Discovering Hybrid World Representations with Co-Evolving Foundation Models

Jiajun Wu[1], **Yunzhi Zhang**[1], **Hong-Xing Yu**[1], **Joy Hsu**[1], **Jiayuan Mao**[2]

[1]Stanford University
[2]University of Pennsylvania

Abstract

This perspective article discusses an emerging research direction: to what extent can foundation models yield usable structure for modeling the physical world? We offer a Markovian formulation of structured world models and outline the notion of multi-level hybrid world representations that support compositional structure. We then review and suggest possible discovery paradigms, spanning distillation, interaction-driven continual learning, and ensemble learning.

Introduction

Foundation models encode broad statistical priors but leave unclear how much explicit, executable structure—states, operators, constraints—can be extracted and refined for perception, reasoning, and control in the physical world. Purely parametric approaches capture correlations yet obscure causality; purely symbolic approaches offer clarity but struggle with coverage and robustness. An emerging research question is how far foundation model priors can be pushed toward a structured world model.

In this perspective article, we first outline a definition of structured, Markovian world models, as well as the idea of multi-level hybrid world representations across forms, domains, and modalities. We then discuss paradigms that discover these world models and representations. While most existing approaches have focused on using foundation models for various forms of supervision in learning and optimization, we highlight the possibility of co-evolving foundation models through interaction, as well as integrating multiple foundation model experts. Altogether, this paradigm aims to discover structured world models that are expressive and flexible, generalize across tasks, leverage domain knowledge when available, remain interpretable, editable, and auditable, and have the potential to identify new generative rules of the physical world.

Numerous papers fall into this emerging trend. This perspective article is not intended to be a survey; thus, we refer to only a few representative papers to illustrate the idea, which admittedly are biased toward the authors' work.

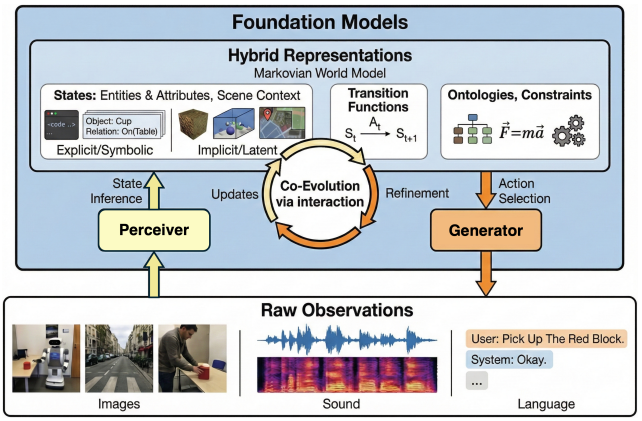

Figure 1: An emerging research direction is to discover structured, Markovian world models with hybrid representations from foundation models, which involves the co-evolution of the representations and models via interaction.

Goal of Discovery—Structured World Models

We assume that the physical world has compositional abstractions and is Markovian. The structured world models—our goal to be discovered from observations and foundation models, and refined through co-evolution—include

- States, including entities and their attributes, as well as the scene context;
- Transition functions that are probabilistic and capture how states change conditioned on actions.

Depending on the tasks, we may also want to discover

- Relations and ontologies among state representations;
- Constraints and physical laws that contract a subset of representations, transition rules, or both.

There are also perceivers and generators—neither part of the structured world model in our definition, but both bridging state representations to raw observations (See Figure 1). They may be in various forms. Perceivers are now typically multi-modal foundation models, though modality-specific encoders are still used. Generators may range from foundation models, to traditional physics engines, to differentiable or neural renderers or simulators, to text-to-image/video/4D generative models.

Hybrid World Representations

We present hybrid world representations—capturing heterogeneity across multiple levels and dimensions of abstraction.

Form. Its hybrid form mixes explicit, interpretable tokens, such as natural language and code (Tang, Key, and Ellis 2024; Oswald et al. 2024), with implicit, latent features. Examples include the Scene Language (Zhang et al. 2025a) for visual generation, and neuro-symbolic planning (Mao et al. 2022; Han et al. 2024; Liu et al. 2024b). Here,

- Explicit tokens (discrete language/code) directly link to the foundation model's reasoning space, keeping learned knowledge interpretable and anchored to learned priors.
- Implicit features (continuous latent vectors) provide expressivity for capturing complex structures, such as geometry, texture, and physics.

Domains. At the middle level, the representation is hybrid as it captures, collectively but disjointly, attributes from different domains. For example, a hybrid representation for object appearance and physics, sharing the same form across domains, has enabled flexible world modeling and action-conditioned prediction across diverse object matters (Li et al. 2023b, 2024b, 2025; Liu et al. 2024a; Gao et al. 2025).

Modalities. At the lower level, even if in the same form (e.g., all explicit) and the same domain, representations may still be hybrid to fuse the complementary strengths of each representation (Song, Song, and Huang 2020).

Intuitively, we argue that these multi-level hybrid representations strike a bias-variance trade-off, enabling the learning of a generalizable, structured world model from limited observations and sample-efficient interactions.

Paradigm of Discovery

The discovery paradigm comprises foundation-model-guided learning (sometimes called distillation), continual learning through interaction, and ensemble learning with multiple expert foundation models. Soon, multi-stage discovery paradigms that fuse these threads will likely emerge.

Foundation-model-guided learning

The most direct way to extract a structured world model from a foundation model is to use the latter as supervision, broadly defined. This includes prompting, gradient-based optimization (including distillation sampling), reinforcement learning, and beyond.

Prompting queries a pretrained foundation model (e.g., a large language model or vision language model) using real-world observations (e.g., language, images) in an inference-only way. The produced structure might be in diverse forms, spanning semantic programs for words (Hsu et al. 2025), policies (Liang et al. 2023), motion plans (Singh et al. 2023), sentences (Surís, Menon, and Vondrick 2023; Gupta and Kembhavi 2023) to object and scene layouts (Ritchie et al. 2023; Hu et al. 2024; Sun et al. 2025), as well as hybrid representations that combine heterogeneous pre-trained foundation models (Ganeshan et al. 2024; Wong et al. 2024; Zhang et al. 2025a).

Gradient-based optimization assumes a differentiable generative engine that translates the state representations into raw data format, such as images and language. The generation results are assessed based on their likelihood using a foundation model, and the gradients are used to refine the representations. Often, the real-world counterparts of the generation output are missing; in such cases, foundation models can be used to augment and complete real-world observations (Zhang et al. 2024; Zhao et al. 2025; Wu et al. 2025), or sampling methods can be employed to compute gradients despite the gap (Poole et al. 2023; Sargent et al. 2024).

Reinforcement learning may be used when the foundation models to be distilled, or any modules such as the generators (which themselves can be foundation models), are not differentiable (Ahn et al. 2022; Wang et al. 2025a,b).

Continual learning through interaction

Beyond distillation from foundation models guided by passive observations (Nottingham et al. 2023; Brahman et al. 2024; Li et al. 2024a), we may benefit from continual learning to refine the discovered worlds and, possibly, the foundation models themselves through interactions with the real world. Recall that our structured world model includes the state representations and transition rules. After initializing them via foundation-model-guided learning, as described above, the continual learning paradigm, including an iterative cycle of perception, interaction, and symbolic abstraction, will better leverage the commonsense knowledge from foundation models. The interactive, continual learning process is driven by two objectives: (1) discovering state representations and transition rules that best explain the real-world observations, and (2) jointly optimizing the decision-making policy that builds on the learned structured world model to achieve task reward (Guan et al. 2023; Tang, Key, and Ellis 2024; Zhu and Simmons 2024).

Interactive learning is mutually beneficial. Our structured world model and its hybrid representations continue to improve through the interpretation of interaction results by foundation models; at the same time, new knowledge gained from interactions is fed back into foundation models, enabling their continued pre-training for better compression, summarization, and future reasoning. This establishes a co-evolving loop in which both world models and foundation models become more capable.

Ensemble learning with multiple experts

The discovery paradigms are not restricted to a single foundation model. Even more so than ever, it appears to be critical to fuse the knowledge of heterogeneous experts trained on disparate datasets.

Such fusion can be straightforward for inference-only discovery processes—one expert, or foundation model, can be prompted to invoke others sequentially or hierarchically (Li et al. 2023a), potentially in a probabilistic way (Zhang et al. 2025b). For reinforcement learning, rewards from individual experts may also just be combined (Wong et al. 2024). Integration becomes technically more challenging for gradient-

based optimization, which requires careful designs to ensure that each module is (approximately) differentiable (Hsu et al. 2023; Mao, Tenenbaum, and Wu 2026).

An emerging idea is reinforcement learning (RL) with tool orchestration. Given a query, the planner, which itself is likely a foundation model, samples a candidate program that specifies which executors, or experts, to call and how to compose their outputs (Schick et al. 2023). The program trace is optimized based on propagated rewards using policy-gradient methods. This way, both the planner and executors are optimized jointly under the same RL framework, yielding a form of multi-agent co-evolution: planners adapt their strategies as hybrid world representations evolve, while experts adapt to the distribution of queries they are invoked to solve. This loop allows the system to refine the structured-world model collaboratively across modules.

Advantages of Our Formulation

We highlight a few unique advantages of our formulation.

Leveraging the MDP structure of the physical world. Our representation is natively aligned with Markov decision processes: it separates states (entities, attributes, and scene context) from actions and transition functions, making the causal structure explicit and learnable, offering action-conditioned prediction "for free". Within the MDP framework, hybrid world representations provide additional flexibility. For example, in robotic manipulation, states can be parameterized as a combination of object poses as continuous vectors and symbolic state descriptions, e.g., "on", "left to". The transition can be modeled as probabilistic outcomes of actions based on the hybrid representation.

Generalizing through compositional abstractions. Abstraction and compositionality collectively enable systematic generalization. Abstraction allows knowledge to extend beyond particular instances—for example, physical rules apply to new objects through object-based abstraction. Compositional abstractions allow individual pieces of knowledge to be used in new situations and combined to form new knowledge. Together with the Markovian assumption, they enable a compact (and therefore efficiently learnable) description of the world's states and transition rules.

Incorporating domain knowledge when available. The hybrid form makes it easy to inject priors at the right level: constraints and conservation laws can be expressed as explicit predicates or program fragments; learned simulators and renderers can enforce soft inductive biases; and expert modules (such as vision, physics, and language) can be composed as callable executors. Because explicit tokens remain human-interpretable and implicit features remain differentiable, the system supports both rule-driven and gradient-based updates, allowing principled fusion of curated knowledge with data-driven learning.

Interfacing with humans intuitively. By grounding part of the state in language and code, and by explicitly interacting with foundation models, the model exposes a natural interface for inspection, editing, and debugging. Practitioners can prompt to propose structure, edit symbolic fragments to test hypotheses, and trace program-like executions to understand failures.

Unveiling causality for scientific discovery. Our formulation treats world modeling as a continual, generative, and inherently causal process of uncovering latent states and transitions, yielding representations that are compact, explanatory, and increasingly predictive. This perspective not only drives self-improvement but also exposes candidate generative rules of the world, providing a principled path toward machine-aided scientific discovery.

Conclusion

Creating AI systems that interact with the physical world and make real-world decisions hinges on robust and generalizable structured world models. Our formulation of structured world model discovery offers flexibility to both where inductive bias originates and how it is validated. It provides an interpretable, editable, and data-efficient interface, while remaining amenable to co-evolve with expressive foundation models.

Acknowledgments

This work is in part supported by AFOSR YIP FA9550-23-1-0127, ONR N00014-23-1-2355, ONR YIP N00014-24-1-2117, ONR MURI N00014-22-1-2740, ONR MURI N00014-24-1-2748, NSF RI #2211258 and RI #2338203, and the Stanford Institute for Human-Centered AI (HAI). The creation of the figure was assisted by foundation models through interaction based on hybrid representations.

References

Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Ho, D.; Hsu, J.; Ibarz, J.; Ichter, B.; Irpan, A.; Jang, E.; Ruano, R. J.; Jeffrey, K.; Jesmonth, S.; Joshi, N. J.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Lee, K.-H.; Levine, S.; Lu, Y.; Luu, L.; Parada, C.; Pastor, P.; Quiambao, J.; Rao, K.; Rettinghouse, J.; Reyes, D.; Sermanet, P.; Sievers, N.; Tan, C.; Toshev, A.; Vanhoucke, V.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yan, M.; and Zeng, A. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *Conference on Robot Learning (CoRL)*.

Brahman, F.; Bhagavatula, C.; Pyatkin, V.; Hwang, J. D.; Li, X. L.; Arai, H. J.; Sanyal, S.; Sakaguchi, K.; Ren, X.; and Choi, Y. 2024. PlaSma: Making Small Language Models Better Procedural Knowledge Models for (Counterfactual) Planning. In *International Conference on Learning Representations (ICLR)*.

Ganeshan, A.; Huang, R. Y.; Xu, X.; Jones, R. K.; and Ritchie, D. 2024. ParSEL: Parameterized Shape Editing with Language. *ACM Transactions on Graphics (TOG)*, 43(6): 1–14.

Gao, Y.; Yu, H.-X.; Zhu, B.; and Wu, J. 2025. FluidNexus: 3D Fluid Reconstruction and Prediction from a Single Video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guan, L.; Valmeekam, K.; Sreedharan, S.; and Kambhampati, S. 2023. Leveraging Pre-trained Large Language Models to Construct and Utilize World Models for Model-based Task Planning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Gupta, T.; and Kembhavi, A. 2023. Visual Programming: Compositional Visual Reasoning Without Training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Han, M.; Zhu, Y.; Zhu, S.-C.; Wu, Y. N.; and Zhu, Y. 2024. InterPreT: Interactive Predicate Learning from Language Feedback for Generalizable Task Planning. In *Robotics: Science and Systems (RSS)*.

Hsu, J.; Mao, J.; Tenenbaum, J.; and Wu, J. 2023. What's Left? Concept Grounding with Logic-Enhanced Foundation Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hsu, J.; Mao, J.; Tenenbaum, J. B.; Goodman, N. D.; and Wu, J. 2025. What Makes a Maze Look like a Maze? In *International Conference on Learning Representations (ICLR)*.

Hu, Z.; Iscen, A.; Jain, A.; Kipf, T.; Yue, Y.; Ross, D. A.; Schmid, C.; and Fathi, A. 2024. SceneCraft: An LLM Agent for Synthesizing 3D Scene as Blender Code. In *International Conference on Machine Learning (ICML)*.

Li, M.; Zhao, S.; Wang, Q.; Wang, K.; Zhou, Y.; Srivastava, S.; Gokmen, C.; Lee, T.; Li, L. E.; Zhang, R.; Liu, W.; Liang, P.; Fei-Fei, L.; Mao, J.; and Wu, J. 2024a. Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Li, S.; Du, Y.; Tenenbaum, J. B.; Torralba, A.; and Mordatch, I. 2023a. Composing Ensembles of Pre-trained Models via Iterative Consensus. In *International Conference on Learning Representations (ICLR)*.

Li, X.; Qiao, Y.-L.; Chen, P. Y.; Jatavallabhula, K. M.; Lin, M.; Jiang, C.; and Gan, C. 2023b. PAC-NeRF: Physics Augmented Continuum Neural Radiance Fields for Geometry-Agnostic System Identification. In *International Conference on Learning Representations (ICLR)*.

Li, Z.; Tucker, R.; Snavely, N.; and Holynski, A. 2024b. Generative Image Dynamics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Z.; Yu, H.-X.; Liu, W.; Yang, Y.; Herrmann, C.; Wetzstein, G.; and Wu, J. 2025. WonderPlay: Dynamic 3D Scene Generation from a Single Image and Actions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Liang, J.; Huang, W.; Xia, F.; Xu, P.; Hausman, K.; Ichter, B.; Florence, P.; and Zeng, A. 2023. Code as Policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Liu, S.; Ren, Z.; Gupta, S.; and Wang, S. 2024a. PhysGen: Rigid-Body Physics-Grounded Image-to-Video Generation. In *European Conference on Computer Vision (ECCV)*.

Liu, W.; Nie, N.; Zhang, R.; Mao, J.; and Wu, J. 2024b. BLADE: Learning Compositional Behaviors from Demonstration and Language. In *Conference on Robot Learning (CoRL)*.

Mao, J.; Lozano-Pérez, T.; Tenenbaum, J.; and Kaelbling, L. 2022. PDSketch: Integrated Domain Programming, Learning, and Planning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Mao, J.; Tenenbaum, J. B.; and Wu, J. 2026. Neuro-Symbolic Concepts. *Communications of the ACM (CACM)*.

Nottingham, K.; Ammanabrolu, P.; Suhr, A.; Choi, Y.; Hajishirzi, H.; Singh, S.; and Fox, R. 2023. Do Embodied Agents Dream of Pixelated Sheep: Embodied Decision Making using Language Guided World Modelling. In *International Conference on Machine Learning (ICML)*.

Oswald, J.; Srinivas, K.; Kokel, H.; Lee, J.; Katz, M.; and Sohrabi, S. 2024. Large Language Models as Planning Domain Generators. In *International Conference on Automated Planning and Scheduling (ICAPS)*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2023. DreamFusion: Text-to-3D Using 2D Diffusion. In *International Conference on Learning Representations (ICLR)*.

Ritchie, D.; Guerrero, P.; Jones, R. K.; Mitra, N. J.; Schulz, A.; Willis, K. D.; and Wu, J. 2023. Neurosymbolic Models for Computer Graphics. *Computer Graphics Forum (CGF)*, 42(2): 545–568.

Sargent, K.; Li, Z.; Shah, T.; Herrmann, C.; Yu, H.-X.; Zhang, Y.; Chan, E. R.; Lagun, D.; Fei-Fei, L.; Sun, D.; and Wu, J. 2024. ZeroNVS: Zero-Shot 360-Degree View Synthesis from a Single Real Image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Singh, I.; Blukis, V.; Mousavian, A.; Goyal, A.; Xu, D.; Tremblay, J.; Fox, D.; Thomason, J.; and Garg, A. 2023. ProgPrompt: Generating Situated Robot Task Plans using Large Language Models. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Song, C.; Song, J.; and Huang, Q. 2020. HybridPose: 6D Object Pose Estimation Under Hybrid Representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sun, C.; Han, J.; Deng, W.; Wang, X.; Qin, Z.; and Gould, S. 2025. 3D-GPT: Procedural 3D Modeling with Large Language Models. In *International Conference on 3D Vision (3DV)*.

Surís, D.; Menon, S.; and Vondrick, C. 2023. ViperGPT: Visual Inference via Python Execution for Reasoning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Tang, H.; Key, D.; and Ellis, K. 2024. WorldCoder, a Model-Based LLM Agent: Building World Models by Writing Code and Interacting with the Environment. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, K.; Zhang, P.; Wang, Z.; Gao, Y.; Li, L.; Wang, Q.; Chen, H.; Lu, Y.; Yang, Z.; Wang, L.; et al. 2025a. VAGEN: Reinforcing World Model Reasoning for Multi-Turn VLM Agents. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, Z.; Wang, K.; Wang, Q.; Zhang, P.; Li, L.; Yang, Z.; Jin, X.; Yu, K.; Nguyen, M. N.; Liu, L.; Gottlieb, E.; Lu, Y.; Cho, K.; Wu, J.; Fei-Fei, L.; Wang, L.; Choi, Y.; and Li, M. 2025b. RAGEN: Understanding Self-Evolution in LLM Agents via Multi-Turn Reinforcement Learning. *arXiv preprint:2504.20073*.

Wong, L.; Mao, J.; Sharma, P.; Siegel, Z. S.; Feng, J.; Korneev, N.; Tenenbaum, J. B.; and Andreas, J. 2024. Learning Grounded Action Abstractions from Language. In *International Conference on Learning Representations (ICLR)*.

Wu, R.; Gao, R.; Poole, B.; Trevithick, A.; Zheng, C.; Barron, J. T.; and Holynski, A. 2025. CAT4D: Create Anything in 4D with Multi-View Video Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, T.; Yu, H.-X.; Wu, R.; Feng, B. Y.; Zheng, C.; Snavely, N.; Wu, J.; and Freeman, W. T. 2024. PhysDreamer: Physics-Based Interaction with 3D Objects via Video Generation. In *European Conference on Computer Vision (ECCV)*.

Zhang, Y.; Li, Z.; Zhou, M.; Wu, S.; and Wu, J. 2025a. The Scene Language: Representing Scenes with Programs, Words, and Embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Y.; Murtuza-Lanier, C.; Li, Z.; Du, Y.; and Wu, J. 2025b. Product of Experts for Visual Generation. *arXiv preprint:2506.08894*.

Zhao, Y.; Lin, C.-C.; Lin, K.; Yan, Z.; Li, L.; Yang, Z.; Wang, J.; Lee, G. H.; and Wang, L. 2025. GenXD: Generating Any 3D and 4D Scenes. In *International Conference on Learning Representations (ICLR)*.

Zhu, F.; and Simmons, R. 2024. Bootstrapping Cognitive Agents with a Large Language Model. In *AAAI Conference on Artificial Intelligence (AAAI)*.