# Large Language and Vision-Language Models as World Models: A Survey

**Azmine Toushik Wasi**[1,2†], **Sadia Tasneem Meem**[1,2],
**Riashat Islam**[3], and **Md Rizwan Parvez**[4]
[1]Computational Intelligence and Operations Laboratory
[2]Shahjalal University of Science and Technology
[3]Microsoft AI Frontiers [4]Qatar Computing Research Institute
[†]**Correspondence:** azmine32@student.sust.edu

## Abstract

Recent advances have reframed Artificial Intelligence (AI) around world models, internal simulators that predict environmental dynamics and support general intelligence. This review surveys Large Language Models (LLMs) and Vision–Language Models (VLMs) not merely as text or vision processors, but as predictive systems capable of modeling state, causality, and dynamics. We organize the literature with a dual taxonomy: a theoretical taxonomy covering pixel-space generative simulators, latent-space predictive architectures, neuro-symbolic and code-coupled models, autoregressive foundation models, and hybrid mixtures of world models; and an application taxonomy covering autonomous driving, embodied robotics, interactive simulation and gaming, and social and agentic simulation. Synthesizing recent works, we conclude that although autoregressive transformers exhibit emergent topological and causal reasoning, robust physical world modeling requires moving beyond next-token prediction toward hybrid architectures that integrate generative diffusion models with symbolic grounding.

## 1 Introduction

The dominant deep learning paradigm has been guided by the scaling hypothesis, which links increased model size and data to improved performance (Minaee et al., 2025; Núñez-Molina et al., 2025). However, as access to high-quality data saturates, scaling alone appears insufficient for open-ended, interactive settings that demand anticipation of environment dynamics rather than pattern completion. This limitation has renewed interest in *world models* (WMs), which learn internal predictive representations to support simulation, planning, and counterfactual reasoning (Zhang et al., 2025a).

Formally, a world model estimates environment dynamics through transition distributions of the form $P(s_{t+1} \mid s_t, a_t)$, enabling agents to evaluate futures before acting (Richens et al., 2025). While such models were historically confined to explicit simulators or low-dimensional reinforcement learning settings, recent advances in Transformers and diffusion models now allow these dynamics to be learned directly from high-dimensional text, image, and video streams (Feng et al., 2025c). This shift raises fundamental questions: *Do LLMs and VLMs merely encode surface correlations, or do they induce structured internal representations of state, causality, and dynamics? What architectural and representational choices distinguish genuine world modeling from reactive sequence prediction?*

In this review, we address these questions through a synthesis of recent work on LLM- and VLM-based world models, grounded in the background of § 2 and § D and the unified framework of § 3. We introduce a dual taxonomy: a *theoretical taxonomy* (§4) that categorizes models by representation and learning paradigm, and an *application taxonomy* (§5) that organizes deployments across autonomous driving, robotics, gaming, and social simulation. Throughout the survey, we treat a system as an LLM/VLM-based world model only when a language or vision-language foundation model supplies or materially shapes the predictive state and dynamics used for simulation, planning, or counterfactual reasoning. Systems in which an LLM/VLM serves only as a text interface or perception module do not meet this definition; non-LLM/VLM systems discussed below are identified as architectural contrast cases. This separation of representational foundations from deployment contexts enables systematic comparison across architectures while analyzing real-world instantiation and evaluation. Our synthesis indicates that although transformers exhibit emergent world modeling abilities, robust physical reasoning requires mechanisms beyond next-token prediction, highlighting challenges such as state persistence, physical consistency, and counterfactual reasoning.



*Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026*, pages 1582–1614
November 6-10, 2026 ©2026 Association for Computational Linguistics

## 2 Background

**Foundations.** World models originate in control theory and cognitive science, where intelligent behavior arises from internal simulation rather than direct stimulus–response mappings (Zhang et al., 2025a). In modern AI, this corresponds to learning environment dynamics that support planning and decision-making under uncertainty (Richens et al., 2025; Núñez-Molina et al., 2025). Unlike standard LLM agents that map observations directly to actions, world models introduce a latent simulation space for evaluating alternative futures before execution (Richens et al., 2025). This separation between prediction and action motivates world modeling as a foundation for general intelligence (Thomas, 2025), reframing intelligence as inference over dynamics and counterfactual evaluation.

**Emergence, Evidence, and the Shift Toward Multimodal World Modeling.** Debate persists over whether large pretrained models learn genuine internal representations. The *stochastic parrot* hypothesis argued that LLMs primarily reproduce surface statistics without grounded understanding (Bender et al., 2021). However, theoretical results show that neural networks can recover latent generative variables under suitable inductive biases and multi-task training (Zhang et al., 2025b). Empirical studies similarly demonstrate that transformers trained on structured tasks encode explicit environment states, with causal interventions confirming their influence on predictions (Li et al., 2024b; Yuan and Søgaard, 2025; Hazineh et al., 2023). Related analyses report unsupervised grounding and alignment with environmental structure (Yuan and Søgaard, 2025; Spies et al., 2025). Because text alone incompletely captures physical processes, recent work emphasizes multimodal systems, particularly Vision–Language Models, as stronger substrates for physical world modeling (Feng et al., 2025c). Benchmarks such as WM-ABench reveal persistent gaps in motion prediction, physical consistency, and disentangled reasoning, highlighting the need for explicit latent state dynamics beyond next-token prediction (Gao et al., 2025a; Thomas, 2025; Yue et al., 2025).

## 3 A Unified Mathematical Framework

A world model can be formalized as a predictive system for a partially observable controlled dynamical process:

$$s_{t+1} \sim P(s_{t+1} \mid s_t, a_t), \quad o_t \sim P(o_t \mid s_t), \quad (1)$$

where $s_t \in \mathcal{S}$ denotes latent state, $a_t \in \mathcal{A}$ an intervention, and $o_t \in \mathcal{O}$ the observation. The transition kernel encodes causal dynamics while the observation model maps latent variables to measurable signals. Because $s_t$ is unobserved, inference occurs under partial observability.

**Internal State Representation.** World models introduce an internal state $z_t \in \mathcal{Z}$ approximating sufficient statistics of past observations and actions:

$$z_t = f_\theta(o_{\leq t}, a_{<t}). \quad (2)$$

The estimator $f_\theta$ performs belief updating, producing representations that preserve predictive information while discarding irrelevant detail. Representations may be deterministic or stochastic, continuous or discrete, and structured or unstructured, including embeddings, latent variables, token sequences, graphs, predicates, or program states. This general interface unifies neural, probabilistic, and symbolic formulations.

**Dynamics Prediction.** A defining component is a transition operator in representation space:

$$z_{t+1} \sim g_\theta(z_t, a_t). \quad (3)$$

The dynamics model $g_\theta$ captures temporal evolution and encodes inductive biases such as continuity, causality, and controllability. It enables multi-step rollouts, counterfactual evaluation, and planning by simulating trajectories under hypothetical interventions.

**Decoding and Grounding.** Some models include decoders that map latent states to observations or structured variables:

$$\hat{o}_{t+1} = h_\theta(z_{t+1}) \quad \text{or} \quad \hat{s}_{t+1} = \psi_\theta(z_{t+1}). \quad (4)$$

The observation decoder $h_\theta$ supports reconstruction objectives common in generative modeling, while $\psi_\theta$ maps latent states to interpretable quantities or symbolic abstractions. Other approaches rely primarily on latent predictive consistency, treating decoding as auxiliary.

Together, inference, transition, and optional decoding define a modular framework encompassing diverse architectures. Pixel-space models optimize likelihood in observation space, latent-space models emphasize predictive sufficiency and efficiency, neuro-symbolic systems enforce structured consistency, and hybrid models integrate these principles within unified predictive architectures (§ C).

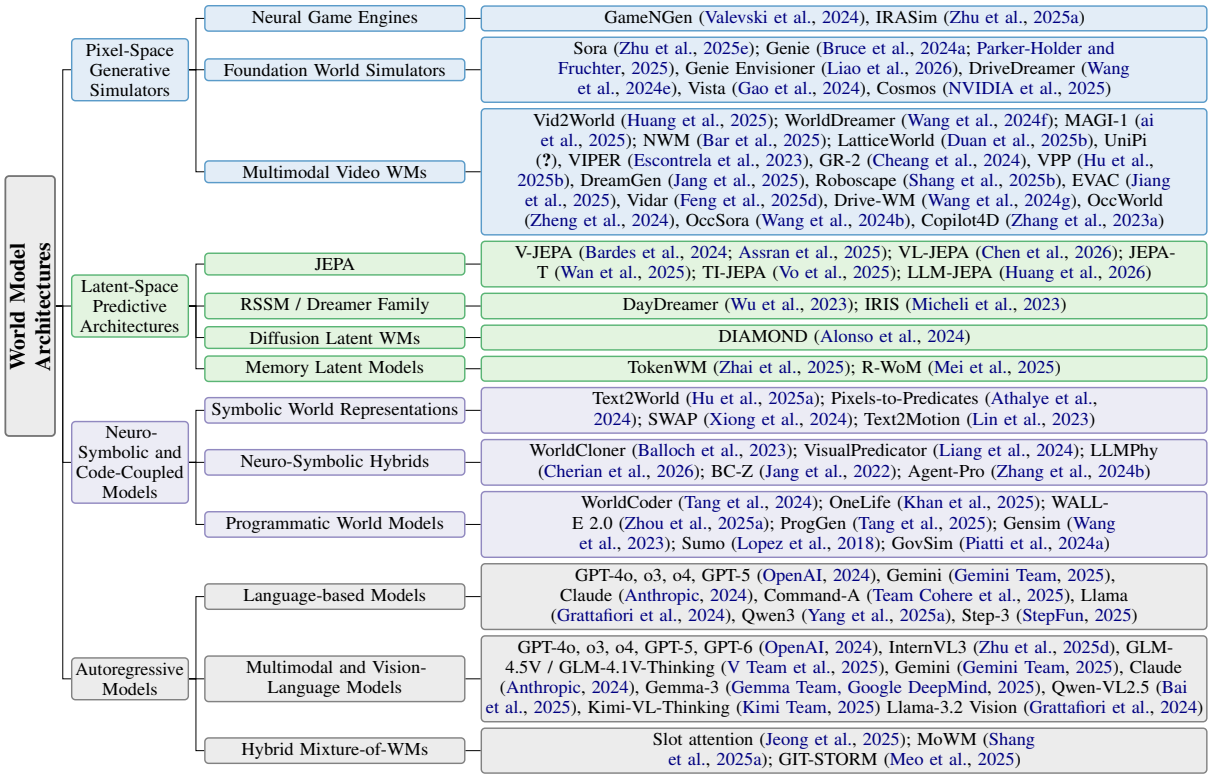



Figure 1: Theoretical taxonomy of world-model architectures organized by representation and prediction paradigm: (A) pixel-space generative simulators, (B) latent-space predictive architectures, (C) neuro-symbolic and code-coupled models, (D) autoregressive foundation models, and (E) hybrid mixtures of world models.

## 4 Theoretical Taxonomy

We classify world-model approaches by how they represent state and predict its evolution. The five categories are pixel-space generative simulators, latent-space predictive architectures, neuro-symbolic and code-coupled models, autoregressive foundation models, and hybrid mixtures of world models. The first four describe a dominant representational or predictive mechanism; hybrid mixtures couple two or more such mechanisms. Figure 1 summarizes the taxonomy. This classification supports comparison of state representation, action control, causal grounding, interpretability, and long-horizon prediction.

### 4.1 Pixel-Space Generative Simulators

Pixel-space world models equate environmental simulation with high-fidelity video generation: if a model can predict future frames conditioned on past observations and actions, it implicitly approximates underlying dynamics. This paradigm is enabled by diffusion models and large autoregressive video transformers that encode large-scale visual priors from internet-scale data (Huang et al., 2025; Feng et al., 2025c). Vid2World advances this direction by *causalizing* diffusion objectives, converting pretrained video generators into action-conditioned simulators (Huang et al., 2025). However, benchmarks such as WorldPrediction reveal a persistent gap between photorealistic rollout quality and long-horizon semantic reasoning, indicating that visual fidelity does not guarantee procedural competence (Chen et al., 2025a). Some systems in this subsection, including GameNGen, are not LLM/VLM-based; we retain them as architectural contrast cases for comparing foundation-model approaches with dedicated world models.

■ **Neural Game Engines.** GameNGen represents a full replacement of a symbolic game engine with a diffusion-based simulator trained on gameplay trajectories (Valevski et al., 2024). Through noise augmentation and iterative denoising, it mitigates autoregressive drift and sustains long rollouts. This demonstrates that structured symbolic state, including health, inventory, and agent logic, can be implicitly embedded in continuous generative weights, albeit with high computational overhead. *Neural game engines suggest that large-scale generative models can internalize discrete state transitions, but efficiency and controllability constraints*

*limit direct transfer to real-time embodied settings.*

■ **Foundation World Simulators.** Foundation World Models such as Sora and Genie aim to extend video simulation beyond domain-specific tasks. Sora's spacetime-patch representation improves visual coherence and object permanence, yet still exhibits failures in physical causality (Zhu et al., 2025e). Genie introduces unsupervised latent action discovery, transforming passive video corpora into interactive environments without labeled controls (Bruce et al., 2024a; Parker-Holder and Fruchter, 2025). Cosmos (NVIDIA et al., 2025) introduces an open-source, video-tokenized general-purpose simulator designed specifically to be fine-tuned into customized digital twins for physical AI setups and Newton benchmark further formalizes evaluation of long-term consistency and physical invariants in such systems (Campbell, 2025). *These models indicate that scaling and architectural restructuring improve temporal coherence, but robust physical reasoning requires explicit mechanisms for action grounding and invariant preservation.*

■ **Multimodal Video World Modeling.** Recent systems integrate text, vision, and action into unified simulators. WorldDreamer (Wang et al., 2024f), NWM (Bar et al., 2025), MAGI-1 (ai et al., 2025), and Open-Sora variants (Li et al., 2025e) emphasize scalable action-conditioned prediction, while PhyT2V enhances physical plausibility via external reasoning modules (Xue et al., 2024). Systems such as WorldGen (Wang et al., 2025a), Aether (Zhu et al., 2025c), and LatticeWorld (Duan et al., 2025b) extend simulation into structured 3D and interactive environments. Together, they illustrate the growing convergence of generative modeling with agent-centric design principles. *Multimodal simulators represent a transition from passive video prediction to action-aware world modeling, yet their effectiveness depends on how explicitly they encode persistent state, causal structure, and evaluable reasoning constraints.*

**Reflection.** Pixel-space models have evolved from pure video generation to action-conditioned, multimodal simulation, reflecting closer alignment with world-model principles. However, visual realism alone does not ensure causal reasoning, state persistence, or counterfactual capability. This taxonomy distinguishes methods by their degree of structured state modeling, invariant enforcement, and controllable long-horizon inference, enabling deeper cross-architecture comparison.

### 4.2 Latent-Space Predictive Architectures

Latent-space models prioritize compact, task-relevant representations over explicit pixel prediction, aiming to improve efficiency while preserving predictive sufficiency. By operating in representation space, they enable faster imagination, lower latency, and more scalable planning. JEPA operationalizes this principle by predicting future representations instead of raw observations (Jain, 2025; Huang et al., 2026). Video-JEPA extends this idea through large-scale spatiotemporal masking, encouraging object-centric and motion-aware abstractions rather than texture reconstruction (Huang et al., 2026). These abstractions substantially reduce computational cost, making latent predictive models particularly attractive for robotics and embodied control (Aguilera, 2025). Some latent world models discussed here, including RSSM-based systems and DIAMOND, are not LLM/VLM-based; we include them as contrast cases that clarify the benefits and limits of explicit latent dynamics.

■ **Joint Embedding Predictive Architectures (JEPA).** JEPA define latent predictive models that forecast future representations rather than reconstruct observations, emphasizing predictive sufficiency over pixel fidelity. Within the scope of LLM/VLMs, relevant variants include V-JEPA (Bardes et al., 2024) and its scalable extensions (Assran et al., 2025), as well as multimodal formulations such as VL-JEPA (Chen et al., 2026), JEPA-T (Wan et al., 2025), TI-JEPA (Vo et al., 2025), and LLM-JEPA (Huang et al., 2026), which integrate visual and language representations within predictive embedding spaces. These models demonstrate that world modeling can be achieved through joint embedding dynamics, enabling efficient long-horizon prediction without autoregressive token generation. However, their reliance on implicit decoding limits its controllability and interpretability, positioning JEPA as a complementary paradigm that benefits

from integration with language-based reasoning for general-purpose WMs. More details in §E.

■ **Dreamer and RSSM-Based Models.** Dreamer and related approaches employ Recurrent State-Space Models (RSSMs) to separate deterministic structure from stochastic uncertainty, enabling *planning in imagination* entirely within latent space (Wu et al., 2023). This decomposition supports highly sample-efficient policy optimization while maintaining uncertainty-aware dynamics. Transformer-based variants such as IRIS replace recurrent architectures with attention mechanisms, extending temporal horizons and enhancing discrete modeling capacity (Micheli et al., 2023). Extensions like DeformNet further demonstrate that RSSM-style models can represent complex deformable objects with effectively infinite-dimensional latent structure (Li et al., 2024a). *These developments show that structured latent dynamics, combined with scalable sequence modeling, can support long-horizon control without explicit pixel-level simulation.*

■ **Diffusion-Based Latent World Models.** DIAMOND revisits the compression assumption, arguing that overly aggressive abstraction may discard subtle but policy-relevant visual signals (Alonso et al., 2024). By performing latent imagination through diffusion rather than variational bottlenecks, it preserves perceptual detail while retaining model-based reinforcement learning capability. This approach illustrates a convergence between pixel-level realism and latent efficiency, suggesting that representation quality and generative fidelity can be jointly optimized. *Diffusion-based latent models bridge perceptual richness and compact dynamics, highlighting that effective world models require balanced compression rather than maximal abstraction.*

■ **Memory Latent Models.** TokenWM enhances recurrent latent architectures with tokenized state representations and attention-based memory, improving scalability and long-horizon reasoning (Zhai et al., 2025). R-WoM further integrates retrieval-augmented procedural knowledge, strengthening performance in extended planning and software-like environments (Mei et al., 2025). These memory mechanisms address limitations in fixed-capacity latent states by enabling structured information accumulation across time. *As temporal horizons expand, persistent memory and retrieval become essential complements to compact latent dynamics, enabling more robust planning and cross-context generalization.*

**Reflection.** In the context of LLM/VLMs as world models, latent-space approaches show that predictive state learning, rather than observation reconstruction, is central to scalable reasoning. Progress depends on structured uncertainty, temporal abstraction, and stable long-horizon imagination. This shift reframes compression as a mechanism for learning causally organized internal dynamics that enable planning and control.

### 4.3 Neuro-Symbolic and Code-Coupled WMs

Symbolic world models represent dynamics as predicates, graphs, or executable programs to improve interpretability and logical consistency. We distinguish symbolic, neuro-symbolic, programmatic, and hybrid systems by the form of their state representation and by how their transition is computed (Table 5). Text2World demonstrates LLM-based synthesis of PDDL-style domains, though consistency often requires agentic refinement (Hu et al., 2025a), while Pixels-to-Predicates enables zero-shot planning via symbolic state extraction (Athalye et al., 2024) and SWAP enforces reliability through entailment-based constraints (Xiong et al., 2024). Neuro-symbolic systems further integrate symbolic transitions with neural perception, as in WorldCloner (Balloch et al., 2023) and VisualPredicator (Liang et al., 2024), while analysis-by-synthesis approaches such as LLMPhy combine LLM priors with differentiable physics (Cherian et al., 2026). Programmatic models including OneLife (Khan et al., 2025), WALL-E 2.0 (Zhou et al., 2025a), React (Yao et al., 2023), WorldCoder (Tang et al., 2024), GIF-MCTS (Dainese et al., 2024), and ProgGen (Tang et al., 2025) encode dynamics as executable code, enabling transparent counterfactual simulation at the cost of increased synthesis complexity.

**Reflection.** These results indicate that increasing structural explicitness improves causal controllability, interpretability, and counterfactual reasoning, but shifts complexity toward representation design and program synthesis. More broadly, they suggest that scalable world modeling will require hybrid architectures that balance symbolic rigor with neural flexibility.

### 4.4 Autoregressive Models

Autoregressive architectures underpin modern LLMs, where next-token prediction induces internal representations of structure, causality, and temporal dynamics. Although developed for language generation, evidence suggests sufficiently large models can encode latent state abstractions resembling simplified world models.

■ **Language-based Models.** Representative text models include GPT-4o, o3, o4, and GPT-5 (OpenAI, 2025, 2024), Gemini (Gemini Team, 2025), Claude (Anthropic, 2024), Command-A (Team Cohere et al., 2025), and open-weight Llama models (Grattafiori et al., 2024). Additional systems include Qwen3 (Yang et al., 2025a) and mixture-of-experts architectures such as Step-3 (StepFun, 2025), which exhibit strong generative reasoning and structured prediction abilities.

■ **Multimodal and Vision-Language Models.** These models form a spectrum from general-purpose systems trained across text, image, audio, or video to systems specialized for visual input with language supervision. Their placement on this spectrum is more informative than a strict boundary: models such as InternVL3 and Qwen-VL2.5 combine visual specialization with broader multimodal capabilities. Representative systems include GPT-4o, GPT-5, GPT-6, Gemini, Claude, InternVL3, GLM-4.5V, Gemma-3, Qwen-VL2.5, Kimi-VL-Thinking, and Llama-3.2 Vision (OpenAI, 2025, 2024; Gemini Team, 2025; Anthropic, 2024; Zhu et al., 2025d; V Team et al., 2025; Gemma Team, Google DeepMind, 2025; Bai et al., 2025; Kimi Team, 2025; Grattafiori et al., 2024). For this survey, their world-modeling relevance depends on whether they track state and predict outcomes across observations, rather than on modality labels alone.

### 4.5 Hybrid Mixtures of World Models

Hybrid mixtures combine distinct state representations or transition mechanisms rather than relying on one representational substrate. Object-centric models use slot attention to decompose scenes into structured entities (Jeong et al., 2025), while MoWM combines pixel and latent dynamics (Shang et al., 2025a) and GIT-STORM couples discrete and continuous state transitions (Meo et al., 2025). In the unified framework, these systems factorize the state as multiple components and exchange information across them through the coupling function $\phi$. This design can preserve perceptual detail while supporting compact prediction or structured reasoning, but it adds coordination and training complexity. The key comparison is therefore whether cross-representation coupling improves controllability and long-horizon consistency enough to justify that cost.

**Reflection.** Together, these autoregressive, multimodal, and vision–language architectures provide an empirical basis for assessing whether next-token prediction can approximate key properties of world models, including latent state tracking, multimodal grounding, and predictive reasoning.

**Reflection on Theoretical Approaches.** Across paradigms, a clear pattern emerges: realism alone does not guarantee understanding, and abstraction alone sacrifices grounding. The field is converging toward hybrid architectures that combine generative perception, latent prediction, symbolic structure, and retrieval-based grounding. World models are no longer monolithic simulators, but modular cognitive systems, each component optimized for a different aspect of reasoning about the future.

## 5 Application Taxonomy

World models serve different purposes across application domains, each with its own operational constraints. We organize the applications discussed in the main text into four domains: autonomous driving, embodied robotics, interactive simulation and gaming, and social and agentic simulation (Figure 2). The taxonomy separates deployment setting from model architecture, allowing a system to have both an architectural category and an application category.

### 5.1 Autonomous Driving

Autonomous driving exemplifies a high-stakes setting where safety, geometric fidelity, and action-conditioned prediction must align. World models in this domain increasingly transition from auxiliary perception modules to core components of planning, validation, and closed-loop simulation.

■ **Generative Video and Scenario Simulation.** GAIA-1 (Hu et al., 2023) introduces large-scale multimodal transformers for generating future driving scenarios from video, text, and actions, enabling rare-event training. DriveDreamer (Zhao et al., 2024) complements this direction using

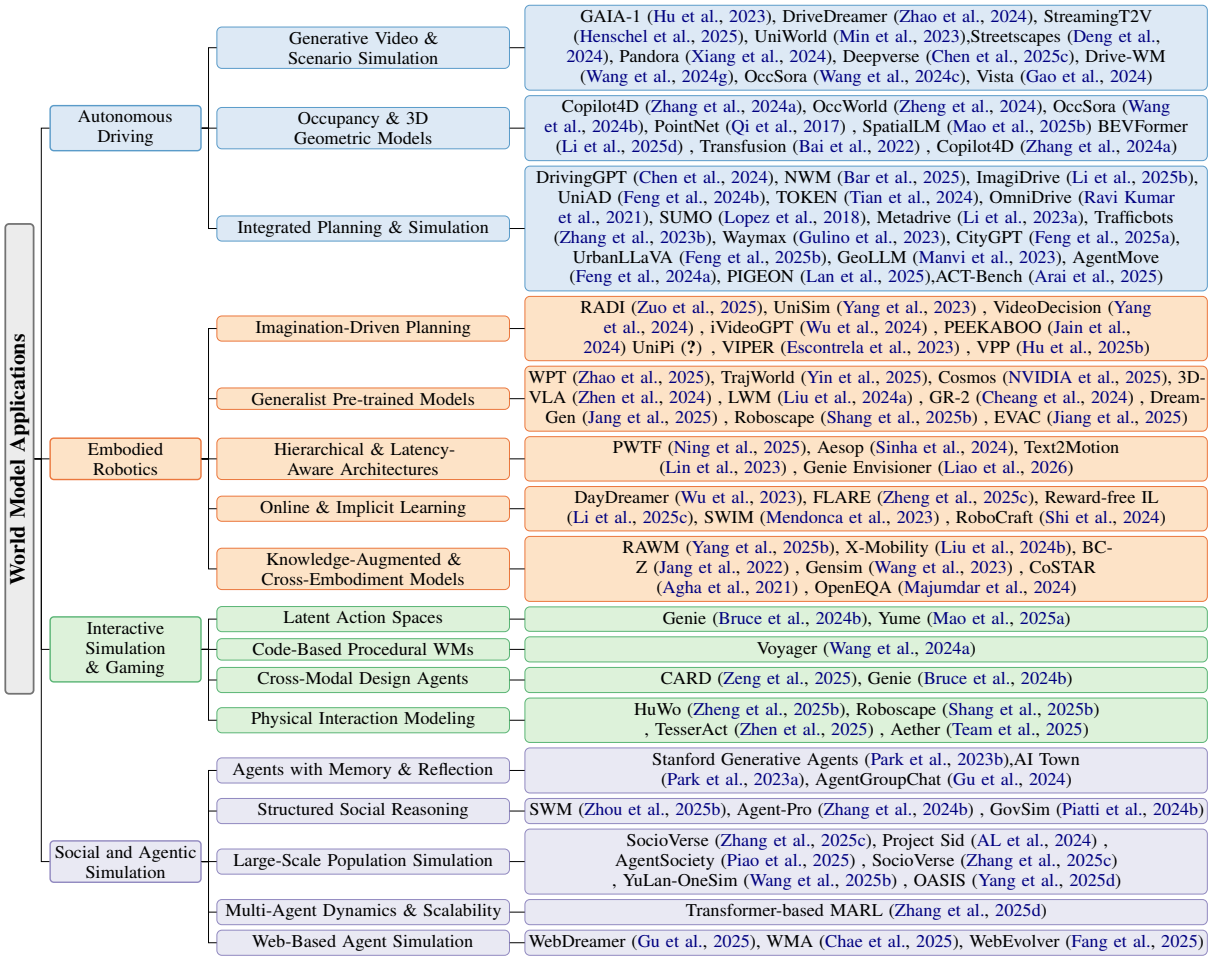



Figure 2: Application taxonomy of world-model applications across four areas: (A) autonomous driving, (B) embodied robotics, (C) interactive simulation and gaming, and (D) social and agentic simulation.

diffusion-based prediction to model structured traffic dynamics.

■ **Occupancy and 3D Geometric Models.** To improve physical plausibility, occupancy-based models such as Copilot4D (Zhang et al., 2024a), OccWorld (Zheng et al., 2024), and OccSora (Wang et al., 2024b) predict future 3D voxel states, directly supporting motion planning under geometric constraints (Ding et al., 2025).

■ **Integrated Planning and Simulation.** DrivingGPT (Chen et al., 2024) and ImagiDrive (Li et al., 2025b) unify trajectory generation and scenario simulation within autoregressive frameworks, emphasizing iterative, action-conditioned refinement, while NWM (Bar et al., 2025) enable dynamic constraint evaluation by simulating and ranking trajectories. In terms of evaluation, ACT-Bench (Arai et al., 2025) establishes rigorous evaluation protocols focusing on how faithfully simulators follow motion instructions, underscoring the shift from visual realism to actionable accuracy.

**Reflection.** Across approaches, the field is shifting from visually plausible rollout generation toward geometry-aware, action-conditioned simulation. The central analytical progression is from pixel fidelity to structured state prediction, where robustness is determined less by realism and more by consistency under intervention and long-horizon planning.

### 5.2 Embodied Robotics

Robotics imposes strict latency, safety, and generalization constraints, requiring world models that support real-time decision-making in physically grounded environments.

■ **Imagination-Driven Planning.** RADI (Zuo et al., 2025) uses LLM-based simulation of action outcomes with self-reflection to prevent costly failures in safety-critical tasks.

■ **Generalist Pre-trained Models.** WPT (Zhao et al., 2025) and TrajWorld (Yin et al., 2025) scale across embodiments via large datasets and flexible

architectures, enabling cross-task transfer beyond fixed pipelines. Cosmos (NVIDIA et al., 2025) further democratizes cross-task transfer by providing open-weight digital twins that can be efficiently fine-tuned to solve complex physical problems across diverse downstream environments.

■ **Hierarchical and Latency-Aware Designs.** PWTF (Ning et al., 2025) and Aesop (Sinha et al., 2024) decouple fast reactive control from slower deliberative reasoning to satisfy real-time constraints.

■ **Online and Implicit Learning.** DayDreamer (Wu et al., 2023), FLARE (Zheng et al., 2025c), and reward-free imitation methods (Li et al., 2025c) demonstrate latent-space adaptation and predictive alignment without heavy supervision.

■ **Knowledge-Augmented and Cross-Embodiment Models.** RAWM (Yang et al., 2025b) and X-Mobility (Liu et al., 2024b) incorporate retrieval and zero-shot transfer, advancing toward universal robotic representations.

**Reflection.** Robotics highlights that world models must operate under strict temporal and safety constraints. The dominant trend is hierarchical integration, where compact latent prediction, memory, and retrieval are combined with fast control loops, revealing that scalability in robotics depends on latency-aware structure rather than model size alone.

### 5.3 Interactive Simulation and Gaming

Interactive simulation positions world models as programmable engines for perception, physics, and user-driven generation. Genie (Bruce et al., 2024a) and Yume (Mao et al., 2025a) learn latent action spaces from video to enable controllable environment generation, while code-based approaches (Voyager (Wang et al., 2024a)) show that executable programs can outperform pixel prediction in structured settings. Cross-modal agents like CARD integrate text, geometry, and domain knowledge for spatial design (Zeng et al., 2025), and physical interaction models including HuWo (Zheng et al., 2025b) and Roboscape (Shang et al., 2025b) explicitly model dynamics in environments.

**Reflection.** WMs increasingly function as controllable generative engines rather than passive predictors, shifting to unified simulations where latent actions, code, and multimodal inputs jointly govern environment evolution.

### 5.4 Social and Agentic Simulation

Social world models focus on multi-agent interaction, memory, and emergent collective behavior. Generative Agents integrate memory, reflection, and planning to produce coherent social dynamics (Park et al., 2023b), while structured approaches such as SWM incorporate relational and theory-of-mind representations to improve interaction modeling (Zhou et al., 2025b). At scale, SocioVerse enables population-level simulation with millions of agents for societal forecasting (Zhang et al., 2025c), and transformer-based decentralized architectures address scalability in multi-agent reinforcement learning through structured aggregation of local dynamics (Zhang et al., 2025d). Extensions to digital environments, including WebDreamer (Gu et al., 2025), WMA (Chae et al., 2025), and WebEvolver (Fang et al., 2025), demonstrate world modeling in web-based ecosystems.

**Reflection.** Social simulation works show that world models increasingly encode not only physical dynamics but also beliefs, memory, and coordination across agents. More broadly, they indicate a convergence between world modeling and multi-agent cognitive systems, where scalability depends on structured interaction and persistent agent-level state.

Despite domain differences, world models converge toward multi-modality, action conditioning, memory, and structured state for closed-loop reasoning. Progress increasingly depends on persistent state, constraint enforcement, and intervention simulation rather than perceptual realism. These trends indicate that next-generation LLMs and VLMs as World Models will center on explicit simulation loops, scalable representations, and evaluation based on causal fidelity.

## 6 Discussion

### 6.1 Current Challenges

World models face several challenges that limit reliable deployment in embodied systems. **Inference latency** remains a constraint, as large LLMs and diffusion simulators often exceed real-time requirements; robotic control may require sub-100 ms responses while cloud inference can exceed 500 ms (Gao et al., 2025b). **Hallucination and counterfactual reasoning** present another difficulty: planning requires generating hypothetical states, yet these must remain physically and envi-

ronmentally consistent (Yang et al., 2025c). **Data scarcity** further limits supervised scaling because embodied datasets are costly; recent work therefore learns latent dynamics from large unlabeled video corpora, such as Genie-style approaches (Bruce et al., 2024a), though aligning latent actions with real-world causality remains unresolved. **Security and safety** risks also arise when world models control physical agents, as trojan and backdoor attacks can manipulate robotic policies, motivating security-aware training and deployment (Wang et al., 2024d). Lastly, **Evaluation** remains underdeveloped: standard pixel-level metrics fail to capture causal, temporal, and action-grounded reasoning, motivating semantics-aware evaluators and relational benchmarks. More details in Appendix F.

### 6.2 Reflections and Future Outlook

Progress requires joint advances across architecture, learning efficiency, robustness, and evaluation. **Hybrid Architectures** that integrate symbolic constraints with neural generative models can enhance physical consistency while preserving counterfactual flexibility. **Efficient Learning** through large-scale self-supervised training on unlabeled video data is essential to address data bottlenecks and improve generalization. **Security-Aware Design** must embed adversarial robustness and backdoor resistance into training and deployment to ensure trustworthy operation. **Multimodal Evaluation** based on relational, temporal, and counterfactual benchmarks is necessary to measure true world-model fidelity and guide systematic improvement. Together, these directions position world models as scalable, secure, and rigorously evaluated systems for embodied intelligence and complex interactive environments. More details in Appendix G.

## 7 Conclusion

World models have evolved from theoretical constructs to systems enabling causal reasoning across domains. Pixel-space, latent predictive, neuro-symbolic and code-coupled, autoregressive, and hybrid mixture approaches offer different trade-offs in state representation, controllability, and predictive fidelity. Future progress lies in unifying these paradigms into foundation world models for coherent reasoning about environments. While autoregressive transformers exhibit emergent causal structure, robust modeling requires moving beyond next-token prediction toward hybrid architectures integrating generative methods with symbolic grounding for improved interpretability and effectiveness.

## Limitations

While this review comprehensively covers recent advances in large language models (LLMs) and vision-language models (VLMs) as world models, the rapid pace of AI research means some of the latest developments may not be fully captured. However, by distilling broad trends and identifying critical challenges, this work provides a structured conceptual framework to unify diverse strands of research. The comparative, cross-domain approach highlights shared opportunities and domain-specific constraints, offering a foundation for future integrative advances.

As a review paper, this work focuses on synthesis and conceptual organization rather than new experiments, empirical validation, benchmarking studies, or technical implementations. Consequently, experimental comparisons, quantitative evaluations, and system-level performance analyses are outside the scope of this article. Our goal is to provide theoretical clarity and structured taxonomy rather than introduce new datasets, algorithms, or benchmark results.

## References

Ali Agha, Kyohei Otsu, Benjamin Morrell, David D. Fan, Rohan Thakker, Angel Santamaria-Navarro, Sung-Kyun Kim, Amanda Bouman, Xianmei Lei, Jeffrey Edlund, Muhammad Fadhil Ginting, Kamak Ebadi, Matthew Anderson, Torkom Pailevanian, Edward Terry, Michael Wolf, Andrea Tagliabue, Tiago Stegun Vaquero, Matteo Palieri, Scott Tepsuporn, Yun Chang, Arash Kalantari, Fernando Chavez, Brett Lopez, Nobuhiro Funabiki, Gregory Miles, Thomas Touma, Alessandro Buscicchio, Jesus Tordesillas, Nikhilesh Alatur, Jeremy Nash, William Walsh, Sunggoo Jung, Hanseob Lee, Christoforos Kanellakis, John Mayo, Scott Harper, Marcel Kaufmann, Anushri Dixit, Gustavo Correa, Carlyn Lee, Jay Gao, Gene Merewether, Jairo Maldonado-Contreras, Gautam Salhotra, Maira Saboia Da Silva, Benjamin Ramtoula, Yuki Kubo, Seyed Fakoorian, Alexander Hatteland, Taeyeon Kim, Tara Bartlett, Alex Stephens, Leon Kim, Chuck Bergh, Eric Heiden, Thomas Lew, Abhishek Cauligi, Tristan Heywood, Andrew Kramer, Henry A. Leopold, Chris Choi, Shreyansh Daftry, Olivier Toupet, Inhwan Wee, Abhishek Thakur, Micah Feras, Giovanni Beltrame, George Nikolakopoulos, David Shim, Luca Carlone, and Joel Burdick. 2021. Nebula: Quest for robotic autonomy in challenging environments; team costar at the darpa subterranean challenge.

Frank Morales Aguilera. 2025. The crisis of scaling and the architectural turn to world models: The jepa mandate. Accessed: 2025-12-20.

Sand. ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yuqiao Li. 2025. Magi-1: Autoregressive video generation at scale. *Preprint*, arXiv:2505.13211.

Altera. AL, Andrew Ahn, Nic Becker, Stephanie Carroll, Nico Christie, Manuel Cortes, Arda Demirci, Melissa Du, Frankie Li, Shuying Luo, Peter Y Wang, Mathew Willows, Feitong Yang, and Guangyu Robert Yang. 2024. Project sid: Many-agent simulations toward ai civilization. *Preprint*, arXiv:2411.00114.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J. Storkey, Tim Pearce, and François Fleuret. 2024. Diffusion for world modeling: Visual details matter in atari. *ArXiv*, abs/2405.12399.

Anthropic. 2024. Model Card Addendum: Claude 3.5 Haiku and Upgraded Claude 3.5 Sonnet. Accessed: 2025-09-11.

Hidehisa Arai, Keishi Ishihara, Tsubasa Takahashi, and Yu Yamaguchi. 2025. ACT-bench: Towards action controllable world models for autonomous driving. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. *Preprint*, arXiv:2301.08243.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. 2025. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *Preprint*, arXiv:2506.09985.

Ashay Athalye, Nishanth Kumar, Tom Silver, Yichao Liang, Jiuguang Wang, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. 2024. From pixels to predicates: Learning symbolic world models via pretrained vision-language models.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. 2022. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. *Preprint*, arXiv:2203.11496.

Randall Balestriero and Yann LeCun. 2025. Lejepa: Provable and scalable self-supervised learning without the heuristics. *Preprint*, arXiv:2511.08544.

Jonathan C. Balloch, Zhiyu Lin, Robert William Wright, Xiangyu Peng, Mustafa Hussain, Aarun Srinivas, Julia Kim, and Mark O. Riedl. 2023. Neuro-symbolic world models for adapting to open world novelty. *ArXiv*, abs/2301.06294.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. 2025. Navigation world models. *Preprint*, arXiv:2412.03572.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. 2024. V-JEPA: Latent video prediction for visual representation learning.

Adrien Bardes, Jean Ponce, and Yann LeCun. 2023. Mc-jepa: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *Preprint*, arXiv:2307.12698.

Jai Bardhan, Radhikesh Agrawal, Abhiram Tilak, Cyrin Neeraj, and Subhadip Mitra. 2025. Hep-jepa: A foundation model for collider physics using joint embedding predictive architecture. *Preprint*, arXiv:2502.03933.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

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. 2024a. Genie: Generative interactive environments. *Preprint*, arXiv:2402.15391.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris

Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. 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. 2024b. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*.

Spruce Campbell. 2025. Newton - a small benchmark for interactive foundation world models. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Hyungjoo Chae, Namyoung Kim, Kai Tzu iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. 2025. Web agents with world models: Learning and leveraging environment dynamics in web navigation. *Preprint*, arXiv:2410.13232.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. 2024. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *Preprint*, arXiv:2410.06158.

Delong Chen, Willy Chung, Yejin Bang, Ziwei Ji, and Pascale Fung. 2025a. Worldprediction: A benchmark for high-level world modeling and long-horizon procedural planning. In *ICML 2025 Workshop on Assessing World Models*.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Yejin Bang, Allen Bolourchi, Yann LeCun, and Pascale Fung. 2026. Vl-jepa: Joint embedding predictive architecture for vision-language. *Preprint*, arXiv:2512.10942.

Dengsheng Chen, Jie Hu, Xiaoming Wei, and Enhua Wu. 2025b. Denoising with a joint-embedding predictive architecture. *Preprint*, arXiv:2410.03755.

Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, and Tong He. 2025c. Deepverse: 4d autoregressive video generation as a world model. *Preprint*, arXiv:2506.01103.

Yuntao Chen, Yu-Quan Wang, and Zhao-Xiang Zhang. 2024. Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 26890–26900.

Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. 2026. Llmphy: Parameter-identifiable physical reasoning combining large language models and physics engines. *Preprint*, arXiv:2411.08027.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2019. Textworld: A learning environment for text-based games.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. 2025. Emu3.5: Native multimodal models are world learners. *Preprint*, arXiv:2510.26583.

Nicola Dainese, Matteo Merler, Minttu Alakuijala, and Pekka Marttinen. 2024. Generating code world models with large language models guided by monte carlo tree search. *Preprint*, arXiv:2405.15383.

Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. 2024. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, New York, NY, USA. Association for Computing Machinery.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. 2025. Emerging properties in unified multimodal pretraining. *Preprint*, arXiv:2505.14683.

Jingtao Ding, Yunke Zhang, Yu Shang, Jie Feng, Yuheng Zhang, Zefang Zong, Yuan Yuan, Hongyuan Su, Nian Li, Jinghua Piao, Yucheng Deng, Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. Understanding world or predicting future? a comprehensive survey of world models. *Preprint*, arXiv:2411.14499.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. 2025a. Worldscore: A unified evaluation benchmark for world generation. *Preprint*, arXiv:2504.00983.

Yinglin Duan, Zhengxia Zou, Tongwei Gu, Wei Jia, Zhan Zhao, Luyi Xu, Xinzhu Liu, Yenan Lin, Hao Jiang, Kang Chen, and Shuang Qiu. 2025b. Latticeworld: A multimodal large language model-empowered framework for interactive complex world generation. *Preprint*, arXiv:2509.05263.

Amartya Dutta, Kazi Sajeed Mehrab, Medha Sawhney, Abhilash Neog, Mridul Khurana, Sepideh Fatemi, Aanish Pradhan, M. Maruf, Ismini Lourentzou, Arka Daw, and Anuj Karpatne. 2025. Open world scene graph generation using vision language models. In *ICML 2025 Workshop on Assessing World Models*.

Sofiane Ennadir, Siavash Golkar, and Leopoldo Sarra. 2025. Joint embeddings go temporal. *Preprint*, arXiv:2509.25449.

Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee,

Danijar Hafner, and Pieter Abbeel. 2023. Video prediction models as rewards for reinforcement learning. *Preprint*, arXiv:2305.14343.

Tianqing Fang, Hongming Zhang, Zhisong Zhang, Kaixin Ma, Wenhao Yu, Haitao Mi, and Dong Yu. 2025. Webevolver: Enhancing web agent self-improvement with coevolving world model. *Preprint*, arXiv:2504.21024.

Jie Feng, Yuwei Du, Jie Zhao, and Yong Li. 2024a. Agentmove: A large language model based agentic framework for zero-shot next location prediction.

Jie Feng, Tianhui Liu, Yuwei Du, Siqi Guo, Yuming Lin, and Yong Li. 2025a. Citygpt: Empowering urban spatial cognition of large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 591–602, New York, NY, USA. Association for Computing Machinery.

Jie Feng, Shengyuan Wang, Tianhui Liu, Yanxin Xi, and Yong Li. 2025b. Urbanllava: A multi-modal large language model for urban intelligence with spatial reasoning and understanding. *Preprint*, arXiv:2506.23219.

Kaituo Feng, Changsheng Li, Dongchun Ren, Ye Yuan, and Guoren Wang. 2024b. On the road to portability: Compressing end-to-end motion planner for autonomous driving. *Preprint*, arXiv:2403.01238.

Tuo Feng, Wenguan Wang, and Yi Yang. 2025c. A survey of world models for autonomous driving. *Preprint*, arXiv:2501.11260.

Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. 2025d. Vidar: Embodied video diffusion model for generalist manipulation. *Preprint*, arXiv:2507.12898.

Qiyue Gao, Xinyu Pi, Kevin Liu, Junrong Chen, Ruolan Yang, Xinqi Huang, Xinyu Fang, Lu Sun, Gautham Kishore, Bo Ai, Stone Tao, Mengyang Liu, Jiaxi Yang, Chao-Jung Lai, Chuanyang Jin, Jiannan Xiang, Benhao Huang, David Danks, Hao Su, Tianmin Shu, Ziqiao Ma, Lianhui Qin, and Zhiting Hu. 2025a. Do vision-language models have internal world models? towards an atomic evaluation. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. 2024. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yifang Gao, Wei Luo, Xuye Wang, Shunshun Zhang, and Patrick Goh. 2025b. Lamars: Large language model-based anticipation mechanism acceleration in real-time robotic systems. *IEEE Access*, 13:3864–3880.

Gemini Team. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team, Google DeepMind. 2025. Gemma 3 technical report. Accessed: 2025-09-11.

Hafez Ghaemi, Eilif Muller, and Shahab Bakhtiari. 2026. seq-jepa: Autoregressive predictive learning of invariant-equivariant world models. *Preprint*, arXiv:2505.03176.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan

Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yu Gu, Kai Zhang, Yuting Ning, Boyuan Zheng, Boyu Gou, Tianci Xue, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, Huan Sun, and Yu Su. 2025. Is your llm secretly a world model of the internet? model-based planning for web agents. *Preprint*, arXiv:2411.06559.

Zhouhong Gu, Xiaoxuan Zhu, Haoran Guo, Lin Zhang, Yin Cai, Hao Shen, Jiangjie Chen, Zheyu Ye, Yifei Dai, Yan Gao, Yao Hu, Hongwei Feng, and Yanghua Xiao. 2024. Agentgroupchat: An interactive group chat simulacra for better eliciting emergent behavior. *Preprint*, arXiv:2403.13433.

Pierre Guetschel, Thomas Moreau, and Michael Tangermann. 2024. S-jepa: towards seamless cross-dataset transfer through dynamic spatial attention. In *9th Graz Brain-Computer Interface Conference*, Graz, Austria.

Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, John D. Co-Reyes, Rishabh Agarwal, Rebecca Roelofs, Yao Lu, Nico Montali, Paul Mougin, Zoey Yang, Brandyn White, Aleksandra Faust, Rowan McAllister, Dragomir Anguelov, and Benjamin Sapp. 2023. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Preprint*, arXiv:2310.08710.

Shangmin Guo, Omar Darwiche Domingues, Raphaël Avalos, Aaron Courville, and Florian Strub. 2025. World modelling improves language model agents.

Dean S. Hazineh, Zechen Zhang, and Jeffery Chiu. 2023. Linear latent world models in simple transformers: A case study on othello-gpt. *Preprint*, arXiv:2310.07582.

Xiangteng He, Shunsuke Sakai, Shivam Chandhok, Sara Beery, Kun Yuan, Nicolas Padoy, Tatsuhito Hasegawa, and Leonid Sigal. 2026. Dseq-jepa: Discriminative sequential joint-embedding predictive architecture. *Preprint*, arXiv:2511.17354.

Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, Kevin Lin, William Yang Wang, Lijuan Wang, and Xin Eric Wang. 2024. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *Preprint*, arXiv:2406.08407.

Mariya Hendriksen, Tabish Rashid, David Bignell, Raluca Georgescu, Abdelhak Lemkhenter, Katja Hofmann, Sam Devlin, and Sarah Parisot. 2025. Adapting vision-language models for evaluating world models. In *ICML 2025 Workshop on Assessing World Models*.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2025. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *Preprint*, arXiv:2403.14773.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. 2023. Gaia-1: A generative world model for autonomous driving. *Preprint*, arXiv:2309.17080.

Mengkang Hu, Tianxing Chen, Yude Zou, Yuheng Lei, Qiguang Chen, Ming Li, Yao Mu, Hongyuan Zhang, Wenqi Shao, and Ping Luo. 2025a. Text2world: Benchmarking large language models for symbolic world model generation. *ArXiv*, abs/2502.13092.

Naiwen Hu, Haozhe Cheng, Yifan Xie, Shiqi Li, and Jihua Zhu. 2024. 3d-jepa: A joint embedding predictive architecture for 3d self-supervised representation learning. *Preprint*, arXiv:2409.15803.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. 2025b. Video prediction policy: A generalist robot policy with predictive visual representations. *Preprint*, arXiv:2412.14803.

Hai Huang, Yann LeCun, and Randall Balestriero. 2026. LLM-JEPA: Large language models meet joint embedding predictive architectures. In *The Fourteenth International Conference on Learning Representations*.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. 2025. Vid2world: Crafting video diffusion models to interactive world models. *ArXiv*, abs/2505.14357.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyürek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. 2025. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *Preprint*, arXiv:2405.09605.

Anil Jain. 2025. Jepa: Lecun's path towards more human-like ai. Medium; accessed 20 December 2025.

Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. 2024. Peekaboo: Interactive video generation via masked-diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8079–8088. IEEE.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. 2022. Bc-z: Zero-shot task generalization with robotic imitation learning. In *conference on Robot Learning*, pages 991–1002. PMLR.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loïc Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. 2025. Dreamgen: Unlocking generalization in robot learning through video world models. In *9th Annual Conference on Robot Learning*.

Youngjoon Jeong, Junha Chun, Soonwoo Cha, and Taesup Kim. 2025. Object-centric world model for language-guided manipulation. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Yuxin Jiang, Shengcong Chen, Siyuan Huang, Liliang Chen, Pengfei Zhou, Yue Liao, Xindong He, Chiming Liu, Hongsheng Li, Maoqing Yao, and Guanghui Ren. 2025. Enerverse-ac: Envisioning embodied environments with action condition.

András Kalapos and Bálint Gyires-Tóth. 2024. Cnn-jepa: Self-supervised pretraining convolutional neural networks using joint embedding predictive architecture. *2024 International Conference on Machine Learning and Applications (ICMLA)*, pages 1111–1114.

Zaid Khan, Archiki Prasad, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. 2025. One life to learn: Inferring symbolic world models for stochastic environments from unguided exploration. *Preprint*, arXiv:2510.12088.

Sehun Kim. 2024. Learning general representation of 12-lead electrocardiogram with a joint-embedding predictive architecture. *Preprint*, arXiv:2410.08559.

Kimi Team. 2025. Kimi-vl technical report. *Preprint*, arXiv:2504.07491.

Xiaochong Lan, Jie Feng, Yizhou Sun, Chen Gao, Jiahuan Lei, Xinleishi Xinleishi, Hengliang Luo, and Yong Li. 2025. Open-set living need prediction with large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5454–5472.

Chenchang Li, Zihao Ai, Tong Wu, Xiaosa Li, Wenbo Ding, and Huazhe Xu. 2024a. Deformnet: Latent space modeling and dynamics prediction for deformable object manipulation. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14770–14776.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. 2025a. Imagine while reasoning in space: Multimodal visualization-of-thought. In *Forty-second International Conference on Machine Learning*.

Jingyu Li, Bozhou Zhang, Xin Jin, Jiankang Deng, Xiatian Zhu, and Li Zhang. 2025b. Imagidrive: A unified imagination-and-planning framework for autonomous driving. *Preprint*, arXiv:2508.11428.

Kenneth Li. 2023. Do large language models learn world models or just surface statistics? *The Gradient*. Accessed: 2025-12-20.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024b. Emergent world representations: Exploring a sequence model trained on a synthetic task. *Preprint*, arXiv:2210.13382.

Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. 2023a. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3461–3475.

Shangzhe Li, Zhiao Huang, and Hao Su. 2025c. Reward-free world models for online imitation learning. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Wei-Jang Li, Wei Yang, Tianpeng Liu, Yuenan Hou, Yuxuan Li, Zhen Liu, Yongxiang Liu, and Li Liu. 2023b. Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Yixia Li, Hongru WANG, Jiahao Qiu, Zhenfei Yin, Dongdong Zhang, Cheng Qian, Zeping Li, Xiaoteng Ma, Guanhua Chen, and Heng Ji. 2026. From word to world: Can large language models be implicit text-based world models? In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. 2025d. Bevformer: Learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036.

Zongyue Li, Sikuan Yan, Yunpu Ma, Yusong Li, Xing Lyu, and Matthias Schubert. 2025e. BEYOND SINGLE-STEP: MULTI-FRAME ACTION-CONDITIONED VIDEO GENERATION FOR REINFORCE- MENT LEARNING ENVIRONMENTS. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Yichao Liang, Nishanth Kumar, Hao Tang, Adrian Weller, Joshua B. Tenenbaum, Tom Silver, João Henriques, and Kevin Ellis. 2024. Visualpredicator: Learning abstract world models with neuro-symbolic predicates for robot planning. *ArXiv*, abs/2410.23156.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Si Liu, Jianlan Luo, Liliang Chen, Shuicheng YAN, Maoqing Yao, and Guanghui Ren. 2026. Genie envisioner: A

unified world foundation platform for robotic manipulation. In *The Fourteenth International Conference on Learning Representations*.

Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. 2023. Text2motion: From natural language instructions to feasible plans. *Autonomous Robots*, 47(8):1345–1365.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024a. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*.

Wei Liu, Hui Zhao, Chenran Li, Joydeep Biswas, Billy Okal, Pulkit Goyal, Yan Chang, and Soha Pouya. 2024b. X-mobility: End-to-end generalizable navigation via world modeling. *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7569–7576.

Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wiessner. 2018. Microscopic traffic simulation using sumo. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2575–2582.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498.

Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2023. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. 2025a. Yume: An interactive world generation model. *Preprint*, arXiv:2507.17744.

Yongsen Mao, Junhao Zhong, Chuan Fang, Jia Zheng, Rui Tang, Hao Zhu, Ping Tan, and Zihan Zhou. 2025b. Spatiallm: Training large language models for structured indoor modeling. *Preprint*, arXiv:2506.07491.

Kai Mei, Jiang Guo, Shuaichen Chang, Mingwen Dong, Dongkyu Lee, Xing Niu, and Jiarong Jiang. 2025. R-wom: Retrieval-augmented world model for computer-use agents. *Preprint*, arXiv:2510.11892.

Russell Mendonca, Shikhar Bahl, and Deepak Pathak. 2023. Structured world models from human videos. *arXiv preprint arXiv:2308.10901*.

Cristian Meo, Mircea Tudor Lică, Zarif Ikram, Akihiro Nakano, Vedant Shah, Aniket Rajiv Didolkar, Dianbo Liu, Anirudh Goyal, and Justin Dauwels. 2025. Masked generative priors improve world models sequence modelling capabilities. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Vincent Micheli, Eloi Alonso, and François Fleuret. 2023. Transformers are sample-efficient world models. In *The Eleventh International Conference on Learning Representations*.

Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. 2023. Uniworld: Autonomous driving pre-training via world models. *arXiv preprint arXiv:2308.07234*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large language models: A survey. *Preprint*, arXiv:2402.06196.

Shentong Mo and Shengbang Tong. 2024. Connecting joint-embedding predictive architecture with contrastive self-supervised learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. 2026. Causal-jepa: Learning world models through object-level latent interventions. *Preprint*, arXiv:2602.11389.

Chuanruo Ning, Kuan Fang, and Wei-Chiu Ma. 2025. Prompting with the future: Open-world model predictive control with interactive digital twins. *Preprint*, arXiv:2506.13761.

NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. 2025. Cosmos world foundation model platform for physical ai. *Preprint*, arXiv:2501.03575.

Carlos Núñez-Molina, Vicenç Gómez, and Hector Geffner. 2025. From next token prediction to (strips) world models – preliminary results. *Preprint*, arXiv:2509.13389.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Openai o3 and o4-mini system card. System Cards. Accessed: 2025-09-11.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023a. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023b. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Jack Parker-Holder and Shlomi Fruchter. 2025. Genie 3: A new frontier for world models. Accessed: 2025-12-20.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *Preprint*, arXiv:2502.08691.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024a. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, 37:111715–111759.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024b. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Preprint*, arXiv:2404.16698.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Preprint*, arXiv:1612.00593.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, LEI BAI, Wanli Ouyang, and Ruimao Zhang. 2025. Worldsimbench: Towards video generation models as world simulators.

Yifu Qiu, Zheng Zhao, Waylon Li, Yftah Ziser, Anna Korhonen, Shay B. Cohen, and Edoardo M. Ponti. 2026a. Self-improving world modelling with latent actions. *Preprint*, arXiv:2602.06130.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Shay B. Cohen, and Edoardo M. Ponti. 2026b. Can vlms predict future states? bootstrapping world models from inverse dynamics. *Preprint*, arXiv:2506.06006.

Yuping Qiu, Rui Zhu, and Ying cong Chen. 2025. Improving joint embedding predictive architecture with diffusion noise. *Preprint*, arXiv:2507.15216.

Varun Ravi Kumar, Senthil Yogamani, Hazem Rashed, Ganesh Sitsu, Christian Witt, Isabelle Leang, Stefan Milz, and Patrick Mäder. 2021. Omnidet: Surround view cameras based multi-task visual perception network for autonomous driving. *IEEE Robotics and Automation Letters*, 6(2):2830–2837.

Jonathan Richens, Tom Everitt, and David Abel. 2025. General agents need world models. In *Forty-second International Conference on Machine Learning*.

Alain Riou, Stefan Lattner, Gaëtan Hadjeres, Michael Anslow, and Geoffroy Peeters. 2024. Stem-jepa: A joint-embedding predictive architecture for musical stem compatibility estimation. *Preprint*, arXiv:2408.02514.

Ayumu Saito, Prachi Kudeshia, and Jiju Poovvancheri. 2025. Point-jepa: A joint embedding predictive architecture for self-supervised learning on point cloud. *Preprint*, arXiv:2404.16432.

Yu Shang, Yangcheng Yu, Xin Zhang, Xin Jin, Haisheng Su, Wei Wu, and Yong Li. 2025a. Mowm: Mixture-of-world-models for embodied planning via latent-to-pixel feature modulation. *Preprint*, arXiv:2509.21797.

Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. 2025b. Roboscape: Physics-informed embodied world model. *Preprint*, arXiv:2506.23135.

Haochen Shi, Huazhe Xu, Zhiao Huang, Yunzhu Li, and Jiajun Wu. 2024. Robocraft: Learning to see, simulate, and shape elasto-plastic objects in 3d with graph networks. *The International Journal of Robotics Research*, 43(4):533–549.

Rohan Sinha, Amine Elhafsi, Christopher Agia, Matthew Foutter, Edward Schmerling, and Marco Pavone. 2024. Real-time anomaly detection and reactive planning with large language models. *Preprint*, arXiv:2407.08735.

Alexander F Spies, William Edwards, Michael Ivanitskiy, Adrians Skapars, Tilman Räuker, Katsumi Inoue, Alessandra Russo, and Murray Shanahan. 2025. Transformers use causal world models in maze-solving tasks. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

StepFun. 2025. Step-3 is large yet affordable: Model-system co-design for cost-effective decoding. *Preprint*, arXiv:2507.19427.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. 2025. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8406–8416.

Hao Tang, Kevin Ellis, Suhas Lohit, Michael J. Jones, and Moitreya Chatterjee. 2025. Programmatic video prediction using large language models. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Hao Tang, Darren Key, and Kevin Ellis. 2024. Worldcoder, a model-based llm agent: Building world models by writing code and interacting with the environment. *Advances in Neural Information Processing Systems*, 37:70148–70212.

Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. 2025. Aether: Geometric-aware unified world modeling. *Preprint*, arXiv:2503.18945.

Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt Bobkin, Adi Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier, Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru, Pierre Clavier, Henry Conklin, Lucas Crawhall-Stein, Devon Crouse, Andres Felipe Cruz-Salinas, Ben Cyrus, Daniel D'souza, Hugo Dalla-Torre, John Dang, William Darling, Omar Darwiche Domingues, Saurabh Dash, Antoine Debugne, Théo Dehaze, Shaan Desai, Joan Devassy, Rishit Dholakia, Kyle Duffy, Ali Edalati, Ace Eldeib, Abdullah Elkady, Sarah Elsharkawy, Irem Ergün, Beyza Ermis, Marzieh Fadaee, Boyu Fan, Lucas Fayoux, Yannis Flet-Berliac, Nick Frosst, Matthias Gallé, Wojciech Galuba, Utsav Garg, Matthieu Geist, Mohammad Gheshlaghi Azar, Ellen Gilsenan-McMahon, Seraphina Goldfarb-Tarrant, Tomas Goldsack, Aidan Gomez, Victor Machado Gonzaga, Nithya Govindarajan, Manoj Govindassamy, Nathan Grinsztajn, Nikolas Gritsch, Patrick Gu, Shangmin Guo, Kilian Haefeli, Rod Hajjar, Tim Hawes, Jingyi He, Sebastian Hofstätter, Sungjin Hong, Sara Hooker, Tom Hosking, Stephanie Howe, Eric Hu, Renjie Huang, Hemant Jain, Ritika Jain, Nick Jakobi, Madeline Jenkins, JJ Jordan, Dhruti Joshi, Jason Jung, Trushant Kalyanpur, Siddhartha Rao Kamalakara, Julia Kedrzycki, Gokce Keskin, Edward Kim, Joon Kim, Wei-Yin Ko, Tom Kocmi, Michael Kozakov, Wojciech Kryściński, Arnav Kumar Jain, Komal Kumar Teru, Sander Land, Michael Lasby, Olivia Lasche, Justin Lee, Patrick Lewis, Jeffrey Li, Jonathan Li, Hangyu Lin, Acyr Locatelli, Kevin Luong, Raymond Ma, Lukáš Mach, Marina Machado, Joanne Magbitang, Brenda Malacara Lopez, Aryan Mann, Kelly Marchisio, Olivia Markham, Alexandre Matton, Alex McKinney, Dominic McLoughlin, Jozef Mokry, Adrien Morisot, Autumn Moulder, Harry Moynehan, Maximilian Mozes, Vivek Muppalla, Lidiya Murakhovska, Hemangani Nagarajan, Alekhya Nandula, Hisham Nasir, Shauna Nehra, Josh Netto-Rosen, Daniel Ohashi, James Owers-Bardsley, Jason Ozuzu, Dennis Padilla, Gloria Park, Sam Passaglia, Jeremy Pekmez, Laura Penstone, Aleksandra Piktus, Case Ploeg, Andrew Poulton, Youran Qi, Shubha Raghvendra, Miguel Ramos, Ekagra Ranjan, Pierre Richemond, Cécile Robert-Michon, Aurélien Rodriguez, Sudip Roy, Sebastian Ruder, Laura Ruis, Louise Rust, Anubhav Sachan, Alejandro Salamanca, Kailash Karthik Saravanakumar, Isha Satyakam, Alice Schoenauer Sebag, Priyanka Sen, Sholeh Sepehri, Preethi Seshadri, Ye Shen, Tom Sherborne, Sylvie Shang Shi, Sanal Shivaprasad, Vladyslav Shmyhlo, Anirudh Shrinivason, Inna Shteinbuk, Amir Shukayev, Mathieu Simard, Ella Snyder, Ava Spataru, Victoria Spooner, Trisha Starostina, Florian Strub, Yixuan Su, Jimin Sun, Dwarak Talupuru, Eugene Tarassov, Elena Tommasone, Jennifer Tracey, Billy Trend, Evren Tumer, Ahmet Üstün, Bharat Venkitesh, David Venuto, Pat Verga, Maxime Voisin, Alex Wang, Donglu Wang, Shijian Wang, Edmond Wen, Naomi White, Jesse Willman, Marysia Winkels, Chen Xia, Jessica Xie, Minjie Xu, Bowen Yang, Tan Yi-Chern, Ivan Zhang, Zhenyu Zhao, and Zhoujie Zhao. 2025. Command a: An enterprise-ready large language model. *Preprint*, arXiv:2504.00698.

Hugo Thimonier, José Lucas De Melo Costa, Fabrice Popineau, Arpad Rimmel, and Bich-Liên Doan. 2025. T-jepa: Augmentation-free self-supervised learning for tabular data. *Preprint*, arXiv:2410.05016.

Graison Thomas. 2025. World models: The next leap beyond llms. Accessed: 2025-12-20.

Ran Tian, Boyi Li, Xinshuo Weng, Yuxiao Chen, Edward Schmerling, Yue Wang, Boris Ivanovic, and Marco Pavone. 2024. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. *Preprint*, arXiv:2407.00959.

Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, Naila Murray, Marjan Ghazvininejad, Mike Lewis, Nicolas Ballas, Amir Bar, Michael Rabbat, Jakob Verbeek, Luke Zettlemoyer, Koustuv Sinha, Yann LeCun, and Saining Xie. 2026. Beyond language modeling: An exploration of multimodal pretraining. *Preprint*, arXiv:2603.03276.

Ludovic Tuncay, Etienne Labbé, Emmanouil Benetos, and Thomas Pellegrini. 2025. Audio-jepa: Joint-embedding predictive architecture for audio representation learning. *Preprint*, arXiv:2507.02915.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu

Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. 2024. Diffusion models are real-time game engines. *Preprint*, arXiv:2408.14837.

Khang H. N. Vo, Duc P. T. Nguyen, Thong Nguyen, and Tho T. Quan. 2025. Ti-jepa: An innovative energy-based joint embedding strategy for text-image multimodal systems. *Preprint*, arXiv:2503.06380.

Aleksandar Vujinovic and Aleksandar Kovacevic. 2025. Act-jepa: Novel joint-embedding predictive architecture for efficient policy representation learning. *Preprint*, arXiv:2501.14622.

Siheng Wan, Zhengtao Yao, Zhengdao Li, Junhao Dong, Yanshu Li, Yikai Li, Linshan Li, Haoyan Xu, Yijiang Li, Zhikang Dong, Huacan Wang, and Jifeng Shen. 2025. Jepa-t: Joint-embedding predictive architecture with text fusion for image generation. *Preprint*, arXiv:2510.00974.

Dilin Wang, Hyunyoung Jung, Tom Monnier, Kihyuk Sohn, Chuhang Zou, Xiaoyu Xiang, Yu-Ying Yeh, Di Liu, Zixuan Huang, Thu Nguyen-Phuoc, Yuchen Fan, Sergiu Oprea, Ziyan Wang, Roman Shapovalov, Nikolaos Sarafianos, Thibault Groueix, Antoine Toisoul, Prithviraj Dhar, Xiao Chu, Minghao Chen, Geon Yeong Park, Mahima Gupta, Yassir Azziz, Rakesh Ranjan, and Andrea Vedaldi. 2025a. Worldgen: From text to traversable and interactive 3d worlds. *Preprint*, arXiv:2511.16825.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024a. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*.

Lei Wang, Heyang Gao, Xiaohe Bo, Xu Chen, and Ji-Rong Wen. 2025b. Yulan-onesim: Towards the next generation of social simulator with large language models. In *Workshop on Scaling Environments for Agents*.

Lening Wang, Wenzhao Zheng, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, and Jiwen Lu. 2024b. Occsora: 4d occupancy generation models as world simulators for autonomous driving. *ArXiv*, abs/2405.20337.

Lening Wang, Wenzhao Zheng, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, and Jiwen Lu. 2024c. Occsora: 4d occupancy generation models as world simulators for autonomous driving. *Preprint*, arXiv:2405.20337.

Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. 2023. Gensim: Generating robotic simulation tasks via large language models. *arXiv preprint arXiv:2310.01361*.

Xianlong Wang, Hewen Pan, Hangtao Zhang, Minghui Li, Shengshan Hu, Ziqi Zhou, Lulu Xue, Aishan Liu, Yunpeng Jiang, Leo Yu Zhang, and Xiaohua Jia. 2024d. Robot collapse: Supply chain backdoor attacks against vlm-based robotic manipulation.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. 2024e. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pages 55–72. Springer.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. 2024f. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *ArXiv*, abs/2401.09985.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. 2024g. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. 2024. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. 2023. Daydreamer: World models for physical robot learning. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 2226–2240. PMLR.

Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. 2024. Pandora: Towards general world model with natural language actions and video states.

Siheng Xiong, Ali Payani, Yuan Yang, and Faramarz Fekri. 2024. Deliberate reasoning in language models as structure-aware planning with an accurate world model. In *Annual Meeting of the Association for Computational Linguistics*.

Siheng Xiong, Ali Payani, Yuan Yang, and Faramarz Fekri. 2025. Deliberate reasoning in language models as structure-aware planning with an accurate world model. *Preprint*, arXiv:2410.03136.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. 2024. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18826–18836.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Chang Yang, Xinrun Wang, Qinggang Zhang, Qi Jiang, and Xiao Huang. 2025b. Efficient integration of external knowledge to LLM-based world models via retrieval-augmented generation and reinforcement learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9484–9501, Suzhou, China. Association for Computational Linguistics.

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. 2023. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6.

Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. 2024. Video as the new language for real-world decision making. *arXiv preprint arXiv:2402.17139*.

Shuai Yang, Qi Yang, Luoxi Tang, Jeremy Blackburn, and Zhaohan Xi. 2025c. On the eligibility of llms for counterfactual reasoning: A decompositional study. *Preprint*, arXiv:2505.11839.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Konisberg, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun Wang, Bernard Ghanem, Huchuan Lu, Wanli Ouyang, Yu Qiao, Philip Torr, and Jing Shao. 2025d. OASIS: Open agents social interaction simulations on a large scale.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Shaofeng Yin, Jialong Wu, Siqiao Huang, Xingjian Su, Xu He, Jianye HAO, and Mingsheng Long. 2025. Trajectory world models for heterogeneous environments. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Yifei Yuan and Anders Søgaard. 2025. Revisiting the othello world model hypothesis. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Jingtong Yue, Ziqi Huang, Zhaoxi Chen, Xintao Wang, Pengfei Wan, and Ziwei Liu. 2025. Simulating the visual world with artificial intelligence: A roadmap. *Preprint*, arXiv:2511.08585.

Pengyu Zeng, Jun Yin, Miao Zhang, Yuqin Dai, Jizhizi Li, Zhanxiang Jin, and Shuai Lu. 2025. Card: Cross-modal agent framework for generative and editable residential design. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9315–9330.

Guangyao Zhai, Xingyuan Zhang, and Nassir Navab. 2025. Recurrent world model with tokenized latent states. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. 2023a. Copilot4d: Learning unsupervised world models for autonomous driving via discrete diffusion. *arXiv preprint arXiv:2311.01017*.

Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. 2024a. Copilot4d: Learning unsupervised world models for autonomous driving via discrete diffusion. *Preprint*, arXiv:2311.01017.

Peng-Fei Zhang, Ying Cheng, Xiaofan Sun, Shijie Wang, Fengling Li, Lei Zhu, and Heng Tao Shen. 2025a. A step toward world models: A survey on robotic manipulation. *Preprint*, arXiv:2511.02097.

Tianren Zhang, Guanyu Chen, and Feng Chen. 2025b. When do neural networks learn world models? In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li, Yueting Zhuang, and Weiming Lu. 2024b. Agent-pro: Learning to evolve via policy-level reflection and optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5348–5375, Bangkok, Thailand. Association for Computational Linguistics.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, Guanying Li, Ling Yan, Yao Hu, Siming Chen, Yu Wang, Xuanjing Huang, Jiebo Luo, Shiping Tang, Libo Wu, Baohua Zhou, and Zhongyu Wei. 2025c. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *Preprint*, arXiv:2504.10157.

Yang Zhang, Chenjia Bai, Bin Zhao, Junchi Yan, Xiu Li, and Xuelong Li. 2025d. Decentralized transformers with centralized aggregation are sample-efficient multi-agent world models. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. 2023b. Trafficbots: Towards world models for autonomous driving simulation and motion prediction. *Preprint*, arXiv:2303.04116.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. 2024. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *Preprint*, arXiv:2403.06845.

Yi Zhao, Aidan Scannell, Yuxin Hou, Tianyu Cui, Le Chen, Dieter Büchler, Arno Solin, Juho Kannala, and Joni Pajarinen. 2025. Generalist world model pre-training for efficient reinforcement learning. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 2024. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. 2025. Tesseract: Learning 4d embodied world models. *Preprint*, arXiv:2504.20995.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. 2025a. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *Preprint*, arXiv:2503.21755.

Han Zheng, Yi Cheng, Hang Liu, Linqi Ye, and Houde Liu. 2025b. Huwo: Building physical interaction world models for humanoid robot locomotion. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. 2025c. Flare: Robot learning with implicit world modeling. *ArXiv*, abs/2505.15659.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. 2024. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European conference on computer vision*, pages 55–72. Springer.

Siyu Zhou, Tianyi Zhou, Yijun Yang, Guodong Long, Deheng Ye, Jing Jiang, and Chengqi Zhang. 2025a. Wall-e 2.0: World alignment by neurosymbolic learning improves world model-based llm agents. *Preprint*, arXiv:2504.15785.

Xuhui Zhou, Jiarui Liu, Akhila Yerukola, Hyunwoo Kim, and Maarten Sap. 2025b. Social world models. *Preprint*, arXiv:2509.00559.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. 2025a. Irasim: A fine-grained world model for robot manipulation. *Preprint*, arXiv:2406.14540.

Haoran Zhu, Zhenyuan Dong, Kristi Topollai, Beiyao Sha, and Anna Choromanska. 2025b. Self-supervised representation learning with joint embedding predictive architecture for automotive lidar object detection. *Preprint*, arXiv:2501.04969.

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. 2025c. Aether: Geometric-aware unified world modeling. *ArXiv*, abs/2503.18945.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025d. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Bohan Li, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, Yang You, Zhaoxiang Zhang, Dawei Zhao, Liang Xiao, Jian Zhao, Jiwen Lu, and Guan Huang. 2025e. Is sora a world simulator? a comprehensive survey on general world models and beyond. *Preprint*, arXiv:2405.03520.

Dongqi Zuo, Zheng CHEN, Chuan Zhou, Yandong Guo, Xiao He, and Mingming Gong. 2025. RADI: LLMs as world models for robotic action decomposition and imagination. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

## A Motivation and Justification of the Dual Taxonomy

We introduce a dual taxonomy to disentangle representational design from deployment context, reflecting the core observation that advances in model architecture do not map directly onto downstream performance. World models differ primarily in how they represent state and dynamics, yet they are evaluated across heterogeneous application settings with distinct constraints, data regimes, and safety requirements. Separating these dimensions enables a structured comparison of learning paradigms while independently analyzing real-world instantiation, evaluation protocols, and system-level integration.

The **Theoretical Taxonomy** organizes models by representational substrate and predictive mechanism. *Pixel-Space Generative Simulators* model dynamics directly in observation space, emphasizing reconstruction fidelity and visual realism. *Latent-Space Predictive Architectures* learn compact state representations with explicit transition operators, prioritizing predictive sufficiency and multi-step rollout consistency. *Neuro-Symbolic and Code-Coupled World Models* integrate structured representations or executable programs to enhance causal traceability, logical constraints, and counterfactual simulation. *Hybrid Mixture-of-World-Models* combine complementary inductive biases, such as object-centric modules, discrete-continuous state mixing, or parallel pixel-latent pathways, to balance abstraction and expressivity. Finally, *Autoregressive Models* are included as a foundational paradigm, as next-token prediction can implicitly induce predictive structure and latent state tracking, providing a baseline for evaluating whether explicit world-model components are necessary or beneficial.

The **Application Taxonomy** categorizes deployment domains according to operational requirements rather than model architecture. *Autonomous Driving* and *Embodied Robotics* impose strict real-time, safety-critical, and physical consistency constraints. *Interactive Simulation and Gaming* emphasizes controllable environment generation, long-horizon coherence, and user interaction. *Social and Agentic Simulation* focuses on multi-agent reasoning, behavioral modeling, and strategic interaction. *Medical and Biological Systems* require causal interpretability, uncertainty calibration, and decision support under high-stakes conditions. By organizing applications separately from model design, we avoid conflating architectural capability with domain-specific performance.

Together, the dual taxonomy provides a two-dimensional analytical framework: one axis captures how world models are constructed, and the other captures where they are deployed. This structure supports cross-cutting comparison, clarifies shared methodological foundations across domains, and highlights gaps between theoretical modeling advances and application-level robustness. It thereby enables systematic analysis of representational trade-offs, evaluation standards, and scalability challenges in Large Language and Vision-Language Models as World Models.

## B Survey Methodology

This survey follows a structured multi-stage protocol to ensure coverage, consistency, and reproducibility in the study of Large Language and Vision-Language Models as World Models. The goal is to construct a representative corpus, systematically classify contributions, and synthesize findings along the dual taxonomy introduced in the main text.

**Stage 1: Data Collection.** We collected candidate papers from arXiv, ACL Anthology, DBLP, Scopus, IEEE Xplore, SpringerLink, ScienceDirect, NeurIPS, ICML, ICLR, CVPR, ICCV, ECCV, RSS, CoRL, AAAI, IJCAI, and related AI and robotics venues (see Table 1). Searches were conducted using combinations of keywords including `world model`, `predictive dynamics`, `latent state modeling`, `next-token prediction`, `autoregressive simulation`, `diffusion-based simulation`, `neuro-symbolic reasoning`, `object-centric representation`, `multimodal world modeling`, `embodied AI`, `counterfactual reasoning`, and `causal inference`. The initial query phase yielded approximately 1624 candidate papers.

**Stage 2: Deduplication and Metadata Cleaning.** All retrieved records were merged and standardized. Duplicate entries, version overlaps, withdrawn papers, non-English manuscripts, inaccessible files, and workshop abstracts without full technical content were removed. After this stage, 1348 papers remained.

**Stage 3: Title Screening.** Titles were screened for explicit relevance to world models, environ-

Table 1: Primary sources used for collecting candidate papers on world models and related paradigms.

| Source Type | Venues / Repositories |
|---|---|
| Preprint Servers | arXiv, bioRxiv |
| Indexing Databases | DBLP, Scopus, Web of Science |
| Digital Libraries | IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect |
| NLP Venues | ACL Anthology, EMNLP, NAACL |
| ML Conferences | NeurIPS, ICML, ICLR |
| Vision Conferences | CVPR, ICCV, ECCV |
| Robotics Venues | RSS, CoRL, ICRA, IROS |
| General AI Conferences | AAAI, IJCAI |
| AI Journals | JMLR, TMLR, IEEE Transactions on Artificial Intelligence, Other Machine Learning Journals |
| Vision & Robotics Journals | TPAMI, IJCV, RA-L, Autonomous Robots |

ment dynamics, latent state learning, predictive simulation, multimodal modeling, neuro-symbolic reasoning, or counterfactual inference. Clearly unrelated works were excluded. After this stage, 572 papers remained.

**Stage 4: Abstract Screening.** Abstracts were reviewed to confirm the presence of a modeling contribution, including explicit representation learning, transition modeling, structured inference, or evaluation of predictive dynamics. Purely application-focused studies without methodological novelty were removed. After this stage, 344 papers remained.

**Stage 5: Full-Text Eligibility Assessment.** The remaining papers were evaluated in full text. Inclusion required explicit modeling of environment state, dynamics, or structured predictive reasoning, rather than downstream task performance alone. Papers lacking formal architectural or representational contributions were excluded. After this stage, 167 papers remained.

**Stage 6: LLM and VLM Relevance Verification.** We assessed whether each paper directly studied a language or vision-language foundation model, an autoregressive foundation model, or a generative model integrated with a language or vision backbone in a world-modeling role. We retained a paper when the model contributed to predictive state representation, environment dynamics, simulation, planning, or counterfactual prediction. We excluded papers in which an LLM/VLM served only as a text interface or perception module, and general world-model papers without a qualifying foundation-model connection. This final relevance check reduced the 167 papers passing full-text assessment to 137 papers for analysis (Table 2). Non-LLM/VLM systems discussed elsewhere are explicitly treated as architectural contrast cases, not as members of the 137-paper LLM/VLM corpus.

Table 2: Summary of paper selection process.

| Stage | Papers |
|---|---|
| Initial retrieval | 1,624 |
| Deduplication and metadata cleaning | 1,348 |
| Title screening | 572 |
| Abstract screening | 344 |
| Full-text eligibility | 167 |
| LLM/VLM relevance verification | 137 |

**Stage 7: Theoretical Coding.** Each paper was categorized according to the theoretical taxonomy: pixel-space generative simulators, latent-space predictive architectures, neuro-symbolic or code-coupled world models, hybrid mixture-of-world-models, and autoregressive models as a foundational baseline. Classification was based on primary representational mechanism rather than application domain.

**Stage 8: Application Mapping.** We additionally assigned each work to one or more application domains: autonomous driving, embodied robotics, interactive simulation and gaming, social and agentic simulation, and medical and biological systems. When models spanned multiple domains, all relevant contexts were recorded to preserve analytical completeness.

**Curatorial Principles.** To ensure analytical neutrality, papers were selected based on methodological contribution rather than reported benchmark performance. Emphasis was placed on works introducing new representations, transition operators, training objectives, or evaluation protocols. Disagreements in categorization were resolved through consensus review.

**Synthesis Strategy.** Analysis proceeded along the dual taxonomy to decouple representation from application. This structure enabled controlled com-

parison across architectural families while examining domain-specific constraints independently. Quantitative trends were summarized descriptively, and qualitative insights were synthesized to highlight common design patterns, trade-offs, and open challenges in next-generation world models.

### B.1 Rationale and Analytical Role of the Dual Taxonomy

The dual taxonomy is introduced as an organizational and analytical framework rather than an evaluative benchmark. Its purpose is to structure a rapidly growing and technically heterogeneous literature on Large Language and Vision-Language Models as World Models by separating representational design choices from deployment contexts. This separation clarifies that architectural mechanisms and application domains address different scientific questions and should not be conflated in comparative assessment.

The **Theoretical Taxonomy** organizes models according to their core representational and learning principles. Pixel-Space Generative Simulators prioritize explicit observation reconstruction and high-fidelity synthesis, typically operating in observable space. Latent-Space Predictive Architectures shift prediction into compressed embedding spaces, emphasizing state abstraction, temporal coherence, and scalability. Neuro-Symbolic and Code-Coupled World Models integrate structured representations, predicates, or executable programs to enhance causal consistency and interpretability. Autoregressive Models, including foundation LLMs and VLMs, are treated as implicit world models when next-token or next-token-like objectives induce internal predictive state representations. This theoretical grouping highlights how different modeling substrates encode dynamics, manage uncertainty, and support counterfactual simulation.

The **Application Taxonomy** categorizes deployments by environment type and interaction modality. Autonomous Driving emphasizes real-time perception, trajectory prediction, and safety-constrained decision-making. Embodied Robotics focuses on closed-loop control, action-conditioned prediction, and physical interaction. Interactive Simulation and Gaming examines controllable generative environments and latent action spaces. Social and Agentic Simulation addresses multi-agent reasoning, memory, and emergent collective behavior. This application-oriented structure enables comparison of how similar architectural principles manifest across distinct operational constraints and evaluation regimes.

Together, the dual taxonomy separates representation from deployment, allowing independent analysis of modeling paradigms and task environments. This design supports systematic comparison across architectural families while avoiding over-interpretation of performance differences that may stem from evaluation setting rather than intrinsic modeling capability. By making structural assumptions explicit, the taxonomy facilitates identification of shared challenges, such as state persistence, long-horizon consistency, and scalable multimodal grounding, across both theoretical and applied lines of work.

## C Adoption of the Unified Mathematical Framework Across World Modeling Paradigms

Under the unified framework introduced above, different world modeling paradigms can be understood as instantiations of the same underlying predictive structure, differing primarily in their choice of representation space, training objective, and constraints imposed on the transition operator. We outline how the base formulation adapts to each major class of world models.

The formalism in Section 3 provides a common interface for comparing otherwise different systems. In the subsections below, we specify each paradigm by its state space, transition operator, action interface, and decoder where one is present. Pixel-space models use observations as state; latent models predict compact representations; neuro-symbolic and programmatic systems constrain state or transitions with structured representations; hybrid systems couple multiple state spaces; and autoregressive models implement an implicit transition over tokenized context. These instantiations make the taxonomy analytically comparable rather than a list of model families.

### C.1 Pixel-Space Generative World Models

Pixel-space generative simulators correspond to the setting in which the internal representation coincides with the observation space:

$$\mathcal{Z} \equiv \mathcal{O}. \tag{5}$$

These models directly learn the predictive distribution:

$$P(o_{t+1} \mid o_{\leq t}, a_t), \tag{6}$$

typically using autoregressive transformers or diffusion-based generative processes. Although their internal computation often relies on latent variables, tokenized patches, or VAE-compressed spaces, their defining characteristic is that the prediction target is the future observation itself.

From the unified perspective, such models implicitly marginalize over latent environment states:

$$P(o_{t+1} \mid o_{\le t}, a_t) = \int P(o_{t+1} \mid s_{t+1}) P(s_{t+1} \mid s_t, a_t)\, ds, \quad (7)$$

without explicitly representing $s_t$ or a compact latent $z_t$. As a result, world dynamics are encoded implicitly in the generative weights and are evaluated through the coherence, physical plausibility, and temporal consistency of predicted observations. This design favors perceptual fidelity and generality, but provides limited inductive bias toward compact, persistent state representations.

### C.2 Latent-Space Predictive Architectures

Latent world models introduce an explicit information bottleneck by learning a compact internal state:

$$z_t \in \mathbb{R}^d, \quad d \ll \dim(\mathcal{O}), \quad (8)$$

and defining dynamics directly in latent space. A common training objective reconstructs future observations through a decoder:

$$\mathcal{L} = \mathbb{E}\big[\ell\big(h_\theta(g_\theta(z_t, a_t)), o_{t+1}\big)\big], \quad (9)$$

where the latent transition model $g_\theta$ captures environment dynamics and $h_\theta$ provides grounding to observation space.

Representation-predictive approaches, such as Joint Embedding Predictive Architectures (JEPA), instead avoid explicit reconstruction and optimize:

$$\mathcal{L}_{\text{JEPA}} = \ell\big(g_\theta(z_t, a_t), \text{stopgrad}(z_{t+1})\big), \quad (10)$$

encouraging $z_t$ to retain only information that is predictive of future latent states. Across variants, the central principle is predictive sufficiency: the learned latent state should preserve the information necessary for accurate long-horizon prediction and planning, while discarding observation-level detail that is irrelevant to dynamics.

### C.3 Neuro-Symbolic and Code-Coupled WMs

Neuro-symbolic and programmatic approaches constrain the representation space to structured domains:

$$\mathcal{Z} \in \{\mathcal{G}, \mathcal{P}, \mathcal{C}\}, \quad (11)$$

where $\mathcal{G}$ denotes graph-based representations, $\mathcal{P}$ logical predicates or relational states, and $\mathcal{C}$ executable programs or simulators. In this setting, state transitions are governed by:

$$z_{t+1} = T(z_t, a_t), \quad (12)$$

where the transition operator $T$ enforces logical consistency, physical laws, or procedural constraints. Neural components are often used to infer or ground symbolic states, while dynamics are explicitly structured and low-entropy. This design yields interpretable and verifiable world models, at the cost of reduced flexibility and higher modeling overhead.

### C.4 Hybrid and Mixture-of-World-Models

Hybrid architectures combine multiple representations to balance perceptual realism, abstraction, and constraint satisfaction. These models maintain a set of interacting internal states:

$$z_t = \left\{ z_t^{\text{pixel}}, z_t^{\text{latent}}, z_t^{\text{symbolic}} \right\}, \quad (13)$$

with coupled transition dynamics:

$$z_{t+1}^{(i)} = g_\theta^{(i)}\big(z_t^{(i)}, a_t, \phi(z_t^{(-i)})\big), \quad (14)$$

where $\phi$ denotes cross-representation information exchange. Each component captures complementary structure: pixel-level states preserve perceptual detail, latent states support efficient planning and long-horizon prediction, and symbolic states impose explicit constraints.

From a probabilistic standpoint, hybrid models can be interpreted as factorized approximations to the true transition operator $P(s_{t+1} \mid s_t, a_t)$, with each factor encoding a distinct inductive bias. Training typically optimizes a multi-objective loss that jointly enforces perceptual accuracy, predictive consistency, and structural validity.

### C.5 Autoregressive Models

Autoregressive architectures learn a conditional distribution $p_\theta(x_t \mid x_{<t})$ over discrete tokens. Although designed for language generation, there is now substantial mechanistic evidence that this objective induces internal representations encoding state, causality, and temporal dynamics, provided the training distribution is sufficiently structured.

■ **Mechanistic Connection to the Transition Operator.** Within the unified framework of § 3, next-token prediction can be interpreted as learning an

implicit transition operator $\hat{g}_\theta$ over a tokenized state space. Formally, if we identify the token sequence $x_{<t}$ as a discretized encoding of latent state $z_t$, then the causal self-attention computation

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \quad (15)$$

aggregates historical context into a key-value memory and writes a new state vector, approximating the recurrent update $z_{t+1} \leftarrow g_\theta(z_t, a_t)$. Each transformer layer refines this approximation, and the full forward pass corresponds to a multi-step latent rollout compressed into a single feedforward pass. The causal mask enforces the temporal ordering required by the Markov condition $P(s_{t+1} \mid s_t, a_t)$; without it, the model would be unable to respect the arrow of time.

This interpretation is not merely definitional. Probing studies demonstrate that trained LLMs encode structured latent states in their residual stream. The Othello-GPT experiment (Li, 2023) shows that a GPT-class model trained only on legal move sequences—without any board representation in the input—develops linear representations of board occupancy in intermediate layers, representations that support causal intervention: patching in a modified board state causes the model to predict moves consistent with the patched state rather than the observed sequence. Similar linear probing evidence is reported for more complex environments (Li et al., 2024b; Hazineh et al., 2023). Text-based interactive fiction environments such as TextWorld (Côté et al., 2019) similarly reveal that LLMs trained on sequential game transcripts track inventory state, room topology, and object affordances in a distributed but linearly decodable manner.

■ **The Linguistic-Correlation vs. Physical-Causality Gap.** Despite this evidence, a critical limitation remains: autoregressive objectives optimize for token-level prediction accuracy, not for causal fidelity. The model learns correlational structure present in training data; if that data reflects physical regularities, these regularities are captured, but confounded associations are equally reinforced. This asymmetry produces a systematic failure mode: models predict plausible-sounding next tokens that violate physical invariants such as mass conservation, object permanence, or counterfactual consistency (Núñez-Molina et al., 2025). The transition operator $\hat{g}_\theta$ approximated by attention is therefore best understood as a *statistical* transition operator rather than a *causal* one. Addressing this gap requires either causal pretraining objectives, explicit grounding to verifiable world states, or integration with symbolic constraints—motivating the hybrid architectures discussed in § 4.

Pretraining on unified image-text-action corpora partially closes this gap. Native unified VLMs trained jointly on visual generation and language modeling exhibit emergent behaviors—including consistent object tracking across frames and predictive depth of scene dynamics—not present when either modality is trained separately (Deng et al., 2025; Tong et al., 2026; Cui et al., 2025). These models use a single autoregressive objective over interleaved visual and linguistic tokens, making them closer to the state-tracking formulation in § C. LLM-based agents augmented with explicit world-model modules further reduce hallucination and improve long-horizon task success (Guo et al., 2025; Qiu et al., 2026b; Li et al., 2026; Qiu et al., 2026a; Li et al., 2025a).

### C.5.1 Cross-Paradigm Comparison

Table 3 summarizes how the four main paradigms differ along six dimensions central to world modeling: representation controllability, causal grounding, long-horizon consistency, interpretability, compute cost, and planning support. We use H, M, and L as qualitative summaries of the capabilities supported by the cited mechanisms and evaluations, not as measured scores. Controllability refers to whether actions or interventions can directly alter predicted state; causal grounding refers to whether transitions are tied to explicit causal structure or intervention tests; long-horizon consistency refers to evidence across multi-step rollouts; interpretability refers to whether state and transition rules can be inspected; compute cost describes relative inference or training demands; and planning support refers to demonstrated use of predicted states for action selection. We assign a rating only when the cited literature provides relevant evidence, and use a range when results vary across implementations. These ratings summarize heterogeneous studies and should not be read as a controlled benchmark comparison.

Pixel-space models achieve the highest perceptual fidelity, but their dynamics are implicit in generative weights: they offer no explicit state variable $z_t$ for planning or counterfactual intervention. Latent predictive architectures trade fidelity for a compact, explicit $z_t$ with a differentiable transition operator $g_\theta$, enabling multi-step rollouts, but their la-

Table 3: Qualitative comparison of world-model paradigms across six analytical dimensions. Ratings are relative: H = high, M = medium, L = low.

| Paradigm | Controllability | Causal Grounding | Long-Horizon | Interpretability | Compute | Planning |
|---|---|---|---|---|---|---|
| Pixel-space generative | L | L | L | L | H | L |
| Latent predictive (JEPA/RSSM) | M | M | M–H | L | M | M–H |
| Neuro-symbolic / programmatic | H | H | H | H | L–M | H |
| Autoregressive (LLM/VLM) | M | L–M | M | L | H | M |

tent space is learned without causal supervision and remains opaque to symbolic manipulation. Neuro-symbolic and programmatic systems provide the highest controllability and interpretability because their transition operators $T(z_t, a_t)$ are explicitly structured by predicates, graphs, or programs; the cost is reduced scalability and increased design overhead. Autoregressive LLMs and VLMs occupy an intermediate position: their self-attention mechanism implicitly encodes state transitions, and probing studies confirm structured internal representations, but their training objective (token-level cross-entropy) does not enforce causal invariance, creating the linguistic-correlation vs. physical-causality gap discussed in § C.5. Across all paradigms, realism alone does not guarantee understanding, and abstraction alone sacrifices grounding. The field is therefore converging toward hybrid systems that combine generative perception, latent prediction, symbolic structure, and retrieval-based grounding.

### C.6 Combining Supervised and Unsupervised Learning in World Models

Modern world models often integrate supervised and unsupervised learning to balance scalability, generalization, and task-specific guidance. Let $\mathbf{x}_{1:T}$ denote a sequence of observations and $\mathbf{a}_{1:T}$ the corresponding actions. Unsupervised objectives typically aim to learn latent dynamics $\mathbf{z}_t$ by predicting future representations:

$$\mathcal{L}_{\text{unsup}} = \sum_{t=1}^{T-1} \| f_\theta(\mathbf{z}_t, \mathbf{a}_t) - \mathbf{z}_{t+1} \|^2, \tag{16}$$

where $f_\theta$ is the transition function in latent space. This can be learned from unlabeled video, text, or multimodal streams without requiring task-specific labels.

Supervised signals, when available, complement unsupervised dynamics by aligning latent states $\mathbf{z}_t$ with interpretable variables or task outcomes $y_t$:

$$\mathcal{L}_{\text{sup}} = \sum_{t=1}^{T} \ell(g_\phi(\mathbf{z}_t), y_t), \tag{17}$$

where $g_\phi$ maps latent states to supervised targets and $\ell(\cdot)$ is a suitable loss (e.g., cross-entropy or MSE).

The combined objective becomes:

$$\mathcal{L} = \mathcal{L}_{\text{unsup}} + \lambda \, \mathcal{L}_{\text{sup}}, \tag{18}$$

with $\lambda$ controlling the balance between unsupervised dynamics learning and task-guided supervision. This framework naturally generalizes to reinforcement or hybrid neuro-symbolic settings, where $\mathcal{L}_{\text{sup}}$ may encode policy rewards, symbolic constraints, or verification signals.

By explicitly combining unsupervised representation learning with complementary supervised objectives, world models can scale from raw observations while maintaining task relevance, interpretability, and planning capabilities.

## D Large Language and Vision-Language Models as World Models

### D.1 Conceptual Definition

In this survey, an LLM/VLM-based world model is a system in which a language or vision-language foundation model supplies or materially shapes an internal predictive state, a transition prediction, or a simulation used for planning or counterfactual reasoning. A model does not qualify solely because it describes a world, answers questions about an image, or converts instructions into actions. When an LLM/VLM provides only perception, language grounding, or planning around a separate dynamics model, we label that role explicitly rather than treating the whole system as an LLM/VLM transition model.

### D.2 How LLMs and VLMs Function as World Models

Autoregressive LLMs learn next-token prediction over large corpora, which induces internal represen-

tations capturing statistical regularities, temporal dependencies, and latent structure. When extended to multimodal inputs, VLMs incorporate visual tokens or cross-attention mechanisms, enabling joint modeling of perceptual observations and language. In interactive settings, these models can condition on prior observations, actions, and external memory, thereby forming a closed-loop predictive system. Through large-scale pretraining and instruction alignment, they acquire the ability to represent entities, relations, and state transitions, which are essential components of internal simulation.

From an architectural perspective, world-model behavior emerges when the model approximates a transition operator in latent space. In language-only settings, this transition corresponds to sequential prediction over symbolic state descriptions. In multimodal settings, it extends to joint state representations that integrate textual, visual, and sometimes temporal signals. Diffusion-based generative models integrated with language or vision backbones further contribute by modeling structured data distributions, supporting stochastic simulation and counterfactual sampling.

### D.3 Mathematical Connection to World Models

As described in Section 3, formally, a world model can be described as a learned approximation of environment dynamics:

$$p_\theta(s_{t+1} \mid s_t, a_t), \tag{19}$$

where $s_t$ denotes the latent state and $a_t$ denotes actions. In foundation-model-based systems, $s_t$ is not explicitly defined but is implicitly represented in hidden activations, embedding spaces, or multimodal token sequences. Autoregressive LLMs can be interpreted as modeling:

$$p_\theta(x_t \mid x_{<t}), \tag{20}$$

which corresponds to learning a transition over a tokenized state representation. When extended to multimodal inputs, the model approximates:

$$p_\theta(x_t \mid x_{<t}, v), \tag{21}$$

where $v$ represents visual or other sensory context. Under this framing, the model acts as a latent simulator, mapping histories to predictive distributions over future observations.

The connection to classical world models lies in three properties. First, state persistence arises from recurrent context windows and memory mechanisms. Second, counterfactual reasoning corresponds to conditional generation under alternative prompts or actions. Third, stochastic sampling from the learned distribution enables internal simulation of multiple future trajectories. Thus, LLMs and VLMs can be viewed as implicit probabilistic simulators operating in high-dimensional embedding spaces.

### D.4 Scope and Relevance Criteria

Because the scope of this survey is *Large Language and Vision-Language Models as World Models*, we introduced an additional relevance check to ensure alignment with foundation-model-based approaches. Works were retained only if they directly involved LLMs, multimodal LLMs, Vision-Language Models, autoregressive foundation models, diffusion-based generative models integrated with language or vision backbones, foundational strategies derived from LLM/VLM development paradigms, or architectures explicitly extending these systems toward world modeling. This criterion ensures conceptual coherence with the survey title and prevents dilution with general world modeling literature that does not rely on large-scale foundation models.

We distinguish four roles when coding systems: backbone, where the LLM/VLM supplies the predictive state or transition; planner, where it plans over a separate world model; grounding module, where it provides perception or language grounding to a separate dynamics model; and contrast case, where a non-LLM/VLM system is included only for architectural comparison.

### D.5 Implications for World Modeling Research

Viewing LLMs and VLMs as world models reframes evaluation, architecture design, and training objectives. It shifts emphasis from surface-level generation quality toward latent state consistency, temporal coherence, and action-conditioned prediction. It also motivates hybrid systems that combine autoregressive reasoning, multimodal grounding, symbolic constraints, and generative modeling. Within this unified perspective, foundation models serve as scalable approximators of environment dynamics, forming a bridge between representation learning, simulation, and interactive intelligence.

## E More Details on Joint Embedding Predictive Architectures (JEPA)

Joint Embedding Predictive Architectures (JEPA) define a class of latent predictive models that learn to forecast future representations rather than reconstruct raw observations. By operating in embedding space, JEPA avoids pixel-level generation and instead optimizes predictive sufficiency, enabling scalable learning of structured, semantically meaningful state abstractions. This formulation emphasizes invariance, temporal coherence, and object-level reasoning, positioning JEPA as a foundational paradigm within latent-space world modeling. Table 4 outlines different JEPA approaches.

■ **Core Vision Architectures.** I-JEPA (Assran et al., 2023) introduces masked prediction over large semantic regions, learning image representations by inferring abstract embeddings instead of pixels. V-JEPA (Bardes et al., 2024) extends this principle to video, using spatiotemporal masking to capture motion dynamics and object persistence from unlabeled data, while V-JEPA 2 (Assran et al., 2025) and V-JEPA 2-AC scale this approach to large corpora and incorporate action-conditioned prediction. Variants such as C-JEPA (Mo and Tong, 2024) integrate regularization strategies to stabilize embeddings, Causal-JEPA (Nam et al., 2026) uses counterfactual reasoning, DSeq-JEPA (He et al., 2026) introduces sequential attention-based prediction, and MC-JEPA (Bardes et al., 2023) jointly models motion and semantic content. CNN-JEPA (Kalapos and Gyires-Tóth, 2024) adapts the framework to convolutional architectures, demonstrating that predictive embedding learning is not restricted to transformer backbones.

■ **Multimodal and Language Integration.** JEPA has been extended to multimodal settings where vision and language are jointly modeled. VL-JEPA (Chen et al., 2026) replaces autoregressive text generation with direct prediction of continuous language embeddings, reducing inference latency while preserving semantic grounding. Related frameworks such as JEPA-T (Wan et al., 2025) and TI-JEPA (Vo et al., 2025) unify text and visual representations through cross-modal attention, while LLM-JEPA (Huang et al., 2026) integrates large language models with predictive embedding objectives, bridging autoregressive reasoning and latent world modeling.

■ **Sequential, Control, and Embodied Settings.** Beyond static perception, JEPA variants model temporal and action-conditioned dynamics. ACT-JEPA (Vujinovic and Kovacevic, 2025) combines self-supervised learning with imitation to predict structured action-observation sequences, enabling policy-relevant representations. TS-JEPA (Ennadir et al., 2025) and T-JEPA (Thimonier et al., 2025) extend the framework to time-series forecasting under noise and irregular sampling, while seq-JEPA (Ghaemi et al., 2026) learns invariant and equivariant sequence representations for future state prediction. These approaches demonstrate the applicability of JEPA to control, trajectory modeling, and embodied decision-making.

■ **Domain-Specific Extensions.** The predictive embedding paradigm generalizes across modalities and scientific domains. ECG-JEPA (Kim, 2024) and Signal-JEPA (Guetschel et al., 2024) apply JEPA to physiological signals for tasks such as arrhythmia detection, while HEP-JEPA (Bardhan et al., 2025) models particle interactions in high-energy physics. Audio-JEPA (Tuncay et al., 2025) and Stem-JEPA (Riou et al., 2024) extend the framework to sound and music, and 3D-JEPA (Hu et al., 2024) and Point-JEPA (Saito et al., 2025) operate on spatial point clouds. In autonomous systems, AD-L-JEPA (Zhu et al., 2025b) models LiDAR-based spatial dynamics, and SAR-JEPA (Li et al., 2023b) addresses radar signal prediction under noise.

■ **Theoretical and Generative Extensions.** Recent work refines the theoretical foundations of JEPA. LeJEPA (Balestriero and LeCun, 2025) introduces principled regularization to prevent embedding collapse, replacing heuristic constraints with formal guarantees. Hybrid variants such as D-JEPA (Chen et al., 2025b) and N-JEPA (Qiu et al., 2025) integrate diffusion or flow-based generative processes with predictive embedding learning, combining representation-level prediction with generative modeling capabilities.

Across these variants, JEPA consistently trades pixel-level fidelity for structured latent prediction, improving scalability, robustness, and abstraction. The framework demonstrates that world modeling can emerge from predictive representation learning without explicit reconstruction, though challenges remain in aligning latent dynamics with physical causality and long-horizon consistency.

Table 4: Overview of Joint Embedding Predictive Architectures (JEPA) variants and their key characteristics.

| Category | Model Variants | Key Contribution |
|---|---|---|
| Core Vision Architectures | I-JEPA (Assran et al., 2023), V-JEPA (Bardes et al., 2024), V-JEPA 2 (Assran et al., 2025), V-JEPA 2-AC | Masked embedding prediction for images and videos; spatiotemporal reasoning and action conditioning |
| | C-JEPA (Mo and Tong, 2024), Causal-JEPA (Nam et al., 2026), DSeq-JEPA (He et al., 2026), MC-JEPA (Bardes et al., 2023), CNN-JEPA (Kalapos and Gyires-Tóth, 2024) | Stabilization, causal inference, sequential attention, motion-content modeling, and CNN adaptation |
| Multimodal & Language Integration | VL-JEPA (Chen et al., 2026), JEPA-T (Wan et al., 2025), TI-JEPA (Vo et al., 2025), LLM-JEPA (Huang et al., 2026) | Joint vision-language embeddings, non-autoregressive text prediction, and integration with LLM reasoning |
| Sequential, Control & Embodied Settings | ACT-JEPA (Vujinovic and Kovacevic, 2025), TS-JEPA (Ennadir et al., 2025), T-JEPA (Thimonier et al., 2025), seq-JEPA (Ghaemi et al., 2026) | Action-conditioned prediction, time-series modeling, invariant sequence learning, and policy-relevant representations |
| Domain-Specific Extensions | ECG-JEPA (Kim, 2024), Signal-JEPA (Guetschel et al., 2024), HEP-JEPA (Bardhan et al., 2025) | Physiological signals and scientific data modeling (e.g., healthcare, physics) |
| | Audio-JEPA (Tuncay et al., 2025), Stem-JEPA (Riou et al., 2024) | Audio and music representation learning |
| | 3D-JEPA (Hu et al., 2024), Point-JEPA (Saito et al., 2025), AD-L-JEPA (Zhu et al., 2025b), SAR-JEPA (Li et al., 2023b) | 3D spatial reasoning, LiDAR dynamics, and radar signal prediction |
| Theoretical & Generative Extensions | LeJEPA (Balestriero and LeCun, 2025), D-JEPA (Chen et al., 2025b), N-JEPA (Qiu et al., 2025) | Theoretical guarantees, collapse prevention, and integration with diffusion/flow-based generative modeling |

## F Extended Challenges in World Models

World models face a range of structural and system-level challenges that limit their reliability, scalability, and deployment in real-world settings.

■ **Cross-Domain Analysis.** Despite surface differences, the four domains share a set of transferable modeling assumptions and reveal a set of domain-specific bottlenecks.

*Transferable assumptions.* Action conditioning, structured latent state, and memory mechanisms improve performance across all four domains. Latent rollouts computed by RSSM-style models in robotics transfer directly to gaming and driving when the action space is compatible. Self-supervised pretraining objectives from JEPA—spatiotemporal masking and predictive embedding—generalize from autonomous driving sensors to robotic manipulation and even social agent representation.

*Universal evaluation criteria.* State-tracking accuracy under counterfactual interventions, long-horizon rollout divergence rate, and physical invariant preservation (mass, occlusion, object permanence) are meaningful metrics across all domains. Current benchmarks fragment evaluation along domain-specific lines, which obscures these shared dimensions and makes cross-architecture comparison unreliable.

*Domain-specific bottlenecks.* Autonomous driving is bottlenecked by geometric fidelity and real-time latency under sensor noise. Embodied robotics is bottlenecked by action-conditioned data scarcity and sim-to-real transfer. Interactive gaming is bottlenecked by controllable long-horizon coherence in open-ended environments. Social simulation is bottlenecked by theory-of-mind modeling and multi-agent scalability beyond hundreds of agents.

These observations suggest that the most productive near-term direction is developing *shared evaluation protocols* that measure state-tracking and causal consistency independently of perceptual realism, enabling meaningful cross-domain and cross-architecture comparison. These shared criteria support cross-domain comparison, while

Table 5: Structural grouping of symbolic, neuro-symbolic, programmatic, and hybrid world models according to their state representation and transition mechanism.

| Subgroup | State representation | Transition mechanism | Examples |
|---|---|---|---|
| Symbolic world representations | Predicates or planning domains such as PDDL | An LLM or vision language model may generate, ground, or refine the symbolic representation; transitions are defined by learned or specified symbolic operators and domain rules | Text2World, Pixels-to-Predicates, SWAP |
| Neuro-symbolic hybrids | Predicates, graphs, or symbolic knowledge grounded or complemented by neural models | Neural or language models infer symbolic state information, while symbolic operators, learned rules, or executable constraints structure prediction and planning | WorldCloner, VisualPredicator, WALL-E 2.0 |
| Programmatic and simulator-coupled world models | Executable programs, programmatic laws, or simulator parameters | A model synthesizes or updates executable code or parameters, and program execution or an external simulator computes state transitions | WorldCoder, OneLife, ProgGen, LLMPhy |
| Hybrid representation mixtures | Multiple coupled representations, such as pixel, latent, and object-factorized states | Information from complementary representation spaces is combined or exchanged for prediction, planning, or control | Object-centric world models, MoWM |

the application-specific bottlenecks explain why a single benchmark or architecture rating cannot stand in for performance in every deployment setting. Medical and biological systems remain an underexplored direction in this survey rather than a fifth application domain developed in the main taxonomy.

■ **Inference Latency.** Large LLMs and diffusion-based simulators often fail to meet real-time constraints required in embodied systems. Robotic control typically requires sub-100 ms response times, whereas cloud-based inference pipelines can exceed 500 ms (Gao et al., 2025b). Although compression, distillation, and edge-cloud co-design reduce latency, maintaining predictive fidelity under strict timing constraints remains unresolved.

■ **Hallucination and Counterfactual Consistency.** World models must generate hypothetical future states for planning, making controlled hallucination necessary. The challenge lies in ensuring physical plausibility and environmental consistency while preserving generative flexibility (Yang et al., 2025c). Current approaches lack principled mechanisms to constrain counterfactual rollouts under real-world dynamics.

■ **Data Scarcity and Action Alignment.** Supervised embodied datasets are limited due to high collection cost and safety constraints. Unsupervised approaches such as Genie-style video-based learning infer latent dynamics from large-scale unlabeled corpora (Bruce et al., 2024a), but aligning inferred latent actions with true causal interventions remains an open problem.

■ **Security and Safety.** When world models are deployed in physical systems, vulnerabilities such as trojan and backdoor attacks can induce unsafe or adversarial behaviors. Vision-language robotic systems are particularly susceptible to covert manipulation, highlighting the need for robust, security-aware training and deployment protocols (Wang et al., 2024d).

■ **Evaluation Limitations.** Standard pixel-level or likelihood-based metrics fail to capture causal reasoning, temporal coherence, and action-conditioned prediction. Semantics-aware evaluators such as UNIVERSE (Hendriksen et al., 2025) and relational benchmarks like Open-World Scene Graph Generation (Dutta et al., 2025) provide improved assessment. Broader benchmarks including EWOK (Ivanova et al., 2025) and MMWorld (He et al., 2024) reveal persistent gaps in multimodal world understanding, motivating multidimensional evaluation protocols.

■ **Benchmark Fragmentation.** Existing evaluation suites remain fragmented across modalities and tasks. Systems such as WorldSimBench (Qin et al., 2025), WorldScore (Duan et al., 2025a), VBench (Huang et al., 2024), VBench-2.0 (Zheng

et al., 2025a), and T2V-CompBench (Sun et al., 2025) focus on specific aspects of generative or simulation quality, but lack unified standards for assessing state tracking, causality, and long-horizon prediction.

■ **Long-Horizon Consistency.** Maintaining coherent predictions over extended temporal horizons remains difficult. Error accumulation in latent rollouts leads to drift, degraded state representations, and failure in planning scenarios, particularly in partially observable environments.

■ **State Persistence and Memory.** World models require persistent internal representations to track entities, objects, and interactions over time. Current architectures often rely on limited context windows or implicit memory, which constrains their ability to maintain consistent world states across long sequences.

■ **Multimodal Alignment.** Integrating vision, language, and action modalities introduces alignment challenges, as representations learned from different data sources may not share consistent semantics or temporal structure. Misalignment degrades both prediction accuracy and interpretability.

■ **Scalability vs. Interpretability Trade-off.** Highly scalable neural architectures often sacrifice interpretability, while structured or symbolic approaches improve transparency at the cost of increased design complexity. Balancing these competing objectives remains a central challenge in world model development.

## G Reflections and Future Outlook

Advancing world models requires coordinated progress across representation, learning, robustness, and evaluation, particularly in the context of Large Language and Vision-Language Models.

### G.1 Reflections and Future Outlook

■ **Causal pretraining objectives.** Current autoregressive objectives minimize token-level cross-entropy, which is a necessary but not sufficient condition for causal world modeling. A concrete open question is: *Can an auxiliary objective—e.g., do-calculus consistency loss or counterfactual invariance regularization—be integrated into standard next-token training at scale without prohibitive compute overhead?* The trade-off is between added gradient computation and the degree to which the learned transition operator $\hat{g}_\theta$ generalizes under intervention rather than under distributional shift alone.

■ **Latent state standardization.** Pixel-space models, JEPA variants, and autoregressive LLMs all learn internal state representations, but these are incompatible across architectures, preventing modular composition. A concrete direction is defining a *world state interface*: a shared latent format with documented dimensionality, temporal resolution, and grounding protocol, enabling a JEPA encoder to supply $z_t$ to an LLM-based planner. The trade-off is between representational expressivity and interoperability.

■ **Evaluation beyond pixel metrics.** The most critical near-term bottleneck is evaluation. Existing benchmarks measure generation quality (FID, SSIM) or task success rates, neither of which captures state-tracking accuracy or causal consistency under counterfactual interventions. Developing an *intervention benchmark*—where ground-truth outcomes of $do(a_t)$ operations are known—is a prerequisite for distinguishing genuine world modeling from correlational sequence prediction. The trade-off is between benchmark construction cost and the specificity of what the benchmark measures.

■ **Scalable neuro-symbolic grounding.** Programmatic world models achieve the highest controllability and interpretability but do not scale to open-domain visual inputs. A concrete research direction is *differentiable predicate induction*: learning symbolic transition rules jointly with neural perception so that the structured transition operator $T(z_t, a_t)$ is discoverable from data rather than hand-specified. The trade-off is between the expressivity of learned predicates and the difficulty of gradient propagation through discrete symbolic structures.

■ **Hybrid Architectures.** Future systems are likely to integrate neural generative models with symbolic or structured constraints to improve causal consistency and controllability. Neuro-symbolic approaches and code-coupled models demonstrate that combining learned representations with explicit structure enhances interpretability and planning reliability, while preserving the flexibility of generative models (Xiong et al., 2025; Tang et al., 2024).

■ **Scalable Self-Supervised Learning.** Efficient learning from large-scale unlabeled data remains central to overcoming data bottlenecks. Video-based self-supervised approaches, including generative interactive environments such as Genie, show that latent dynamics can be learned without explicit supervision, enabling scalable acquisition of

world knowledge (Bruce et al., 2024a). Continued progress in multimodal pretraining is expected to further improve generalization and transfer.

■ **Robustness and Security-Aware Design.** As world models are deployed in safety-critical settings, robustness to adversarial perturbations and backdoor attacks becomes essential. Empirical evidence of vulnerabilities in vision-language robotic systems highlights the need for security-aware training, verification, and deployment pipelines (Wang et al., 2024d). Incorporating robustness objectives into model design remains an open research direction.

■ **Multimodal and Causal Evaluation.** Reliable assessment requires evaluation frameworks that capture temporal coherence, causal reasoning, and counterfactual validity. Emerging benchmarks such as UNIVERSE (Hendriksen et al., 2025), OpenWorld Scene Graph Generation (Dutta et al., 2025), EWOK (Ivanova et al., 2025), and MMWorld (He et al., 2024) highlight limitations of existing metrics and motivate multidimensional evaluation protocols.

■ **Long-Horizon Prediction and Memory.** Improving long-horizon consistency requires architectures that maintain persistent state representations and mitigate error accumulation during rollout. Advances in latent state modeling and memory mechanisms are critical for stable multi-step prediction and planning.

■ **Integration with Foundation Models.** A key direction is the tighter integration of world modeling objectives with large-scale foundation models, particularly LLMs and VLMs. While autoregressive transformers exhibit emergent reasoning capabilities, extending them with explicit dynamics modeling remains necessary for robust physical and interactive reasoning (Thomas, 2025).

Collectively, these directions position world models as increasingly scalable, robust, and systematically evaluated systems, capable of supporting embodied intelligence and complex multimodal interaction in real-world environments.