for extracting low-level visual features and the state part that produces values of physical variables. The latent vector size is the same for the baseline and the modified models. Figures 3A and 3C show that Principle 1 significantly reduces the MSE for longer horizons, highlighting the stability that comes from physical interpretability.

Principle 2: We specify a function g that shifts the lunar lander's position, and a corresponding function g that shifts the latent state. For cart pole, we shift both the rotation and position, with corresponding changes made to the latent state. Figures 3A and 3C show that Principle 2 reduces prediction error across all prediction horizons, confirming the value of equivariance. While this principle improves performance on lunar lander, it has less of an effect on the cart pole. We hypothesize that this principle benefits more complex and partially observable systems.

Principle 3: Here we train world models in semi- and weakly-supervised settings: (1) only static information (position, angle) is supervised, while dynamic (velocity) is unknown; (2) velocity is estimated from positions/angles, adding supervision through physical knowledge. Figures 3B and 3D show that weak physical supervision improves prediction quality at all prediction horizons.

Principle 4: We partition the original cartpole and lunar lander images into three parts with SAM, training three smaller decoders for each and combining them as shown in Figure 4 in the Appendix.

The partitioned generator inputs are the exact physical states, while the baseline approach's uninterpretable latents. Our partitioning reduces the baseline's parameters by *27.7%* while keeping a comparable reconstruction quality remains comparable, as per Table 1. Though based on simple environments, these standard benchmarks help isolate each principle's effect. We plan to extend validation to more complex domains like 3D navigation and visual robotic manipulation.

5. Future Research Directions

A. Extracting Physical Knowledge from Foundation Models. It is difficult for humans to externalize their implicit knowledge of the physical world (Trager et al., 2023; Xu et al., 2024). Having absorbed humanity-scale data patterns, large language models are promising sources of implicit and plausible physical knowledge. We intend to investigate how to extract candidate dynamics templates, invariances, and equivariances. An important step is validating the candidate information (e.g., via open datasets) before incorporating it into the world model training.

B. Physically Aligned Multimodality. Reliable multimodal world models are urgently needed in many autonomous systems (Gupta et al., 2024; Zheng et al., 2025). However, the consistency of predicted modalities has been a challenge for learned representations (Lu et al., 2024b). We suggest the use of physically meaningful representations in making image and LiDAR predictions consistent on real-world datasets such as nuPlan (Caesar et al., 2022) and Waymo Open (Sun et al., 2020).

C. Interpretable Uncertainty in World Models. Commonplace uncertainty quantification techniques for deep learning models struggle to express the uncertainty in the terms relevant to the application domain (Gal and Ghahramani, 2016; Kendall and Gal, 2017). Traditional Bayesian approaches and ensemble methods often focus on model uncertainty but fail to capture the structured uncertainty inherent in physical systems (Zhang et al., 2019). In contrast, uncertainty estimation within physically meaningful latent representations allows for more interpretable and actionable uncertainties. We consider it fruitful to develop an uncertainty quantification method based on distributions over physically meaningful latent states and partitioned outputs, which can facilitate robust decision-making and improve reliability in downstream tasks (Depeweg et al., 2018).

D. Unified Training Pipeline. We outlined several training objectives and supervision strategies for world models. However, when their combinations are used, the convergence and stability of training remain elusive (Sener and Koltun, 2018). We recommend developing an automated training pipeline that will combine and tune different losses to ensure reliable training (Li et al., 2018).

E. Integrating World Models into Classical Autonomy. Physically meaningful states enable high-performance components of world models to serve as state estimators, trajectory predictors, and verification models (Mao and Ruchkin, 2024). This allows combining the previously incompatible first-principles and end-to-end learning models. We intend to improve the performance of classic autonomy tasks with world-model components while preserving their reliability and verifiability.

Acknowledgments

The authors thank Vedansh Maheshwari, Mrinall Umasudhan, Rohith Reddy Nama, Sukanth Sundaran, and Liam Cade McGlothlin for their experiments with neural representations. This research is supported in part by the NSF Grant CCF-2403616. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF) or the United States Government.

References

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International conference on machine learning*, pages 159–168. PMLR, 2018.

Matthias Althoff. An introduction to cora 2015. In *Proc. of the workshop on applied verification for continuous and hybrid systems*, pages 120–151, 2015.

Yasa Baig, Helena R. Ma, Helen Xu, and Lingchong You. Autoencoder neural networks enable low dimensional structure analyses of microbial growth dynamics. *Nature Communications*, 14 (1):7937, December 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-43455-0. URL https://www.nature.com/articles/s41467-023-43455-0. Publisher: Nature Publishing Group.

Jonathan C. Balloch, Zhiyu Lin, Xiangyu Peng, Mustafa Hussain, Aarun Srinivas, Robert Wright, Julia M. Kim, and Mark O. Riedl. Neuro-Symbolic World Models for Adapting to Open World Novelty. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 2848–2850, Richland, SC, May 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-9432-1.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation World Models, December 2024. URL http://arxiv.org/abs/2412.03572. arXiv:2412.03572 [cs].

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016. URL https://arxiv.org/abs/1606.01540.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative Interactive Environments, February 2024. URL http://arxiv.org/abs/2402.15391. arXiv:2402.15391 [cs].

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 113(15):3932–3937, April 2016. ISSN 1091-6490. doi: 10.1073/pnas.1517384113.

Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. NuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles, February 2022. URL http://arxiv.org/abs/2106.11810. arXiv:2106.11810 [cs].

Feiyang Cai, Chuchu Fan, and Stanley Bak. Scalable surrogate verification of image-based neural network control systems using composition and unrolling. *arXiv preprint arXiv:2405.18554*, 2024.

Boyuan Chen, Kuang Huang, Sunand Raghupathi, Ishaan Chandratreya, Qiang Du, and Hod Lipson. Automated discovery of fundamental variables hidden in experimental data. *Nature Computational Science*, 2(7):433–442, July 2022. ISSN 2662-8457. doi: 10.1038/s43588-022-00281-6.

URL https://www.nature.com/articles/s43588-022-00281-6. Publisher: Nature Publishing Group.

Chang Chen, Yi-Fu Wu, Jaesik Yoon, and Sungjin Ahn. TransDreamer: Reinforcement Learning with Transformer World Models, November 2024. URL http://arxiv.org/abs/2202.09481. arXiv:2202.09481 [cs].

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

Eric Crawford and Joelle Pineau. Exploiting Spatial Invariance for Scalable Unsupervised Object Tracking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3684–3692, April 2020. ISSN 2374-3468. doi: 10.1609/aaai.v34i04.5777. URL https://ojs.aaai.org/index.php/AAAI/article/view/5777. Number: 04.

Fei Deng, Junyeong Park, and Sungjin Ahn. Facing Off World Model Backbones: RNNs, Transformers, and S4. In *Proc. of NeurIPS 2023*, November 2023. doi: 10.48550/arXiv.2307.02064. arXiv:2307.02064 [cs].

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pages 1184–1193. PMLR, 2018.

Zihan Ding, Amy Zhang, Yuandong Tian, and Qinqing Zheng. Diffusion world model, 2024.

Daniel J. Fremont, Tommaso Dreossi, Shromona Ghosh, Xiangyu Yue, Alberto L. Sangiovanni-Vincentelli, and Sanjit A. Seshia. Scenic: a language for scenario specification and scene generation. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI 2019, pages 63–78, New York, NY, USA, June 2019. Association for Computing Machinery. ISBN 978-1-4503-6712-7. doi: 10.1145/3314221.3314633. URL https://doi.org/10.1145/3314221.3314633.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Arpit Garg, Hao-Tien Lewis Chiang, Satomi Sugaya, Aleksandra Faust, and Lydia Tapia. Comparison of Deep Reinforcement Learning Policies to Formal Methods for Moving Obstacle Avoidance. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3534–3541, November 2019. doi: 10.1109/IROS40897.2019.8967945. URL https://ieeexplore.ieee.org/document/8967945. ISSN: 2153-0866.

Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. WorldGPT: Empowering LLM as Multimodal World Model, September 2024. URL http://arxiv.org/abs/2404.18202. arXiv:2404.18202 [cs].

Yuang Geng, Souradeep Dutta, and Ivan Ruchkin. Bridging dimensions: Confident reachability for high-dimensional controllers, 2024.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/26cd8ecadce0d4efd6cc8a8725cbd1f8-Abstract.html.

Christian Gumbsch, Noor Sajid, Georg Martius, and Martin V. Butz. Learning Hierarchical World Models with Adaptive Temporal Abstractions from Discrete Latent Dynamics. July 2023. URL https://openreview.net/forum?id=5qappsbO73r.

Tarun Gupta, Wenbo Gong, Chao Ma, Nick Pawlowski, Agrin Hilmkil, Meyer Scetbon, Marc Rigter, Ade Famoti, Ashley Juan Llorens, Jianfeng Gao, Stefan Bauer, Danica Kragic, Bernhard Schölkopf, and Cheng Zhang. The Essential Role of Causality in Foundation World Models for Embodied AI, April 2024. URL http://arxiv.org/abs/2402.06665. arXiv:2402.06665.

David Ha and Jürgen Schmidhuber. Recurrent World Models Facilitate Policy Evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning Latent Dynamics for Planning from Pixels. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2555–2565. PMLR, May 2019. URL https://proceedings.mlr.press/v97/hafner19a.html. ISSN: 2640-3498.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning Behaviors by Latent Imagination, March 2020. URL http://arxiv.org/abs/1912.01603. arXiv:1912.01603 [cs].

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models, 2022.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering Diverse Domains through World Models, April 2024. URL http://arxiv.org/abs/2301.04104. arXiv:2301.04104 [cs].

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with Language Model is Planning with World Model, October 2023. URL http://arxiv.org/abs/2305.14992. arXiv:2305.14992 [cs].

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. November 2016. URL https://openreview.net/forum?id=Sy2fzU9gl.

Geoffrey E. Hinton. Learning Distributed Representations of Concepts. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 8(0), 1986. URL https://escholarship.org/uc/item/79w838g1.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A Generative World Model for Autonomous Driving, September 2023. URL http://arxiv.org/abs/2309.17080. arXiv:2309.17080 [cs].

Masha Itkina and Mykel Kochenderfer. Interpretable Self-Aware Neural Networks for Robust Trajectory Prediction. August 2022. URL https://openreview.net/forum?id=fnaMlJbRc4t.

Jindong Jiang, Sepehr Janghorbani, Gerard de Melo, and Sungjin Ahn. SCALOR: Generative World Models with Scalable Object Representations, March 2020. URL http://arxiv.org/abs/1910.02384. arXiv:1910.02384 [cs].

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

Sydney M. Katz, Anthony L. Corso, Christopher A. Strong, and Mykel J. Kochenderfer. Verification of Image-Based Neural Network Controllers Using Generative Models. *Journal of Aerospace Information Systems*, 19(9):574–584, 2022. ISSN 1940-3151. doi: 10.2514/1.I011071. URL https://doi.org/10.2514/1.I011071. Publisher: American Institute of Aeronautics and Astronautics _eprint: https://doi.org/10.2514/1.I011071.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. Active World Model Learning with Progress Curiosity, July 2020. URL http://arxiv.org/abs/2007.07853. arXiv:2007.07853 [cs].

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural Relational Inference for Interacting Systems, June 2018. URL http://arxiv.org/abs/1802.04687.

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

Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A World Model for Indoor Navigation, August 2021. URL http://arxiv.org/abs/2105.08756. arXiv:2105.08756 [cs].

Victor Kolev, Rafael Rafailov, Kyle Hatch, Jiajun Wu, and Chelsea Finn. Efficient Imitation Learning with Conservative World Models, August 2024. URL http://arxiv.org/abs/2405.13193. arXiv:2405.13193 [cs].

Adam R. Kosiorek, Hyunjik Kim, Ingmar Posner, and Yee Whye Teh. Sequential Attend, Infer, Repeat: Generative Modelling of Moving Objects, November 2018. URL http://arxiv.org/abs/1806.01794. arXiv:1806.01794 [cs].

Jannik Kossen, Karl Stelzner, Marcel Hussing, Claas Voelcker, and Kristian Kersting. Structured Object-Aware Physics Prediction for Video Modeling and Planning, February 2020. URL http://arxiv.org/abs/1910.02425. arXiv:1910.02425 [cs].

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

Qifeng Li, Xiaosong Jia, Shaobo Wang, and Junchi Yan. Think2Drive: Efficient Reinforcement Learning by Thinking in Latent World Model for Quasi-Realistic Autonomous Driving (in CARLA-v2), July 2024. URL http://arxiv.org/abs/2402.16720. arXiv:2402.16720 [cs].

Yichao Liang, Nishanth Kumar, Hao Tang, Adrian Weller, Joshua B. Tenenbaum, Tom Silver, João F. Henriques, and Kevin Ellis. VisualPredicator: Learning Abstract World Models with Neuro-Symbolic Predicates for Robot Planning, October 2024. URL http://arxiv.org/abs/2410.23156. arXiv:2410.23156 [cs].

Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, and Sungjin Ahn. Improving Generative Imagination in Object-Centric World Models, October 2020. URL http://arxiv.org/abs/2010.02054. arXiv:2010.02054 [cs].

Ori Linial, Neta Ravid, Danny Eytan, and Uri Shalit. Generative ODE modeling with known unknowns. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, pages 79–94, New York, NY, USA, April 2021. Association for Computing Machinery. ISBN 978-1-4503-8359-2. doi: 10.1145/3450439.3451866. URL https://dl.acm.org/doi/10.1145/3450439.3451866.

Wenliang Liu, Wei Xiao, and Calin Belta. Learning Robust and Correct Controllers from Signal Temporal Logic Specifications Using BarrierNet. 2023. doi: 10.48550/ARXIV.2304.06160. URL https://arxiv.org/abs/2304.06160. Publisher: arXiv Version Number: 1.

Juanwu Lu, Wei Zhan, Masayoshi Tomizuka, and Yeping Hu. Towards Generalizable and Interpretable Motion Prediction: A Deep Variational Bayes Approach. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 4717–4725. PMLR, April 2024a. URL https://proceedings.mlr.press/v238/lu24a.html. ISSN: 2640-3498.

Zhen Lu, Imran Afridi, Hong Jin Kang, Ivan Ruchkin, and Xi Zheng. Surveying neuro-symbolic approaches for reliable artificial intelligence of things. *Journal of Reliable Intelligent Environments*, July 2024b. ISSN 2199-4676. doi: 10.1007/s40860-024-00231-1. URL https://doi.org/10.1007/s40860-024-00231-1.

Haoyu Ma, Jialong Wu, Ningya Feng, Chenjun Xiao, Dong Li, Jianye Hao, Jianmin Wang, and Mingsheng Long. HarmonyDream: Task Harmonization Inside World Models. June 2024. URL https://openreview.net/forum?id=x0yIaw2fgk.

Yanbing Mao, Yuliang Gu, Lui Sha, Huajie Shao, Qixin Wang, and Tarek Abdelzaher. Phy-Taylor: Partially Physics-Knowledge-Enhanced Deep Neural Networks via NN Editing. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):447–461, January 2025. ISSN 2162-2388. doi: 10.1109/TNNLS.2023.3325432. URL https://ieeexplore.ieee.org/document/10297119. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

Zhenjiang Mao and Ivan Ruchkin. Towards Physically Interpretable World Models: Meaningful Weakly Supervised Representations for Visual Trajectory Prediction, December 2024. URL http://arxiv.org/abs/2412.12870. arXiv:2412.12870 [cs].

Zhenjiang Mao, Siqi Dai, Yuang Geng, and Ivan Ruchkin. Zero-shot Safety Prediction for Autonomous Robots with Foundation World Models. In *Back to the Future: Robot Learning Going Probabilistic Workshop (co-located with ICRA 2024)*, March 2024a. doi: 10.48550/arXiv.2404.00462. URL http://arxiv.org/abs/2404.00462. arXiv:2404.00462 [cs] version: 1.

Zhenjiang Mao, Carson Sobolewski, and Ivan Ruchkin. How Safe Am I Given What I See? Calibrated Prediction of Safety Chances for Image-Controlled Autonomy. In *Proc. of the Annual Conference on Learning for Dynamics and Control (L4DC)*, 2024b. doi: 10.48550/arXiv.2308.12252. URL http://arxiv.org/abs/2308.12252. arXiv:2308.12252 [cs].

Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured World Models from Human Videos, August 2023. URL http://arxiv.org/abs/2308.10901. arXiv:2308.10901 [cs].

Yan Miao, Georgios Fainekos, Bardh Hoxha, Hideki Okamoto, Danil Prokhorov, and Sayan Mitra. From Dashcam Videos to Driving Simulations: Stress Testing Automated Vehicles against Rare Events, January 2025. URL http://arxiv.org/abs/2411.16027. arXiv:2411.16027 [cs].

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are Sample-Efficient World Models. 2023. URL https://openreview.net/forum?id=vhFu1Acb0xb.

Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. UniWorld: Autonomous Driving Pre-training via World Models, August 2023. URL http://arxiv.org/abs/2308.07234. arXiv:2308.07234 [cs].

Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, Liping Jing, Yiming Nie, and Bin Dai. DriveWorld: 4D Pre-trained Scene Understanding via World Models for Autonomous Driving, May 2024. URL http://arxiv.org/abs/2405.04390. arXiv:2405.04390 [cs].

Akihiro Nakano, Masahiro Suzuki, and Yutaka Matsuo. Interaction-Based Disentanglement of Entities for Object-Centric World Models. September 2022. URL https://openreview.net/forum?id=JQc2VowqCzz.

Viet Dung Nguyen, Zhizhuo Yang, Christopher L. Buckley, and Alexander Ororbia. R-AIF: Solving Sparse-Reward Robotic Tasks from Pixels with Active Inference and World Models, September 2024. URL http://arxiv.org/abs/2409.14216. arXiv:2409.14216 [cs].

Masashi Okada and Tadahiro Taniguchi. DreamingV2: Reinforcement Learning with Discrete World Models without Reconstruction, March 2022. URL http://arxiv.org/abs/2203.00494. arXiv:2203.00494 [cs].

Jung Yeon Park, Ondrej Biza, Linfeng Zhao, Jan Willem van de Meent, and Robin Walters. Learning Symmetric Embeddings for Equivariant World Models, June 2022. URL http://arxiv.org/abs/2204.11371. arXiv:2204.11371 [cs].

Corina S Păsăreanu, Ravi Mangal, Divya Gopinath, Sinem Getir Yaman, Calum Imrie, Radu Calinescu, and Huafeng Yu. Closed-loop analysis of vision-based autonomous systems: A case study. In *International conference on computer aided verification*, pages 289–303. Springer, 2023.

Elise van der Pol, Thomas Kipf, Frans A. Oliehoek, and Max Welling. Plannable Approximations to MDP Homomorphisms: Equivariance under Actions, February 2020. URL http://arxiv.org/abs/2002.11963. arXiv:2002.11963 [cs].

Alexander Popov, Alperen Degirmenci, David Wehr, Shashank Hegde, Ryan Oldja, Alexey Kamenev, Bertrand Douillard, David Nistér, Urs Muller, Ruchi Bhargava, Stan Birchfield, and Nikolai Smolyanskiy. Mitigating Covariate Shift in Imitation Learning for Autonomous Vehicles Using Latent Space Generative World Models, September 2024. URL http://arxiv.org/abs/2409.16663. arXiv:2409.16663 [cs].

Rudra PK Poudel, Harit Pandya, and Roberto Cipolla. Contrastive unsupervised learning of world model with invariant causal features. *arXiv preprint arXiv:2209.14932*, 2022.

M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, February 2019. ISSN 0021-9991. doi: 10.1016/j.jcp.2018.10.045. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

Marc Rigter, Tarun Gupta, Agrin Hilmkil, and Chao Ma. AVID: Adapting Video Diffusion Models to World Models, November 2024. URL http://arxiv.org/abs/2410.12822. arXiv:2410.12822 [cs].

Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based World Models Are Happy With 100k Interactions. 2023. URL https://openreview.net/forum?id=TdBaDGCpjly.

Steindor Saemundsson, Alexander Terenin, Katja Hofmann, and Marc Deisenroth. Variational Integrator Networks for Physically Structured Embeddings. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 3078–3087. PMLR, June 2020. URL https://proceedings.mlr.press/v108/saemundsson20a.html. ISSN: 2640-3498.

Selma Saidi, Dirk Ziegenbein, Jyotirmoy V. Deshmukh, and Rolf Ernst. Autonomous Systems Design: Charting a New Discipline. *IEEE Design & Test*, 39(1):8–23, February 2022. ISSN 2168-2364. doi: 10.1109/MDAT.2021.3128434. Conference Name: IEEE Design & Test.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International conference on machine learning*, pages 8583–8592. PMLR, 2020.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked World Models for Visual Control, May 2023. URL http://arxiv.org/abs/2206.14244. arXiv:2206.14244 [cs].

Vaisakh Shaj, Saleh Gholam Zadeh, Ozan Demir, Luiz Ricardo Douat, and Gerhard Neumann. Multi Time Scale World Models.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Daniil Sorokin and Iryna Gurevych. End-to-end representation learning for question answering with weak supervision. In *Semantic Web Challenges: 4th SemWebEval Challenge at ESWC 2017, Portoroz, Slovenia, May 28-June 1, 2017, Revised Selected Papers*, pages 70–83. Springer, 2017.

Kaustubh Sridhar, Souradeep Dutta, James Weimer, and Insup Lee. Guaranteed Conformance of Neurosymbolic Models to Natural Constraints. In *L4DC 2023*, April 2023. doi: 10.48550/arXiv.2212.01346. URL http://arxiv.org/abs/2212.01346. arXiv:2212.01346 [cs].

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurélien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2443–2451, June 2020. doi: 10.1109/CVPR42600.2020.00252. URL https://ieeexplore.ieee.org/document/9156973. ISSN: 2575-7075.

Ajay Kumar Tanwani, Pierre Sermanet, Andy Yan, Raghav Anand, Mariano Phielipp, and Ken Goldberg. Motion2vec: Semi-supervised representation learning from surgical videos. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2174–2181. IEEE, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Izzeddin Teeti, Salman Khan, Ajmal Shahbaz, Andrew Bradley, Fabio Cuzzolin, and Lud De Raedt. Vision-based intention and trajectory prediction in autonomous vehicles: A survey. In *IJCAI*, pages 5630–5637, 2022.

Ufuk Topcu, Nadya Bliss, Nancy Cooke, Missy Cummings, Ashley Llorens, Howard Shrobe, and Lenore Zuck. Assured Autonomy: Path Toward Living With Autonomous Systems We Can Trust, October 2020. URL http://arxiv.org/abs/2010.14443. arXiv:2010.14443 [cs].

Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear Spaces of Meanings: Compositional Structures in Vision-Language Models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15349–15358, October 2023. doi: 10.1109/ICCV51070.2023.01412. URL https://ieeexplore.ieee.org/document/10377972/. Conference Name: 2023 IEEE/CVF International Conference on Computer Vision (ICCV) ISBN: 9798350307184 Place: Paris, France Publisher: IEEE.

Renukanandan Tumu, Lars Lindemann, Truong Nghiem, and Rahul Mangharam. Physics Constrained Motion Prediction with Uncertainty Quantification. In *Intelligent Vehicles 2023*. arXiv, May 2023. URL http://arxiv.org/abs/2302.01060. arXiv:2302.01060 [cs].

Tongzhou Wang, Simon S. Du, Antonio Torralba, Phillip Isola, Amy Zhang, and Yuandong Tian. Denoised MDPs: Learning World Models Better Than the World Itself, April 2023a. URL http://arxiv.org/abs/2206.15477. arXiv:2206.15477 [cs].

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards Real-world-driven World Models for Autonomous Driving, November 2023b. URL http://arxiv.org/abs/2309.09777. arXiv:2309.09777 [cs].

Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. OccLLaMA: An Occupancy-Language-Action Generative World Model for Autonomous Driving, September 2024. URL http://arxiv.org/abs/2409.03272. arXiv:2409.03272 [cs].

Lionel Wong, Gabriel Grand, Alexander K. Lew, Noah D. Goodman, Vikash K. Mansinghka, Jacob Andreas, and Joshua B. Tenenbaum. From Word Models to World Models: Translating from Natural Language to the Probabilistic Language of Thought, June 2023. URL http://arxiv.org/abs/2306.12672. arXiv:2306.12672 [cs].

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. DayDreamer: World Models for Physical Robot Learning. August 2022. URL https://openreview.net/forum?id=3RBY8fKjHeu.

Weizhe Xu, Mengyu Liu, Oleg Sokolsky, Insup Lee, and Fanxin Kong. LLM-enabled Cyber-Physical Systems: Survey, Research Opportunities, and Challenges. In *International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*, May 2024.

Ziyang Yan, Wenzhen Dong, Yihua Shao, Yuhang Lu, Liu Haiyang, Jingwen Liu, Haozhe Wang, Zhe Wang, Yan Wang, Fabio Remondino, and Yuexin Ma. RenderWorld: World Model with Self-Supervised 3D Label, September 2024. URL http://arxiv.org/abs/2409.11356. arXiv:2409.11356 [cs].

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. pages 9593–9602, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yang_CausalVAE_Disentangled_Representation_Learning_via_Neural_Structural_Causal_Models_CVPR_2021_paper.html.

Tsung-Yen Yang, Tingnan Zhang, Linda Luu, Sehoon Ha, Jie Tan, and Wenhao Yu. Safe Reinforcement Learning for Legged Locomotion. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2454–2461, October 2022. doi: 10.1109/IROS47612.2022.9982038. URL https://ieeexplore.ieee.org/document/9982038. ISSN: 2153-0866.

Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. Invariant Causal Prediction for Block MDPs. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11214–11224. PMLR, November 2020. URL https://proceedings.mlr.press/v119/zhang20t.html. ISSN: 2640-3498.

Dongkun Zhang, Lu Lu, Ling Guo, and George Em Karniadakis. Quantifying total uncertainty in physics-informed neural networks for solving forward and inverse stochastic problems. *Journal of Computational Physics*, 397:108850, 2019.

Haiming Zhang, Ying Xue, Xu Yan, Jiacheng Zhang, Weichao Qiu, Dongfeng Bai, Bingbing Liu, Shuguang Cui, and Zhen Li. An Efficient Occupancy World Model via Decoupled Dynamic Flow and Image-assisted Training, December 2024. URL http://arxiv.org/abs/2412.13772. arXiv:2412.13772 [cs].

Zhipeng Zhao, Bowen Li, Yi Du, Taimeng Fu, and Chen Wang. PhysORD: A Neuro-Symbolic Approach for Physics-infused Motion Prediction in Off-road Driving, October 2024. URL http://arxiv.org/abs/2404.01596. arXiv:2404.01596 [cs].

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. OccWorld: Learning a 3D Occupancy World Model for Autonomous Driving, November 2023. URL http://arxiv.org/abs/2311.16038. arXiv:2311.16038 [cs].

Xi Zheng, Ziyang Li, Ivan Ruchkin, Ruzica Piskac, and Miroslav Pajic. NeuroStrata: Harnessing Neurosymbolic Paradigms for Improved Design, Testability, and Verifiability of Autonomous CPS, February 2025. URL http://arxiv.org/abs/2502.12267. arXiv:2502.12267 [cs].

Weiheng Zhong and Hadi Meidani. PI-VAE: Physics-Informed Variational Auto-Encoder for stochastic differential equations. *Computer Methods in Applied Mechanics and Engineering*, 403:115664, January 2023. ISSN 00457825. doi: 10.1016/j.cma.2022.115664. URL https://linkinghub.elsevier.com/retrieve/pii/S0045782522006193.

Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. Supervised representation learning: Transfer learning with deep autoencoders. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. GaussianWorld: Gaussian World Model for Streaming 3D Occupancy Prediction, December 2024. URL http://arxiv.org/abs/2412.10373. arXiv:2412.10373 [cs].

Appendix

Experimental Details

All experiments used an NVIDIA GeForce RTX 3090 GPU. The source code can be found at https://github.com/trustworthy-engineered-autonomy-lab/piwm-principles.

Principles 1–3. Our world model employs a VAE for encoding/decoding high-dimensional image observations and an LSTM time-series predictor for modeling state transitions in the latent space. The encoder consists of three convolutional layers with increasing feature maps (16, 32, 64) and ReLU activations, downsampling the input image through strided convolutions. The latent representation is parameterized by two fully connected layers (μ and $\log \sigma^2$), each mapping the encoded feature vector to a latent space of 64 dimensions. The decoder reconstructs the input image using a fully connected layer followed by three transposed convolutional layers, producing a three-channel output with a sigmoid activation. The VAE is trained using the Adam optimizer with an initial learning rate of 0.001, incorporating learning rate decay to stabilize convergence.

The input to the LSTM consists of 64-dimensional latent representations extracted by the VAE. The network comprises two LSTM layers with a hidden size of 64, followed by a fully connected output layer mapping to a 64-dimensional output representing the predicted latent state at the next time step. The LSTM predictor is trained using the Adam optimizer with an initial learning rate of 0.001 and also incorporates learning rate decay. The objective is to minimize the MSE between predicted and true latent representations over time.

Principle 4. Our decoder network maps low-dimensional physical state representations to high-dimensional images using a series of transposed convolutional layers. The baseline decoder has one linear layer, two convolutional layers, and a 4-dimensional encoded feature map. Our partitioned decoder only contains one linear layer and one smaller convolutional layer. Using a fully connected layer, the decoder first maps the input state (four-dimensional vector in cartpole; eight-dimensional vector in lunar lander) to a high-dimensional feature space. This produces an intermediate representation of size 3×16×24×24. The image output is further refined through independent transposed convolutional layers, each producing a separate image (three independent layers for each segment image for cartpole and lunar lander). The model is trained using the Adam optimizer with an initial learning rate of 0.001. Training is conducted with mini-batches of size 64, incorporating validation loss tracking to ensure generalization. The loss function is a λ-weighted combination of the reconstruction MSE of the overall reconstructed image and each segmented part.

For the partitioned loss function in Definition 4, the choice of λ plays a crucial role in image generation behavior: (a) If λ is too small (< 0.1), the model fails to separate the three parts, blending "shadows" of the original image into the outputs; (b) If λ is too big (> 0.5), the three parts are completely disconnected, leading to inferior reconstruction quality. Through hyperparameter tuning, we found that setting $\lambda = 0.2$ provides an optimal balance between the quality of the separation and the reconstruction in both case studies.

Additional Illustrations

- Table 2 lists the literature with the interpretability and adherence to the proposed principles.
- Figure 4 shows example observations and their partitioned reconstructions for Principle 4.
- Figure 5 shows the imperfect part-wise reconstruction for inadequate values of λ.

Short Name	Reference	Principle 1	Principle 2	Principle 3	Principle 4	State interp.	Dyn. interp.
WM	(Ha and Schmidhuber, 2018)						
PlaNet	(Hafner et al., 2019)	Weak					Weak
Dreamer	(Hafner et al., 2020)	Weak					Weak
G-SWM	(Lin et al., 2020)	Strong	Weak		Weak	Moderate	Weak
AWM	(Kim et al., 2020)	Weak				Weak	Weak
Plan2Explore	(Sekar et al., 2020)	Weak				Weak	Weak
Pathdreamer	(Koh et al., 2021)		Weak	Weak	Strong	Weak	
DreamerV2	(Hafner et al., 2022)	Weak				Weak	Weak
NSV	(Chen et al., 2022)		Strong			Strong	
DayDreamer	(Wu et al., 2022)	Weak					Weak
DreamingV2	(Okada and Taniguchi, 2022)	Weak				Weak	Weak
SEN	(Park et al., 2022)		Strong			Moderate	
STEDI	(Nakano et al., 2022)	Strong	Moderate			Strong	
DriveDreamer	(Wang et al., 2023b)	Weak		Strong		Weak	
GAIA-1	(Hu et al., 2023)	Strong				Moderate	
IFactor	(Liu et al., 2023)	Moderate				Moderate	
IRIS	(Micheli et al., 2023)	Weak				Weak	
MTS3	(Shaj et al.)	Weak	Weak			Weak	Weak
Denoised MDP	(Wang et al., 2023a)	Moderate				Moderate	
WM2WM	(Wong et al., 2023)			Strong			Moderate
MWM	(Seo et al., 2023)	Weak					Weak
OccWorld	(Zheng et al., 2023)	Strong	Strong			Strong	
RAP	(Hao et al., 2023)						Moderate
S4WM	(Deng et al., 2023)						Moderate
SWIM	(Mendonca et al., 2023)	Moderate	Moderate			Moderate	
TWM	(Robine et al., 2023)	Weak			Weak		
UniWorld	(Min et al., 2023)		Strong	Strong		Strong	
WorldCloner	(Balloch et al., 2023)		Strong			Moderate	Strong
THICK	(Gumbsch et al., 2023)	Weak	Weak				Weak
AVID	(Rigter et al., 2024)			Strong			Moderate
CMIL	(Kolev et al., 2024)			Strong			Weak
DreamerV3	(Hafner et al., 2024)	Weak	Weak			Weak	Weak
DriveWorld	(Min et al., 2024)		Strong			Moderate	Weak
DWM	(Ding et al., 2024)						
GaussianWorld	(Zuo et al., 2024)	Moderate	Strong			Strong	Moderate
Genie	(Bruce et al., 2024)	Moderate				Moderate	
HarmonyWM	(Ma et al., 2024)	Weak					Weak
OccWM	(Zhang et al., 2024)	Moderate	Strong			Strong	Weak
CovWM	(Popov et al., 2024)	Weak	Weak			Weak	Weak
NWM	(Bar et al., 2024)						
OccLLaMA	(Wei et al., 2024)		Strong			Strong	
PIWM	(Mao and Ruchkin, 2024)		Strong	Strong		Moderate	Strong
R-AIF	(Nguyen et al., 2024)		Weak			Weak	Weak
RenderWorld	(Yan et al., 2024)	Strong	Strong			Strong	
Think2Drive	(Li et al., 2024)	Weak					Weak
TransDreamer	(Chen et al., 2024)	Weak				Weak	Weak
VisualPredicator	(Liang et al., 2024)		Strong			Strong	Moderate
WorldGPT	(Ge et al., 2024)	Moderate			Moderate	Moderate	Moderate
Our future vision		Strong	Strong	Strong	Strong	Strong	Strong

Table 2: Review of notable and state-of-the-art world model architectures for adherence to the *4 principles* and their dynamical/state interpretability.

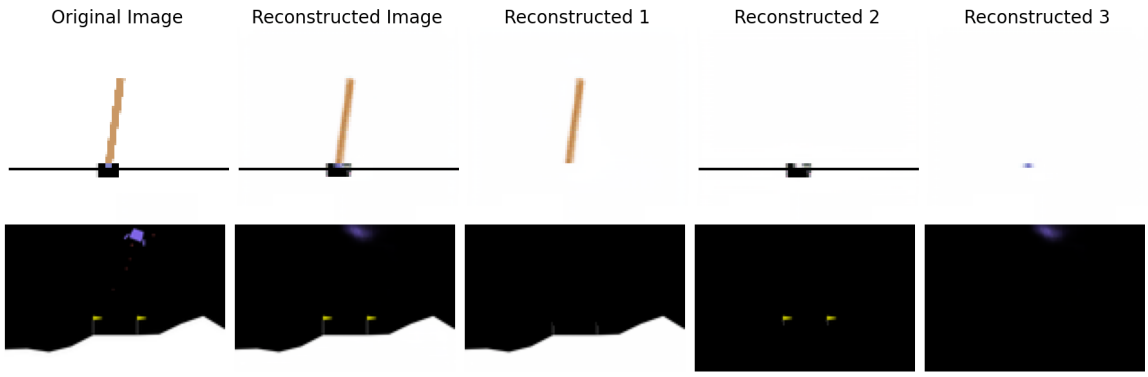

Figure 4: Observations and three reconstructed parts (Principle 4) for the cartpole and lunar lander with $\lambda = 0.2$.

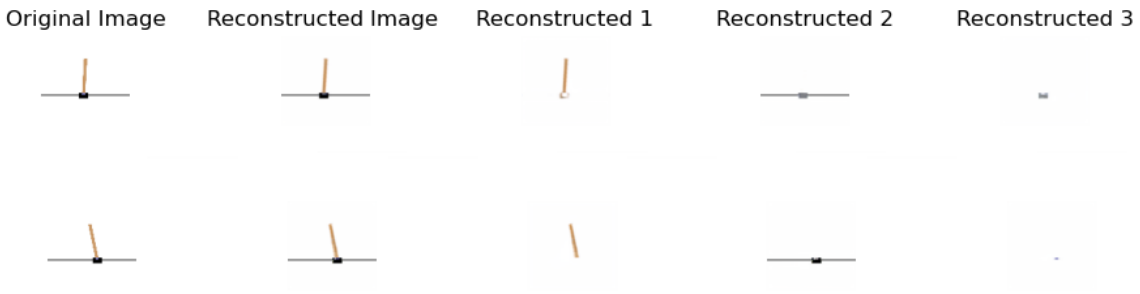

Figure 5: Imperfect reconstruction for the cart pole: the upper row corresponds to $\lambda = 0.01$, while the bottom row corresponds to $\lambda = 0.9$.