Four Principles for Physically Interpretable World Models

Jordan Peper* Zhenjiang Mao* Yuang Geng Siyuan Pan Ivan Ruchkin
Trustworthy Engineered Autonomy (TEA) Lab
University of Florida
Gainesville, Florida, USA
{jpeper, z.mao, yuang.geng, pansiyuan}@ufl.edu, iruchkin@ece.ufl.edu

***Abstract*—As autonomous robots are increasingly deployed in open and uncertain settings, there is a growing need for trustworthy 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) structuring latent spaces according to the physical intent of variables, (2) learning aligned invariant and equivariant representations of the physical world, (3) adapting training to the varied granularity of supervision signals, and (4) partitioning generative outputs to support scalability and verifiability. We experimentally demonstrate the value of each principle on two benchmarks. This paper opens intriguing directions to achieve and capitalize on full physical interpretability in world models.**

***Index Terms*—world models, representation learning, neuro-symbolic AI, trustworthy autonomy**
***Source Repository*—https://github.com/trustworthy-engineered-autonomy-lab/piwm-principles**

I. Introduction

Autonomous robots are increasingly deployed in open and uncertain environments [1], [2] and use high-dimensional observations like high-resolution images and LiDAR scans to perceive these environments. To achieve high performance, their planning and control are often implemented with deep learning methods like reinforcement learning (RL) [3], [4]. Since RL training is sample-inefficient, it is impractical to perform in the real world — leading the controllers to be trained "in the imagination" of *world models* [5], [6].

World models learn to approximate the physical world by predicting future observations based on current observations and actions. Popular neural world models rely on compressing high-dimensional observations into the latent space using an autoencoder. Then these latent values are propagated forward in time based on learned temporal dependencies [7] and decoded into predicted observations. World models can be drastically improved by injecting symbolic physical knowledge into their structure and training process. For example, the approach by [8] automatically discovers physically meaningful variables from raw observations for more stable long-horizon predictions than methods with standard high-dimensional autoencoders, and [9] identifies the governing equations of nonlinear dynamical systems from noisy measurements. In another instance, [10] decompose control into neuro-symbolic predicates powered by vision-language models and predefined control primitives, thereby improving generalization.

A major challenge of modern world models is *their lack of physical interpretability*: usually the latent state cannot be easily mapped to physical quantities (e.g., pose or velocity). This limits their use 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 [11]. 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 [12]. 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, **physically interpretable world models should:**

- **Principle 1:** ... be *structured* according to the physical intent of latent variables.
- **Principle 2:** ... learn *aligned* invariant and equivariant representations of the physical world.
- **Principle 3:** ... adapt their training to the *varied granularity* of supervision signals.
- **Principle 4:** ... *partition* their generative outputs to support scalability and verifiability.

Knowledge gap. Based on our literature survey found in Appendix A, we observe the lack of world model architectures supporting full physical interpretability (as illustrated in Figure 3 and Table II in the appendix). Some existing neuro-symbolic architectures scrape the threshold of physically interpretable dynamics, yet lack fluid state

* First co-authors: equal contribution.

representations. On the other hand, powerful multimodal transformer-based architectures preserve the physical context through 3D occupancy, but their prediction mechanisms are black-box. Bridging this gap is key to transitioning from merely *physically informed* world models to fully *physically interpretable* ones. Our recent work highlights the need to address these open questions [13] and the promise of predictive world models [14] and their foundation-model variants [15].

II. Principles of Physical Interpretability

This section briefly defines each of the four principles. Additional details on how the existing work supports these principles can be found in Appendix B.

A. Principle 1: Structuring the Latent Space

We propose to functionally organize a world model with a modular latent space. Each state in that latent space is a vector z, which contains n distinct representations of a single observation dedicated to unique world model functionalities. Let x represent the world model inputs (e.g., images), and $\mathrm{enc}_i(x) = z_i$ represent the encoder for a particular workflow branch $f_i, i = 1..n$ of the world model, as in Figure 1. Thus, the structured latent space becomes:

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

These functionalities require a human expert to choose task-relevant conceptual priors, which correspond to the physical phenomena being captured by each functionality. For instance, in autonomous driving, the world model's latent space might be organized in three branches: (1) physical dynamics of the vehicle itself and the non-agentic environment, (2) emergent physical/dynamical reasoning arising from interactions with other agents, and (3) residual yet relevant features of the surroundings.

$$\mathcal{L} \propto \sum_i L_i(f_i(\mathrm{enc}_i(x)), x)$$

> Principle 1: Physically interpretable world models should be *structured* according to the physical intent of latent variables.

B. Principle 2: Exploiting Invariances and Equivariances

Deep neural networks have achieved impressive performance due in part to their ability to learn rich *distributed representations* from training data. Rather than simply memorizing examples, these models construct hierarchical feature embeddings that capture patterns in the data to generalize to i.i.d. samples [12]. Nevertheless, training a model to internalize and imagine the world in a human-like manner from scratch is still a nontrivial challenge at best ([16]). 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 the latent space is distorted by spurious correlations or if it effectively encodes the details necessary for discriminating between features that should remain *functionally disentangled*.

Invariance and equivariance relations are one way to address uninterpretability in representation learning models. These terms are often used to classify representations based on their response to observation space transformations. If the representation of x shifts in an expected manner due to a transformation $f(x)$, 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.

We classify representation models 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 II.1 (Equivariant Representation)**.** Let $\mathrm{enc} : X \to Z$ be a map from each observation $x \in X$ to latent representation $z \in Z$. Let $T = \{(f_1, g_1), (f_2, g_2), \dots, (f_n, g_n)\}$ be a finite set of transformation pairs such that $f_i : X \to X$ and $g_i : Z \to Z$. We say that $z = \mathrm{enc}(x)$ is *equivariant* to T if $\forall (f, g) \in T, \mathrm{enc}(f(x)) = g(\mathrm{enc}(x))$. Invariance is a special case of equivariance where $g(z) = z$.

Following Definition II.1, a simple loss function promotes aligned invariance and equivariance:

$$\mathcal{L}_{wm}(x, y) \propto \frac{\lambda}{N} \sum_{(f,g)\in T}^{N} \| \mathrm{enc}(f(x)) - g(\mathrm{enc}(x)) \|_2^2$$

C. Principle 3: Multi-Level Supervision for Representations

World models must bridge the gap between rich observations and physical meaning. This requires adapting the training strategy based on the type and quality of available supervision signals [17], [18]. Instead of treating all supervision signals in the same way, effective world models carefully exploit supervision signals — whether they are precise, coarse, or missing — by selecting the most fitting training method to align with the underlying physical system.

Multi-level supervision tailors the loss functions and training process to the granularity of supervision signals.

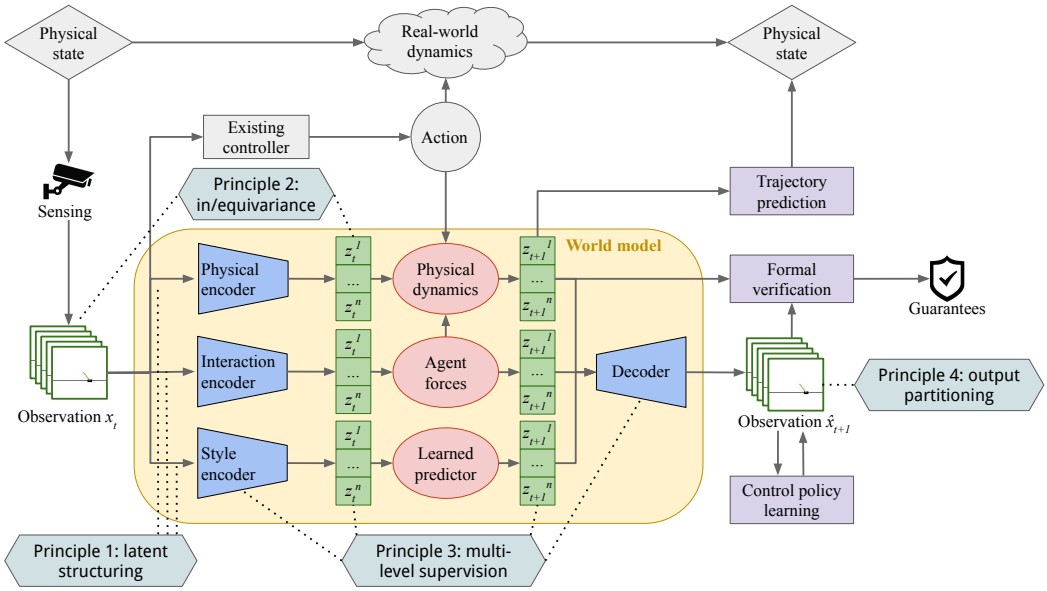

Fig. 1. Overview of physically interpretable world models and four principles.

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 adapt their training to the *granularity* of supervision signals.

Combining Supervision Levels: For a given dataset $\mathcal{D}$ with multiple supervision signals, the training objective must flexibly incorporate relevant terms to ensure that the latent space aligns with the physical world across diverse scenarios. We advocate for using every available supervision opportunity. Given explicit state labels, use direct supervision. When only partial information is available, use weakly supervised constraints to refine the representations. In addition, use self-supervision to extract the sequential physical patterns of observations. Since all these signals come from the physical world, we expect their multi-level integration to improve physical interpretability.

D. Principle 4: Output Space Partitioning for Verifiability

We propose to *partition the generated image* into physically meaningful parts to enable the safety verification of vision-based controllers. Specifically, a world model will contain multiple generators of output signals — each dedicated to its own image region. 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 analysis and execution of world models more parallelizable and scalable. Moreover, when applied to physically interpretable latent states, this principle enables the transfer of verification guarantees to the physical world: the generators will represent the relationship between images and physical states, not uninterpretable latent ones.

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

Definition II.2 (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 $\mathcal{L}_{\text{gen}}$ is:

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

The question of automatic partitioning of world model outputs can be answered by zero-shot approaches like the Segment Anything Model (SAM) [19]. Recently, SAM was used to segment images to improve image and safety prediction [15]. A similar partitioning was used in the action space to scale up the verification of vision-based controllers via multiple low-dimensional approximations [20]. Principle 4 propagates the physical meaning from different parts of the world model (established in Principle 1) to its generative

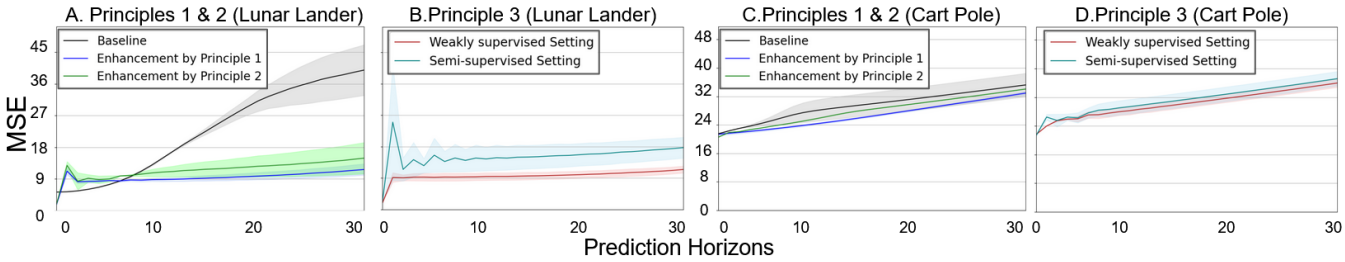

Fig. 2. MSE of physical state prediction across different prediction horizons for Principles 1–3.

outputs, effectively linking the high-dimensional observation space with a lower-dimensional state representation.

III. Experimental Validation

Our experiments 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 measure is the mean squared error (MSE) of state predictions over many prediction horizons.

Two case studies are used for validation: the *Lunar Lander* and *Cart Pole* environments from OpenAI's Gym [21]. We utilize classical models, namely a Variational Autoencoder (VAE) for encoding observations and a Long Short-Term Memory (LSTM) network for temporal prediction. The latent dimensions number is 64 in all experiments. The baseline interpretable world model employs additional linear layers to transform otherwise uninterpretable latents into state values. The experimental details can be found in Appendix C and the online repository.

Principle 1: Here we split the encoder into the image part for extracting visual features and the state part that produces physical variables. The latent vector size is the same for the baseline and the modified models. Figures 2A and 2C show that Principle 1 significantly reduces the MSE for longer horizons, highlighting the stability due to physical grounding.

Principle 2: For the lunar lander, we add a translation to both the observation and position, ensuring the equivariance to translation. For the cart pole, a translation is applied to the cart and a rotation to the pole, with corresponding changes to the latent state. Figures 2A and 2C show that Principle 2 reduces prediction error across all prediction horizons.

Principle 3: Here we have 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 2B and 2D show that weak physical supervision improves prediction at all prediction horizons.

Principle 4: We partition the original cartpole and lunar lander images into three parts with SAM. Then three smaller decoders are trained to generate each image part, which are combined in the end (as illustrated in Figure 4 in the appendix). 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. For both, the latent space has 4 dimensions for the cartpole and 8 for the lunar lander. The partitioned generator inputs are the exact physical states, while the baseline gets uninterpretable latents. Our partitioning reduces *200,259 parameters* in the baseline to only *144,665 parameters*. Despite that, the reconstruction quality remains comparable (see Table I in the appendix)

IV. Future Research Directions

A. Extracting Physical Knowledge from Foundation Models. Having absorbed humanity-scale data patterns, large language models are promising sources of implicit and plausible physical knowledge. We will 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 [22], [23]. However, the consistency of predicted modalities has been a challenge for learned representations [13]. We suggest the use of physically meaningful representations in making image and LiDAR predictions consistent on real-world datasets such as nuPlan [24] and Waymo Open [25].

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 [26], [27]. In contrast, uncertainty estimation within physically meaningful latent representations allows for more interpretable and actionable uncertainties. We suggest developing 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 [28].

D. Unified Training Pipeline. When using combinations of supervisory signals, the convergence and stability of training remain elusive [29]. We recommend designing an automated training pipeline that will combine and tune different losses to ensure reliable training [30].

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 models for verification [31]. We intend to enhance classic autonomy tasks with world-model components to improve their performance while preserving their 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

[1] S. Saidi, D. Ziegenbein, J. V. Deshmukh, and R. Ernst, "Autonomous Systems Design: Charting a New Discipline," *IEEE Design & Test*, vol. 39, no. 1, pp. 8–23, Feb. 2022, conference Name: IEEE Design & Test.

[2] U. Topcu, N. Bliss, N. Cooke, M. Cummings, A. Llorens, H. Shrobe, and L. Zuck, "Assured Autonomy: Path Toward Living With Autonomous Systems We Can Trust," Oct. 2020, arXiv:2010.14443 [cs]. [Online]. Available: http://arxiv.org/abs/2010.14443

[3] T.-Y. Yang, T. Zhang, L. Luu, S. Ha, J. Tan, and W. Yu, "Safe Reinforcement Learning for Legged Locomotion," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2022, pp. 2454–2461, iSSN: 2153-0866. [Online]. Available: https://ieeexplore.ieee.org/document/9982038

[4] A. Garg, H.-T. L. Chiang, S. Sugaya, A. Faust, and L. 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)*, Nov. 2019, pp. 3534–3541, iSSN: 2153-0866. [Online]. Available: https://ieeexplore.ieee.org/document/8967945

[5] D. Ha and J. Schmidhuber, "World Models," Mar. 2018, arXiv:1803.10122 [cs]. [Online]. Available: http://arxiv.org/abs/1803.10122

[6] P. Wu, A. Escontrela, D. Hafner, K. Goldberg, and P. Abbeel, "DayDreamer: World Models for Physical Robot Learning," Jun. 2022, arXiv:2206.14176 [cs]. [Online]. Available: http://arxiv.org/abs/2206.14176

[7] F. Deng, J. Park, and S. Ahn, "Facing Off World Model Backbones: RNNs, Transformers, and S4," Nov. 2023, arXiv:2307.02064 [cs]. [Online]. Available: http://arxiv.org/abs/2307.02064

[8] B. Chen, K. Huang, S. Raghupathi, I. Chandratreya, Q. Du, and H. Lipson, "Automated discovery of fundamental variables hidden in experimental data," *Nature Computational Science*, vol. 2, no. 7, pp. 433–442, Jul. 2022, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s43588-022-00281-6

[9] S. L. Brunton, J. L. Proctor, and J. N. 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*, vol. 113, no. 15, pp. 3932–3937, Apr. 2016.

[10] Y. Liang, N. Kumar, H. Tang, A. Weller, J. B. Tenenbaum, T. Silver, J. F. Henriques, and K. Ellis, "VisualPredicator: Learning Abstract World Models with Neuro-Symbolic Predicates for Robot Planning," Oct. 2024, arXiv:2410.23156 [cs]. [Online]. Available: http://arxiv.org/abs/2410.23156

[11] S. M. Katz, A. L. Corso, C. A. Strong, and M. J. Kochenderfer, "Verification of Image-Based Neural Network Controllers Using Generative Models," *Journal of Aerospace Information Systems*, vol. 19, no. 9, pp. 574–584, 2022, publisher: American Institute of Aeronautics and Astronautics _eprint: https://doi.org/10.2514/1.I011071. [Online]. Available: https://doi.org/10.2514/1.I011071

[12] G. E. Hinton, "Learning Distributed Representations of Concepts," *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 8, no. 0, 1986. [Online]. Available: https://escholarship.org/uc/item/79w838g1

[13] Z. Lu, I. Afridi, H. J. Kang, I. Ruchkin, and X. Zheng, "Surveying neuro-symbolic approaches for reliable artificial intelligence of things," *Journal of Reliable Intelligent Environments*, Jul. 2024. [Online]. Available: https://doi.org/10.1007/s40860-024-00231-1

[14] Z. Mao, C. Sobolewski, and I. 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)*, 2024, arXiv:2308.12252 [cs]. [Online]. Available: http://arxiv.org/abs/2308.12252

[15] Z. Mao, S. Dai, Y. Geng, and I. 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)*, Mar. 2024, arXiv:2404.00462 [cs] version: 1. [Online]. Available: http://arxiv.org/abs/2404.00462

[16] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 2451–2463, https://worldmodels.github.io. [Online]. Available: https://papers.nips.cc/paper/7512-recurrent-world-models-facilitate-policy-evolution

[17] D.-H. 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*, vol. 3, no. 2. Atlanta, 2013, p. 896.

[18] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[19] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[20] Y. Geng, S. Dutta, and I. Ruchkin, "Bridging dimensions: Confident reachability for high-dimensional controllers," 2024.

[21] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016. [Online]. Available: https://arxiv.org/abs/1606.01540

[22] T. Gupta, W. Gong, C. Ma, N. Pawlowski, A. Hilmkil, M. Scetbon, M. Rigter, A. Famoti, A. J. Llorens, J. Gao, S. Bauer, D. Kragic, B. Schölkopf, and C. Zhang, "The Essential Role of Causality in Foundation World Models for Embodied AI," Apr. 2024, arXiv:2402.06665. [Online]. Available: http://arxiv.org/abs/2402.06665

[23] X. Zheng, Z. Li, I. Ruchkin, R. Piskac, and M. Pajic, "NeuroStrata: Harnessing Neurosymbolic Paradigms for Improved Design, Testability, and Verifiability of Autonomous CPS," Feb. 2025, arXiv:2502.12267 [cs]. [Online]. Available: http://arxiv.org/abs/2502.12267

[24] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari, "NuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles," Feb. 2022, arXiv:2106.11810 [cs]. [Online]. Available: http://arxiv.org/abs/2106.11810

[25] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in Perception for Autonomous Driving: Waymo Open Dataset," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 2443–2451, iSSN: 2575-7075. [Online]. Available: https://ieeexplore.ieee.org/document/9156973

[26] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[27] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

[28] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning," in *International conference on machine learning*. PMLR, 2018, pp. 1184–1193.

[29] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," *Advances in neural information processing systems*, vol. 31, 2018.

[30] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *Journal of Machine Learning Research*, vol. 18, no. 185, pp. 1–52, 2018.

[31] Z. Mao and I. Ruchkin, "Towards Physically Interpretable World Models: Meaningful Weakly Supervised Representations for Visual

Trajectory Prediction," Dec. 2024, arXiv:2412.12870 [cs]. [Online]. Available: http://arxiv.org/abs/2412.12870

[32] V. Micheli, E. Alonso, and F. Fleuret, "Transformers are Sample-Efficient World Models," Mar. 2023, arXiv:2209.00588 [cs]. [Online]. Available: http://arxiv.org/abs/2209.00588

[33] J. Robine, M. Höftmann, T. Uelwer, and S. Harmeling, "Transformer-based World Models Are Happy With 100k Interactions," Mar. 2023, arXiv:2303.07109 [cs]. [Online]. Available: http://arxiv.org/abs/2303.07109

[34] D. Hafner, T. P. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," *CoRR*, vol. abs/1811.04551, 2018. [Online]. Available: http://arxiv.org/abs/1811.04551

[35] D. Hafner, T. P. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," *CoRR*, vol. abs/1912.01603, 2019. [Online]. Available: http://arxiv.org/abs/1912.01603

[36] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba, "Mastering atari with discrete world models," 2022.

[37] Z. Ding, A. Zhang, Y. Tian, and Q. Zheng, "Diffusion world model," 2024.

[38] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "DriveDreamer: Towards Real-world-driven World Models for Autonomous Driving," Nov. 2023, arXiv:2309.09777 [cs]. [Online]. Available: http://arxiv.org/abs/2309.09777

[39] C. Min, D. Zhao, L. Xiao, J. Zhao, X. Xu, Z. Zhu, L. Jin, J. Li, Y. Guo, J. Xing, L. Jing, Y. Nie, and B. Dai, "DriveWorld: 4D Pre-trained Scene Understanding via World Models for Autonomous Driving," May 2024, arXiv:2405.04390 [cs]. [Online]. Available: http://arxiv.org/abs/2405.04390

[40] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "OccWorld: Learning a 3D Occupancy World Model for Autonomous Driving," Nov. 2023, arXiv:2311.16038 [cs]. [Online]. Available: http://arxiv.org/abs/2311.16038

[41] C. Min, D. Zhao, L. Xiao, Y. Nie, and B. Dai, "UniWorld: Autonomous Driving Pre-training via World Models," Aug. 2023, arXiv:2308.07234 [cs]. [Online]. Available: http://arxiv.org/abs/2308.07234

[42] Z. Yan, W. Dong, Y. Shao, Y. Lu, L. Haiyang, J. Liu, H. Wang, Z. Wang, Y. Wang, F. Remondino, and Y. Ma, "RenderWorld: World Model with Self-Supervised 3D Label," Sep. 2024, arXiv:2409.11356 [cs]. [Online]. Available: http://arxiv.org/abs/2409.11356

[43] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," Nov. 2016. [Online]. Available: https://openreview.net/forum?id=Sy2fzU9gl

[44] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang, "CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models," 2021, pp. 9593–9602.

[45] J. Lu, W. Zhan, M. Tomizuka, and Y. Hu, "Towards Generalizable and Interpretable Motion Prediction: A Deep Variational Bayes Approach," in *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. PMLR, Apr. 2024, pp. 4717–4725, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v238/lu24a.html

[46] M. Itkina and M. Kochenderfer, "Interpretable Self-Aware Neural Networks for Robust Trajectory Prediction," Aug. 2022. [Online]. Available: https://openreview.net/forum?id=fnaMlJbRc4t

[47] Z. Lin, Y.-F. Wu, S. Peri, B. Fu, J. Jiang, and S. Ahn, "Improving Generative Imagination in Object-Centric World Models," Oct. 2020, arXiv:2010.02054 [cs]. [Online]. Available: http://arxiv.org/abs/2010.02054

[48] A. R. Kosiorek, H. Kim, I. Posner, and Y. W. Teh, "Sequential Attend, Infer, Repeat: Generative Modelling of Moving Objects," Nov. 2018, arXiv:1806.01794 [cs]. [Online]. Available: http://arxiv.org/abs/1806.01794

[49] J. Kossen, K. Stelzner, M. Hussing, C. Voelcker, and K. Kersting, "Structured Object-Aware Physics Prediction for Video Modeling and Planning," Feb. 2020, arXiv:1910.02425 [cs]. [Online]. Available: http://arxiv.org/abs/1910.02425

[50] J. Jiang, S. Janghorbani, G. d. Melo, and S. Ahn, "SCALOR: Generative World Models with Scalable Object Representations," Mar. 2020, arXiv:1910.02384 [cs]. [Online]. Available: http://arxiv.org/abs/1910.02384

[51] E. Crawford and J. Pineau, "Exploiting Spatial Invariance for Scalable Unsupervised Object Tracking," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 3684–3692, Apr. 2020, number: 04. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5777

[52] H. Cui, T. Nguyen, F.-C. Chou, T.-H. Lin, J. Schneider, D. Bradley, and N. Djuric, "Deep Kinematic Models for Kinematically Feasible Vehicle Trajectory Predictions," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, May 2020, pp. 10 563–10 569, iSSN: 2577-087X.

[53] K. Sridhar, S. Dutta, J. Weimer, and I. Lee, "Guaranteed Conformance of Neurosymbolic Models to Natural Constraints," in *L4DC 2023*, Apr. 2023, arXiv:2212.01346 [cs]. [Online]. Available: http://arxiv.org/abs/2212.01346

[54] R. Tumu, L. Lindemann, T. Nghiem, and R. Mangharam, "Physics Constrained Motion Prediction with Uncertainty Quantification," in *Intelligent Vehicles 2023*. arXiv, May 2023, arXiv:2302.01060 [cs]. [Online]. Available: http://arxiv.org/abs/2302.01060

[55] O. Linial, N. Ravid, D. Eytan, and U. Shalit, "Generative ODE modeling with known unknowns," in *Proceedings of the Conference on Health, Inference, and Learning*, ser. CHIL '21. New York, NY, USA: Association for Computing Machinery, Apr. 2021, pp. 79–94. [Online]. Available: https://dl.acm.org/doi/10.1145/3450439.3451866

[56] W. Zhong and H. Meidani, "PI-VAE: Physics-Informed Variational Auto-Encoder for stochastic differential equations," *Computer Methods in Applied Mechanics and Engineering*, vol. 403, p. 115664, Jan. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0045782522006193

[57] Y. Mao, Y. Gu, L. Sha, H. Shao, Q. Wang, and T. Abdelzaher, "Phy-Taylor: Partially Physics-Knowledge-Enhanced Deep Neural Networks via NN Editing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 1, pp. 447–461, Jan. 2025, conference Name: IEEE Transactions on Neural Networks and Learning Systems. [Online]. Available: https://ieeexplore.ieee.org/document/10297119

[58] Y. Baig, H. R. Ma, H. Xu, and L. You, "Autoencoder neural networks enable low dimensional structure analyses of microbial growth dynamics," *Nature Communications*, vol. 14, no. 1, p. 7937, Dec. 2023, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-023-43455-0

[59] J. Balloch, Z. Lin, R. Wright, X. Peng, M. Hussain, A. Srinivas, J. Kim, and M. O. Riedl, "Neuro-Symbolic World Models for Adapting to Open World Novelty," Jan. 2023, arXiv:2301.06294 [cs]. [Online]. Available: http://arxiv.org/abs/2301.06294

[60] Z. Zhao, B. Li, Y. Du, T. Fu, and C. Wang, "PhysORD: A Neuro-Symbolic Approach for Physics-infused Motion Prediction in Off-road Driving," Oct. 2024, arXiv:2404.01596 [cs]. [Online]. Available: http://arxiv.org/abs/2404.01596

[61] S. Greydanus, M. Dzamba, and J. Yosinski, "Hamiltonian Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://papers.nips.cc/paper_files/paper/2019/hash/26cd8ecadce0d4efd6cc8a8725cbd1f8-Abstract.html

[62] A. R. Luria, "The Functional Organization of the Brain," *Scientific American*, vol. 222, no. 3, pp. 66–79, 1970, publisher: Scientific American, a division of Nature America, Inc. [Online]. Available: https://www.jstor.org/stable/24925755

[63] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel, "Neural Relational Inference for Interacting Systems," Jun. 2018. [Online]. Available: http://arxiv.org/abs/1802.04687

[64] 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*, vol. 378, pp. 686–707, Feb. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999118307125

[65] S. Saemundsson, A. Terenin, K. Hofmann, and M. Deisenroth, "Variational Integrator Networks for Physically Structured Embeddings," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR, Jun. 2020, pp. 3078–3087, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v108/saemundsson20a.html

[66] A. Zhang, C. Lyle, S. Sodhani, A. Filos, M. Kwiatkowska, J. Pineau, Y. Gal, and D. Precup, "Invariant Causal Prediction for Block MDPs," in *Proceedings of the 37th International Conference on Machine*

Learning. PMLR, Nov. 2020, pp. 11 214–11 224, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v119/zhang20t.html

[67] E. v. d. Pol, T. Kipf, F. A. Oliehoek, and M. Welling, "Plannable Approximations to MDP Homomorphisms: Equivariance under Actions," Feb. 2020, arXiv:2002.11963 [cs]. [Online]. Available: http://arxiv.org/abs/2002.11963

[68] J. Y. Park, O. Biza, L. Zhao, J. W. v. d. Meent, and R. Walters, "Learning Symmetric Embeddings for Equivariant World Models," Jun. 2022, arXiv:2204.11371 [cs]. [Online]. Available: http://arxiv.org/abs/2204.11371

[69] F. Zhuang, X. Cheng, P. Luo, S. J. Pan, and Q. He, "Supervised representation learning: Transfer learning with deep autoencoders," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[70] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, 2016.

[71] A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy, "Fixing a broken elbo," in *International conference on machine learning*. PMLR, 2018, pp. 159–168.

[72] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International conference on machine learning*. PMLR, 2014, pp. 1278–1286.

[73] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[74] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[75] A. K. Tanwani, P. Sermanet, A. Yan, R. Anand, M. Phielipp, and K. Goldberg, "Motion2vec: Semi-supervised representation learning from surgical videos," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2174–2181.

[76] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.

[77] D. Sorokin and I. 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*. Springer, 2017, pp. 70–83.

[78] R. P. Poudel, H. Pandya, and R. Cipolla, "Contrastive unsupervised learning of world model with invariant causal features," *arXiv preprint arXiv:2209.14932*, 2022.

[79] R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak, "Planning to explore via self-supervised world models," in *International conference on machine learning*. PMLR, 2020, pp. 8583–8592.

[80] I. Teeti, S. Khan, A. Shahbaz, A. Bradley, F. Cuzzolin, and L. De Raedt, "Vision-based intention and trajectory prediction in autonomous vehicles: A survey." in *IJCAI*, 2022, pp. 5630–5637.

[81] M. Althoff, "An introduction to cora 2015," in *Proc. of the workshop on applied verification for continuous and hybrid systems*, 2015, pp. 120–151.

[82] C. S. Păsăreanu, R. Mangal, D. Gopinath, S. Getir Yaman, C. Imrie, R. Calinescu, and H. Yu, "Closed-loop analysis of vision-based autonomous systems: A case study," in *International conference on computer aided verification*. Springer, 2023, pp. 289–303.

[83] F. Cai, C. Fan, and S. Bak, "Scalable surrogate verification of image-based neural network control systems using composition and unrolling," *arXiv preprint arXiv:2405.18554*, 2024.

[84] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning Latent Dynamics for Planning from Pixels," Jun. 2019, arXiv:1811.04551 [cs]. [Online]. Available: http://arxiv.org/abs/1811.04551

[85] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to Control: Learning Behaviors by Latent Imagination," Mar. 2020, arXiv:1912.01603 [cs]. [Online]. Available: http://arxiv.org/abs/1912.01603

[86] K. Kim, M. Sano, J. D. Freitas, N. Haber, and D. Yamins, "Active World Model Learning with Progress Curiosity," Jul. 2020, arXiv:2007.07853 [cs]. [Online]. Available: http://arxiv.org/abs/2007.07853

[87] J. Y. Koh, H. Lee, Y. Yang, J. Baldridge, and P. Anderson, "Pathdreamer: A World Model for Indoor Navigation," Aug. 2021, arXiv:2105.08756 [cs]. [Online]. Available: http://arxiv.org/abs/2105.08756

[88] M. Okada and T. Taniguchi, "DreamingV2: Reinforcement Learning with Discrete World Models without Reconstruction," Mar. 2022, arXiv:2203.00494 [cs]. [Online]. Available: http://arxiv.org/abs/2203.00494

[89] A. Nakano, M. Suzuki, and Y. Matsuo, "Interaction-Based Disentanglement of Entities for Object-Centric World Models," Sep. 2022. [Online]. Available: https://openreview.net/forum?id=JQc2VowqCzz

[90] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "GAIA-1: A Generative World Model for Autonomous Driving," Sep. 2023, arXiv:2309.17080 [cs]. [Online]. Available: http://arxiv.org/abs/2309.17080

[91] Y.-R. Liu, B. Huang, Z. Zhu, H. Tian, M. Gong, Y. Yu, and K. Zhang, "Learning World Models with Identifiable Factorization," Jun. 2023, arXiv:2306.06561 [cs]. [Online]. Available: http://arxiv.org/abs/2306.06561

[92] V. Micheli, E. Alonso, and F. Fleuret, "Transformers are sample-efficient world models," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=vhFu1Acb0xb

[93] V. Shaj, S. G. Zadeh, O. Demir, L. R. Douat, and G. Neumann, "Multi Time Scale World Models."

[94] T. Wang, S. S. Du, A. Torralba, P. Isola, A. Zhang, and Y. Tian, "Denoised MDPs: Learning World Models Better Than the World Itself," Apr. 2023, arXiv:2206.15477 [cs]. [Online]. Available: http://arxiv.org/abs/2206.15477

[95] L. Wong, G. Grand, A. K. Lew, N. D. Goodman, V. K. Mansinghka, J. Andreas, and J. B. Tenenbaum, "From Word Models to World Models: Translating from Natural Language to the Probabilistic Language of Thought," Jun. 2023, arXiv:2306.12672 [cs]. [Online]. Available: http://arxiv.org/abs/2306.12672

[96] Y. Seo, D. Hafner, H. Liu, F. Liu, S. James, K. Lee, and P. Abbeel, "Masked World Models for Visual Control," May 2023, arXiv:2206.14244 [cs]. [Online]. Available: http://arxiv.org/abs/2206.14244

[97] S. Hao, Y. Gu, H. Ma, J. J. Hong, Z. Wang, D. Z. Wang, and Z. Hu, "Reasoning with Language Model is Planning with World Model," Oct. 2023, arXiv:2305.14992 [cs]. [Online]. Available: http://arxiv.org/abs/2305.14992

[98] R. Mendonca, S. Bahl, and D. Pathak, "Structured World Models from Human Videos," Aug. 2023, arXiv:2308.10901 [cs]. [Online]. Available: http://arxiv.org/abs/2308.10901

[99] C. Gumbsch, N. Sajid, G. Martius, and M. V. Butz, "Learning Hierarchical World Models with Adaptive Temporal Abstractions from Discrete Latent Dynamics," Jul. 2023. [Online]. Available: https://openreview.net/forum?id=5qappsbO73r

[100] M. Rigter, T. Gupta, A. Hilmkil, and C. Ma, "AVID: Adapting Video Diffusion Models to World Models," Nov. 2024, arXiv:2410.12822 [cs]. [Online]. Available: http://arxiv.org/abs/2410.12822

[101] V. Kolev, R. Rafailov, K. Hatch, J. Wu, and C. Finn, "Efficient Imitation Learning with Conservative World Models," Aug. 2024, arXiv:2405.13193 [cs]. [Online]. Available: http://arxiv.org/abs/2405.13193

[102] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering Diverse Domains through World Models," Apr. 2024, arXiv:2301.04104 [cs]. [Online]. Available: http://arxiv.org/abs/2301.04104

[103] S. Zuo, W. Zheng, Y. Huang, J. Zhou, and J. Lu, "GaussianWorld: Gaussian World Model for Streaming 3D Occupancy Prediction," Dec. 2024, arXiv:2412.10373 [cs]. [Online]. Available: http://arxiv.org/abs/2412.10373

[104] J. Bruce, M. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, Y. Aytar, S. Bechtle, F. Behbahani, S. Chan, N. Heess, L. Gonzalez, S. Osindero, S. Ozair, S. Reed, J. Zhang, K. Zolna, J. Clune, N. d. Freitas, S. Singh, and T. Rocktäschel, "Genie: Generative Interactive Environments," Feb. 2024, arXiv:2402.15391 [cs]. [Online]. Available: http://arxiv.org/abs/2402.15391

[105] H. Ma, J. Wu, N. Feng, C. Xiao, D. Li, J. Hao, J. Wang, and M. Long, "HarmonyDream: Task Harmonization Inside World

Models," Jun. 2024, arXiv:2310.00344 [cs]. [Online]. Available: http://arxiv.org/abs/2310.00344
[106] H. Zhang, Y. Xue, X. Yan, J. Zhang, W. Qiu, D. Bai, B. Liu, S. Cui, and Z. Li, "An Efficient Occupancy World Model via Decoupled Dynamic Flow and Image-assisted Training," Dec. 2024, arXiv:2412.13772 [cs]. [Online]. Available: http://arxiv.org/abs/2412.13772
[107] A. Popov, A. Degirmenci, D. Wehr, S. Hegde, R. Oldja, A. Kamenev, B. Douillard, D. Nistér, U. Muller, R. Bhargava, S. Birchfield, and N. Smolyanskiy, "Mitigating Covariate Shift in Imitation Learning for Autonomous Vehicles Using Latent Space Generative World Models," Sep. 2024, arXiv:2409.16663 [cs]. [Online]. Available: http://arxiv.org/abs/2409.16663
[108] A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun, "Navigation World Models," Dec. 2024, arXiv:2412.03572 [cs]. [Online]. Available: http://arxiv.org/abs/2412.03572
[109] J. Wei, S. Yuan, P. Li, Q. Hu, Z. Gan, and W. Ding, "OccLLaMA: An Occupancy-Language-Action Generative World Model for Autonomous Driving," Sep. 2024, arXiv:2409.03272 [cs]. [Online]. Available: http://arxiv.org/abs/2409.03272
[110] V. D. Nguyen, Z. Yang, C. L. Buckley, and A. Ororbia, "R-AIF: Solving Sparse-Reward Robotic Tasks from Pixels with Active Inference and World Models," Sep. 2024, arXiv:2409.14216 [cs]. [Online]. Available: http://arxiv.org/abs/2409.14216
[111] Q. Li, X. Jia, S. Wang, and J. Yan, "Think2Drive: Efficient Reinforcement Learning by Thinking in Latent World Model for Quasi-Realistic Autonomous Driving (in CARLA-v2)," Jul. 2024, arXiv:2402.16720 [cs]. [Online]. Available: http://arxiv.org/abs/2402.16720
[112] C. Chen, Y.-F. Wu, J. Yoon, and S. Ahn, "TransDreamer: Reinforcement Learning with Transformer World Models," Nov. 2024, arXiv:2202.09481 [cs]. [Online]. Available: http://arxiv.org/abs/2202.09481
[113] Z. Ge, H. Huang, M. Zhou, J. Li, G. Wang, S. Tang, and Y. Zhuang, "WorldGPT: Empowering LLM as Multimodal World Model," Sep. 2024, arXiv:2404.18202 [cs]. [Online]. Available: http://arxiv.org/abs/2404.18202

APPENDIX

A. State of the Art in World Models

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 [7], [32], [33]. Modern deep-world models were first proposed by [16]: a variational autoencoder (VAE) generated image representations and a recurrent neural network made latent-space predictions. Later work refined both the encoder-decoder architecture and the surrogate dynamics: PlaNet introduced a recurrent state-space model (RSSM) for prediction [34]; Dreamer propagated gradients back through the imagined trajectories to refine latent prediction [35]; and DreamerV2 extended the RSSM to categorical latent variables [36]. More recent research integrates autoregressive transformers with self-attention layers to capture detailed temporal dependencies [33], or diffusion models to mitigate compounding errors [37]. World models have also found success in optimizing planning algorithms for autonomous vehicles in realistic environments: DriveDreamer [38] leverages multi-modal inputs to generate realistic video trajectories for policy optimization; DriveWorld [39], OccWorld [40], UniWorld [41], and RenderWorld [42] forecast detailed 3D occupancy to inform motion planning and control.

Towards interpretable world models. Despite the strong performance of world models, their interpretability remains a major challenge in most frameworks. Early efforts toward disentangling latent variables (i.e., reducing their mutual dependency) include β-VAEs [43] and causal VAEs [44]. This disentanglement strategy is also employed in driving prediction frameworks like GNeVA [45] and ISAP [46]. Under the umbrella of world models, G-SWM [47] investigated a principled modeling framework that inherits interpretable object and context latent separation from various spatial attention approaches [48]–[51]. Interpretability can also manifest in the world model's forward dynamics: [52] propose embedding known dynamics into neural pipelines, while more recent work imposes further physical constraints on system identification [53], motion prediction [54], and ordinary differential equations with learnable parameters [55]–[57]. Incorporating partial knowledge of physics with weak supervision has recently improved both the state and dynamics interpretability [31]. A recent Nature article leveraged the biological alignment of latent representations to predict microbiome community interactions and antibiotic resistance [58]. Neuro-symbolic world models have also begun to emerge: VisualPredicator [10] learns a set of abstract states and high-level actions for strong out-of-distribution generalization, whereas WorldCloner [59] learns symbolic rules to adapt the dynamics to open world novelty.Additional neuro-symbolic work that inspires physically-aligned world models is PhysORD [60] which embeds physical laws into neural models, improving long-horizon motion prediction and interpretability.

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 [10], [31], [47], [59], [61]. 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 drastically improve RL sample efficiency by shrinking the search space to physically feasible solutions.

B. Existing Work Supporting the Principles

Principle 1. Although many facets of human cognition remain a mystery, extensive research indicates that distinct regions of the brain are dedicated to unique functions [62]. For example, early studies of patients with brain injuries revealed the existence of functional systems partially dedicated to meaningful tasks such as spatial awareness, voluntary movement, and sensory processing [62]. While modeling the minutia of human cognition is beyond the scope of this principle, this idea of *functional organization* inspires how world model architectures are increasingly structured.

Recent work in the context of Branch 1 (physical variables) showed that latent representations in world models can be effectively aligned with physical properties [31]. Although

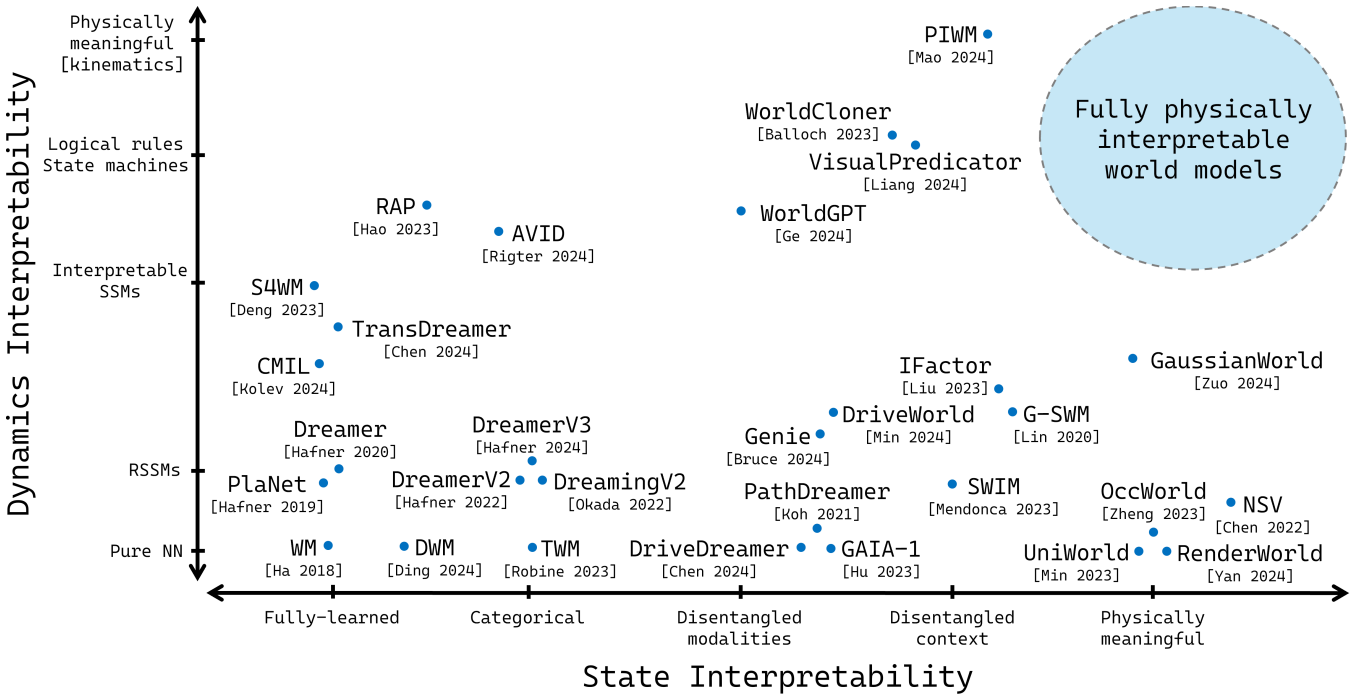

Fig. 3. State-of-the-art world models by the interpretability of their state and dynamics.

modeling agent interactions is a well-explored topic, early work by [63] revealed how physical interactions between agents can be learned through graph neural networks (GNN) in an unsupervised manner. Later advances in learning physical interactions within the context of world models by [47] proposed a method that constructs a separate latent representation using a GNN to capture agent occlusion and interaction relationships. Physics informed neural networks [64], [65] can also improve the physical interpretability of the world model's forward dynamics. For instance, Hamiltonian neural networks [61] learn and adhere to physical conservation laws, leading to impressive generalization. Finally, Branch 3 follows the typical strategy for creating uninterpretable world models and is considered a useful layer to the structured latent space [47].

Latent space structuring has become increasingly prevalent in improving the performance of planning and control. To this end, [45] developed a goal-based neural variational agent (GNeVA) that uses separate polyline embeddings for the agent and the map to achieve interpretable generative motion prediction. Similarly, [46] designed an interpretable car trajectory prediction that integrates three distinct workflow branches: agent states, high-definition maps, and social context.

Principle 2. [66] utilize bisimulation metrics to learn latent representations of roadway obstacles that are invariant to obstacle type, size, and brightness. The dual of avoiding explicitly irrelevant details is explicitly learning relevant details. [67] introduce contrastive loss to enforce action-equivariance on learned representations. However, we still lack ways to effectively incorporate these expert conceptual priors into world models. [68] innovate a symmetric embedding network for world models that learns simple latent space transformations from complex observation space transformations.

In practice, deciding which features the representation should be invariant or equivariant to solely depends on the task at hand and the discretion of the expert engineer. In autonomous driving, consider a recovery component that overrides the default controller during inclement weather. This component's representations of real-time images might change when the sky is cloudy. On the other hand, the world model of the default controller might be invariant to the features of the sky.

Principle 3. Here we discuss the supervision signal types for the latent states and their impact on physical interpretability. *Supervised Learning:* Direct supervision connects latent values with physical states. In many cases, supervision signals are introduced directly into the embeddings to capture key features from labeled data [69]. 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 [70]–[72].
Semi-Supervised Learning: When the labels are only available for some data, semi-supervised techniques can refine representations. Pseudo-labeling (e.g., Mean Teacher [73] and FixMatch [74]) utilizes both labeled and unlabeled data to iteratively improve the latent space. In Motion2Vec [75], 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 [76] stabilize noisy

trajectories in tasks such as autonomous driving. Interval signals as weak supervision can be directly incorporated into the loss function [31] or combined with contrastive learning to reinforce constraints [77].

Self-Supervised Learning: In the absence of labels, contrastive learning [18] aligns latent representations with task-specific similarity metrics (e.g., Euclidean distance or structural similarity). Contrastive world models [78] explicitly employ representation learning losses to map similar states closer in the latent space. Plan2Explore [79] generates self-supervised uncertainty-driven objectives to guide the representations.

Principle 4. Ensuring the safety of vision-based autonomy is a critical and open challenge [20], [80]. The verification of such systems remains difficult due to high-dimensional image inputs: traditional techniques struggle to handle this complexity, making it essential to develop new principles for pre-deployment safety guarantees [81], [82]. A recent approach employs a generative image model to map the physical state to the observed image, in turn fed into a state estimator or controller [11], [83]. Unfortunately, due to uninterpretable latent states, such "verification modulo generative models" does not provide guarantees regarding the physical world. Furthermore, it does not scale to large image sizes.

C. Experimental Details

The state dimensions for the Lunar Lander and Cart Pole are 8 and 4, respectively, reflecting different levels of complexity in achieving interpretability. All experiments were conducted on 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. The architecture details can be found in Section III. Our decoder network maps low-dimensional physical state representations to high-dimensional images using a series of transposed convolutional layers. 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 Equation 1, the choice of λ plays a crucial role in image generation behavior:

- If λ is too small (< 0.1), the model fails to separate the three parts, blending "shadows" of the original image into the outputs.
- 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.

D. Additional Illustrations

- Table I shows the comparison between a unified and partitioned generator.
- 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.
- Table II lists the literature on world models and indicates the extent of state/dynamics interpretability and adherence to the proposed principles.

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 I
MODEL SIZE AND RECONSTRUCTION PERFORMANCE FOR VALIDATING PRINCIPLE 4 WITH $\lambda = 0.2$.

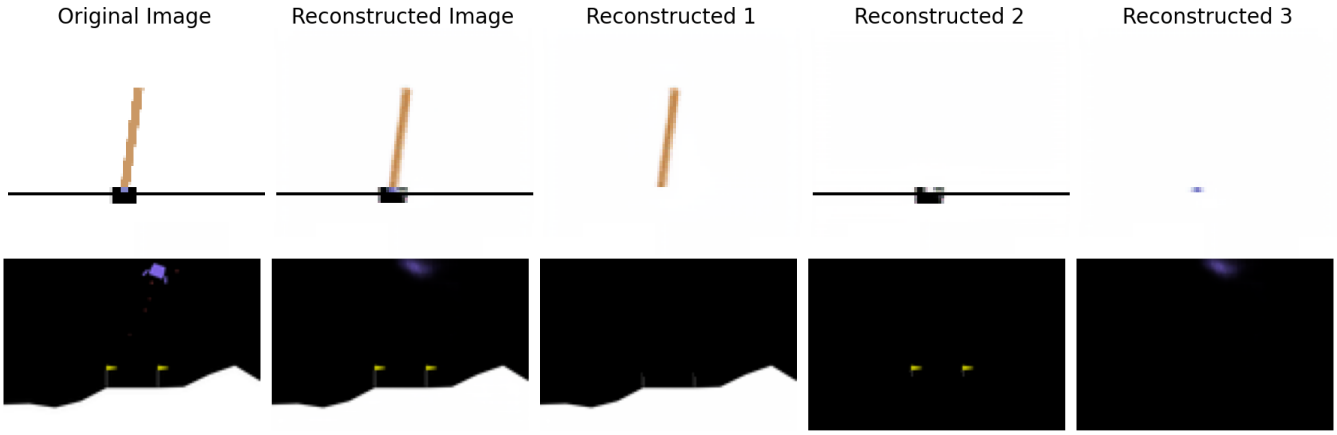

Fig. 4. Observations and three reconstructed parts (Principle 4) for the cartpole and lunar lander with $\lambda = 0.2$.

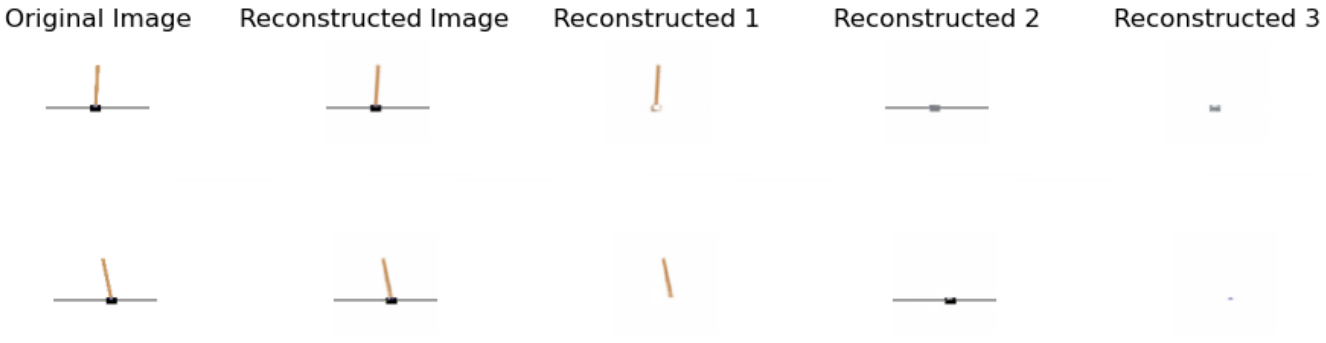

Fig. 5. Imperfect reconstruction for the cart pole: the upper row corresponds to $\lambda = 0.01$, while the bottom row corresponds to $\lambda = 0.9$.

Short Name	Reference	Principle 1	Principle 2	Principle 3	Principle 4	State interp.	Dyn. interp.
WM	[5]						
PlaNet	[84]	Weak					Weak
Dreamer	[85]	Weak					Weak
G-SWM	[47]	Strong	Weak		Weak	Moderate	Weak
AWM	[86]	Weak				Weak	Weak
Plan2Explore	[79]	Weak				Weak	Weak
Pathdreamer	[87]		Weak	Weak	Strong	Weak	
DreamerV2	[36]	Weak				Weak	Weak
NSV	[8]		Strong			Strong	
DayDreamer	[6]	Weak					Weak
DreamingV2	[88]	Weak				Weak	Weak
SEN	[68]		Strong			Moderate	
STEDI	[89]	Strong	Moderate			Strong	
DriveDreamer	[38]	Weak		Strong		Weak	
GAIA-1	[90]	Strong				Moderate	
IFactor	[91]	Moderate				Moderate	
IRIS	[92]	Weak				Weak	
MTS3	[93]	Weak	Weak			Weak	Weak
Denoised MDP	[94]	Moderate				Moderate	
WM2WM	[95]			Strong			Moderate
MWM	[96]	Weak					Weak
OccWorld	[40]	Strong	Strong			Strong	
RAP	[97]						Moderate
S4WM	[7]						Moderate
SWIM	[98]	Moderate	Moderate			Moderate	
TWM	[33]	Weak			Weak		
UniWorld	[41]		Strong	Strong		Strong	
WorldCloner	[59]		Strong			Moderate	Strong
THICK	[99]	Weak	Weak				Weak
AVID	[100]			Strong			Moderate
CMIL	[101]			Strong			Weak
DreamerV3	[102]	Weak	Weak			Weak	Weak
DriveWorld	[39]		Strong			Moderate	Weak
DWM	[37]						
GaussianWorld	[103]	Moderate	Strong			Strong	Moderate
Genie	[104]	Moderate				Moderate	
HarmonyWM	[105]	Weak					Weak
OccWM	[106]	Moderate	Strong			Strong	Weak
CovWM	[107]	Weak	Weak			Weak	Weak
NWM	[108]						
OccLLaMA	[109]		Strong			Strong	
PIWM	[31]		Strong	Strong		Moderate	Strong
R-AIF	[110]		Weak			Weak	Weak
RenderWorld	[42]	Strong	Strong			Strong	
Think2Drive	[111]	Weak					Weak
TransDreamer	[112]	Weak				Weak	Weak
VisualPredicator	[10]		Strong			Strong	Moderate
WorldGPT	[113]	Moderate			Moderate	Moderate	Moderate
Our future vision		Strong	Strong	Strong	Strong	Strong	Strong

TABLE II

REVIEW OF NOTABLE AND STATE-OF-THE-ART WORLD MODEL ARCHITECTURES FOR ADHERENCE TO THE FOUR PRINCIPLES AND THEIR DYNAMICAL/STATE INTERPRETABILITY.