

From Specialist to Generalist: A Comprehensive Survey on World Models

Kai Xu¹(✉), Hang Zhao^{1,2}, Ruizhen Hu³, Yuhang Huang^{1,4}, Ziqiao Zhou^{1,2}, Wancheng Feng¹, Li Yi⁵, Sida Peng⁶, Xing Liu⁷, Zihao Liu⁷, Jin Zhang⁷, Chenyang Zhu⁴, Renjiao Yi⁴, Qin Zou², Bo Du², and Haibin Yu¹

© The Author(s)

Abstract World models endow artificial agents with the internal predictive capabilities necessary to anticipate future states so as to act purposefully. While theoretical results underscore the necessity of world models for general tasking capability, implementing them involves navigating complex challenges in high-dimensional dynamics and compounding errors over long horizons. Currently, no existing approach simultaneously attains both precision and generalization, creating a divide between specialist and generalist models. In this survey, we systematically review the rapidly evolving field of world models through a distinct lens: Specialist versus Generalist. Unlike existing reviews, we frame the literature as a technical continuum spanning explicit physics-based priors to implicit data-driven learning. A key insight from our analysis is that despite the evolutionary trend toward generalist architectures, specialist and generalist paradigms are destined to coexist. We demonstrate that this persistence stems from a fundamental trade-off between the high precision required for control tasks and the broad adaptability needed for open-ended environments. By critically analyzing the strengths, limitations, and practical applications across this spectrum, we identify the open challenges hindering widespread deployment and propose a research roadmap to reconcile accuracy with transferability—a synergy essential for realizing Artificial General Intelligence (AGI).

Keywords World Model; Simulation; World Foundation Model; Physical Intelligence; Artificial General Intelligence

1 Introduction

World models [1] provide an agent with internal predictive capabilities of how the world evolves, enabling them to anticipate future states under control inputs and hence to act purposefully. The concept can trace back to the dynamics model of a system being controlled in the control theory, where a mathematical model of the system's dynamics can

be used either for direct planning or for learning a control policy. The model can usually be obtained with system identification [2] through fitting observed pairs of action and state. World models are also closely related to internal mental models in cognitive science [3], which refer to the predictive mechanisms that humans and animals estimate the consequences of their own actions. In robotics and artificial intelligence, world models [4–9] have been employed to enable agents to learn from experience, plan actions, and generalize across tasks.

Although world models and their related concepts have long existed, the past few years have witnessed a remarkable resurgence of interest. Several converging trends contribute to this: the rapid development of model-based reinforcement learning (RL) [10], the availability of large-scale real-world data [11–14], and advances in learnable architectures capable of capturing complex dynamics [15, 16]. Progress in foundation models [17] has further underscored the predictive modeling at scale [18]. A notable milestone is the emergence of Sora [19], which generates diverse and physically consistent video content conditioned from textual descriptions, sparking tremendous enthusiasm across the community. Together, these developments suggest that approximating real-world dynamics is entering a new stage of maturity, and the way toward open-world agents is beginning to come into view [20].

World models offer both the necessity and the effective-

- 1 Institute of AI for Industries, Chinese Academy of Sciences, Nanjing, 211135, China. E-mail: kxu@iaii.ac.cn (✉).
- 2 Wuhan University, Wuhan, 430072, China.
- 3 Shenzhen University, Shenzhen, 518060, China.
- 4 National University of Defense Technology, Changsha, 410073, China.
- 5 Tsinghua University, Beijing, 100084, China.
- 6 Zhejiang University, Hangzhou, 310027, China.
- 7 Northwestern Polytechnical University, Xi'an, 710072, China.

Manuscript received: 2025-01-01; accepted: 2025-01-01

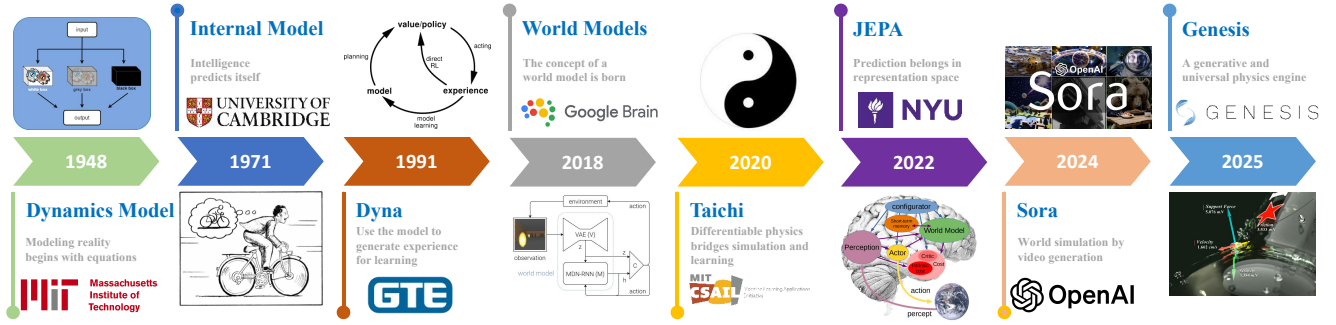


Fig. 1 The significant progress in the field of world models.

ness required for artificial intelligence. Existing theoretical results establish that any agent that achieves competent performance for a sufficiently diverse set of goal-directed tasks must have learned an accurate predictive model of its environment [20, 21], demonstrating the necessity. Beyond this, world models are adequately effective. World models only need task-agnostic experience while providing a reusable predictive environment in which policies can be trained for diverse tasks and goals [22–24], saving the collection of high-quality, task-specific demonstrations [25] which are extremely costly in the real world. Moreover, world models usually generalize better than policies learned directly from imitating demonstrations [20]. Therefore, world models can synthesize additional trajectories beyond what is observed, enabling more sample-efficient policy learning compared with model-free approaches [8, 26].

Despite their importance, learning effective world models remains challenging. It must reason over high-dimensional observations while precisely capturing the complex underlying dynamics that govern real-world systems. Difficulties also arise during model inference: long-horizon rollouts tend to compound errors and drift away from the true dynamics [27]; agents frequently encounter out-of-distribution states during exploration [5] and hence generalizing across environments remains challenging [28]. To contextualize these problems, we examine the full spectrum of world model research, from the classical system identification [29] for specialist world modeling to the emerging world foundation models (WFMs) towards generalist [19, 30], as shown in Fig. 1. A comprehensive survey of the literature reveals a key insight: while there is a clear evolutionary trend from specialist to generalist world models, these paradigms are destined to coexist. This persistence stems from their distinct roles and characteristics in facilitating decision-making. Specialist models achieve high precision within narrow domains but face challenges in construction and transferability, rendering them ideally suited for precise control tasks. Generalist models, in contrast,

offer superior adaptability across environments but currently compromise on fine-grained accuracy. As a result, they are best suited for tasks with relaxed precision constraints, or otherwise require fine-tuning to ensure effective performance in the target domain.

The ultimate objective of world model research is to reconcile high accuracy with broad transferability, a synergy essential for realizing Artificial General Intelligence (AGI). To accelerate progress toward this vision, we present a comprehensive survey of this rapidly advancing field. While prior reviews have explored world models [31–41], our work distinguishes itself by systematically analyzing the literature through the specific lens of the “specialist versus generalist” paradigm. This perspective reveals a distinct technical continuum, ranging from explicit physics-based modeling to implicit data-driven learning. Accordingly, we critically analyze the respective strengths, limitations, and practical applications of methods across this spectrum. In the end, we identify the key open challenges currently hindering the widespread deployment of world models and propose a roadmap of future research directions to bridge these gaps.

This survey is organized as follows. We formalize the definition of world models in Sec. 2 and discuss the characteristics of specialist and generalist world models, along with their connections to physics-based and data-driven techniques. The essential components of a world model, including state representations, action representations, and the dynamic architectures, are carefully summarized in Sec. 3. Building on these components, Sec. 4 reviews the methodologies for learning both specialist and generalist world models, while their mainstream usages and applications are reviewed in Sec. 5 and Sec. 6, respectively. Finally, Sec. 7 outlines open challenges still to be addressed in the field.

2 Specialist vs. Generalist World Model

As world models become more and more prevalent, their formulations span a wide spectrum and have inspired many

insightful taxonomies [31–41] to discuss their discrepancies. We adopt a conceptual division between *Specialist* and *Generalist* world models to clarify their differing design principles and characteristics. In this section, we first outline the definitions of world models, then provide a detailed comparison between specialists and generalists, as well as introduce their connections to physics-based and data-driven techniques.

2.1 Concept and Definition of World Model

World models provide an internal representation of how the environment evolves in response to actions. Formally, it can be expressed as:

$$\mathbf{s}_{t+1} \sim \mathcal{W}_{\theta}(\mathbf{s}_{t-h:t}, a_t), \quad (1)$$

Here, state \mathbf{s}_t encapsulates the environment’s status at time t , action a_t represents the control input, and \mathcal{W}_{θ} denotes the world model parameterized by θ that predicts how the state evolves from time t to $t + 1$ given recent state history $\mathbf{s}_{t-h:t}$ and the current action a_t .

While the above captures the general formulation, world models in practice take diverse forms depending on how states and actions are represented and how dynamics are modeled, spanning from classical physics simulations to fully learning-based predictive systems. To provide a unified perspective, we categorize world models along a key axis of specialist versus generalist, as demonstrated in Fig. 2. This distinction centers on their accuracy and scalability, reflecting whether a model is tailored to a narrow set of related environments or designed for broad generalization. The following sections introduce the characteristics of specialist and generalist models.

2.2 Specialist World Model

A class of world models focuses on modeling the dynamics of a single or a narrow set of related environments to guarantee high accuracy, which we refer to as *Specialists*. These models typically rely on targeted priors or dedicated data collection to capture specific environmental behavior. Physics simulation [42, 43] exemplify this category. Although the underlying physical laws are universal, deploying a simulator requires identifying specific instantiations that effectively tailor the model to a particular environment. Learned-based specialists [44, 45] follow a similar philosophy, with their training data, architectures, and parameterizations tightly coupled to the target environment.

The primary advantage of specialists lies in their precision and within-domain generalization [46]. Here, “domain” refers to a closely related set of environments sharing similar dynamics. Their accuracy stems from physical rules or

carefully collected data that thoroughly cover the state-action space, making them excel in scenarios requiring accurate control, such as industrial automation [47–49]. However, this specialization comes at the cost of limited cross-domain applicability; transferring the model to a new environment setting often requires substantial redesign, data recollection, or retraining. As a result, specialists usually lack the rapid adaptation capabilities needed for open-world applications.

2.3 Generalist World Model

There also exists a class of world models that aims to approximate the dynamics of an open-ended distribution of environments [8, 50, 51], which we refer to as generalists. In current practice, generalist models are typically built by training foundation models [17] on large-scale datasets spanning multiple environments, with their scaling laws already verified [18]. Though future generalists may also emerge from alternative formulations, such as physics engines equipped with fast and automated system identification, we focus on data-driven generalists in this survey, as they represent the mainstream direction in recent years. The broad data coverage and scalable architectures of such models together contribute to capturing the common patterns of physical interaction, allowing a unified model to predict dynamics across varied contexts. Such unified world models enable strong generalization and even zero-shot adaptation in the open world.

The strong cross-domain generalization of generalists makes them well-suited for open-world scenarios [52–54]. However, these benefits come with trade-offs. Generalists often sacrifice within-domain precision compared with carefully engineered specialists. These limitations propagate into policy learning and degrade downstream control performance. Ensuring physical consistency across many environments remains challenging. Despite these limitations, generalist world models represent a significant step toward unified, scalable, and adaptable models of the physical world, laying the foundation for more flexible and capable agents.

2.4 Techniques behind Specialist and Generalist

In summary, specialists prioritize high accuracy and within-domain generalization, whereas generalists emphasize scalability and cross-domain generalization [55]. Future progress will likely involve integrating both strengths. To better understand how these capabilities arise, we examine the major technical pathways for building world models. The primary factor distinguishing a specialist from a generalist lies in how they leverage physics prior and data. Models grounded in established physics or trained on narrow, domain-specific

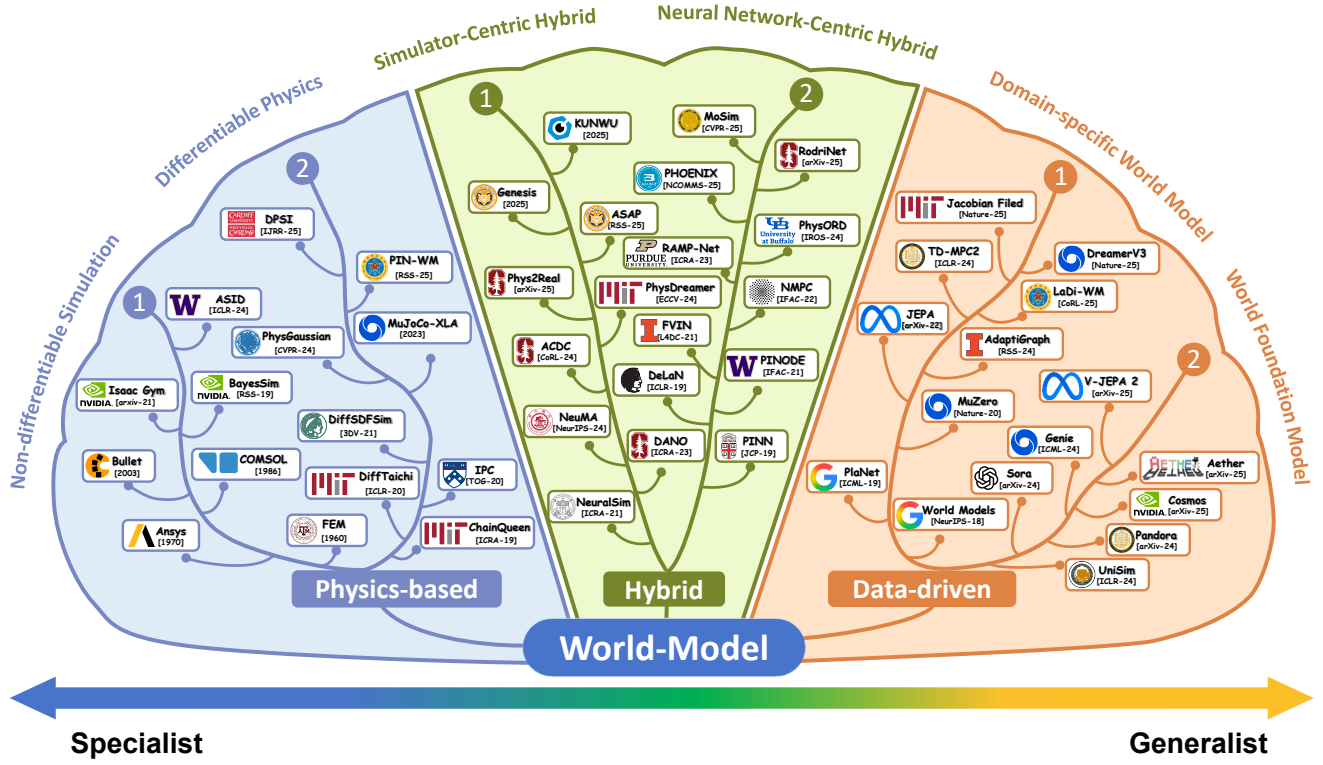


Fig. 2 Taxonomy of world models based on their characteristics as specialists or generalists. Depending on how they incorporate physics priors and data, they can be categorized into physics-based, neural-physics hybrid, and data-driven models. Models closer to the physics-based end behave as high-accuracy specialists, whereas data-driven models function as scalable generalists.

datasets naturally exhibit specialist behavior, while models designed to ingest broad, heterogeneous data at scale tend to function as generalists.

Building on this perspective, we classify existing world models into three major families: physics-based, neural-physics hybrid, and data-driven world models, with the taxonomy illustrated in Fig. 2. Physics-based models rely on explicit physical laws. While these laws are universal, applying them requires precise knowledge of environmental parameters, e.g., mass and friction. Because acquiring these parameters for arbitrary open-world objects is difficult [29], these models are typically constrained to controlled, well-defined domains, making them specialists. Data-driven models learn dynamics directly from observations. While small-scale data-driven models can be specialists, this paradigm uniquely benefits from scaling laws [18]. By ingesting massive, heterogeneous datasets, these models learn implicit representations that bypass the need for explicit parameter identification, allowing them to function as generalists in open-ended environments. Between these two extremes, a line of work lies on neural-physics hybrids, striking a balance by either augmenting physics simulation with neural components [56] or enhancing neural models with physical priors [57].

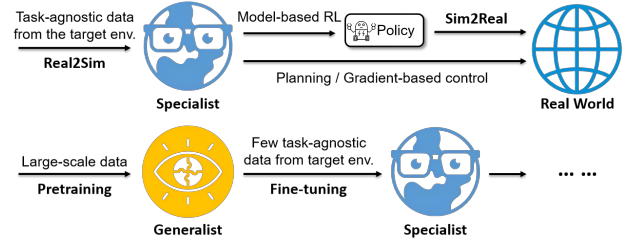


Fig. 3 An example of transferring a generalist world model to a specialist one. A specialist can be obtained by directly training on task-agnostic data from the target environment. The resulting Real2Sim model can then be leveraged for efficient Sim2Real control. Alternatively, it can be obtained by adapting a pre-trained generalist using only a small amount of task-agnostic interaction data.

It is important to note that specialists and generalists do not form a strict dichotomy. With appropriate techniques, the strengths of both can be leveraged to complement each other. For example, to build a high-accuracy specialist for a target environment, one may first train a general WFM and then fine-tune it using a small amount of task-agnostic interaction data [58–62], as illustrated in Fig. 3. This approach leverages the broad adaptability of the generalist model while enhancing predictive precision in a target environment. Moreover, specialists can also advance generalists. For instance, Cosmos [30] enhances the physical plausibility of its WFM

by generating diverse and high-quality data through multiple specialized physics simulators. Together, these complementary pathways suggest that future progress will hinge not on choosing either approach, but on effectively integrating them to pursue the goal of both accuracy and generalization.

3 Representation and Architecture

The key components of world models differ, including state s , action a , and system dynamics \mathcal{W} , which coherently influence whether a model behaves more like a specialist or a generalist. In this section, we discuss world model design across state representations, action representations, and dynamics architectures. We identify two overarching philosophies: *physics-grounded* designs, which explicitly align with physical constraints and provide strong interpretability, and *learning-included* designs, which rely on learning to understand abstracted representations and offer greater scalability.

3.1 State Representation

State representation s defines how the environment is described and perceived, ranging from raw observable variables to high-dimensional latent encodings. Physics-grounded representations preserve explicit physical meaning and are compatible with both physics-based and data-driven world models for accurate reasoning. In contrast, learning-induced representations derive compact abstractions from data, enhancing generalization, but temporarily cannot be understood by physical engines directly.

3.1.1 Physically Interpretable State

Physics-grounded state representations are typically raw observation variables with explicit meaning or hand-crafted data structures extracted from sensory inputs. We introduce several common types, including raw configuration, sensor observations, and 3D geometric representations.

Raw Configuration Raw configuration refers to variables that directly describe the system’s condition, such as positions, velocities, or forces, which are typically compact and low-dimensional. Their key advantage lies in being physically meaningful and compact, allowing physics-based methods to directly model them [63, 64] and enabling data-driven approaches to learn efficiently and reason stably due to the small dimensionality. However, the main limitations are that such informative configurations often require specialized sensors or perceptual modules [65, 66] for acquisition, and sometimes lack the sufficiency to handle more complex tasks which require rich perception for environment understanding.

Sensory Observation Sensory observations are high-dimensional data obtained directly from sensing modalities, such as RGB images, depth maps, and tactile readings. Unlike raw configurations describing the underlying physics in a compact form, sensory observations provide richer semantic information about the environment, are more accessible through common sensors, and are more capable of capturing complex scenes. However, such high-dimensional observations often give rise to other challenges, i.e., increasing the difficulty for physics-based methods to model them directly [67, 68], and requiring data-driven approaches to handle substantial perceptual redundancy [4, 44], making it harder to focus on the underlying physical dynamics.

Geometric Representation Geometric representations explicitly provide the spatial descriptions using structured 3D description formats such as point clouds [69, 70], signed distance fields [71, 72], or meshes [9, 46], built upon raw sensor observations. Geometry is suitable for tasks requiring simulation primitives such as collision detection, contact dynamics, and friction modeling. They also serve as informative inputs for data-driven model learning [73, 74] when spatial reasoning is needed. However, constructing geometric representations typically requires additional processing or reconstruction algorithms [75, 76], which can be computationally expensive. Moreover, the geometry processing demands high accuracy, as imprecise geometry may compromise physical reasoning and lead to accumulating errors [46].

3.1.2 Learnable Compact State

While interpretable and meaningful, physically grounded states may introduce either insufficiency or redundancy, which are hard to balance. In contrast, learning-based representations derive compact abstractions that help world models to focus on essential dynamics. We discuss several representative types, including object-centric, relational, and latent representations.

Object-Centric Representation Object-centric representations treat a scene as a collection of discrete entities, each with its own state. This decomposition aligns with the compositional structure of the physical world, making it easier for a world model to track, reason about, and predict interactions. It also offers superior compositional generalization compared to global modeling alternatives. Early unsupervised approaches OP3 [77] and O2P2 [78] demonstrate that entity-abstraction enables strong generalization to unseen object counts and configurations in block-stacking tasks. More recently, compositional NeRF-based frameworks [79] combine per-object latent codes with GNN dynamics, yielding stable

long-horizon predictions for multi-object scenes. However, these methods rely on accurate entity discovery and consistent object tracking, limiting reliability in spatially complex scenes, such as occlusion or visual clutter.

Relational Representation Relational representations describe scene dynamics through structured relations, typically instantiated as graphs. Graph Neural Networks (GNNs) [80, 81] are commonly used to perform relational reasoning and dynamics prediction on these graphs. A broad set of works follow this paradigm by representing objects or particles as nodes and using message passing to model their interactions, enabling robust prediction in multi-object [79, 82] and deformable [83, 84] settings. Relational representations depend on the reliability of the constructed graph, and relations inferred from heuristic or weakly supervised cues may result in incorrect or missing edges. When topology changes rapidly, the graph structure becomes harder to maintain consistently, which affects the model’s accuracy.

Latent Representation Latent representations are compact hidden states that encode the task-relevant structure of high-dimensional observations, enabling world models to predict and plan efficiently without operating directly on raw sensory data. From Dreamer’s stochastic latent rollouts [44, 85] to decoder-free models like JEPA [86], latent-space world models abandon pixel reconstruction in favor of compact representations, enabling scalable dynamics learning and efficient planning directly in latent space. Despite its promise, its effectiveness remains constrained by the invariances of the underlying encoder; any information discarded or overly abstracted at the encoding stage directly limits the fidelity of the latent dynamics. Learning a sufficiently powerful and semantically rich representation is itself a major challenge.

3.2 Action Representation

Action representation a defines interaction or control signals that influence future environment states. Some action corresponds to physically executable commands, while the other relies on high-level abstractions. This distinction reflects a trade-off between controllability and flexibility, which we will elaborate on below.

3.2.1 Physically Executable Action

Physically executable actions specify control parameters that an agent can directly understand and respond to. We primarily discuss motion, force/torque, and process parameters.

Motion Parameter Motion parameters refer to control signals that specify the kinematic response of an agent, such as positions and velocities of joints or end-effectors. For physics-based world models, motion parameters naturally integrate with rigid-body or continuum kinematic formulations [63], while for data-driven models, their low dimensionality facilitates efficient learning and stable inference [87]. However, low-level motion commands alone are often insufficient for modeling complex interaction dynamics [88]. For example, when handling fragile objects, relying solely on position control may still generate excessive contact forces, potentially leading to damage despite accurate motion execution.

Force/Torque Parameter Force and torque parameters are action commands that explicitly regulate the interaction forces applied to the environment. Compared with purely motion actions, they enable finer control in contact-rich or compliant manipulation scenarios, such as assembly [89] and polishing [90], where interaction safety is critical and cannot be ensured through kinematics alone. From a modeling perspective, physics-based world models naturally align force/torque commands with analytical dynamics formulations [63, 71], whereas for data-driven approaches, such signals are often scarce and the underlying contacts are difficult to fully capture, making this an ongoing challenge.

Process Parameter Process parameters refer to action commands that regulate operating conditions, typically appearing in manufacturing scenarios. For example, in welding tasks, welding voltage and current parameters directly affect the melt pool dynamics and weld quality [91]; in polishing tasks, polishing pressure and speed parameters will impact material removal rates and surface finish [92]. They often depend on complex, nonlinear, and multi-physics coupling mechanisms and require domain expertise or empirical tuning. From a modeling perspective, physics-based world models offer interpretable formulations when the underlying process physics is well understood, but may struggle with heterogeneous materials or other complexities where physical laws are complex or unknown. Data-driven world models can leverage collected data for approximation, yet typically require large-scale and high-quality datasets, which are often hard to obtain, especially when involving irreversible transformations or high-cost trials.

3.2.2 Learnable Abstract Action

Some world models employ abstract actions that encode high-level intent or latent control patterns rather than direct actuator commands. This enables semantic-level control across diverse

tasks, but also forces the world model to infer meaning from ambiguous or underspecified inputs, making such actions less common and harder to use. We introduce representative forms of textual instructions and latent actions.

Latent Action Latent actions are abstract codes learned from data that are mapped into executable behaviors through learnable decoders. This formulation is often necessary when datasets lack action labels [58, 93], forcing the model to infer a latent that still captures the behavior-controlling factors needed for prediction. Genie [93] learns discrete latent actions from unlabeled Internet videos, enabling action-controllable generation without ground-truth labels. This paradigm quickly gains traction in recent works: WorldEval [94] maps real robot actions into the same latent space for policy rollouts, UniVLA [95] learns task-centric latent actions from heterogeneous embodiment videos, and CoLA [96] apply latent action layers to frozen LLMs to improve controllability. Latent actions enable world models to leverage large-scale unlabeled web video. Despite promising, their abstract nature reduces interpretability and is often heavily compressed, which often limits control granularity [93].

Textual Instruction Textual instructions are semantic control signals expressed in natural language that specify goals or intent for a world model to predict, typically requiring strong language understanding and thus commonly appearing in world models integrated with large language models. Recent models, from language-conditioned video generators like Sora [97] and Goku [98], to interactive systems like Genie [93], demonstrate controllable video rollouts and accurate long-horizon predictions from textual goals. Textual instructions provide a highly flexible and scalable interface for specifying diverse tasks and objectives that are also interpretable to humans. However, natural language is also inherently vague, making the mapping from text to precise low-level dynamics ill-defined and unreliable. Its weak controllability for fine-grained tasks still poses challenges.

3.3 Architecture

The architecture \mathcal{W} of a world model defines how dynamics are captured and predicted, from physics simulation to neural predictors. Some architectures directly build physical rules into the computational structure, while others learn dynamics from data without predefined constraints. We examine corresponding representatives.

3.3.1 Physically Grounded Dynamics

Physically grounded architectures involve simulation in which physical laws are strictly guaranteed. This area has blossomed into a rich and vibrant landscape of methods, and we highlight several representative approaches. Depending on how the physical world is discretized, they can be broadly categorized into grid-based, particle-based, and multibody simulation.

Grid-based Simulation Grid-based simulation discretizes the physical domain into structured grids to solve partial differential equations that govern dynamics. They are particularly effective for modeling continuous fields with high numerical accuracy, such as stress-strain responses, heat transfer, and fluid flows. Classical methods such as the Finite Element Method (FEM) [42] provide a rigorous framework for continuum mechanics, enabling accuracy backed by well-established solvers and material models. Despite strong physical fidelity, they rely on fixed meshes, making it challenging to handle large deformations or topological changes. In addition, the careful meshing, extensive computational resources, and precise material parameterization, also limit the practicality in dynamic or poorly modeled environments.

Particle-based Simulation Particle-based simulation represents scenarios as a set of discrete particles, which overcomes the limitations of grid-based methods in handling deformations. Representative techniques span a spectrum of accuracy-efficiency trade-offs. The Material Point Method (MPM) [99] combines particle and grid representations to support large-strain and fracture phenomena, while Smoothed Particle Hydrodynamics (SPH) [100] provides a mesh-free alternative for fluid simulation, enabling flexible modeling of free-surface flows. For real-time applications, Position-Based Dynamics (PBD) [101] offers a constraint-driven approximation that sacrifices strict physical accuracy for numerical stability. Despite their flexibility, particle-based approaches may require complex kernel or constraint tuning. Besides, simulations can be computationally expensive at high resolutions; otherwise, accuracy degradation may occur.

Multibody Simulation Multibody simulation models the world as collections of rigid or compliant bodies subject to kinematic and contact constraints. To accurately handle contact and friction behaviors, the Linear Complementarity Problem (LCP) [102] provides a principled framework for enforcing non-penetration and friction laws. More recently, the Incremental Potential Contact (IPC) framework [103] further improves physical robustness by ensuring energy

consistency and mitigating geometric artifacts in complex contact scenarios. Multibody simulators are typically efficient, scalable, and easily integrated with control and planning pipelines, making them widely used in robotics for interactive simulation [104, 105]. However, despite recent advances such as IPC that extend support to deformable bodies, multibody simulators are fundamentally optimized for rigid systems and remain less suited for complex continuum phenomena, such as fluid, granular matter, or phase change.

3.3.2 Learning-based Dynamics Model

Physically grounded dynamics demands expertise and hand-crafted pipelines, limiting scalability in complex or partially understood systems. In contrast, learning-based approaches adopt a more flexible perspective, representing system evolution through learned probabilistic functions. This enables richer model forms and broader applicability across heterogeneous environments. We introduce several representatives.

Gaussian Process A Gaussian Process (GP) is a non-parametric Bayesian model that defines a distribution over functions, enabling it to predict system dynamics while quantifying uncertainty [106]. Instead of parameter learning, GPs assume that function values follow a joint Gaussian distribution governed by a kernel. Their strong data efficiency and principled uncertainty estimation make them attractive for low-sample or analytically tractable model-based control. PILCO [107] demonstrates the strength of GPs by learning system dynamics with extremely low data requirements and uncertainty-aware policy optimization. In robotics, GPs have been used for tactile surface modeling [108], providing a simple yet flexible baseline for perception-driven reconstruction. Despite their flexibility, GPs rely heavily on kernel design and struggle to model high-dimensional dynamics, limiting their use in complex environments.

State Space Model State-space models (SSMs) [109] describe a dynamical system by defining a hidden state space and modeling how this state evolves over time and produces observations, typically paired with neural latent representations. RSSM [4] further augments SSM with a recurrent deterministic backbone and stochastic latent variables. RSSM has since been widely adopted in vision-based control [6–8, 45, 85], where accurate latent-space dynamics extrapolation improves sample efficiency and final performance across diverse tasks. Their main limitation is that learned SSMs depend on the latent dynamics; when facing distribution shift or unseen situations, their predictions can drift quickly, limiting reliability in open environments.

Transformer Dynamics Model Transformers are sequence models built on self-attention [110], enabling them to capture long-range dependencies, handle variable-length inputs, and integrate information across modalities. Their strong parallel computation, scalable capacity, and unified modeling of vision, language, and actions make them particularly well-suited for world models, which must reason over long temporal horizons and heterogeneous sensory streams. Recent studies show that a Transformer Dynamics Model (TDM) can adapt to new environments with only a few samples while maintaining accurate predictions [55]. IRIS [111] and TWM [112] have demonstrated that TDMs achieve state-of-the-art sample efficiency on Atari 100k. TransDreamer [113] handles complex visual control tasks requiring long-horizon reasoning. TDMs’ unified handling of multiple modalities and the scalability with data and capacity make them especially suited for learning generalist world models [93, 114–116].

Diffusion World Model Diffusion models [117, 118] are generative models that synthesize data by reversing a gradual noising process, enabling stable training and high-quality generation. Their ability to model complex, multi-modal distributions makes them particularly attractive for world modeling [119]. Across recent studies, diffusion world models consistently demonstrate remarkable performance in long-horizon prediction and downstream control. DWM [120] improves offline RL returns through accurate multi-step prediction in a single pass. DIAMOND [16] reaches new state-of-the-art Atari 100k scores, showing that diffusion rollouts translate directly into stronger control policies. DISTR [119] achieves markedly better continual RL performance by replaying high-return trajectories. However, the generation process of diffusion models typically requires multiple steps of gradual denoising, making the process time-consuming and computationally expensive. This limits their applicability in real-time scenarios.

Large Language Model Large Language Models (LLMs) are large-scale neural networks trained on massive corpora of text, and also on images and videos, that acquire broad semantic knowledge [121]. LLMs can serve as powerful pretrained backbones for learning world models, providing rich priors about real-world dynamics even before task-specific training. In robotics, Genie Envisioner [122] and Cosmos [51] leverage large embodied datasets to equip LLMs with physical interaction priors, enabling zero-shot prediction of complex manipulation and navigation dynamics from language or image prompts, pointing toward scalable generalist world

models. The limitation of such training fashion is that training and deploying these models require significant computational resources. Besides, hallucination phenomena [123] in large models may also affect the accuracy of physical predictions.

4 The Continuum of World Modeling: From Physics-based to Data-driven Method

Based on the definition and core components of world models, we discuss how models are constructed or learned through leveraging physical prior and data. As illustrated in Fig. 2, we organize existing approaches from physics-based, to neural-physics hybrid, and ultimately data-driven paradigms, reflecting the shift from high-fidelity, expert-engineered specialists toward more scalable and adaptable generalists. We next discuss the core features of these paradigms.

4.1 Physics-based World Model

Physics-based world models derive system dynamics directly from physical laws, emphasizing fidelity, interpretability, and reliability. These models are particularly suitable for high-precision scenario modeling where domain knowledge is well understood. Based on their learnability, we categorize them into non-differentiable simulation frameworks and differentiable physics systems.

4.1.1 Non-Differentiable Simulation

Non-differentiable simulation is grounded in numerical integration of physical laws using discrete, event-driven procedures such as collision handling and contact resolution. These operations introduce discontinuities and non-smooth updates, making the simulation inherently non-differentiable, typically constructed by hand-engineered solvers that enforce physical constraints using rule-based computations. Such carefully crafted simulators offer high accuracy but require substantial domain expertise, making them the most representative of specialist world models. We review the corresponding methods and platforms, including both engineering-oriented numerical simulators and general-purpose physics engines, as illustrated in Fig. 4. Building on these established simulators, we also discuss gradient-free optimization techniques that make such models better aligned with real-world dynamics.

Engineering-Oriented Numerical Simulation The earliest and most mature form of computational world modeling originates from engineering-oriented numerical simulation. Their development was primarily driven by industrial demands, where physical experiments are costly, irreversible, or dangerous. These tools typically model physical processes using partial differential equations and solve them through

numerical discretization techniques such as the Finite Element Method (FEM) [42], which provides the mathematical foundation for analyzing continuum mechanics and multi-physics systems. Building on these numerical principles, high-fidelity commercial platforms such as Ansys [124], Abaqus [125], and COMSOL [126], open source toolkits like FEniCS [127] and deal.II [128], enable accurate simulation of structural deformation, fracture mechanics, heat transfer, fluid flow, and coupled physical fields. More recently, NVIDIA Omniverse [129] extends these numerical foundations into GPU-accelerated digital twins, showing how classical solvers can be combined with scalable GPU computation and interactive scene representations to support modern industrial simulation. By explicitly encoding physical laws and solving them with robust numerical schemes, these simulators achieve exceptionally high accuracy and reliability, making them indispensable in engineering applications.

However, engineering-oriented numerical simulation also comes with limitations. Constructing these simulation models typically follows an expert-driven workflow, involving mesh discretization, system identification, solver configuration, and physically informed interpretation of the results. Such workflows require substantial domain knowledge, exhibiting long iteration cycles and limited adaptability. When deployed in dynamic or poorly characterized environments, they struggle to generalize without re-modeling and re-identification. Moreover, high-fidelity numerical solvers are computationally intensive and often require large-scale computing resources to achieve convergence, making them impractical for real-time or iterative applications.

General-Purpose Simulation Engine Compared with engineering-oriented numerical simulators, general-purpose simulation aims to provide faster and more flexible physics modeling for broader applications such as robotics, animation, and virtual reality. Instead of ultra high-fidelity, these engines emphasize computational efficiency, ease of use, and seamless integration with learning-based pipelines. Representative simulators include Bullet [130], MuJoCo [131], and SAPIEN [132]. These engines primarily support rigid-body dynamics, relying on efficient formulations such as LCP-based contact solvers [71] or constraint-based Newton-Euler dynamics [133] to achieve real-time performance, while deformable bodies, fluids, and other complex continua are typically handled only through simplified approximations. More recently, GPU-accelerated simulators, such as Brax [134] and Isaac Gym [43], further push the boundary of efficiency by enabling thousands of environments to run in parallel on a single GPU.

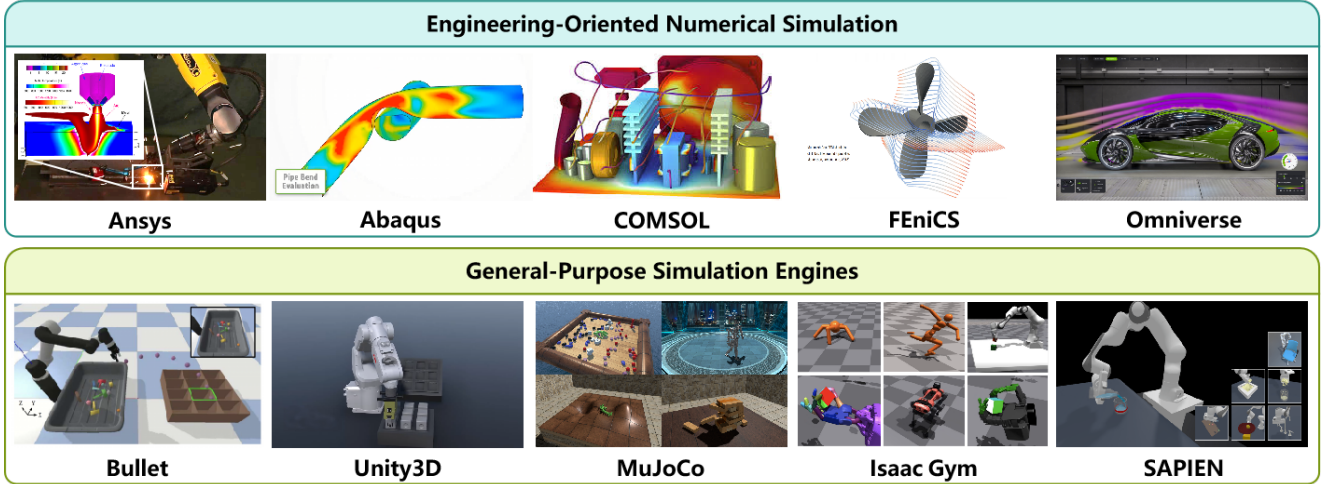


Fig. 4 Examples of two typical categories of physics simulation platforms. Engineering-oriented numerical simulators offer highly precise descriptions of physical processes as well as solutions, but they are difficult to construct, require substantial domain expertise and involve high computational cost. In contrast, general-purpose simulation engines prioritize computational efficiency and ease of use, making them better suited for rapid iteration and integration with learning-based methods, though their physical accuracy is correspondingly reduced.

This not only accelerates the physics computation but also dramatically boosts data sample parallelism for RL, making the policy training [135, 136] more efficient.

The efficiency of general-purpose simulation engines comes at the cost of physical accuracy. To achieve real-time computation, these systems rely on simplified physical assumptions and highly optimized numerical solvers. As a result, contact handling, material properties, and other physical effects are often approximated, which can lead to noticeable discrepancies from real-world behavior. This mismatch manifests as the well-known sim2real gap [137, 138], requiring substantial domain randomization [46, 139] to improve robustness when deploying policies to physical robots.

Learning Dynamics without Differentiability Although regular simulators usually lack analytical gradients, they can still be adapted to match real-world dynamics by treating the simulator as a black-box model and identifying its physical parameters through gradient-free optimization by comparing simulated trajectories with real observations. BayesSim [140] is an early example, framing simulator calibration as likelihood-free Bayesian inference, and it infers a posterior over simulation parameters from real and simulated state-action trajectories. ASID [141] uses an inaccurate simulator itself to design informative exploration policies, executes them on the real robot, and then updates the simulation parameters based on the collected data with Cross-Entropy Method (CEM) [142]. Baumeister et al. [66] introduce an incremental learning framework that continuously refines the

simulator as new real-world data comes in, also adopting CEM for optimization.

Together, these methods show that non-differentiable simulators can be calibrated, bridging part of the sim2real gap and enabling more accurate downstream planning and control. Despite improved ability to match real-world behavior, learning dynamics without differentiability still face inherent limitations. Gradient-free optimization is often sample-inefficient and computationally expensive, especially as the parameter dimension grows [141]. Besides, lacking differentiability prevents the direct use of gradient-based control [143, 144]. These models remain specialists where the simulator structure closely matches the target system.

4.1.2 Differentiable Physics

Recently, the rise of differentiable simulation [15, 145] offers a more effective way for dynamics learning, enabling gradients to flow through simulation for efficient system identification. Achieving differentiability typically requires smooth computational pipelines and carefully designed numerical operators, still more common in general-purpose simulators. We focus on approaches that provide gradients through only physics and methods that simultaneously differentiate Geometry, Appearance, and Physics (GAP) to enable joint simulation optimization. The learning distinction among non-differentiable, physics-only differentiable, and GAP-level differentiable simulations is illustrated in Fig. 5.

Physics-Only Differentiability Incorporating physical priors is essential to learning accurate and generalizable world

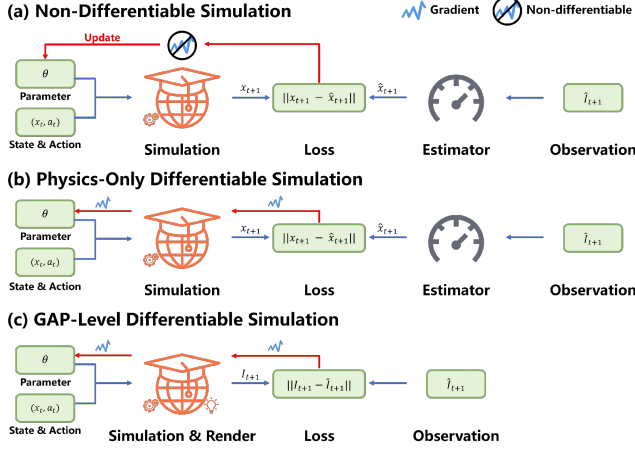


Fig. 5 Different ways of simulation methods that approach real-world dynamics. Requiring sensors or state-estimation modules to convert real-world observations into physical state, (a) non-differentiable simulators fit dynamics via black-box optimization on estimated states and (b) physics-differentiable methods identify parameters by backpropagating errors. (c) GAP-level differentiability incorporates differentiable geometry, appearance, and physics, enabling system identification directly from image-level supervision, removing the need for state estimation and improving scalability.

models [57, 146], where differentiable physics serves as almost a perfect fit. Although the development is relatively recent [15, 145], it has rapidly become a transformative force beyond traditional non-differentiable engines. The breakthrough came with ChainQueen [145], which is the first high-performance differentiable simulator based on MPM [99]. Soon after, DiffTaichi [15] presents a domain-specific language that compiles differentiable simulations to GPUs, allowing complex dynamical systems to be optimized through automatic differentiation. Beyond parallelism, memory efficiency is another key factor limiting large-scale simulation. Quantized simulation addresses this by representing physical states with low-precision formats, enabling substantially larger simulations under fixed memory budgets. A Taichi-based compiler system [147] demonstrates that aggressive quantization can be integrated into general-purpose simulators with minimal performance loss, allowing high-resolution simulations to scale across GPUs and even resource-constrained devices. AutoQuantizer [148] further automates this process by formulating quantization design as a constrained optimization problem, using automatic differentiation to balance simulation error and memory compression. Following, researchers began extending differentiability to a broader class of simulation algorithms [63, 103, 149–151], systematically analyzing the differentiability of contact and friction processes, and exploring strategies such as explicit automatic differentiation and implicit differentiation to handle non-smooth operations in physical simulation. Most recently, differentiable physics

has evolved from individual studies to integrated frameworks. Systems such as Brax [134], Dojo [152], Warp [153], and MuJoCo XLA [154] have emerged, supporting unified differentiable simulations accelerated by GPU computation.

Learning a physics-based world model can be viewed as finding a set of parameters so that the model’s outputs align closely with the observations of the target system. Since most behaviors are described by physics priors, such world models only require a small amount of task-agnostic data [46, 155] to drive accurate dynamics learning and cover the entire state-action space of the target system. The parameters to be identified depend on the type of objects in the scene, such as mass, inertia, and friction for rigid bodies [46], while Young’s modulus, Poisson’s ratio, and density for deformable objects [156]. The loss function is designed to measure the discrepancy between simulated trajectories and observed data, often using metrics such as mean-squared error over observed trajectories [65]. Note that system identification is inherently ill-posed, as multiple parameter sets can explain the same observations [157]. The core criterion for evaluating such models is therefore the accuracy of their predicted dynamics.

We introduce several advanced practices of system identification using differentiable physics, as well as their successful combinations for subsequent model-based control. de Avila Belbute-Peres et al. [63] propose a differentiable simulator constructed upon 2D physics, which analytically differentiates the optimal solution of LCP. This approach was soon adopted by Song et al. [65] to learn the mass and friction distribution of objects to plan actions for sliding objects toward desired targets. DREAM [158] leverages estimated state representations and adopts a differentiable Brax physics engine [134] to infer object masses and driving Gaussian-based scene representations to learn robot control policies. RSR [159] introduces a dual-loop framework combining system identification and policy training, using real-world data to minimize physics loss via gradient descent and an adaptive InfoGap loss to encourage informative exploration, reducing the sim2real gap. Further research [156, 160] extends system identification to soft body by computing the Chamfer Distance between real and simulated point clouds to quantify motion discrepancies. They employ MPM to simulate physics and perform inverse estimation of material parameters such as Young’s modulus and Poisson’s ratio, and further employ RL to enable manipulation of soft bodies such as cloth, ropes, and dough.

Besides identifying physical parameters of the environment, existing works explore estimating parameters of the robot itself, including both physical and control parameters. Lutter et al. [161] identify robots’ link masses and inertia matrices

by minimizing dynamics discrepancies and automatic differentiation, showing that physics-based models outperform black-box approximators in model-based RL. PACE [162] systematically models robot actuators and identifies the robot's intrinsic parameters and corrects for encoder mounting offsets as well as communication and control delays. The resulting sim2real control of legged robots achieves improved energy efficiency and motion accuracy. Degraeve et al. [163] optimize robot control parameters through a differentiable physics engine by minimizing trajectory error.

Physics-only differentiability leverages well-established physical laws to achieve high data efficiency, strong within-domain generalization, and stable long-horizon prediction, making it effective for precise and interpretable control. Despite these strengths, the approach has fundamental limitations. It operates strictly at the state level, reasoning only over positions, velocities, and forces, which prevents it from incorporating richer sensory modalities such as RGB images or depth maps and limits its applicability in visually grounded tasks. In addition, the range of identifiable parameters is narrow and hand-designed. Real-world phenomena such as heterogeneous materials and non-ideal contacts often lie outside the expressive capacity of analytic parameterizations. When the underlying physics formulation is inaccurate or incomplete, differentiability alone cannot compensate for the mismatch, placing a fundamental limit on achievable fidelity.

GAP-Level Differentiability It is highly advantageous to construct a comprehensive representation that jointly models *Geometry*, *Appearance*, and *Physics* (**GAP**), informing intelligent interaction in 3D environments [68, 165]. With the rapid advances in computer graphics [195, 196], jointly differentiable GAP modeling becomes feasible. A differentiable GAP representation enables end-to-end optimization from visual observations to physical parameters, removing the need for separate state estimation modules and improving scalability in approaching real-world dynamics. Tab. 1 summarizes recent advances in fully or partially joint GAP models, followed by a discussion of their differentiable integration.

Beyond physics, computer graphics has already developed a mature foundation for modeling geometry and appearance. Geometry is traditionally modeled using explicit representations such as polygonal meshes [197], point clouds [198], and signed distance fields (SDFs) [199]. Among these, differentiability is mainly achieved through mesh-based differentiable renderers [200–202], which replace discrete rasterization with soft, probabilistic formulations. These methods make pixel colors differentiable with respect to vertex positions, enabling

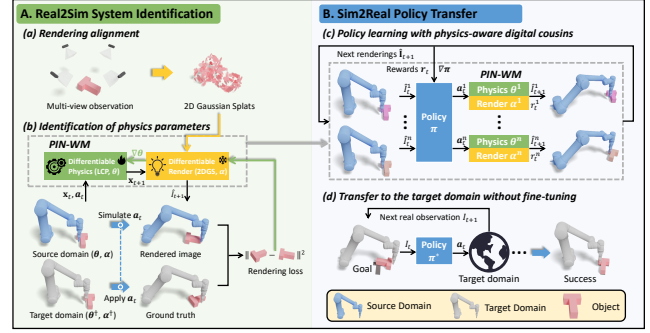


Fig. 6 PIN-WM [46] leverages differentiable rendering and differentiable physics to identify physical parameters from single task-agnostic interaction video data. It further constructs digital cousins to compensate for unmodeled dynamics, enabling the learned policies to transfer directly to real robots without fine-tuning.

gradients to flow to mesh geometry. At the same time, appearance modeling, capturing color, texture, and lighting, is made differentiable through inverse-rendering pipelines that optimize texture maps, BRDF parameters, or lighting models [203–205]. However, separately differentiable approaches for geometry and appearance often lead to inconsistencies between shape and visual effects. These limitations motivate the development of unified representations of neural radiance fields (NeRF) [206] and 3D gaussian splatting (3DGS) [207], which integrate geometry and appearance in continuous volumetric or point-based forms. Notably, the explicit geometric structure in 3DGS makes it particularly well-suited for incorporating physics to describe object interactions and motion. Consequently, a growing body of work has begun integrating 3DGS with differentiable physics engines to reason about complex physical interactions [46, 69, 208], leveraging it as a core building block for joint GAP modeling.

Given the rendering differentiability of joint GAP representations, recent studies have begun directly leveraging image supervision to identify physical parameters, i.e., propagating gradients from rendered images back to physical parameters by minimizing rendering loss [206]. PIN-WM [46] integrates 2DGS [75] to represent object geometry and reconstruct the appearance of target objects, and incorporates the differentiable LCP engine [71] to approximate real dynamics. A physics-aware digital cousin is employed, which randomizes the physical parameters of the target object around the identified values to consider unmodeled dynamics for robust sim2real transfer. Its full pipeline is illustrated in Fig. 6. Besides physical parameters, kinematic parameters [181] such as the joint positions and rotation axes of articulated objects, can also be identified through image supervision. Heiden et al. [67] create digital twins of articulated mechanisms from depth or RGB video, and use RANSAC to infer the types

Table 1 Recent advances in world models that integrate geometry, appearance, and physics uniformly using 3D representations. Env. denotes the environment, including both objects and scenes.

	Method	Geometry	Appearance	Physics	Target	Usage
Rigid body	Zhu et al. [68]	Particle (Known)	PyTorch3D [164]	LCP	Env.	Jointly optimize the geometry, appearance, and physics
	Abou et al. [165]	Particle (3DGS)	3DGS	PBD	Env.	Predict future object states and correct them from observations
	VR-Robo [166]	Mesh (3DGS)	3DGS	Isaac Sim [167]	Env.	Generate photorealistic environments
	SplatMesh [168]	Mesh (3DGS)	3DGS	MuJoCo [131]	Env.	Generate photorealistic environments
	RL-GSBridge [169]	Mesh (Reconstruction)	3DGS	Bullet [170]	Env.	Synchronize the rendering of the simulator with the real world
	DANO [171]	Density field (NeRF)	NeRF	Dojo [172]	Env.	Simulate object motion with density fields
	NeuralSim [56]	Mesh (Known)	N/A	LCP + Neural	Robot	Combine simulation with neural networks to model dynamics
	Robo-GS [9]	Mesh (3DGS)	3DGS	Newton-Euler	Robot	Robot manipulation via motion planning
	Prof. Robot [144]	Particle (3DGS)	3DGS	N/A	Robot	Robot action optimization via differentiable rendering
	DiffGen [173]	Mesh (Known)	Redner [174]	NimblePhysics [175]	Robot	Robot action optimization via differentiable rendering and physics
	DREAM [158]	Mesh (3DGS)	3DGS	Brax [134]	Robot	Robot joint model reconstruction and retargeting
	SplatSim [176]	Particle (3DGS)	3DGS	Bullet [170]	Robot, Env.	Robot manipulation via imitation learning
	GWM [74]	Particle (3DGS)	3DGS	Neural	Robot, Env.	Robot manipulation via imitation learning
	Manigaussian [73]	Particle (Neural)	3DGS	Neural	Robot, Env.	Robot manipulation via model prediction control
	RoboGSim [177]	Mesh (3DGS)	3DGS	Isaac Sim [167]	Robot, Env.	Robot manipulation via reinforcement learning
	ASID [141]	Mesh (Known)	N/A	MuJoCo [131]	Env.	Identify rigid body properties with gradient-free optimization
	PIN-WM [46]	Mesh (Known)	2DGS	LCP	Env.	Identify rigid body properties for robot manipulation
	DREMA [155]	Mesh (3DGS)	3DGS	Bullet [170]	Robot, Env.	Identify rigid body properties for robot manipulation
	Chen et al. [64]	Mesh (Known)	N/A	MuJoCo [131]	Robot	Identify rigid body properties using robot proprioception
	NeRF2Physics [178]	Particle (NeRF)	NeRF	N/A	Env.	Estimate physical parameters from images using LLMs
Articulated object	Phys2Real [179]	Mesh (3DGS)	3DGS	Isaac Sim [167]	Env.	Estimate physical parameters from images using VLMs
	GaussianProperty [180]	Particle (3DGS)	3DGS	N/A	Env.	Estimate physical parameters from images using VLMs
	MonoMobility [181]	Mesh (3DGS)	3DGS	N/A	Env.	Discover motion parts and kinematics from a single-view video
	DexSim2Real ² [182]	Mesh (AIGC)	AIGC	SAPIEN [132]	Env.	Articulated object generation and manipulation
	Heiden et al. [67]	SDF (Reconstruction)	nvdiffrast [183]	TDS [56]	Env.	Determine joint types and physical parameters for articulated objects
	ArtGS [184]	Mesh (3DGS)	3DGS	SAPIEN [132]	Env.	Discover motion parts and kinematics for robot manipulation
	ScrewSplat [185]	Particle (3DGS)	3DGS	N/A	Env.	Joint, smooth optimization of geometry and kinematics
	ArticulatedGS [186]	Mesh (3DGS)	3DGS	Neural	Env.	Learn part shape and appearance while optimizing kinematics
Soft body	Hu et al. [187]	Particle (3DGS)	3DGS	Neural	Robot	Learning a link-level robot model directly from RGB images
	PhysGaussian [188]	Particle (3DGS)	3DGS	MPM	Env.	Integrate dynamics within 3D Gaussians for motion synthesis
	PAC-NeRF [189]	Particle (Sample)	NeRF	MPM	Env.	System identification from videos without known geometry
	NeuMA [69]	Particle (3DGS)	3DGS	MPM + Neural	Env.	System identification from videos without known geometry
	PhysTwin [190]	Particle (3DGS)	3DGS	Warp [153]	Env.	System identification for deformable objects and robot manipulation
	DPSI [156]	Particle (Known)	N/A	MPM	Env.	System identification for deformable objects and robot manipulation
	GenDOM [160]	Grid (Discretization)	N/A	PlasticineLab [191]	Env.	System identification of ropes and cloths for robot manipulation
	Physics3D [192]	Particle (3DGS)	3DGS	MPM	Env.	Learn object physical properties from video generation model
	PhysDreamer [193]	Particle (3DGS)	3DGS	MPM	Env.	Learn object physical properties from video generation model
	Li et al. [194]	N/A	NeRF	Jacobian Field	Robot	Model robot actuation and materials for model prediction control

and parameters of joints connecting rigid bodies and use end-to-end differentiable simulation to estimate kinematic parameters. ScrewSplat [185] integrates 3DGS to reconstruct 3D geometry and segment objects into rigid and movable components. By employing screw theory [209] for continuous parameterization of joint axes, it achieves smooth and end-to-end optimization of both geometric and kinematic components. Research also extends GAP-level differentiability to soft body modeling. PAC-NeRF [189] reconstructs NeRF from multi-view observations and samples particles from the NeRF density field, then couples them with MPM to identify the physical parameters. PhysTwin [190] uses interaction data collected from real robots to identify the physical parameters of soft bodies and train feasible control policies that can be directly deployed on real robots.

Joint GAP representations also facilitate learning proprioception dynamics and control parameters directly from visual

observations. ChainQueen [145] employs MPM to simulate the physical dynamics of soft robots, enabling gradient-based optimization of both controller and material parameters. ∇ Sim [143] introduces a differentiable rendering module, allowing the joint optimization of physical and control parameters through image reconstruction loss, directly enabling learning soft robot control from videos. RISP [210] estimates a quadrotor’s body mass, moment of inertia, and propeller thrust parameters directly from real-world images, enabling the simulated quadrotor’s motion to closely match that observed in real videos. Dr. Robot [211] links a robot’s appearance to its control parameters by modeling canonical geometry with 3DGS and mapping joint angles to surface deformations via skinning techniques. It introduces a learnable function to capture pose-dependent visual changes and optimizes joint angles by minimizing the image reconstruction error. Its follow-up work, Prof. Robot [144], introduces a learnable

collision classifier into the optimization, ensuring that the corrected poses avoid both self-collisions and collisions with the environment.

GAP-level differentiability extends beyond state-level physics with a unified differentiable framework, enabling naturally visually grounded reasoning and reducing reliance on handcrafted perception modules. Despite these advantages, GAP-level methods remain computationally heavy. Joint optimization often becomes prohibitively expensive for large scenes or long sequences. The coupling between rendering and dynamics also produces highly nonconvex objectives, leading to instability or convergence to incorrect solutions.

4.2 Neural-Physics Hybrid World Model

Physics-based world models achieve high fidelity but rely on hand-crafted assumptions, which limit their scalability. Hybrid world models, which combine neural networks with physics priors, provide a promising middle ground. Depending on the primary role of dynamics modeling, they can be mainly classified into simulator-centric hybrid and neural network-centric hybrid. We discuss both techniques next.

4.2.1 Simulator-Centric Hybrid

Simulator-centric hybrid uses neural networks to augment physics engines, improving scalability and flexibility. Enhancements may arise from neural creation, where generative models configure simulation assets and parameters for physics engines, or from neural correction, where neural components are embedded inside the simulation loop to refine physical predictions, as illustrated in Fig. 7.

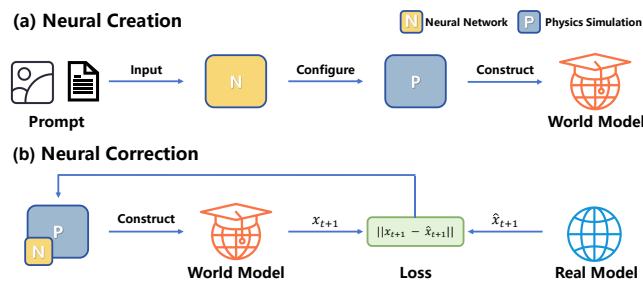


Fig. 7 Simulator-centric hybrids. (a) Neural creation uses generative models to configure simulation. (b) Neural correction embeds learned components inside the simulation to refine physical accuracy.

Neural Creation Neural networks can generate diverse scenes, often conditioned on user instructions, while still relying on the underlying physics engine for dynamics modeling. This design preserves the physical fidelity of classical simulators yet avoids the manual effort of building large numbers of environments. Recently, foundation models [212] have extended GAP representations toward more general and scalable

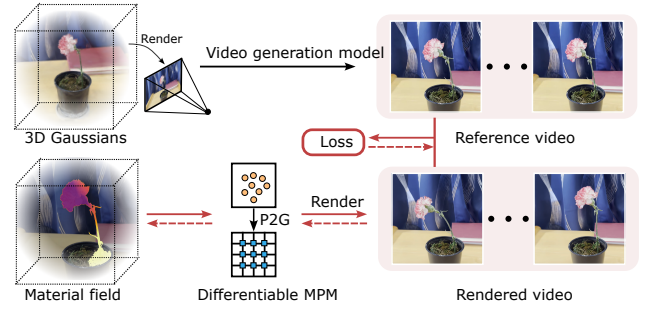


Fig. 8 PhysDreamer [193] estimates materials and dynamics for 3D Gaussian objects by matching differentially rendered MPM simulations to image-to-video-generated motion references.

world modeling frameworks, becoming a key enabler for this direction. For *geometry* and *appearance*, foundation models like TRELIS [213], CAST [214], and Hunyuan3D [215] synthesize 3D content directly from textual or visual prompts. Generative models prioritize semantic richness and diversity guided by learned priors, making them better suited for randomization and data augmentation. ACDC [216] presents an automated pipeline to transform real scenes into diverse and interactive *digital cousin* environments using generative techniques, demonstrating higher transfer success rates compared to purely digital twins [217].

Recent works also leverage generative models for physics modeling. Some works directly infer object physical properties [178–180] from textual or visual descriptions using LLMs or VLMs. Such approaches often rely on the commonsense reasoning of language models to estimate only a coarse range of physical properties. Specifically, LLMPhy [218] performs simulations based on the generated physical parameters and iteratively refines them using error feedback from the simulation results. This method achieves comparable performance to optimization-based methods, but it only supports estimating physical properties for a limited number of objects. Other works leverage video generation models [219] to predict the future dynamics of given image prompts and, by integrating differentiable rendering and physics engines, estimate physical parameters from video priors [192, 193, 220], as PhysDreamer in Fig. 8.

Beyond individual modeling of GAP, generative models are also developed to construct entire simulation scenes. Genesis [221] proposes a generative engine that transforms user-prompted natural language descriptions into simulation scenarios. Genesis integrates various physics solvers and their coupling into a unified framework. This core physics engine is further enhanced by an agent framework [222] that operates at an upper level, aiming towards fully automated data generation for robotics and beyond. Although the concept is still

in its early stages of realization, it represents an exciting and promising direction toward the future of scalable simulation and world modeling. Similarly, KUNWU [223] generates 3D industrial production lines from 2D CAD designs and produces corresponding controllers based on user requirements, forming a complete industrial digital-twin system. This allows industrial tasks to be validated efficiently within low-cost simulated environments, and the control software can be iteratively improved based on the feedback.

Neural-created simulation offers a unique trade-off. By using generative models to create diverse and highly flexible environments, these approaches provide a level of scene variability that conventional simulators cannot achieve without substantial manual engineering. This diversity supports large-scale training and testing under richly varied conditions, while the underlying physics engine still ensures stable and physically plausible dynamics. Compared with purely data-driven world models, externally driven simulation retains grounded physical structure; compared with pure physics simulation, it greatly reduces the cost and expertise to construct large numbers of environments.

Neural Correction Beside creation, neural networks can also be embedded within the simulation loop to correct physical predictions, retaining physical structure while gaining the flexibility to capture hard-to-model dynamics. A prominent class replaces or parameterizes key solver components with neural operators. DANO [171] embeds neural implicit object representations within a differentiable simulator, deriving mass, inertia, and contact forces directly from learned geometry. NeRD [203] goes further by substituting low-level dynamics and contact solvers with neural operators that act as drop-in replacements within traditional engines. These approaches yield hybrid simulators that maintain physical consistency while substantially expanding modeling capacity.

Besides component substitution, another line of work adopts residual correction, where neural networks refine the outputs of physics models rather than replacing them. NeuralSim [56] injects neural residuals into a differentiable rigid-body engine to correct frictional and contact behaviors that classical formulations struggle to model. PAN [224] similarly incorporates residual neural dynamics within a differentiable simulation loop, supporting rapid online adaptation to real-world disturbances. ASAP [225] learns a delta-action residual model to compensate for dynamics mismatch between simulation and real humanoid robots, enabling pre-trained policies to be fine-tuned toward real-world agility. At a deeper level, NeuMA [69] introduces residual learning into constitu-

tive modeling itself: the neural adaptor learns corrections to expert-designed material laws, allowing intrinsic dynamics to be inferred directly from visual observations.

Internally coupled methods intervene directly in the dynamics computation of a simulator, allowing them to correct failure modes or simplifications of analytic physics. These approaches exhibit limitations. Their corrective capacity is ultimately bounded by the structure of the underlying simulator. When the physics formulation is fundamentally misspecified, residual learning or neural module substitution can only partially address the mismatch. Training stability is another challenge, as gradients must propagate through complex and tightly coupled simulation loops. The inclusion of neural components also increases computational overhead and may compromise interpretability, making validation more difficult than in purely analytic simulators.

4.2.2 Neural Network-Centric Hybrid

Neural network-centric hybrids embed physical knowledge into the learning, not only improving data efficiency but also enhancing extrapolation compared to purely neural models. Broadly, methods can be grouped by how physical knowledge is injected, either as hard constraints encoded in the network architecture or as soft constraints imposed through loss functions, as illustrated in Fig. 9.

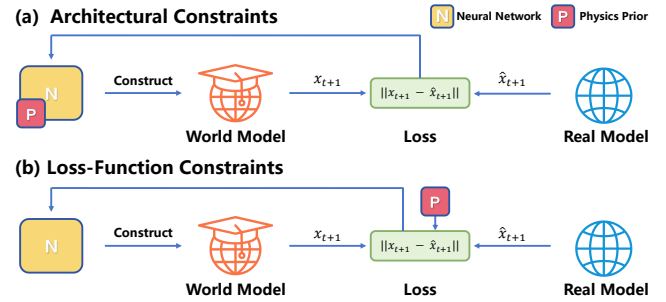


Fig. 9 Neural network-centric hybrids. (a) Architectural constraints encode hard physics by enforcing physical structure directly in the network. (b) Loss-function constraints impose soft physics by penalizing violations during training.

Architectural Constraints End-to-end neural dynamics models can be constructed by shaping their architectures according to physical principles. Rather than simulating physics directly, These networks mirror the governing equations of real mechanical systems, reducing the reliance on simulator construction while preserving strong, physically grounded inductive biases. A major family of structural networks draws from Lagrangian and Hamiltonian mechanics. Deep Lagrangian Networks [226] parameterize kinetic and potential energy with neural networks, ensuring that the resulting

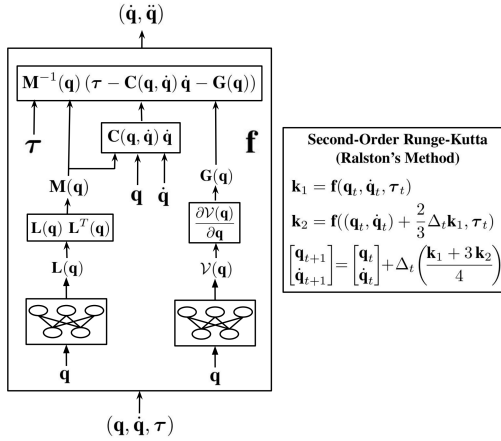


Fig. 10 The Lagrangian Neural Network dynamics model [231], which learns potential energy and a Cholesky-factorized mass matrix to compute accelerations via Lagrangian mechanics, followed by numerical integration to predict the next state.

dynamics obey Lagrange's equations and remain physically plausible even under limited data. Similarly, Hamiltonian Neural Networks [227] enforce canonical Hamiltonian structure, guaranteeing conservation of total energy and time-reversible trajectories. Another influential line of work preserves geometric and symplectic structure through learned integrators. Variational Integrator Networks [228] and their forced variants [229] represent the update rule itself as a neural discretization of the variational principle, yielding long-horizon stability and accurate momentum/energy behavior unavailable to generic recurrent networks. Extensions to Lie groups, such as LieFVIN [230], incorporate the group structure of $SE(3)$ or $SO(3)$ directly into the network architecture, enabling robot dynamics to be modeled with guaranteed geometric consistency. Beyond classical mechanics, researchers have also explored embedding kinematic structure inside the network. Ramesh et al. [231] embed the structure of rigid-body equations directly into the network architecture, as illustrated in Fig. 10, ensuring that learned predictions inherently obey physical laws rather than relying on a purely black-box approximator. RodriNet [232] constructs a learnable Rodrigues operator that generalizes forward kinematics, enabling the network to reason over articulated degrees of freedom while retaining inductive biases unavailable to standard MLPs.

These structure-informed neural dynamics models remain fully differentiable, yet they inherit interpretability and data efficiency from embedded physics structure. Because physical laws are encoded directly within the network architecture, these models often extrapolate more reliably and generalize across tasks with far fewer samples. However, the incorporation of a strict physical structure also introduces limitations. These models require careful architectural design tailored

to specific system families, making them less flexible for complex or multimodal environments. Moreover, fixed physical assumptions can limit expressiveness when real-world phenomena deviate from idealized mechanics. As a result, structure-informed models trade flexibility for fidelity, excelling in well-defined mechanical systems but struggling to scale to diverse, open-world dynamics.

Loss-Function Constraints Another approach to incorporate physical knowledge into neural dynamics is through enforcing loss constraints during training that penalize violations of governing laws, such as PDE/ODE residuals, conservation laws, or energy constraints. The neural network itself remains a flexible function approximator, more easily adapting to specific tasks compared to structure-informed models. A canonical example is physics-informed neural networks (PINNs) [57], which enforce nonlinear PDE residuals within the loss, enabling data-efficient learning even under scarce or noisy observations. Beyond forward prediction, PINNs are also widely used for inverse problems, estimating unknown physical parameters by optimizing them jointly with the network so that the resulting dynamics best satisfy the encoded physical laws. This allows recovering material properties, force fields, or boundary conditions without explicit supervision, a capability particularly valuable when direct measurement is difficult. Later extensions, such as PIN-ODE [233], integrate Lagrangian mechanics into neural ODE training via collocation-based constraints. Lutter et al. [234] introduce a deep network framework based on Lagrangian mechanics, efficiently learning equations of motion while ensuring physical plausibility.

Approaches based on loss function constraints have been shown to be effective in control. RAMP-Net [235] augments the MPC loss with ODE-based physics regularization to enforce robustness under uncertainties. Nicodemus et al. [236] modeled a multi-link manipulator using PINNs, embedding physical knowledge directly into the loss function to enforce the Lagrangian equations of motion. This formulation captures the dynamic relationship between system states and control inputs, enabling accurate trajectory tracking via MPC. MoSim [237] embeds Hamiltonian dynamics into neural ODEs to accurately predict the future physical states of robotic systems. Constraint-based formulations also extend to neuro-symbolic modeling. Works such as PhysORD [238] explicitly embed Euler-Lagrange conservation or friction laws as loss constraints, enabling networks to generalize to highly unstructured off-road environments where purely data-driven models fail. More broadly, constraint-driven models have been

applied to energy conservation [230], dissipation laws [229], and stability or contraction properties for safe control [239], demonstrating the versatility of this paradigm.

Physical constraints reduce data requirements, improve robustness to distribution shifts, and prevent physically implausible predictions. Unlike structure-informed models, which hard-wire mechanics into the architecture and can be restrictive, constraint-based approaches retain full architectural flexibility, allowing the network to capture complex residual dynamics beyond analytic models and offering a practical balance between physical fidelity and expressive power. However, they still require physical laws to be carefully encoded as loss terms, demanding significant domain expertise; this reliance on hand-crafted constraints limits scalability and makes it difficult for such models to generalize across diverse or poorly understood environments.

4.3 Data-Driven World Model

Data-driven world models learn dynamics directly from data without physical priors, relying on diverse and extensive datasets to train expressive architectures that fit observed transitions. With adequate coverage, these models offer strong flexibility to adapt to complex environments. We review these models from domain-specific designs to emerging WFMs.

4.3.1 Domain-Specific World Model

Domain-specific world models learn directly from data collected within a narrow set of related environments distributed in the target domain. Since specific domains allow for rich data collection, these models typically accept fine-grained actions and produce frame-by-frame predictions, enabling precise control. Early approaches emphasize pixel-space prediction, directly forecasting raw sensory streams. Subsequent work shifted toward latent dynamics, compressing observations into compact states that suppress visual redundancy and highlight decision-relevant variability. More recent efforts introduce structured assumptions as inductive biases to simplify dynamics learning. Collectively, this trajectory reflects a move from reproducing observations toward representations organized for control, as illustrated in Fig. 11.

Pixel-Space Prediction A classic line of work models dynamics directly fit high-dimensional pixel observations, mapping current observations and actions to future frames without abstraction. By framing environment prediction as an end-to-end predictive task, these methods rely on the strong approximation capabilities of deep learning architectures to learn visual dynamics purely from observation sequences.

(a) Pixel-Space Prediction



(b) Latent-Space Prediction



(c) Structured Prediction

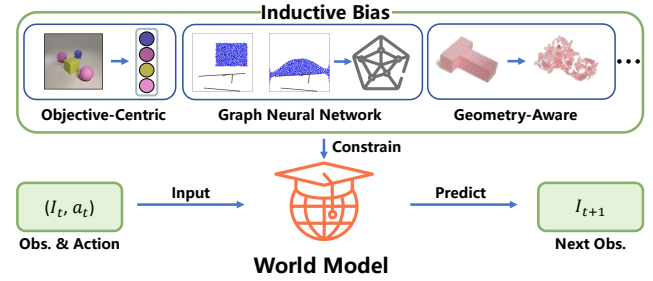


Fig. 11 Domain-specific world models which operate within a narrow family of related environments. (a) Pixel-space prediction models dynamics directly in high-dimensional observations. (b) Latent prediction operates in compact latent spaces to reduce redundancy, while (c) structured prediction introduces inductive biases that simplify dynamics modeling.

Early implementations [240] use CNN-based models to fuse camera images with control signals, predicting future frames to support image-based robot navigation. With the introduction of Transformers, world modeling was recast as a sequence modeling task. Decision Transformer [241] demonstrated that reinforcement learning can be formulated as conditional sequence modeling, naturally accommodating the joint prediction of future states and actions. Building on this perspective, StARformer [242] represents short-horizon state-action-reward tuples and integrates local attention with global temporal modeling to improve long-range consistency in visual control. More recently, diffusion-based world models have gained momentum due to their strong generative fidelity. DIAMOND [16] applies diffusion to Atari environments, showing that retaining fine-grained visual detail is essential for high-performance agents in visually rich tasks. In autonomous driving, Copilot4D [243] applies discrete diffusion to tokenized point clouds and BEV representations for unsupervised learning of 4D world dynamics, while ReSim [244] employs a diffusion transformer to synthesize diverse and safety-critical driving scenarios, providing a controllable evaluator for policy development.

Despite their ability to generate visually compelling predictions, observation-space world models face substantial

limitations in computational efficiency and scalability. Modeling dynamics directly in high dimensional observations incurs high training cost and instability, making real-time reasoning and long-horizon planning challenging. Consequently, such models are often impractical for real-world applications, motivating a shift toward more principled world model architectures that achieve better prediction quality.

Latent-Space Prediction High-dimensional observations can be compressed into compact latent representations, allowing dynamics to be learned directly in an abstract space [86]. This family of approaches is known as latent world models. Rather than reconstructing future states at the pixel level, latent world models focus on capturing the underlying factors of variation that matter most for control and decision-making. This abstraction improves computational efficiency and robustness, reducing error accumulation over long horizons and enabling faster inference. Consequently, latent models are well-suited for complex control tasks. The Dreamer series [8, 44, 62, 85] provides a foundational framework for learning latent dynamics and performing planning through imagination for continuous control. V-JEPA [114] predicts future latent representations directly, prioritizing controllable and predictive structure over raw visual appearance. EgoAgent [116] jointly models embeddings and actions, learning high-level vision-action dynamics from large-scale egocentric video data. Most recently, vision foundation models have been adapted to latent world modeling to exploit powerful pretrained features. DINO-WM [245] leverages DINOv2 [246] features to achieve temporal consistency and geometric awareness without explicit 3D supervision, while LaDi-WM [247] further incorporates CLIP [248] features to enhance semantic understanding.

Despite their efficiency, latent world models face limitations in interpretability and potential information loss. Because dynamics unfold in an abstract feature space, their predictions are harder to verify, and the compression step may discard subtle yet important details needed for precise interaction or safety. Their effectiveness is ultimately bounded by the invariances and biases of the encoder: what is not preserved during encoding cannot be recovered by latent dynamics. Even so, the latent modeling paradigm remains highly promising. Its scalability, robustness, and suitability for long-horizon planning position it as a key direction for future world models, especially as representation learning continues to advance and latent spaces become more aligned with physical dynamics.

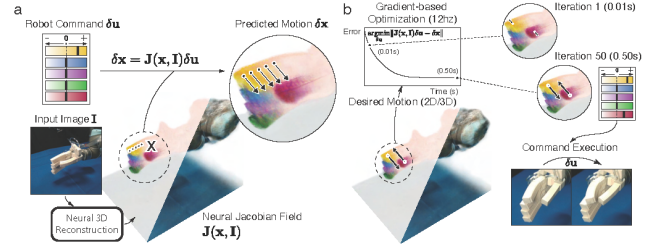


Fig. 12 Neural Jacobian Fields [96], which reconstruct a robot’s 3D geometry and local motion sensitivities from vision and enable closed-loop control by optimizing commands that realize desired 2D or 3D motions at interactive rates.

Structured Prediction Incorporating inductive biases into world model architecture helps align them with the causal nature of the physical world, simplifying dynamics learning and improving generalization. A common approach is object-centric, which decomposes scenes into distinct entities. This factorization allows the model to learn dynamics and interactions separately, making it well-suited for multi-object manipulation and other settings where compositional generalization is crucial. Existing works [249–251] map visual observations into object slots and model their pairwise interactions, enabling improved reasoning and control across diverse downstream tasks. More recently, researchers have incorporated 3D geometric constraints to ensure spatial consistency. Several methods [73, 74, 187] integrate action-conditioned Gaussian propagation into scalable diffusion transformers, achieving precise future scene reconstruction and serving as effective neural simulators for learning manipulation policies, as demonstrated in Fig. 12. To capture more complex physical phenomena, other approaches explicitly model interaction relations using graph-based structures. These methods typically employ GNN backbones in which objects or spatial regions serve as nodes and physical influences are transmitted along edges. Prior work [252–254] represents physical systems as particle or region graphs and propagates dynamics through message passing, enabling accurate long-horizon simulation across varying materials and mechanical structures.

Structured world models benefit from strong inductive biases that enforce spatial and causal consistency. Unlike structure-informed neural dynamics models, which embed physical structure directly into the dynamics, these approaches impose structure only on the input representation, making them easier to apply to domain-specific design choices such as object decompositions or relational graphs. While straightforward to construct, this design introduces limitations. Their generality is constrained because handcrafted structural assumptions that work well within a domain often fail to transfer when scene composition or interaction patterns change. This

stands in contrast to data-driven world models, which aim to learn representations and dynamics directly from diverse observations flexibly.

4.3.2 World Foundation Model

World Foundation Models (WFMs) [30] are large, general-purpose dynamics models trained on broad, heterogeneous data that support zero/few-shot prediction. Since the data spanning diverse domains often lacks detailed action annotations, these models usually take textual descriptions as input and predict future video segments. We review the recent wave of WFMs and summarize their key properties in Tab. 2. Despite that, WFMs still represent the closest existing instantiation of a generalist world model, and have therefore attracted widespread attention. Current WFMs obtain strong and transferable dynamic priors through two major pathways: drawing upon the commonsense and causal knowledge embedded in pretrained large language models, or learning physical dynamics directly from large-scale video data. Their differences are illustrated in Fig. 13.

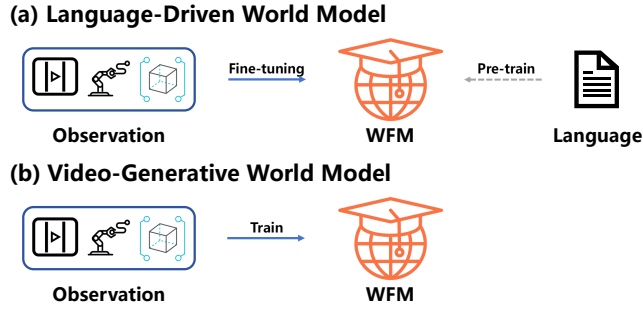


Fig. 13 Different approaches to learning WFMs. They can be developed either by (a) leveraging commonsense knowledge discovered from large text corpora or by (b) learning physical dynamics directly from large-scale video or 3D observation data.

Language-Driven World Model Large language models (LLMs) [121, 270], trained on extensive text corpora, inherently encode rich commonsense and causal knowledge on how objects interact, how actions lead to outcomes, and how the world tends to evolve. Inspired by this, recent work leverages pretrained LLMs as the dynamic backbone of world models, aligning their implicit knowledge with concrete environment states and actions. One type of these world models directly operates in the symbolic or language domain, while the other develops LLM-based video generators, adopting video sequences as the primary output.

For world models that operate entirely in the symbolic or language domain, they represent the environment as text-based states and actions, and adapt a pretrained LLM to predict

how these symbolic states evolve over time. RAP [271] repurposes a pretrained LLM as both a reasoning agent and a world model, combining it with Monte Carlo Tree Search (MCTS) to simulate state transitions in a latent “reasoning space” and to plan over alternative reasoning paths for tasks such as plan generation, mathematical reasoning, and logical inference. Xie et al. [272] explicitly turn LLMs into symbolic world models by fine-tuning two separate models for precondition and effect prediction, which respectively determine whether an action is applicable in a given language-described state and predict the next state after applying the action, enabling chainable action models that support planning. WebDreamer [273] treats a pretrained LLM as a world model of the internet. It adopts the LLM to predict the outcomes of candidate web actions in an imagined browser environment and performs model-based planning over these simulated trajectories, significantly improving the success rate of web agents on complex multi-step tasks. These methods instantiate a world model purely in the language domain, using symbolic state transitions to support complex multi-step reasoning and planning without explicit grounding in pixel dynamics.

For world models that operate in the video domain, the dynamics can still benefit from the commonsense priors of LLMs, even though the modality mismatch requires bridging language and pixel spaces. WorldGPT [274] couples an LLM-based prompt planner with a diffusion video backbone to synthesize temporally coherent long-horizon videos from text and accompanying images, effectively behaving as a video world model. Ge et al. [275] fine-tune a pretrained multimodal LLM on large-scale videos to predict multimodal state transitions and augment it with a lightweight memory module, enabling it to act as a generalist multimodal world simulator and to synthesize instruction-like data for downstream agents. Pandora [266] builds a hybrid autoregressive-diffusion world model that takes free-text actions as input and generates the corresponding future video of the environment. PAN [276] can be viewed as a more general and scalable successor to Pandora, formalizing the same LLM+video world-modeling idea into a GLP architecture and extending it to interactive, long-horizon simulation across diverse domains. These methods realize world models driven by textual commonsense directly in the video space, where the language backbone provides the underlying dynamics through its pretrained priors.

LLM-based world models inherit broad commonsense and causal priors from large-scale pretraining, enabling strong zero/few-shot generalization. Equipped with abstract knowledge about objects, actions, and typical event progressions, they can anticipate plausible outcomes and produce coherent

Table 2 The recent wave of WFMs as well as their key properties.

Model	Training Data	Input	Controllability	Size	Open Source
Sora [19]	Proprietary	Text, Image	Video-level	-	✗
Kling [255]	Proprietary	Text, Image	Video-level	-	✗
Seaweed-7B [256]	Web videos	Text, Image	Video-level	7B	✗
Wan [257]	Web videos	Text, Image	Video-level	1.3B / 14B	✓
HunyuanVideo [258]	Web videos	Text, Image	Video-level	13B	✓
MAGI-1 [259]	Web videos	Text, Image	Video-level	4.5B / 24B	✓
CogVideo-X [260]	Web videos	Text, Image	Video-level	2B / 5B	✓
Open-Sora 2.0 [261]	Web videos	Text, Image	Video-level	1.1B / 11B	✓
Goku [98]	Web videos	Text, Image	Video-level	1B / 2B / 8B	✓
LTX-Video [262]	Web videos	Text, Image, Video	Video-level	2B / 13B	✓
Mochi-Video [263]	Web videos	Text, Image, Video	Video-level	10B	✓
Genie [93]	Web videos	Image, Action	Frame-level	11B	✗
Cosmos [51]	Web videos, Real-world logs, Synthetic data	Text, Image, Video, Action	Frame-level	2B / 4B / 7B / 14B	✓
Unisim [264]	Web videos, Real-world logs, Synthetic data	Video, Action	Frame-level	5.6B	✗
AETHER [265]	Synthetic Data	Text, Camera Trajectory	Video-level	5B	✓
Pandora [266]	Web videos, Synthetic data	Text	Frame-level	7B	✓
iVideoGPT [115]	Web videos, Synthetic data	Video, Action	Frame-level	0.4B	✓
Vista [267]	Real-world logs	Video, Action	Frame-level	2.5B	✓
GAIA-2 [268]	Real-world logs	Driving Configuration	Frame-level	8.4B	✗
V-JEPA 2 [114]	Web videos, Real-world logs	Video, Action	Frame level	1B / 8B	✓
GEN3C [269]	Real-world logs, Synthetic data	Camera	Video-level	2B	✓
Genie Envisioner [122]	Real-world logs	Text, Image, Video	Frame-level	2B	✓

ent symbolic transitions or future video trajectories with far less supervision. However, the provided flexibility also introduces limitations. Since their dynamics arise from language priors rather than grounded physical principles or observation-driven learning, these models may hallucinate or generate physically implausible predictions, particularly in tasks requiring spatial precision or safety guarantees. They also lack robust mechanisms to enforce physical consistency or incorporate corrective feedback from real environments. For video-based variants, bridging language and continuous visual dynamics remains challenging, and large-scale training is computationally demanding.

Video-Generative World Model In contrast to learning WFMs based on language, another WFM directly learns dynamics from large-scale video observations. Benefiting from the scaling law [18], such research learns broad observational distributions and achieves general dynamic prediction capabilities. From an evolutionary perspective, this research lineage has progressed from video generation models to structured unified world representations, then to agent-interactive world models, and finally to foundational platforms.

The emergence of large-scale video generation models has revealed that powerful generative architectures can acquire implicit world priors simply by learning to predict future visual states from massive video-text corpora. Sora [19] demonstrates that high-capacity diffusion-transformer architectures trained on Internet-scale data can learn to generate high-fidelity, long-horizon, and physically consistent videos. Similarly, CogVideoX [260] leverages an expert-transformer

architecture and 3D VAE to produce long, coherent, narrative-rich video sequences, reflecting stronger temporal and physical understanding. Unified image-video foundation models such as Goku [98], and efficient large-scale systems like Wan [257], Seaweed-7B [256], and Open-Sora 2.0 [261] further show that massive or cost-optimized pretraining pipelines naturally endow video generators with stable temporal coherence and emergent physical regularities. Although these models are not explicitly designed as world simulators, their ability to model high-dimensional video distributions effectively encodes general-purpose statistical dynamics of the real world.

Beyond video generation models focusing on visual pixel representations, WFMs also take steps toward more structured, geometry-aware, or semantically informed world representations. GEN3C [269] introduces 3D point-cloud caches to enforce temporally consistent 3D structure and precise camera control, yielding world-consistent video generation. V-JEPA 2 [114], trained on over one million hours of video with minimal robot fine-tuning, shows that large-scale self-supervised predictive learning can produce representations that support video understanding, action anticipation, and zero-shot robotic planning. AETHER [265] further unifies 4D dynamic reconstruction, action-conditioned prediction, and goal-directed visual planning within a geometry-aware multi-task framework, using camera trajectories as global action inputs, as its overview in Fig. 14 illustrates. RLWG [277] introduces a post-training alignment framework that uses verifiable geometric and temporal rewards to ground pre-trained video world models, significantly improving their



Fig. 14 An overview of AETHER [265], showcasing its three core capabilities, including 4D reconstruction, action-conditioned 4D prediction, and visual planning, trained purely on synthetic data yet demonstrated on unseen real-world inputs.

spatial coherence and long-horizon stability for embodied navigation. At the same time, several works model the environment as explicit 3D worlds. Marble [278] is a multimodal world model that reconstructs and generates full 3D worlds from text, images, videos, or coarse 3D layouts, and exports them as Gaussian splats, meshes, or camera-controlled videos for downstream use. Terra [279] proposes a native 3D world model that represents explorable environments in an intrinsic point-latent space and decodes them into 3D Gaussian primitives for joint geometry-appearance modeling with exact multi-view consistency. HunyuanWorld [280] leverages a DiT-based image generator to synthesize panoramic 360° views and then uses these panoramas as world proxies to construct semantically layered 3D meshes and interactive, explorable scenes from text or image inputs. Together, these models move beyond raw video generation to acquire structured, interpretable, and more physically grounded internal world states as well as explicit 3D scene representations, bridging perception, geometry, and prediction.

WFMs have also begun integrating action conditioning to support agent interaction and control within learned worlds. Genie [93] and Genie3 [281] illustrate how autoregressive, action-conditioned video prediction can support long-horizon rollouts and emergent behaviors entirely within learned environments. Vid2World [282] adapts pretrained video diffusion models into causal, autoregressive predictors augmented with action guidance to support robot manipulation and game simulation. Unisim [264] scales this idea to Internet-scale data, learning a universal action-conditioned world model capable

of supporting both high-level vision-language planning and low-level control policies. Genie Envisioner [122] integrates instruction-conditioned video diffusion, action decoding, and neural simulation into a unified framework for scalable robotic manipulation. These works shift world modeling from passive video generation toward interactive predictive environments, where agents can perform actions and learn policies.

As world models gain controllability and structure, new systems aim to support real-world deployment, digital twin construction, and domain-specific simulation. Cosmos [51] presents a full-stack platform for physical AI, offering generative world models, policy models, and tooling for creating customizable digital twins for robotics, autonomous vehicles, and video analytics. In the driving domain, Vista [267] learns a generalizable driving world model capable of long-horizon, high-fidelity rollouts with flexible controllability across commands and trajectories. GAIA-2 [268] enables structured, controllable driving scene generation using latent-diffusion conditioned on ego motion, agent behaviors, environment geometry, and semantics, supporting the simulation of rare or dangerous scenarios. These systems emphasize practice, serving as tools for safety testing, policy learning, and large-scale scenario generation in real-world applications.

World models derived from large-scale physical observations offer advantages in breadth. By training on massive, heterogeneous video and action corpora, they acquire general-purpose predictive priors grounded directly in real-world dynamics rather than curated datasets or textual abstractions. They require no task-specific data collection or environment engineering, and their predictions tend to respect the physical regularities embedded in the observed videos, reducing hallucinations and improving spatiotemporal fidelity. Their scalability further supports diverse scene coverage and robust generalization across domains without retraining. Despite these strengths, the broad scope of these models still faces challenges. Training at scale demands substantial computational resources, large curated datasets, and careful filtering to avoid learned artifacts. Moreover, although grounded in physical observations, these models still lack explicit physical structure, making them vulnerable to distribution shifts and limiting their reliability in safety-critical settings where hard physical guarantees are required.

5 World Model Usage

World models are task-agnostic, allowing them to be efficiently adapted to a wide range of downstream objectives. Their usages span several mainstream categories: they can act as interactive environments that generate rollouts for policy

learning; serve as white-box dynamical systems that enable gradient-based control or policy optimization; function as evaluators that score trajectories or policy behaviors; and operate as generative models capable of producing diverse synthetic data. These categories are summarized in Fig. 15, and each can leverage either specialists or generalists depending on the requirement between accuracy and generalization.

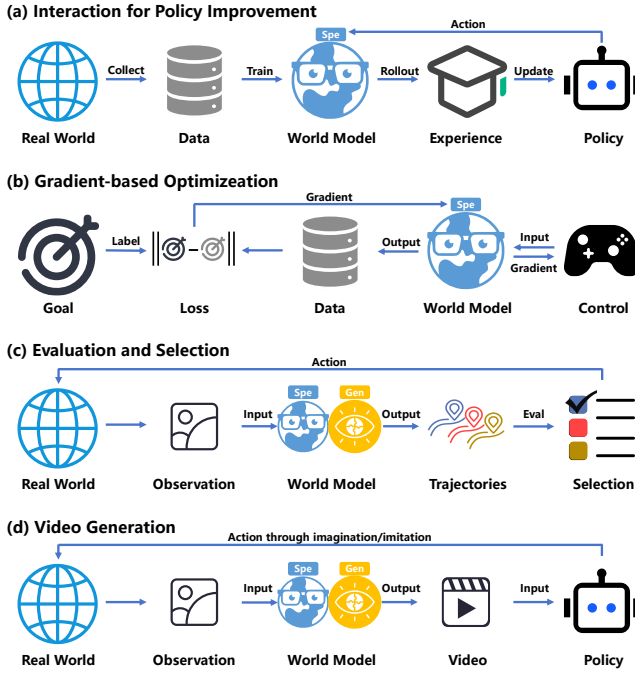


Fig. 15 Main categories of world model usage.

5.1 Interaction for Policy Improvement

Interaction for policy improvement is one of the most common applications of world models. Because these methods require sequential rollouts, they are highly sensitive to compounding prediction errors [27], making model accuracy essential. As a result, policy improvement pipelines typically rely on specialist world models, since the generalist prioritizes broad generalization and applicability over fidelity. The key advantage of interaction via world model interaction lies in sample efficiency. By serving as high-fidelity surrogate environments, world models enable policies to perform extensive trial and error in simulation, reducing the need for costly real-world sampling. This interaction paradigm spans two major settings. In online model-based RL, Dyna-style approaches [283, 284] continually refine the world model with new real data and use synthetic rollouts to enrich the replay buffer. In offline model-based RL [5, 285], the model is learned entirely from a fixed dataset, and policy improvement must remain conservative to avoid exploiting model errors on unseen states.

Together, these settings show how world models enable data amplification when interaction is available and safe policy learning when it is not.

In online model-based RL, agents retain the ability to interact with the real environment, but such interaction is costly and limited. Dyna-style algorithms [283] instantiate this setting by combining real-world experience with imagined rollouts generated by a learned world model. The model serves as a surrogate environment that augments the sparse real transitions with inexpensive simulated ones, allowing the agent to perform extensive trial-and-error in imagination while gradually refining the model from real feedback. ME-TRPO [287] introduces neural network ensembles [288] to capture model uncertainty, using disagreement across models to define trust regions that prevent policies from exploiting inaccurate predictions. Building on these insights, MBPO [284] establishes a modern template for data-efficient online RL. Rather than relying on long-horizon simulated trajectories that accumulate errors, MBPO generates short branched rollouts from real states stored in the replay buffer. These synthetic transitions are then used to train a model-free learner, effectively decoupling the sample complexity of model learning from that of policy learning. As a result, MBPO achieves asymptotic performance comparable to model-free methods while requiring orders of magnitude fewer real-world samples.

Subsequent work continues to refine the Dyna-style paradigm, addressing challenges in stability and exploration. Dedieu et al. [289] incorporate Transformer-based dynamics models with a Dyna warmup strategy, gradually introducing synthetic rollouts to avoid policy collapse in early training. Moving beyond passive augmentation, MoGE [290] proposes exploration augmentation, leveraging a generative model to synthesize high-value or rarely visited states; these synthetic experiences populate an exploratory replay buffer that jointly trains the policy and value functions alongside real data. However, the assumption that synthetic rollouts always benefit learning has been critically re-evaluated. Barkley et al. [291] demonstrate that Dyna-style systems may severely overfit to model-generated artifacts, causing performance degradation in complex domains relative to model-free baselines. This highlights a key vulnerability of online model-based RL: without explicit uncertainty quantification and safeguards against model exploitation, synthetic data can harm rather than help policy improvement.

While online model-based RL assumes intermittent access to real interaction, many practical scenarios, such as robotics, industrial manufacturing, and autonomous driving, often offer only fixed datasets, either due to safety constraints, cost, or

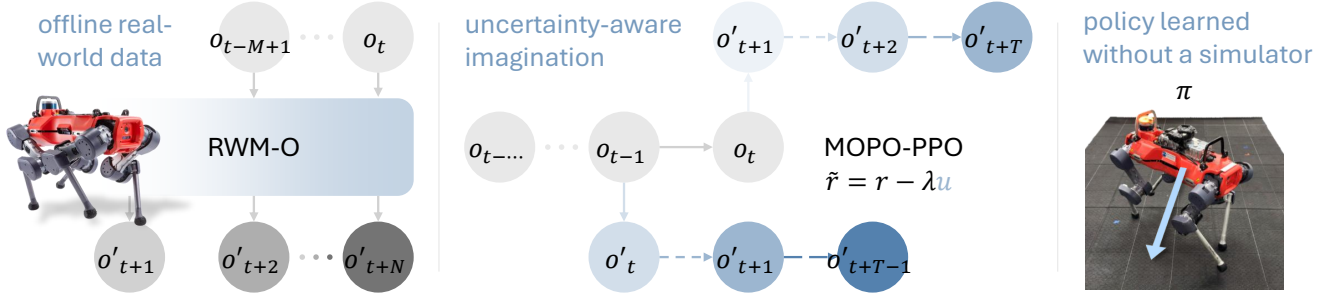


Fig. 16 RWM-O [286] learns long-horizon dynamics and epistemic uncertainty from offline real-world data, which enables uncertainty-aware model rollouts that penalize unreliable predictions and support safe policy learning without a simulator.

irreversibility of online interaction. In these offline settings, the agent must learn exclusively from a static dataset, and imagined rollouts cannot be corrected by real-world feedback. When real-world interaction is strictly prohibited, the primary challenge becomes preventing the policy from exploiting model errors that produce hallucinated high rewards. Offline model-based RL addresses this through uncertainty quantification and pessimism, constraining the policy to remain within regions of the state space where the model is reliable. MOREL [292] constructs a Pessimistic MDP: epistemic uncertainty is estimated via a model ensemble, and transitions with high disagreement are terminated or heavily penalized, creating a safety barrier that discourages the policy from entering unknown regions. MOPO [5] adopts a softer variant of this principle, subtracting an uncertainty-based penalty from predicted rewards so that the policy optimizes a conservative lower bound on return. Extending these ideas to high-dimensional visual control, LOMPO [293] introduces pessimism directly in the latent space. By estimating uncertainty through latent ensemble variance rather than reconstruction, it avoids the burden of image prediction and enables robust offline learning from visual observations.

As pessimism-based frameworks matured, a complementary line of work began addressing their practical limitations in computational cost, scalability, and adaptability. ROSMO [294] reduces reliance on heavy ensemble models by showing that variance from a single probabilistic predictor can serve as an effective uncertainty proxy. COMBO [285] removes the need to explicitly estimate uncertainty at every step by regularizing the value function on model-generated transitions, implicitly inducing conservative behavior. MAPLE [295] tackles distributional heterogeneity by inferring task-specific context, improving robustness under diverse datasets. To address the short-sightedness of pessimism in long-horizon tasks, TempDATA [296] introduces temporal distance-aware augmentation, enabling policies to bridge distant states and better handle sparse rewards.

Because offline agents never interact with the real environment during training, a central challenge is ensuring that policies remain reliable when deployed outside the distribution of the logged dataset. Offline world models trained purely on static data often capture only a narrow subset of real-world dynamics; as a result, even subtle physical mismatches can lead to significant performance degradation at deployment. NeoRL [297] highlights how narrow, conservative datasets can cause offline RL methods to fail dramatically at deployment, reinforcing the need for explicit mechanisms that prevent model exploitation. To mitigate this gap, SPiDR [298] introduces a pessimistic domain-randomization strategy for zero-shot transfer. Interpreting high variance across randomized physics parameters as a safety violation, it restricts the policy to behaviors that remain safe under worst-case dynamics. Complementing this direction, RWM-O [286] learns a high-fidelity probabilistic neural simulator directly from offline real-world trajectories, as demonstrated in Fig. 16. Through strict uncertainty penalties and ensemble-based epistemic modeling, RWM-O enables offline-trained agents to acquire robust policies without any physical interaction, narrowing the policy transfer gap despite the static data.

Although interaction-based policy learning has traditionally relied on specialist world models, whose high fidelity is necessary to avoid compounding errors, recent work has begun exploring whether generalist models can also support policy improvement. A central goal in this direction is to distill universal kinematic priors that transfer across heterogeneous embodiments. TrajWorld [299] exemplifies this trend by introducing a unified trajectory tokenization scheme that accommodates robots with differing state-action dimensions. Pre-training a trajectory transformer on diverse robotic datasets enables the model to capture embodiment-invariant kinematic structure. Instead of learning dynamics from scratch, the policy queries the world model’s internal knowledge to achieve zero-shot or few-shot transfer to previously unseen robot morphologies.

5.2 Gradient-based Optimization

Gradient-based optimization provides a principled mechanism for control when a world model is both differentiable and sufficiently accurate. In such settings, the model acts as a smooth dynamical system linking actions to future states, allowing gradients to propagate through imagined trajectories and casting decision-making as a continuous optimization problem. This capability enables efficient refinement of actions and policies through backpropagation. Because predictive accuracy is essential for meaningful gradients, these methods are predominantly applicable to specialist world models whose dynamics are tailored to a particular robot or environment.

A central question is where the gradients come from. Differentiable physics simulators [71, 143, 145, 300] derive gradients analytically from physical equations, explicitly modeling geometry, contact, and force propagation. The resulting sensitivities retain clear semantic meaning with respect to actuation, making them particularly effective for manipulation and soft-body control, as the visuomotor control results of ∇Sim [143] illustrated in Fig. 17. In contrast, learned neural dynamics models, such as PILCO [107], Dreamer [44], and physics-informed neural world models [231], provide differentiability with greater flexibility, enabling gradient-based planning in high-dimensional visual settings. Although these learned gradients lack explicit physical semantics, they support long-horizon optimization and policy improvement entirely in imagination.

When combined with differentiable world models, classical trajectory optimization methods can integrate naturally with modern learning systems. Techniques such as iLQG [301] and DDP [302] refine action sequences by locally linearizing the dynamics and quadratically approximating the cost-to-go. Recent developments extend these ideas to nonlinear, contact-rich domains: DiffTOP [303] embeds differentiable trajectory optimization directly as a parameterized policy, while D3P [304] adapts DDP-style updates to neural dynamics, yielding stable long-horizon behavior. Differentiable physics engines have similarly been shown to accelerate controller optimization [163], underscoring the value of embedding physical structure within the gradient pathway.

Reparameterized gradients and stochastic value gradients form the basis of a mature line of research in model-based reinforcement learning. SVG [305] provides a unified formulation in which the Bellman equation is expressed as a differentiable function of noise, enabling unbiased policy gradients through learned dynamics. IVG [306] extends this idea to latent models, learning transferable policies from imagined rollouts. Model-augmented actor-critic methods [307]

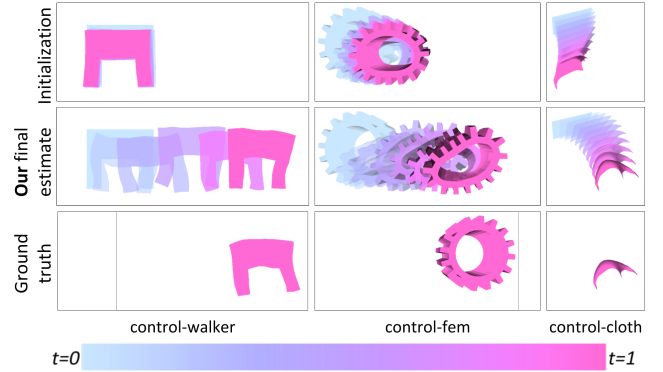


Fig. 17 ∇Sim [143] enables visuomotor control by providing differentiable gradients for optimizing soft-body actuation or initial velocities to match target images across diverse tasks.

and physics-informed neural dynamics [231] further stabilize long-horizon credit assignment by combining model gradients with learned value functions. Recent advances [308] demonstrate that hybridizing model-based value gradients with soft value estimation yields state-of-the-art results even in complex humanoid control tasks.

Because purely gradient-driven optimization is sensitive to nonconvex landscapes and model inaccuracies, a complementary line of work explores hybrid planners that interleave gradient descent with population-based search or amortized policy learning. Combining CEM with gradient refinement [309] improves global exploration while preserving fast local convergence. Gradient-based MPC with learned models [310] achieves performance comparable to sampling-based MPC in low-data regimes. D3P [304] further demonstrates how initializing optimization with a policy network and applying conservative gradient updates mitigates compounding model errors during long-horizon rollouts. These hybrid strategies illustrate a growing consensus: global search for exploration, gradients for precision. Despite their promise, gradient-based planners face inherent limitations. Imperfect world models can yield misleading gradients, particularly over long horizons; many methods alleviate this by anchoring optimization with value functions or receding-horizon updates. Differentiable simulators and trajectory optimization layers introduce substantial computational cost, motivating amortized policies [44] and hybrid planning [309] to reduce inference-time overhead.

5.3 Evaluation and Selection

World models serve as computational surrogates for the real environment, enabling agents to simulate the consequences of candidate actions and select those likely to yield desirable outcomes without real-world risk or cost. This ability

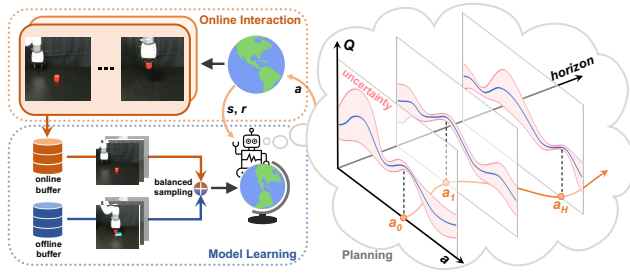


Fig. 18 Pretrain the world model on offline robot data, then finetune it with online rollouts generated via model-based planning, using model uncertainty to guide effective exploration [315].

to mentally try before acting corresponds to model-based planning [311], in which a predictive model is used to forecast short-horizon outcomes and choose actions accordingly. Because it depends only on local predictions rather than long, recursively imagined rollouts, the accuracy requirements are comparatively mild, allowing both specialist and generalist world models to support efficient evaluation and selection.

Early progress in model-based planning relied on specialist world models trained to capture accurate dynamics within restricted environments. These efforts begin with low-dimensional control. PETS [312] establishes a baseline in low-dimensional state spaces by employing probabilistic ensembles to capture epistemic uncertainty. It enables the CEM planner to evaluate and select actions that are robust to model bias. To scale planning to high-dimensional visual inputs, PlaNet [4] introduces latent space planning. It utilizes an RSSM to compress visual observations into a compact latent space, allowing the CEM planner to evaluate thousands of candidate sequences in parallel without pixel-level reconstruction. MuZero [313] pushed this abstraction further by predicting only task-relevant quantities, i.e., value, policy, and reward, and integrating them with MCTS, achieving superhuman planning across logical and visual domains. More recent work has turned to planning efficiency in high-dimensional action spaces. POPLIN [314] integrates policy networks into the planning loop, functioning not as direct controllers but as proposal priors to guide the search. To address the domain shift from offline data to real-world dynamics, Feng et al. [315] propose a few-shot adaptation framework, as shown in Fig. 18. By introducing test-time behavior regularization based on epistemic uncertainty, it allows the planner to safely adapt a pre-trained model to novel real-world environments with minimal interaction. Online Agent (OA) [316] addresses generalization in non-stationary environments by employing online convex optimization, enabling continual adaptation through planning to mitigate catastrophic forgetting.

Despite their prowess in specific domains, specialist plan-

ning models require extensive training from scratch for each new task. This limitation drives the shift toward generalist planning, where pre-trained foundation models enable zero-shot planning across diverse environments. Schubert et al. [55] train a Transformer dynamics model across a wide range of control tasks, leveraging Model Predictive Path Integral (MPPI) [317] to plan directly with this universal predictor to achieve zero-shot generalization to unseen robot embodiments without any fine-tuning. DINO-WM [245] builds a world model directly on pre-trained DINOv2 features and selects actions by minimizing the feature distance to a goal state. V-JEPA 2 [114] learns abstract semantic dynamics from millions of hours of internet video and uses CEM to reach goals in this latent manifold, achieving zero-shot robotic manipulation without task-specific rewards or adaptation. Sobal et al. [22, 318] show that under suboptimal data or long-horizon tasks, latent-dynamics planning consistently outperforms model-free policy learning, indicating that for reward-free, open-ended environments, predictive world-model evaluation offers a more reliable path to generalization than direct policy optimization. FLIP [319] tackles heterogeneous robot embodiments by treating optical flow as a universal action interface, enabling cross-morphology transfer between visual goals and physical control. Pushing universality further, ScaleZero [320] unifies discrete games, continuous control, and text environments into a single world model, showing that large-scale evaluation and selection can support multi-task generalization under one parameter set.

Finally, the paradigm of evaluation and selection has transcended the granularity of individual trajectories to encompass the holistic assessment of full policies. WorldEval [94] establishes the foundational viability of this paradigm by addressing the challenge of heterogeneous action spaces. It maps diverse policy outputs to a unified latent control space, enabling a single video world model to evaluate various robotic agents and demonstrating a strong correlation between virtual success rates and real-world performance. WorldGym [321] encapsulates the world model into an open-ended generative benchmark. It focuses on evaluating generalization by synthesizing OOD tasks via image editing and language modification, verifying that world models can faithfully preserve the relative ranking of policies across different checkpoints and model sizes. Tseng et al. [322] introduce a fully automated evaluation pipeline powered by video-language models. By using video foundation models for high-fidelity rollouts and VLMs as reward judges, the system removes the need for human supervision and demonstrates that generative world models can operate as a scalable, reproducible, and automated

proving ground for generalist robots.

5.4 Video Generation

The world model usages discussed so far primarily rely on frame-by-frame prediction. In parallel, video-generation world models offer a complementary capability: they synthesize future image sequences rather than isolated ones, providing richer temporal context for downstream decision-making. These models can support control in two principal ways, by enabling imagination, where generated rollouts guide long-horizon planning, and by enabling imitation, where videos serve as supervisory signals from which executable actions are inferred. Both specialist and generalist world models can serve as a bridge between video prediction and behavior learning. We next examine how video-based imagination and imitation contribute to the decision making.

Imagination-based world models generate hypothetical future trajectories that allow agents to reason about consequences before acting. Early imagination models were tightly coupled to specific environments. I2A [323] interprets imagined rollouts as auxiliary features rather than being used for explicit planning. Subsequent models, including IDM [324] and Choreographer [325], enable multi-step latent rollouts that improve sample efficiency for visuomotor control. Structured recurrent world models such as Hieros [326] and R2I [327] extend temporal reasoning via long-horizon recurrence.

As the community shifted toward multi-task agents, world models began integrating broader generative priors. Diffusion-based models [247, 328, 329], leverage high-capacity generative backbones to produce high-fidelity future predictions, grounding imagination in broad visual statistics rather than environment-specific dynamics. In parallel, language-conditioned imagination systems [325, 330, 331] introduce reusable behavioral primitives and abstract goal conditioning, allowing imagination to generalize beyond individual tasks. RoboDreamer [332] composes unseen object-action combinations into novel imagined task sequences, demonstrating systematic compositionality. DreMa [155] integrates explicit 3D scene representations with physics priors to construct editable digital twins that generalize from minimal demonstrations. LS-Imagine [333], addressing open-world settings such as MineDojo [334], employs jumpy, goal-conditioned latent transitions to support efficient long-horizon exploration. Generalist video world models such as Gen2Act [335], and recent video-language simulators [274] further expand this capability into multimodal, open-ended domains.

Recent work extends beyond future prediction to direct policy supervision: imagined or real videos serve as target trajec-

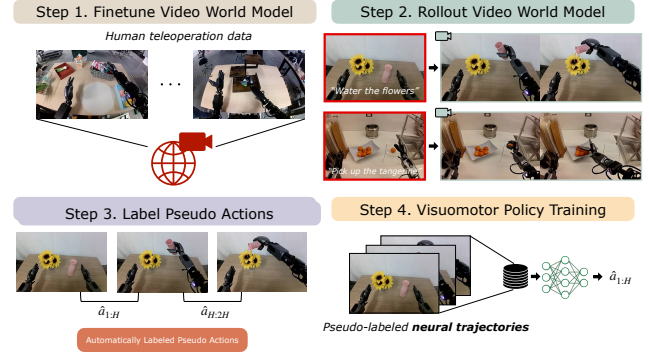


Fig. 19 DREAMGEN [58] fine-tunes a video world model on teleoperated data, generates instruction-conditioned video rollouts, infers pseudo-actions to form neural trajectories, and trains visuomotor policies on these synthetic trajectories.

tories from which actions are inferred. Early video-imitation approaches focused on learning actions from generated or observed videos by inferring latent control variables that explain visual transitions. DreamGen [58] exemplifies this direction: a video world model is first fine-tuned on teleoperated demonstrations, then used to roll out instruction-conditioned future videos, as illustrated in Fig. 19. An inverse-dynamics module recovers pseudo-actions from these imagined trajectories, producing dense neural supervision that trains visuomotor policies without additional real-world data. PlaySlot [336] employs an object-centric video model to infer latent actions that align with predicted object motions, supporting imitation from unstructured play data. ATM [337] predicts future motions of arbitrary points in video frames, allowing the resulting dense trajectories to serve as precise supervision for visuomotor control even when task-specific action labels are sparse. This decouples policy learning from explicit video generation quality while still exploiting generalist video priors. VLP [331] generates high-level visual plans from natural language, after which a low-level controller extracts executable actions, illustrating how video imagination and imitation can operate hierarchically.

Recent systems increasingly treat video prediction as implicit inverse dynamics, merging imagination and imitation into a single framework. ViPRA [338] aligns latent actions with predicted video dynamics, while VPP [328] uses diffusion-based video prediction as a surrogate inverse-dynamics model, decoding robot actions directly from intermediate video features. AnyPos [339] leverages a semantic video generator as a perceptual prior and learns high-precision inverse dynamics from a small amount of task-agnostic action data, illustrating a promising hybrid between generalist video models and specialist control adaptation. Across both imagination and imitation, video-generative world models have

expanded from narrow, specialist predictors into increasingly general and compositional tools for control. Although these models lack explicit physical grounding, their large-scale visual priors, temporal coherence, and capacity to synthesize diverse futures make them powerful engines for scalable policy development.

6 Application

World models are now used across diverse domains such as robotic manipulation, locomotion, navigation, and game playing. Because each domain places different demands on accuracy, generalization, and scalability, world models may function either as precise predictive engines for control or as broad generative models for large-scale learning. We review how world models are applied in these specific domains.

6.1 Robotic Manipulation

Manipulation tasks require reasoning about geometry, appearance, and the physical consequences of interaction [340, 341]. World models with stable multi-step prediction are therefore essential for data-efficient learning and reliable control. Specialist world models emphasize high accuracy for a specific robot and workspace, supporting precise manipulation in structured settings such as industrial automation. In contrast, generalist models provide cross-object and cross-task priors, enabling flexible manipulation in open and diverse environments, such as those encountered by service robots. Representative applications of each paradigm are shown in Fig. 20, and we discuss them in detail below.

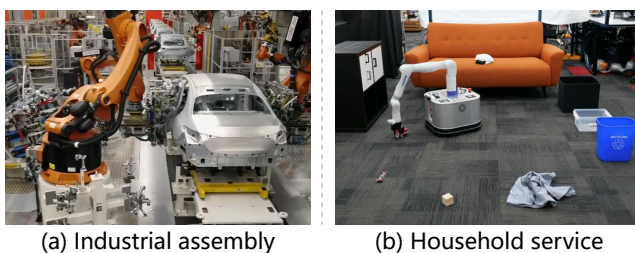


Fig. 20 For (a) industrial manipulation which demands high precision, specialist world models with high-fidelity predictions are essential. In contrast, (b) service robots [342] operating over diverse objects benefit more from generalists which provide broad priors.

6.1.1 Specialist for Precise Manipulation

For precise manipulation, high-fidelity world models tailored to a specific robot and environment are essential. In industrial settings [47], tasks such as welding [343, 344], assembly [345, 346], and packing [347–352] demand accurate dynamics to satisfy tight control tolerances. Similarly,

in on-orbit servicing (OOS), specialist world models support satellite attitude control, docking, and debris removal, where micro-gravity dynamics and strict safety constraints require reliable prediction. Recent OOS systems employ tailored physical simulators and model-based learning frameworks [353–356], highlighting the central role of specialist world models in high-precision space robotics.

In vision-based robotic manipulation, policies were primarily guided by 2D image observations, where robots need world models learned to interpret raw pixels and predict future visual states for planning and control. E2C [357] uses a variational latent model with locally linear dynamics to predict image trajectories and perform optimal control directly from pixels. MoDem [358] improves sample efficiency of visual model-based RL by integrating demonstrations through policy pretraining, targeted exploration, and demonstration oversampling. DayDreamer [45] applies Dreamer [85] directly on real robots for online learning of manipulation from raw images with minimal interaction. TWIST [359] achieves efficient sim2real transfer by distilling a state-trained teacher world model into a vision-based student using domain-randomized image datasets. For image-based world models, robots must implicitly infer spatial relationships from appearance cues, which limits robustness in physically complex scenarios.

It is acknowledged that robust learning can benefit from richer world representations that jointly model 3D geometry, physics, and visual observations [46, 68, 165]. DREAM [158] builds a differentiable real2sim2real pipeline using Gaussian Splat digital twins to identify physical parameters and learn force-aware manipulation policies with accurate sim2real consistency. PIN-WM [46] learns physics-informed 3D rigid-body world models from visual observations using differentiable physics and Gaussian Splatting, with physics-aware randomization to enhance sim2real robustness in non-prehensile tasks. Li et al. [194] learn a visuomotor Jacobian field directly from single-camera video, enabling control of soft and articulated robots without prior models. By modeling spatial structure and contact mechanics rather than only visual features, these approaches yield more reliable motion prediction, stronger physical consistency, and improved sim2real transfer. This shift from purely image-driven reasoning toward geometry-aware world representations marks an important step toward robust, high-precision manipulation.

6.1.2 Generalist for Flexible Manipulation

In contrast to specialist models, generalist world models aim to support manipulation in open and diverse environments. By learning broad priors, they capture common structure across objects, tasks, and embodiments, enabling robots to operate in

previously unseen settings with minimal supervision. Rather than modeling any single system with high fidelity, generalist approaches prioritize flexible and transferable representations of dynamics, allowing one model to plan, predict, and act across a wide spectrum of manipulation scenarios.

Generalist world models influence robot decision-making through various mechanisms that together enable flexible manipulation in diverse, open-ended environments. At the foundation, many models provide action-conditioned visual prediction, allowing robots to look ahead into possible futures and choose actions whose imagined outcomes best satisfy task objectives. Systems such as UniSim [264] and FLIP [319] support these long-horizon visual rollouts, letting agents reason directly over predicted scene evolution. Beyond full-frame prediction, some generalist world models contribute by offering motion and affordance cues derived from large-scale video data. V-JEPA 2 [114] and VidMan [60] extract dense motion fields or affordance maps, supplying control-relevant structure without simulating complete visual futures.

Building on these predictive and perceptual capabilities, another line of work uses world models to synthesize demonstrations and imagined trajectories that bootstrap policy learning. Systems such as DREAMGEN [58] and RoboDreamer [332] generate diverse video rollouts enriched with pseudo-actions or compositional primitives, expanding the behavioral coverage available for training and enabling policies to generalize well beyond the robot's real-world dataset. Finally, a generalist can also support goal-directed planning in latent space, often enhanced by language grounding. Methods like LUMOS [330], WorldVLA [360], and PIVOT-R [361] predict task progress in compact latent representations and use language to define goals or waypoints, enabling coherent long-horizon planning across many tasks and visual conditions. Some models, such as DyWA [362], introduce dynamics adaptation, ensuring robust decision-making even when physical properties shift across environments.

6.2 Locomotion

Locomotion typically refers to the movement of legged robots, such as quadrupeds and bipeds, as they walk, run, or traverse uneven terrain, requiring reliable prediction of rapidly changing dynamics under multiple contacts. These tasks demand maintaining balance, regulating contact forces, and adapting to varying ground conditions. Successful locomotion, therefore, hinges on accurate near-future prediction, where world models serve as internal forward models to support stable, real-time control. Specialist models provide high-precision dynamics tailored to a specific robot, whereas generalist



Fig. 21 Precise tasks, like (a) steadily holding the beer [363], demand accurate physical prediction, which requires a specialist world model. In contrast, (b) completing diverse tasks in complex environments [54] calls for a generalist with stronger generalization.

models offer broader adaptability across terrains and morphologies, as illustrated in Fig. 21. This section reviews their respective roles in complex motion control.

6.2.1 Specialist for Stable Locomotion

For legged locomotion, building specialist world models tailored to specific robot morphologies can ensure safety and stability. The foundational research in robust locomotion relies primarily on proprioception to handle disturbances. These specialist models focus on inferring ground reaction forces and estimating world states solely from internal sensors to reject external perturbations. For instance, Sun et al. [364] utilize a world model estimator to reconstruct world states from proprioceptive history. By treating reconstruction as an auxiliary task, this approach forces the model to learn robust latent representations, enhancing the policy's adaptability and disturbance rejection capabilities in unstructured environments without reliance on potentially unreliable visual feeds.

However, traversing complex geometries requires the agent to anticipate terrain changes, driving the evolution toward visual-guided specialist models. These frameworks integrate exteroceptive inputs (e.g., depth or RGB cameras) to construct local environment models for look-ahead planning. The Puppeteer [365] introduces a hierarchical framework for visual-driven humanoids, employing a layered structure to predict whole-body dynamics from visual inputs, enabling coordinated motion in tasks requiring full-body integration. Similarly, WMP [366] focuses on legged robots across diverse

terrains, fusing proprioceptive and visual perceptions to build environmental world models. This multi-model integration allows the policy to anticipate obstacles, facilitating seamless sim2real generalization by aligning the modeled terrain with real-world perception.

To further address extreme environmental conditions and sensor uncertainties, recent advancements have incorporated physics-informed biases and denoising mechanisms into world models. Rather than treating dynamics as a black box, these methods explicitly model the underlying physical laws to enhance robustness. DWL [367] specializes in handling challenging surfaces like snow and stairs by emphasizing denoising objectives within the world model learning process. This allows the model to distill clean physical laws from noisy sensory data, enabling RL to achieve robust gait stability. Furthermore, HuWo [368] builds physics-informed interaction world models designed to capture intricate contact dynamics between the robot and the environment, rather than merely predicting visual changes, thereby ensuring reliable balance in complex interaction settings.

While specialist world models have mastered low-level stability in specific environments, they are fundamentally constrained by the “one-morphology-one-model” paradigm, where learned dynamics are tightly coupled to the physical parameters of a single robot, preventing zero-shot transfer across diverse embodiments. Furthermore, these models typically focus on geometric traversability while neglecting high-level semantic understanding, restricting agents to simple locomotion rather than complex, cross-task behaviors like mobile manipulation or open-ended navigation. This inability to bridge the gap between robust dynamic control and high-level intelligence highlights the necessity of evolving toward generalist world models, which aim to unify diverse morphologies and integrate cognitive reasoning with physical control for truly versatile, multi-task mobility.

6.2.2 Generalist for Adaptive Skills

Many locomotion settings demand adaptability across environments and tasks. In outdoor exploration, rescue, or household deployment, a controller may face new ground conditions and tasks that extend beyond simple walking. Meeting this need requires generalist world models that generalize dynamics rather than memorize a single system, support transfer instead of per-robot retraining, and integrate perception with high-level reasoning.

The primary challenge for generalists is to establish foundation models adaptable to heterogeneous embodiments. TrajWorld [299] facilitates world model pre-training in heterogeneous environments, capturing universal trajectory dynamics

across diverse sensors and actuators to enhance transfer capabilities to unseen robot forms. In the realm of humanoids, Lei et al. [369] leverage the massive AMASS human motion dataset to train differentiable neural network dynamics models. By imitating and coordinating complex motions via unified motion capture, this approach demonstrates that learning universal kinematic priors enables a single model to orchestrate WBC across varying joint configurations, reducing reliance on robot-specific data.

To execute adaptive skills in dynamic settings, generalist models must extract robust physical laws from noisy visual inputs. HRSSM [370] introduces a mixed recursive state-space approach utilizing spatio-temporal masking and dual simulation principles. These mechanisms force the model to learn latent dynamic representations robust to visual distractions and suitable for complex visual control. To address the sim2real gap, Li et al. [87] propose a neural network simulator trained on real-world data for quadruped robots. Acting as a high-fidelity digital twin, this generalist simulator facilitates robust model-based policy optimization, enabling zero-shot deployment on physical hardware. This highlights a shift where data-driven neural simulators are beginning to supersede physics engines for robust policy transfer.

The ultimate goal of generalist models is to integrate low-level dynamics with high-level cognition for complex interaction planning. Ego-VCP [371] employs a learning-based world model for humanoid contact planning, utilizing a single extensible model trained on offline data to achieve multi-task generalization and adaptability to OOD scenarios. At a broader scale, EgoAgent [116] introduces a unified predictive agent that models egocentric video and 3D human motion as interleaved sequences, jointly learning perception, dynamics, and action forecasting to understand daily activities. Moreover, D²PO [372] augments Large Vision-Language Models (VLMs) for embodied task planning. By jointly optimizing state prediction and action selection, it leverages the commonsense reasoning of VLMs to achieve efficient interactions in dynamic physical environments.

Generalist models demonstrate immense potential for adaptive skills by synthesizing cross-domain data and cognitive reasoning. However, this versatility often comes at the cost of data hunger and high inference latency. Compared to specialist models, generalist frameworks may exhibit lower peak performance in specific high-frequency control tasks and often struggle to meet the millisecond-level precision required in strict industrial scenarios without task-specific fine-tuning.

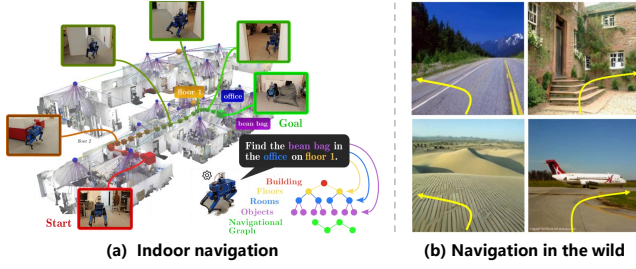


Fig. 22 For (a) navigation in a structured indoor scenario, specialist world models provide accurate scene descriptions for efficient path planning [373]. For (b) open-world navigation, generalists generate richer scenes, enabling safer and more adaptive decisions [374].

6.3 Navigation

World models have become central to embodied navigation, as they enable agents to anticipate how sensory observations and spatial states will evolve under different egomotion actions that change their position or orientation, thereby supporting tasks such as scene exploration, destination-reaching, and long-horizon planning. Within navigation tasks, the distinction between specialist and generalist settings largely stems from the structure and variability of the environment. Indoor navigation typically falls under the specialist category, as its environments are structurally constrained, exhibit limited dynamic variation, and rely on a narrow set of sensor modalities. These properties allow world models to specialize in precise predictions tailored to stable indoor layouts. By contrast, autonomous driving exemplifies a generalist domain, where environments vary dramatically in appearance, traffic behavior, weather, and geographical style, and where perception must integrate heterogeneous modalities such as cameras, LiDAR, radar, and HD maps, demanding world models that generalize across scenes, tasks, and behaviors. Representative applications of each paradigm are shown in Fig. 22.

6.3.1 Specialist for Constrained Environments

For navigation in constrained settings such as indoor scenarios, the primary requirement is reliable spatial understanding for short-range prediction and path planning. Because layouts are stable and variation is limited, models can depend on explicit geometric representations and structured rendering. This makes specialist world models effective since they are tuned for a narrow class of environments and deliver precise rollouts.

Before the emergence of learned world models, explicit simulators played a central role in this category by encoding the world explicitly via pre-defined geometry, rendering, and physics modules. Representative systems include Gibson [375], RoboTHOR [376], and Habitat [377], which provide photorealistic indoor environments for embodied navigation,

as well as CARLA [378], MetaDrive [379], and ScenarioNet [380], which explicitly model urban driving scenes with rich weather and road conditions for prediction and closed-loop evaluation. More recently, VR-Robo [166] extends this paradigm through a real-to-sim2real framework that reconstructs interactive 3D scenes from real-world imagery, while HUGSIM [381] further enhances realism with real-time 3D Gaussian Splatting and supports fully interactive human-autonomy testing.

In navigation settings, 3D reconstruction can also be viewed as a form of world model, as they explicitly recover scene geometry and appearance from visual observations. Once a scene is reconstructed, the model can directly render visual feedback corresponding to egomotion actions, allowing an agent to perceive how observations change as it moves through the environment. Representative approaches include Neural-Recon [382], which performs online dense scene reconstruction from monocular video, and ManhattanSDF [383], which leverages structural priors to recover accurate indoor geometry and semantics for view synthesis. Recently, ReconFusion [384] incorporates diffusion-based novel view synthesis as a prior to regularize 3D reconstruction, enabling consistent visual rollouts from sparse observations. Beyond single-agent settings, multi-robot reconstruction frameworks [385] and scalable neural mapping systems such as MIPS-Fusion [386] further integrate scene reconstruction with agent motion. Once a scene has been reconstructed, navigation reduces to efficient geometric planning [387, 388] where agents can compute collision-free paths, evaluate alternative routes, and localize themselves entirely within a compact spatial representation. While these methods provide high-fidelity within individual scenes, their reliance on per-scene reconstruction definitely limits cross-environment generalization.

6.3.2 Generalist for Open World

For open-world navigation, such as autonomous driving, the main challenge is generalizing across diverse visual styles, traffic patterns, and environmental conditions. These variations make hand-crafted structure insufficient, pushing models to learn scene evolution directly from large-scale, multimodal experience. As a result, generalist world models emphasize cross-scene training and action-conditioned prediction, offering better adaptability to real-world uncertainty.

Video-based world models aim to simulate future driving scenes in pixel space, learning environment dynamics directly from visual data. Unlike explicit simulators, these models learn implicit representations of scene evolution, enabling controllable video rollouts conditioned on past observations

and driving actions. For example, DriveDreamer [389] employs diffusion-based generation to synthesize long-horizon driving videos with controllable layouts and conditions, enabling scalable production of diverse and realistic training data. GAIA-1 [390] extends video generation into controllable simulation by conditioning predicted frames on steering, throttle, and other driving inputs, effectively visualizing how the world would evolve under specific actions. DrivingGPT [391] further unifies video generation and action planning in a multimodal autoregressive framework that predicts both future frames and the driving actions required to achieve them, marking a shift toward integrated, action-aware world modeling for autonomous driving. Recently, Bar et al. [392] employed action-conditioned video generation to simulate future observations and defined a perceptual similarity-based energy function that optimizes action trajectories toward a specified goal, showing how visual imagination can be combined with model predictive control to guide navigation decisions.

Occupancy-based world models represent the driving environment in volumetric or grid formats (e.g., 3D/4D voxels or BEV occupancy grids) and learn to model how that representation changes over time. Such representations capture geometry and temporal dynamics in a form that can be directly consumed by downstream planning and control modules, enabling the models to reason explicitly about free space, occlusions, and interactions. For instance, Occupancy Flow Fields [393] formulates motion forecasting as joint prediction of future occupancy and flow, producing temporally consistent, planner-ready world states.

LiDAR-based world models extend generative modeling into the point-cloud domain, capturing fine-grained 3D geometry and motion that are difficult to represent in images or occupancy grids. By directly learning the distribution of LiDAR scans, these models can synthesize realistic sensor data, complete missing structures, and even forecast future point-cloud evolution. For instance, LiDARGen [394] pioneered score-based generative modeling of point clouds to produce physically consistent and diverse driving scenes, serving as a realistic data source for training perception models. Building upon this, Copilot4D [243] learns a unified discrete latent world model that autoregressively forecasts future LiDAR frames, demonstrating that generative modeling of 4D point-cloud sequences can support action-aware scene prediction in autonomous driving.

6.4 Game Engine

Game environments are a natural fit for world models: they are rich, diverse, and allow unlimited interaction, making

them ideal for training and evaluating learned simulators. In this domain, world models function focuses on efficient and scalable scene creation. For specialists, the goal is to replicate a specific game engine with high fidelity, while generalists aim to create versatile game worlds that can adapt to various environments, as illustrated in Fig. 23.

6.4.1 Specialist for Consistent Replication

Game environments can provide virtually unlimited, low-cost interaction data, making them an ideal testbed for learning and evaluating world models as well as model-based reinforcement learning algorithms [85, 395]. In this context, specialist world models face the rigorous constraint of consistency; they must adhere to exact game mechanics and respond instantly to player actions over long horizons. In these closed-loop environments, consistent observation replication is indispensable, as even minor hallucinations or physics violations can break the gameplay loop and disrupt agent planning.

Research in this area focuses on precisely modeling the visual dynamics of game environments, replacing traditional rendering pipelines with neural networks while preserving pixel-level details crucial for RL agents. DIAMOND [16] leverages diffusion models as interactive world simulators specifically for Atari games. By prioritizing visual fidelity, it captures intricate details that enable agents to be trained directly on high-resolution frame predictions. Pushing the boundaries of stability, GameNGen [396] demonstrates that diffusion models can function as fully real-time game engines for complex FPS environments like DOOM. Through autoregressive prediction conditioned on extended action-frame histories, it achieves a breakthrough in long-term stability, allowing for playable simulations that maintain environmental consistency over thousands of frames without diverging from the ground-truth game logic.

The complexity of game environments continues to escalate, moving from fixed-perspective arcade games to open-ended 3D worlds, requires models to handle vastly expanded state spaces and complex spatial interactions. This drives the evolution toward transformer-based architectures capable of processing multimodal tokens for sandbox environments. MineWorld [395] addresses the challenges of the Minecraft sandbox, utilizing a vision-action autoregressive Transformer to tokenize diverse game scenes. By enabling scalable, action-conditioned generation, it supports real-time human interaction in a non-stationary 3D world. To further enhance generation quality and controllability in such complex settings, Matrix-Game [397] introduces a massive 17B-parameter foundation model specifically for Minecraft-like environments. By employing an image-to-world diffusion



Fig. 23 (a) Specialist world models faithfully replicate a specific game engine, providing a low-cost sampling environment for validating learning algorithms [62]. (b) Generalist models generate diverse content to support broader gameplay [281].

paradigm, it scales up the capacity to model intricate spatial dynamics, ensuring high-fidelity and temporally coherent simulations that align with user controls.

While specialist world models have achieved remarkable success in replacing traditional engines within their respective domains, they are constrained by transferability. These frameworks tend to overfit to the idiosyncratic visual styles, physics rules, and logic of a single environment, making them brittle to any distributional shift. For instance, a representation trained on Atari games may fail to generalize to 3D FPS titles like Minecraft. This inability to bridge diverse domains highlights the necessity of evolving toward generalist world models, capable of distilling universal physical laws and interaction logic across heterogeneous virtual worlds.

6.4.2 Generalist for Diverse Content

Unlike specialist engines confined to the mechanics of a single title, generalist world models aspire to serve as universal neural simulators. These frameworks aim to decouple simulation rules from specific game assets, utilizing massive heterogeneous datasets to achieve broad applicability across diverse virtual styles and, ultimately, real-world scenarios. This evolution is driven by the need for scalable, real-time engines capable of not just replaying existing games, but synthesizing novel interactive experiences.

The first step toward this universality is establishing a scalable foundation model capable of handling multi-domain dynamics within a unified architecture. Matrix-Game 2.0 [398] exemplifies this by positioning itself as an open-source neural game engine trained on approximately 1200 hours of diverse data, ranging from GTA5 driving to Minecraft exploration. By leveraging a 1.8B-parameter Diffusion Transformer [399] with a streaming inference architecture, it proves that a single model can maintain high-fidelity visuals and temporal consistency across vastly different physics and rendering

styles. Crucially, its ability to process frame-level inputs at 25 FPS demonstrates that generalist models can achieve the real-time responsiveness required for interactive media. NitroGen [400] leverages 40000 hours of gameplay across more than 1000 titles to learn a unified mapping between visual observations and their action-conditioned evolution, providing a domain-agnostic representation of how different worlds respond to player inputs.

The ultimate frontier for generalist world models is not merely simulating what exists, but generating novel content through the disentanglement of style and logic. GameFactory [401] explicitly decouples game styles from action controls. Instead of rigid imitation, it allows for cross-domain creation, for instance, applying the action logic of Minecraft to a photorealistic open-world scene. By treating the world model as a generative canvas where dynamics and aesthetics are compositional, GameFactory enables the zero-shot creation of entirely new playable games, transitioning the role of world models from passive simulators to active creative tools.

7 Open Challenge and Future Direction

Across this survey, we traced world models from specialist to generalist and from physics-driven to data-driven methods. Despite substantial progress, no existing approach can simultaneously achieve high physical accuracy, broad generalization, and scalable learning. Which path will ultimately lead to truly universal world models remains an exciting and open question. We summarize the major challenges revealed across the discussed techniques and give our perspective on future directions, aiming to provide inspiration for future research toward scalable and reliable world models.

7.1 Compounding Error for Long-Term Prediction

Compounding error in long-horizon prediction is one of the core challenges in world-model research [27]. Even when a model produces accurate short-term dynamics, small local errors inevitably accumulate over time, gradually pushing the predicted trajectory away from physically plausible states. This compounding effect arises across all model families, whether physics-based, hybrid, or purely data-driven. Because many downstream applications depend on stable multi-step rollouts, uncontrolled error accumulation fundamentally limits the reliability and scalability of current world models.

The source of compounding error lies in the accumulation of prediction inaccuracies during multi-step rollouts. Various learning-based approaches attempt to soften this effect by supervising longer rollouts to avoid repeatedly recycling intermediate predictions [402], by incorporating backward dynamics to reduce sensitivity to early errors [403], or by replacing stepwise losses with trajectory-level distribution matching to curb the growth of long-horizon error [404]. How well such techniques carry over to settings where the underlying dynamics come from explicit physical simulation remains unclear. In practice, they can ease but not eliminate instability, and their impact is ultimately bounded by the fidelity of the world model itself, making continued improvements in predictive accuracy [405] a central direction for mitigating long-term drift. A complementary line of work addresses the issue at the policy level by estimating uncertainty in predicted rollouts and using it to decide when the model should not be trusted [5, 406, 407], providing another practical means of tolerating error while world models remain imperfect.

7.2 Active World Modeling

A world model ideally relies on diverse data, yet exhaustively covering high-dimensional state-action spaces is prohibitively expensive. A practical solution is to actively target informative samples, prioritizing regions of high model uncertainty and collecting data to reduce it [141, 408, 409]. ASID [141] exemplifies this idea by learning an exploration policy that acquires real-world data for system identification (Fig. 24), allowing the model to expose blind spots and evolve through targeted acquisition. Despite its promise, active world modeling faces two challenges. First, exploration strategies based on prediction error are easily misled by white noise [410], mistaking uninformative random fluctuations for valuable signals, which leads to wasted samples and misguided attention. Second, learning progress or information gain is difficult to estimate reliably in high-dimensional, dynamic environments:

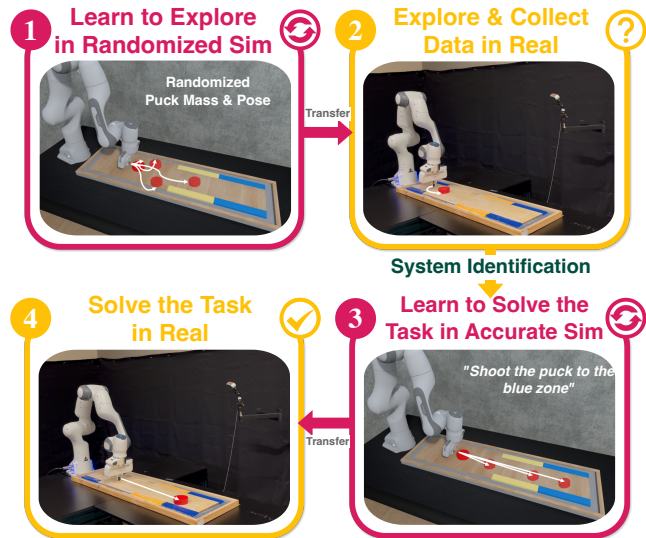


Fig. 24 ASID [141] uses an actively learned exploration policy to collect informative real-world data for system identification, allowing accurate simulator refinement and reliable sim2real transfer.

the model cannot accurately distinguish gaps that are genuinely improvable, nor maintain an effective balance between exploration and exploitation.

Recent advances suggest several scalable paths. To address the white noise problem, current research focuses on designing mechanisms that separate epistemic uncertainty from aleatoric randomness, preventing agents from fixating on inherently unlearnable dynamics. Techniques such as learning-progress monitoring [411] or Bayesian surprise within latent dynamics models [412] explicitly reward improvements in predictive accuracy, ensuring attention is directed toward genuinely learnable state transitions rather than stochastic fluctuations. To balance exploration and exploitation, existing approaches combine extrinsic rewards with intrinsic signals derived from uncertainty or learning progress [413], gradually annealing the intrinsic weight as the world model becomes more confident. More principled formulations rely on epistemic-only uncertainty via ensembles [414] or posterior sampling [415], enabling exploration to decay automatically as posterior uncertainty contracts.

7.3 Efficient Multi-Physics World Model

Many real-world scenarios inherently involve coupled multi-physics phenomena. For instance, welding simultaneously entails heat transfer, material plasticity, and fluid-like behavior of the molten pool. Yet most existing world models focus on a single or limited set of physical regimes, making cross-domain responses challenging to capture. Classical multi-physics solvers [124–126] remain the most mature solutions. However, applying them to real-world environments typically

requires heavy system identification and extensive domain-specific modeling. Moreover, their computational cost is extremely high and severely limits their applicability. Such limitations are particularly pronounced in scenarios requiring rapid adaptation to new conditions or configurations, such as flexible manufacturing [47]. Data-driven world models have advantages in computational efficiency and adaptability, but face challenges that multi-physics datasets are expensive to collect and difficult to annotate, hindering unified and accurate dynamics prediction.

To achieve efficient world modeling while reducing data requirements and domain-specific engineering, neural-physics hybrid approaches offer a promising direction. A most promising direction is physics-informed neural network (PINN) [57], which integrates physical priors through PDE constraints and can be naturally combined with existing PDE-based models to support the modeling of complex physical scenarios [416–418]. Compared with classical numerical solvers, PINNs save the requirement to generate meshes and hand-crafted discretizations, enabling end-to-end learning to solve PDE-constrained problems with GPU acceleration. This often leads to faster turnaround times, especially when repeatedly resolving similar PDEs under changing boundary conditions. Another important advantage of PINNs is their low data requirement: they can learn from sparse and partial observations [419], while PDE constraints enforce global physical consistency. This property makes PINNs particularly attractive in multi-physics domains where data acquisition is costly or incomplete, yet reliable field-level predictions are still required. In principle, PDE-constrained learning can represent multi-physics systems, but enforcing coupled PDEs within a learning framework remains difficult. Strongly coupled multi-physics processes often exhibit stiffness and require complex interface or boundary conditions. Embedding such constraints directly into the optimization objective can lead to unstable training and slow convergence. As a result, achieving robust and tightly coupled multi-physics world models remains an open research challenge.

7.4 Online System Identification

Precise world modeling offers clear advantages for high-accuracy task execution [66]. However, such models are typically constructed offline, as optimizing them often requires substantial data or computation time. If a world model could instead adapt quickly during deployment, its practical applicability would be greatly expanded. For instance, when a robot encounters an empty bottle on a table, it is beneficial to identify key physical properties with only a few interactions to

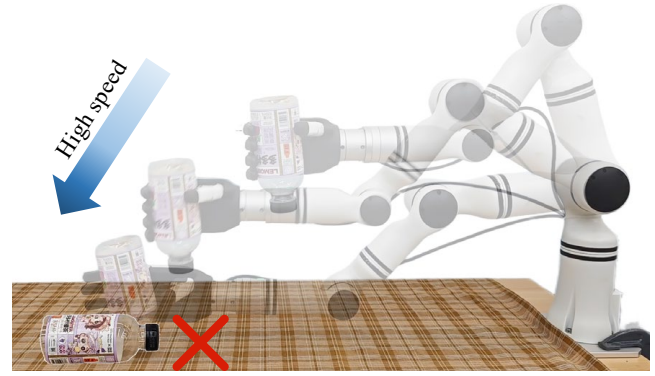


Fig. 25 A robot placing a bottle upright may fail if it relies only on geometry and overlooks that the bottle is hollow, leading to incorrect assumptions about mass distribution. Successful placement requires inferring the bottle’s physical properties through interaction.

support downstream manipulation tasks such as fast and stable placement, which are highly sensitive to underlying physical parameters; otherwise, the robot may drop or topple the bottle due to inaccurate dynamics prediction, as illustrated in Fig. 25. These use cases motivate the need for efficient online system identification, which aims to rapidly infer environment dynamics under minimal interaction budgets.

From an optimization perspective, accelerating online system identification depends on enabling the world model to efficiently adapt to new dynamic distributions. A good initialization and an effective adaptation mechanism are central to this demand. For physics-based methods, this may involve quickly estimating object-level physical priors, potentially augmented by commonsense reasoning from video language models [180]. Efficient physics-oriented parameter estimation algorithms can then refine these priors with limited data. For data-driven approaches, the key is to expose the dynamics model to diverse environments during training so that it can meta-learn a generalizable latent representation applicable across a wide range of scenarios. At test time, techniques such as test-time training [420] can be developed to allow the world model to continually adjust its latent dynamics in response to newly observed behaviors.

7.5 3D World Model

Most recent work focuses on world models that predict 2D frames [30, 257, 266], but such representations break down when spatial reasoning or physical interaction is required. In contrast, 3D world models offer explicit structure and physical grounding, supporting more faithful perception, forecasting, and decision-making [46, 68, 165]. The main barrier is data. Large-scale, high-quality 3D supervision remains difficult to acquire: real scenes require costly multi-view capture and calibration, while synthetic assets demand

extensive manual authoring. As a result, progress lags behind 2D approaches, which benefit from effectively unlimited web-scale video. Improving data availability may require simulation-generated scenes, multi-camera data collection, or generative augmentation to expand 3D coverage without prohibitive cost.

A promising path toward 3D world models is shifting from supervised training to generative [214, 421] and structured [422, 423] representations. Modern 3D generators can already synthesize diverse content [215, 424], and these results can be organized into 3D structures that interface naturally with physics engines [192, 425, 426], offering a direct route to physically meaningful environments. BANG [421] uses part-level “exploded dynamics” as a generative prior, enabling objects to be decomposed into coherent components with controllable spatial separation, effectively turning 3D generation into structure recovery and functional abstraction. CAST [214] extends this idea from objects to scenes, demonstrating how structured generation can recover occluded elements and assemble a physically plausible scene. DexSim2Real² [182] couples structure with interaction, actively building a part-aware digital twin of articulated objects from real observations, then exploiting this explicit model for sampling-based control. Looking ahead, progress will likely depend on more controllable generation that produces accurate, efficient, and general representations, enabling applications such as precision manufacturing.

7.6 AI Agents for Simulation

AI agents [222] offer a new avenue for constructing real-world dynamics. Learning-based world models still struggle with the precision and interpretability required for domains such as high-end manufacturing, whereas classical simulation remains accurate but labor-intensive to build and adapt. Agentic systems bridge this gap by using generative models to propose physically meaningful simulation engines, enabling controllable dynamic environments. Recent efforts illustrate this paradigm: Genesis [221] integrates an agent that configures scenes and delegates physical verification to a simulator, as demonstrated in Fig. 26; while FoamPilot [427] uses an LLM to set up, execute, and interpret CFD simulations for complex fire dynamics.

Despite this promise, AI agents may hallucinate implausible geometries or material properties or generate configurations that violate physical constraints, or accumulate errors when interacting autonomously with the simulator [428]. An effective approach is to validate generated simulation scenarios and use the running results to guide self-correction in the

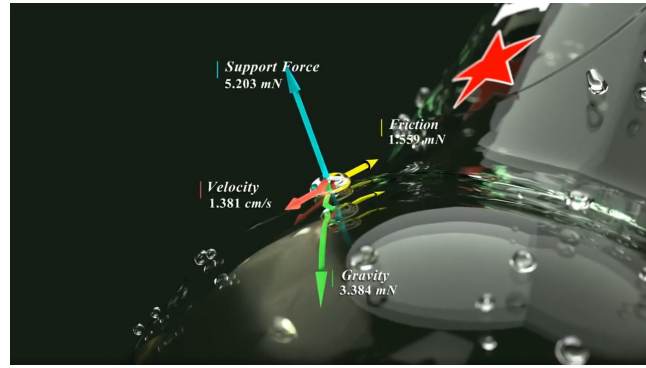


Fig. 26 Genesis [221] employs an AI agent to automatically generate simulation scenarios. Given the prompt “A water droplet falls onto a beer bottle and slowly slides down its surface”, it simulates and visualizes the droplet’s forces and velocity.

generative model. KUNWU [223] synthesizes 3D industrial production lines from 2D sketches, generates control logic according to user instructions, and iteratively refines its outputs based on feedback when the simulated control software reveals failures. Looking forward, enhancing both the generative accuracy and the self-correction mechanisms of AI agents for accurate and robust simulation will be crucial for deploying them in high-stakes simulation domains, such as industrial manufacturing and aerospace, where robustness, physical fidelity, and reliability are paramount.

7.7 Object-Centric World Model

Object-centric world models explicitly decompose environments into discrete entities with persistent identities, naturally enabling compositional generalization. However, constructing such structured representations from raw sensory data remains challenging. Reliable object segmentation and consistent identity binding across viewpoints, occlusions, and appearance changes are often fragile. Moreover, relational reasoning must jointly capture rigid contacts, non-rigid deformations, and articulated part interactions, while remaining robust to varying object counts and topologies. Mobility analysis further complicates this process by requiring the model to distinguish fixed background elements, articulated structures, and freely moving objects. Together, these intertwined perception and reasoning challenges make fully explicit object-centric world models difficult to learn and scale in complex real-world settings.

A promising path forward is to treat modern visual foundation models as front-end perception supervisors, rather than asking a world model to discover objects from scratch. Large-scale segmenters such as SAM2 [429], trackers like Track-Anything [430], and 6D pose estimators such as FoundationPose [431] can already maintain instance identities

through long temporal windows, strong appearance variation, and partial occlusion, capabilities that historically required carefully engineered pipelines. Instead of relying on fragile unsupervised binding, world models can freeze these perception modules and operate directly on stable object slots, allowing learning to focus on dynamics, interaction, and prediction rather than front-end detection. The same trend is emerging in 3D perception. Models like SAM3D [432] provide reliable spatial proposals from images, giving world models access to geometric priors, coarse part structure, and physical scale, which are essential for reasoning about contact or articulation. With such front-ends in place, world models can be trained under explicit relational supervision. This suggests a broader shift: rather than treating object-centric modeling as a standalone discovery problem, future systems may adopt a hybrid foundation-model pipeline and the world model is optimized for predictive consistency. Such a division of labor may finally make scalable object-centric world models viable in real-world environments.

7.8 Hierarchical World Model

Hierarchical world models offer substantial theoretical promise for advancing autonomous intelligence by facilitating multi-level abstraction that parallels human cognition. In complex, dynamic environments, they enable the decomposition of problems into nested layers, ranging from low-level physical dynamics to high-level agentic intentions and conceptual strategies, thereby supporting long-term planning, counterfactual inference, and adaptive decision-making across diverse scales. For instance, they mitigate the limitations of flat architectures, such as video-generative models or joint embedding predictors, which often underperform in open-ended tasks requiring temporal hierarchies, multi-agent interactions, or socially grounded reasoning. Through hierarchical prediction structures, these models address representation collapse in continuous embeddings and enhance stability by integrating discrete concepts to capture deeper semantics, including action intentions and causal chains.

A key avenue for advancing hierarchical world models lies in adaptive, nested architectures like PAN (Physical, Agentic, and Nested) [433], which leverages multi-layer latent predictions to systematically decompose simulations into actionable components, as demonstrated in Fig. 27. PAN commences with a sensory encoder that maps multimodal inputs to hierarchical states, blending discrete tokens for abstract reasoning with continuous embeddings for perceptual fidelity, supported by VQ-VAE-style vocabularies that

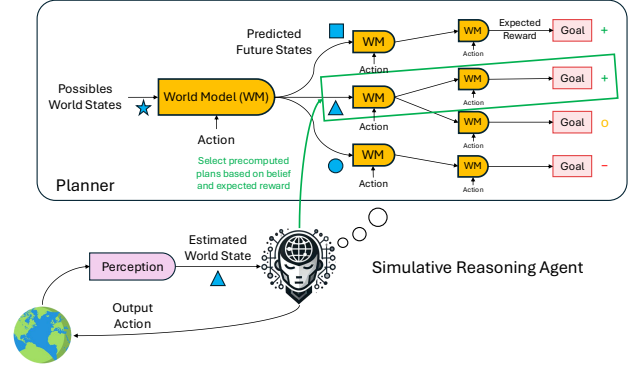


Fig. 27 A simulative reasoning agent powered by the PAN world model [433], which selects actions by querying precomputed, belief-conditioned simulations to enable efficient, long-horizon, goal-directed planning beyond reactive policies or online MPC.

preserve fine-grained distinctions. The world model backbone, augmented by large language models for semantic guidance and diffusion predictors for generative accuracy, forecasts states layer-by-layer through a learned switch for task-adaptive component selection, ensuring spatio-temporal flexibility from physical concreteness to agentic abstraction. Looking ahead, incorporating pretrained foundation models for initial layer bootstrapping will further improve training efficiency and generalization, paving the way for robust deployment in diverse real-world contexts.

7.9 Recover Actions from Passive Video Data

The goal of world models is to capture dynamics across the full state-action space, which in turn demands large-scale and diverse experience. A promising direction is to leverage web-scale videos, which provide rich visual coverage of everyday physical interactions. However, videos typically lack explicit action annotations. As a result, world models trained purely on Internet videos cannot produce correct action-conditioned predictions, making them difficult to apply directly to control. A world model must infer how interventions change the world, not merely what the next frame looks like. Therefore, a natural next step is to recover controllable action representations from passive visual data.

To make world models trained on passive videos actionable, recent works explore latent actions. Genie [93] demonstrates this idea at scale: it learns a spatiotemporal tokenizer and an autoregressive dynamics model, then introduces a lightweight latent action interface that allows users to intervene frame by frame, despite never seeing ground-truth action labels. Latent control softens the annotation bottleneck, but the resulting embeddings lack physical interpretability, making precise force or motion specification difficult and often too low-dimensional to express complex manipulation. Another

direction is inverse dynamics learning [58], which predicts the action that could transform one observed state into the next. This offers a more explicit grounding than latent actions, but the formulation is inherently ill-posed since many actions may lead to similar visual transitions. Besides, inverse dynamics pipelines remain vulnerable to compounding error: as the model rolls forward in time, small prediction deviations accumulate, eventually diverging from original trajectories. As a result, there is still no reliable method for turning Internet-scale video into stable, action-conditioned world models, and developing mechanisms that maintain physical plausibility over long horizons remains an open challenge.

7.10 Physical Consistency in WFM

Recent progress in WFMs [19] has demonstrated impressive generation capabilities across diverse dynamic scenarios. Yet these models still frequently exhibit hallucinated physical behaviors [434], such as violations of object permanence, inconsistent contact, and unrealistic fluid-rigid interactions, making physical consistency a central challenge. An example is shown in Fig. 28, where a cup on a table suddenly jumps, violating physical laws. Such errors limit the applicability of WFMs in downstream tasks that require reliable dynamics. Yet constraints [15, 57, 231] on both model architecture and outputs make these approaches difficult to integrate directly into WFMs, which rely on foundation model designs and output high-dimensional video outputs.

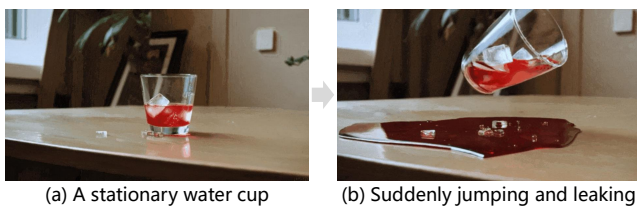


Fig. 28 World foundation models can still exhibit physical hallucinations [435], such as a cup on a table suddenly jumping, limiting their use in tasks like robotic control.

Novel training designs are needed to instill physical consistency of WFMs. On the architectural side, WFMs may incorporate physics-aligned inductive biases to improve predictive stability, such as multi-view encoders that compress observations into informative latents with 3D geometric priors [436, 437]. On the data side, WFMs fundamentally rely on large, diverse, and physically valid dynamic trajectories. This requires generating or collecting vast amounts of physically consistent video data at low cost. Cosmos [51] synthesizes large volumes of simulation scenes to cover a wide distribution of physical dynamics. PhysCtrl [438] uses simulations to produce large-scale examples of how different materials

move under forces, then trains a diffusion model to imitate these dynamics in the form of 3D point trajectories. These learned trajectories are later used as control signals for a video generator, enabling visually realistic motion without explicit simulation. Another promising direction is to leverage a generalist model to automatically generate physics-based specialist models [192, 193] for targeted scenarios, potentially using techniques with differentiable physics and rendering pipelines [46]. These specialists can then produce a physically accurate video that serves as corrective supervision for the generalist, enabling a closed-loop process where the generalist can continually improve itself.

8 Conclusion

World models have progressed from early system identification to modern large-scale generative architectures capable of learning complex dynamics from multimodal experience. This survey examined that evolution through a unified lens, emphasizing how different forms of world models approach the fundamental challenges of accuracy and generalization in complex environments. By positioning existing methods along the continuum from specialist to generalist and from physics-based to data-driven, we delineate the core design trade-offs: specialist models offer high-fidelity, controllable environments but remain narrow in scope, whereas generalist models provide broad coverage and flexibility at the cost of reduced accuracy, interpretability, and reliability.

We review the essential components that define contemporary world models, including state abstraction, action parameterization, and dynamic architectures, and discussed how these components are composed into systems ranging from physics simulation to world foundation models. We further summarized practical usages across policy improvement, gradient-based optimization, evaluation and selection, and video generation. Across major applications, including manipulation, locomotion, navigation, and game engines, world models have shown their effectiveness in enabling agents to perform increasingly diverse tasks with reduced reliance on explicit supervision. Yet, achieving both high accuracy and broad generalization remains a central open challenge. Compounding prediction error, data inefficiency, and physical inconsistency still constrain its deployment. We hope this survey provides clear conceptual structure and technical grounding for the field, offering guidance for future research toward world models that are simultaneously scalable, physically grounded, and generalizable, ultimately bringing us closer to artificial general intelligence.

References

- [1] Ha D, Schmidhuber J. World models. *arXiv preprint arXiv:1803.10122*, 2018, 2(3).
- [2] Wiener N. Cybernetics or Control and Communication in the Animal and the Machine, 2019.
- [3] Marr D. A theory of cerebellar cortex. *The Journal of physiology*, 1969, 202(2): 437–470.
- [4] Hafner D, Lillicrap T, Fischer I, Villegas R, Ha D, Lee H, Davidson J. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, 2019, 2555–2565.
- [5] Yu T, Thomas G, Yu L, Ermon S, Zou JY, Levine S, Finn C, Ma T. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 2020, 33: 14129–14142.
- [6] Mendonca R, Bahl S, Pathak D. Structured World Models from Human Videos. In *Robotics: Science and Systems*, 2023.
- [7] Hansen N, Su H, Wang X. TD-MPC2: Scalable, Robust World Models for Continuous Control. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.
- [8] Hafner D, Pasukonis J, Ba J, Lillicrap T. Mastering diverse control tasks through world models. *Nature*, 2025: 1–7.
- [9] Lou H, Liu Y, Pan Y, Geng Y, Chen J, Ma W, Li C, Wang L, Feng H, Shi L, et al.. Robo-gs: A physics consistent spatial-temporal model for robotic arm with hybrid representation. In *IEEE International Conference on Robotics and Automation*, 2025.
- [10] Luo FM, Xu T, Lai H, Chen XH, Zhang W, Yu Y. A survey on model-based reinforcement learning. *Science China Information Sciences*, 2024, 67(2): 121101.
- [11] Vuong Q, Levine S, Walke HR, Pertsch K, Singh A, Doshi R, Xu C, Luo J, Tan L, Shah D, et al.. Open x-embodiment: Robotic learning datasets and rt-x models. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023.
- [12] Zhou Y, Wang Y, Zhou J, Chang W, Guo H, Li Z, Ma K, Li X, Wang Y, Zhu H, et al.. OmniWorld: A Multi-Domain and Multi-Modal Dataset for 4D World Modeling. *arXiv preprint arXiv:2509.12201*, 2025.
- [13] Bu Q, Cai J, Chen L, Cui X, Ding Y, Feng S, He X, Huang X, et al.. Agibot world colosseum: A large-scale manipulation platform for scalable and intelligent embodied systems. In *IEEE International Conference on Intelligent Robots and Systems*, 2025.
- [14] Chen Z, Liu T, Zhuo L, Ren J, Tao Z, Zhu H, Hong F, Pan L, Liu Z. 4dnex: Feed-forward 4d generative modeling made easy. *arXiv preprint arXiv:2508.13154*, 2025.
- [15] Hu Y, Anderson L, Li TM, Sun Q, Carr N, Ragan-Kelley J, Durand F. DiffTaichi: Differentiable Programming for Physical Simulation. In *International Conference on Learning Representations*, 2020.
- [16] Alonso E, Jelley A, Micheli V, Kanervisto A, Storkey AJ, Pearce T, Fleuret F. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 2024, 37: 58757–58791.
- [17] Bommasani R. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [18] Pearce T, Rashid T, Bignell D, Georgescu R, Devlin S, Hofmann K. Scaling Laws for Pre-training Agents and World Models. In *International Conference on Machine Learning*, 2025.
- [19] Brooks T, Peebles B, Holmes C, DePue W, Guo Y, Jing L, Schnurr D, Taylor J, Luhman T, Luhman E, Ng C, Wang R, Ramesh A. Video generation models as world simulators, 2024.
- [20] Richens J, Everitt T, Abel D. General agents need world models. In *International Conference on Machine Learning*, 2025.
- [21] Zhang T, Chen G, Chen F. When do neural networks learn world models? In *International Conference on Machine Learning*, 2025.
- [22] Sobal V, Zhang W, Cho K, Balestriero R, Rudner TG, LeCun Y. Stress-Testing Offline Reward-Free Reinforcement Learning: A Case for Planning with Latent Dynamics Models. In *7th Robot Learning Workshop: Towards Robots with Human-Level Abilities*.
- [23] Xu J, Tian Y, Ma P, Rus D, Sueda S, Matusik W. Prediction-guided multi-objective reinforcement learning for continuous robot control. In *International Conference on Machine Learning*, 2020, 10607–10616.
- [24] Zhi H, Chen P, Zhou S, Dong Y, Wu Q, Han L, Tan M. 3DFlowAction: Learning Cross-Embodiment Manipulation from 3D Flow World Model. *arXiv preprint arXiv:2506.06199*, 2025.
- [25] Lin F, Hu Y, Sheng P, Wen C, You J, Gao Y. Data Scaling Laws in Imitation Learning for Robotic Manipulation. In *International Conference on Learning Representations*, 2025.
- [26] Wang T, Bao X, Clavera I, Hoang J, Wen Y, Langlois E, Zhang S, Zhang G, Abbeel P, Ba J. Benchmarking model-based reinforcement learning. *arXiv preprint arXiv:1907.02057*, 2019.
- [27] Lambert N, Pister K, Calandra R. Investigating compounding prediction errors in learned dynamics models. *arXiv preprint arXiv:2203.09637*, 2022.
- [28] Xiong X, Mu N, Xie R, Yang S, Wang Y, Wang L, Luan Y, Li S, Xu S, Yang Y, et al.. MrCoM: A Meta-Regularized World-Model Generalizing Across Multi-Scenarios. *arXiv preprint arXiv:2511.06252*, 2025.
- [29] Ljung L. System identification. In *Signal Analysis and Prediction*, 1998.
- [30] Ren X, Lu Y, Cao T, Gao R, Huang S, Sabour A, Shen T, Pfaff T, Wu JZ, Chen R, et al.. Cosmos-Drive-Dreams: Scalable Synthetic Driving Data Generation with World Foundation Models. *arXiv preprint arXiv:2506.09042*, 2025.
- [31] Ding J, Zhang Y, Shang Y, Zhang Y, Zong Z, Feng J, Yuan Y, Su H, Li N, Sukiennik N, et al.. Understanding world or



- predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 2024.
- [32] Cao Y, Lu J, Huang Z, Shen Z, Zhao C, Hong F, Chen Z, Li X, Wang W, Liu Y, et al.. Reconstructing 4d spatial intelligence: A survey. *arXiv preprint arXiv:2507.21045*, 2025.
- [33] Liu D, Zhang J, Dinh AD, Park E, Zhang S, Mian A, Shah M, Xu C. Generative physical ai in vision: A survey. *arXiv preprint arXiv:2501.10928*, 2025.
- [34] Kong L, Yang W, Mei J, Liu Y, Liang A, Zhu D, Lu D, Yin W, Hu X, Jia M, et al.. 3d and 4d world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025.
- [35] Li X, He X, Zhang L, Liu Y. A Comprehensive Survey on World Models for Embodied AI. *arXiv preprint arXiv:2510.16732*, 2025.
- [36] Long X, Zhao Q, Zhang K, Zhang Z, Wang D, Liu Y, Shu Z, Lu Y, Wang S, Wei X, et al.. A Survey: Learning Embodied Intelligence from Physical Simulators and World Models. *arXiv preprint arXiv:2507.00917*, 2025.
- [37] Feng T, Wang W, Yang Y. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025.
- [38] Lin M, Wang X, Wang Y, Wang S, Dai F, Ding P, Wang C, Zuo Z, Sang N, Huang S, et al.. Exploring the evolution of physics cognition in video generation: A survey. *arXiv preprint arXiv:2503.21765*, 2025.
- [39] Xie N, Tian Z, Yang L, Zhang XP, Guo M, Li J. From 2D to 3D Cognition: A Brief Survey of General World Models. *arXiv preprint arXiv:2506.20134*, 2025.
- [40] Xiang K, Zhang TJ, Huang Y, He J, Liu Z, Tang Y, Zhou R, Luo L, Wen Y, Chen X, et al.. Aligning Perception, Reasoning, Modeling and Interaction: A Survey on Physical AI. *arXiv preprint arXiv:2510.04978*.
- [41] Ma Y, Feng K, Hu Z, Wang X, Wang Y, Zheng M, He X, Zhu C, Liu H, He Y, et al.. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025.
- [42] Reddy JN. An introduction to the finite element method. *New York*, 1993, 27(14).
- [43] Makoviychuk V, Wawrzyniak L, Guo Y, Lu M, Storey K, Macklin M, Hoeller D, Rudin N, Allshire A, Handa A, State G. Isaac Gym: High Performance GPU Based Physics Simulation For Robot Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- [44] Hafner D, Lillicrap T, Ba J, Norouzi M. Dream to Control: Learning Behaviors by Latent Imagination. In *International Conference on Learning Representations*, 2020.
- [45] Wu P, Escontrela A, Hafner D, Abbeel P, Goldberg K. Daydreamer: World models for physical robot learning. In *Conference on Robot Learning*, 2023, 2226–2240.
- [46] Li W, Zhao H, Yu Z, Du Y, Zou Q, Hu R, Xu K. Pin-wm: Learning physics-informed world models for non-prehensile manipulation. *arXiv preprint arXiv:2504.16693*, 2025.
- [47] Kai X, Hang Z, Ruizhen H, Min Y, Hao L, Hui Z, Haibin Y. Embodied Intelligence for Flexible Manufacturing: A Survey. *ROBOT*, 2025, 47(4): 581–624.
- [48] Zhao H, Xu J, Yu K, Hu R, Zhu C, Du B, Xu K. Deliberate planning of 3D bin packing on packing configuration trees. *The International Journal of Robotics Research*: 02783649251380619.
- [49] Tang B, Akinola I, Xu J, Wen B, Handa A, Wyk KV, Fox D, Sukhatme GS, Ramos F, Narang Y. AutoMate: Specialist and Generalist Assembly Policies over Diverse Geometries. In *Robotics: Science and Systems*, 2024.
- [50] Zhang Z, Chen R, Ye J, Sun Y, Wang P, Pang J, Li K, Liu T, Lin H, Yu Y, et al.. WHALE: Towards Generalizable and Scalable World Models for Embodied Decision-making. *arXiv preprint arXiv:2411.05619*, 2024.
- [51] Agarwal N, Ali A, Bala M, Balaji Y, Barker E, Cai T, Chattopadhyay P, Chen Y, Cui Y, Ding Y, et al.. Cosmos world foundation model platform for physical ai. *IEEE Conference on Computer Vision and Pattern Recognition*, 2025.
- [52] Wu J, Chong W, Holmberg R, Prasad A, Gao Y, Khatib O, Song S, Rusinkiewicz S, Bohg J. TidyBot++: An Open-Source Holonomic Mobile Manipulator for Robot Learning. In *Conference on Robot Learning*, 2024.
- [53] Fu Z, Zhao TZ, Finn C. Mobile ALOHA: Learning Bimanual Mobile Manipulation using Low-Cost Whole-Body Teleoperation. In *Conference on Robot Learning*, 2024.
- [54] Miki T, Lee J, Hwangbo J, Wellhausen L, Koltun V, Hutter M. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science robotics*, 2022, 7(62): eabk2822.
- [55] Schubert I, Zhang J, Bruce J, Bechtle S, Parisotto E, Riedmiller M, Springenberg JT, Byravan A, Hasenclever L, Heess N. A generalist dynamics model for control. *arXiv preprint arXiv:2305.10912*, 2023.
- [56] Heiden E, Millard D, Coumans E, Sheng Y, Sukhatme GS. NeuralSim: Augmenting differentiable simulators with neural networks. In *IEEE International Conference on Robotics and Automation*, 2021.
- [57] Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 2019.
- [58] Jang J, Ye S, Lin Z, Xiang J, Bjorck J, Fang Y, Hu F, Huang S, Kundalia K, Lin YC, et al.. DreamGen: Unlocking Generalization in Robot Learning through Video World Models. *arXiv preprint arXiv:2505.12705*, 2025.
- [59] Feng Y, Tan H, Mao X, Xiang C, Liu G, Huang S, Su H, Zhu J. Vidar: Embodied Video Diffusion Model for Generalist Manipulation. *arXiv preprint arXiv:2507.12898*, 2025.
- [60] Wen Y, Lin J, Zhu Y, Han J, Xu H, Zhao S, Liang X. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. *Advances in Neural Information Processing Systems*, 2024, 37: 41051–41075.
- [61] Liang J, Liu R, Ozguroglu E, Sudhakar S, Dave A, Tokmakov

- P, Song S, Vondrick C. Dreamitate: Real-World Visuomotor Policy Learning via Video Generation. In *8th Annual Conference on Robot Learning*, 2024.
- [62] Hafner D, Yan W, Lillicrap T. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.
- [63] de Avila Belbute-Peres F, Smith K, Allen K, Tenenbaum J, Kolter JZ. End-to-end differentiable physics for learning and control. *Advances in Neural Information Processing Systems*, 2018.
- [64] Chen PY, Liu C, Ma P, Eastman J, Rus D, Randle D, Ivanov Y, Matusik W. Learning Object Properties Using Robot Proprioception via Differentiable Robot-Object Interaction. In *IEEE International Conference on Robotics and Automation*, 2025, 5997–6004.
- [65] Song C, Boularias A. Learning to Slide Unknown Objects with Differentiable Physics Simulations. In *Robotics: Science and Systems*, 2020.
- [66] Baumeister F, Mack L, Stueckler J. Incremental Few-Shot Adaptation for Non-Prehensile Object Manipulation Using Parallelizable Physics Simulators. In *IEEE International Conference on Robotics and Automation*, 2025, 15394–15400.
- [67] Heiden E, Liu Z, Vineet V, Coumans E, Sukhatme GS. Inferring articulated rigid body dynamics from rgb-d video. In *IEEE International Conference on Intelligent Robots and Systems*, 2022.
- [68] Zhu Y, Xiang T, Dollar AM, Pan Z. One-Shot Real-to-Sim via End-to-End Differentiable Simulation and Rendering. *IEEE Robotics and Automation Letters*, 2025.
- [69] Cao J, Guan S, Ge Y, Li W, Yang X, Ma C. NeuMA: Neural Material Adaptor for Visual Grounding of Intrinsic Dynamics. In *Advances in Neural Information Processing Systems*, 2024.
- [70] Huang S, Chen Q, Zhang X, Sun J, Schwager M. ParticleFormer: A 3D Point Cloud World Model for Multi-Object, Multi-Material Robotic Manipulation. In *Conference on Robot Learning*, 2025.
- [71] Strecke M, Stueckler J. DiffSDFSim: Differentiable rigid-body dynamics with implicit shapes. In *International Conference on 3D Vision*, 2021.
- [72] Zhu X, Ke J, Xu Z, Sun Z, Bai B, Lv J, Liu Q, Zeng Y, Ye Q, Lu C, et al.. Diff-lfd: Contact-aware model-based learning from visual demonstration for robotic manipulation via differentiable physics-based simulation and rendering. In *Conference on Robot Learning*, 2023, 499–512.
- [73] Lu G, Zhang S, Wang Z, Liu C, Lu J, Tang Y. Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation. In *European Conference on Computer Vision*, 2025.
- [74] Lu G, Jia B, Li P, Chen Y, Wang Z, Tang Y, Huang S. GWM: Towards Scalable Gaussian World Models for Robotic Manipulation. *International Conference on Computer Vision*, 2025.
- [75] Huang B, Yu Z, Chen A, Geiger A, Gao S. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH*, 2024.
- [76] Wang P, Liu L, Liu Y, Theobalt C, Komura T, Wang W. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *Advances in Neural Information Processing Systems*, 2021, 34: 27171–27183.
- [77] Veerapaneni R, Co-Reyes JD, Chang M, Janner M, Finn C, Wu J, Tenenbaum J, Levine S. Entity abstraction in visual model-based reinforcement learning. In *Conference on Robot Learning*, 2020, 1439–1456.
- [78] Janner M, Levine S, Freeman WT, Tenenbaum JB, Finn C, Wu J. Reasoning about physical interactions with object-oriented prediction and planning. *arXiv preprint arXiv:1812.10972*, 2018.
- [79] Driess D, Huang Z, Li Y, Tedrake R, Toussaint M. Learning multi-object dynamics with compositional neural radiance fields. In *Conference on Robot Learning*, 2023, 1755–1768.
- [80] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Attention Networks. In *International Conference on Learning Representations*, 2018.
- [81] Corso G, Stark H, Jegelka S, Jaakkola T, Barzilay R. Graph neural networks. *Nature Reviews Methods Primers*, 2024, 4(1): 17.
- [82] Manuelli L, Li Y, Florence P, Tedrake R. Keypoints into the Future: Self-Supervised Correspondence in Model-Based Reinforcement Learning. In *Conference on Robot Learning*, volume 155, 2021, 693–710.
- [83] Ma X, Hsu D, Lee WS. Learning latent graph dynamics for visual manipulation of deformable objects. In *IEEE International Conference on Robotics and Automation*, 2022, 8266–8273.
- [84] Zhang K, Li B, Hauser K, Li Y. AdaptiGraph: Material-Adaptive Graph-Based Neural Dynamics for Robotic Manipulation. In *Robotics: Science and Systems*, 2024.
- [85] Hafner D, Lillicrap TP, Norouzi M, Ba J. Mastering Atari with Discrete World Models. In *International Conference on Learning Representations*, 2021.
- [86] LeCun Y. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 2022, 62(1): 1–62.
- [87] Li C, Krause A, Hutter M. Robotic world model: A neural network simulator for robust policy optimization in robotics. *arXiv preprint arXiv:2501.10100*, 2025.
- [88] Zhi P, Li P, Yin J, Jia B, Huang S. Learning a Unified Policy for Position and Force Control in Legged Loco-Manipulation. In *Conference on Robot Learning*, 2025, 652–669.
- [89] Chen W, Zeng C, Liang H, Sun F, Zhang J. Multimodality driven impedance-based sim2real transfer learning for robotic multiple peg-in-hole assembly. *IEEE Transactions on Cybernetics*, 2023, 54(5): 2784–2797.
- [90] Lakshminarayanan S, Kana S, Mohan DM, Manyar OM, Then D, Campolo D. An adaptive framework for robotic polishing based on impedance control. *The International Journal of*



- Advanced Manufacturing Technology*, 2021, 112(1): 401–417.
- [91] Chen C, Xiao R, Chen H, Lv N, Chen S. Prediction of welding quality characteristics during pulsed GTAW process of aluminum alloy by multisensory fusion and hybrid network model. *Journal of Manufacturing Processes*, 2021, 68: 209–224.
- [92] Alt B, Stöckl F, Müller S, Braun C, Raible J, Alhasan S, Rettig O, Ringle L, Katic D, Jäkel R, et al.. Robogrind: Intuitive and interactive surface treatment with industrial robots. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, 2140–2146.
- [93] Bruce J, Dennis MD, Edwards A, Parker-Holder J, Shi Y, Hughes E, Lai M, Mavalankar A, Steigerwald R, Apps C, et al.. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.
- [94] Li Y, Zhu Y, Wen J, Shen C, Xu Y. WorldEval: World Model as Real-World Robot Policies Evaluator. *arXiv preprint arXiv:2505.19017*, 2025.
- [95] Bu Q, Yang Y, Cai J, Gao S, Ren G, Yao M, Luo P, Li H. Learning to Act Anywhere with Task-centric Latent Actions. In *Robotics: Science and Systems*, 2025.
- [96] Jia C, Li Z, Wang P, Li YC, Hou Z, Dong Y, Yu Y. Controlling Large Language Model with Latent Actions. *arXiv preprint arXiv:2503.21383*, 2025.
- [97] Zhu Z, Wang X, Zhao W, Min C, Deng N, Dou M, Wang Y, Shi B, Wang K, Zhang C, et al.. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024.
- [98] Chen S, Ge C, Zhang Y, Zhang Y, Zhu F, Yang H, Hao H, Wu H, Lai Z, Hu Y, et al.. Goku: Flow based video generative foundation models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025, 23516–23527.
- [99] Sulsky D, Chen Z, Schreyer HL. A particle method for history-dependent materials. *Computer Methods in Applied Mechanics and Engineering*, 1994, 118(1-2): 179–196.
- [100] Gingold RA, Monaghan JJ. Smoothed particle hydrodynamics: theory and application to non-spherical stars. *Monthly Notices of the Royal Astronomical Society*, 1977.
- [101] Müller M, Heidelberger B, Teschner M, Gross M. Meshless deformations based on shape matching. *ACM Transactions on Graphics*, 2005, 24(3): 471–478.
- [102] Baraff D. Fast contact force computation for nonpenetrating rigid bodies. In *SIGGRAPH*, 1994, 23–34.
- [103] Li M, Ferguson Z, Schneider T, Langlois TR, Zorin D, Panozzo D, Jiang C, Kaufman DM. Incremental potential contact: intersection-and inversion-free, large-deformation dynamics. *ACM Transactions on Graphics*, 2020, 39(4): 49.
- [104] Yu T, Quillen D, He Z, Julian R, Hausman K, Finn C, Levine S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, 2020, 1094–1100.
- [105] James S, Ma Z, Arrojo DR, Davison AJ. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 2020, 5(2): 3019–3026.
- [106] Wang J, Hertzmann A, Fleet DJ. Gaussian process dynamical models. *Advances in Neural Information Processing Systems*, 2005, 18.
- [107] Deisenroth M, Rasmussen CE. PILCO: A model-based and data-efficient approach to policy search. In *International Conference on Machine Learning*, 2011, 465–472.
- [108] Yi Z, Calandra R, Veiga F, van Hoof H, Hermans T, Zhang Y, Peters J. Active tactile object exploration with gaussian processes. In *IEEE International Conference on Intelligent Robots and Systems*, 2016, 4925–4930.
- [109] Lu C, Schroecker Y, Gu A, Parisotto E, Foerster J, Singh S, Behbahani F. Structured state space models for in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 2023, 36: 47016–47031.
- [110] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017, 30.
- [111] Micheli V, Alonso E, Fleuret F. Transformers are Sample-Efficient World Models. In *International Conference on Learning Representations*, 2023.
- [112] Robine J, Höftmann M, Uelwer T, Harmeling S. Transformer-based World Models Are Happy With 100k Interactions. In *International Conference on Learning Representations*, 2023.
- [113] Chen C, Yoon J, Wu YF, Ahn S. TransDreamer: Reinforcement Learning with Transformer World Models. In *Deep RL Workshop NeurIPS 2021*, 2021.
- [114] Assran M, Bardes A, Fan D, Garrido Q, Howes R, Muckley M, Rizvi A, Roberts C, Sinha K, Zholus A, et al.. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.
- [115] Wu J, Yin S, Feng N, He X, Li D, Hao J, Long M. ivideopt: Interactive videopts are scalable world models. *Advances in Neural Information Processing Systems*, 2024, 37: 68082–68119.
- [116] Chen L, Wang Y, Tang S, Ma Q, He T, Ouyang W, Zhou X, Bao H, Peng S. EgoAgent: A Joint Predictive Agent Model in Egocentric Worlds. *arXiv preprint arXiv:2502.05857*, 2025.
- [117] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020, 33: 6840–6851.
- [118] Song J, Meng C, Ermon S. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, 2021.
- [119] Chen F, Han F, Guan C, Yuan L, Zhang Z, Yu Y, Zhang Z. Stable Continual Reinforcement Learning via Diffusion-based Trajectory Replay. In *ICLR 2024 Workshop on Generative Models for Decision Making*, 2024.
- [120] Ding Z, Zhang A, Tian Y, Zheng Q. Diffusion world model: Future modeling beyond step-by-step rollout for offline reinforcement learning. *arXiv preprint arXiv:2402.03570*, 2024.

- [121] Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, et al.. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [122] Liao Y, Zhou P, Huang S, Yang D, Chen S, Jiang Y, Hu Y, Cai J, Liu S, Luo J, et al.. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.
- [123] Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, et al.. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2025, 43(2): 1–55.
- [124] Stolarski T, Nakasone Y, Yoshimoto S. Engineering analysis with ANSYS software, 2018.
- [125] Barbero EJ. Finite element analysis of composite materials using Abaqus®, 2023.
- [126] Multiphysics C. Comsol multiphysics, 2014.
- [127] Logg A, Mardal KA, Wells G. Automated solution of differential equations by the finite element method: The FEniCS book, 2012, 84.
- [128] Bangerth W, Hartmann R, Kanschat G. deal. II—a general-purpose object-oriented finite element library. *ACM Transactions on Mathematical Software (TOMS)*, 2007, 33(4): 24–es.
- [129] NVIDIA. Omniverse Platform for OpenUSD. <https://www.nvidia.com/en-us/omniverse/>, 2021.
- [130] Coumans E, Bai Y. PyBullet, a Python module for physics simulation for games, robotics and machine learning. <http://pybullet.org>, 2016–2019.
- [131] Todorov E, Erez T, Tassa Y. Mujoco: A physics engine for model-based control. In *IEEE International Conference on Intelligent Robots and Systems*, 2012.
- [132] Xiang F, Qin Y, Mo K, Xia Y, Zhu H, Liu F, Liu M, Jiang H, Yuan Y, Wang H, et al.. Sapien: A simulated part-based interactive environment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, 11097–11107.
- [133] Featherstone R. Rigid body dynamics algorithms, 2008.
- [134] Freeman CD, Frey E, Raichuk A, Girgin S, Mordatch I, Bachem O. Brax - A Differentiable Physics Engine for Large Scale Rigid Body Simulation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.
- [135] Chen T, Tippur M, Wu S, Kumar V, Adelson E, Agrawal P. Visual dexterity: In-hand reorientation of novel and complex object shapes. *Science Robotics*, 2023, 8(84): eadc9244.
- [136] Gu X, Wang YJ, Chen J. Humanoid-gym: Reinforcement learning for humanoid robot with zero-shot sim2real transfer. *arXiv preprint arXiv:2404.05695*, 2024.
- [137] Peng XB, Andrychowicz M, Zaremba W, Abbeel P. Sim-to-real transfer of robotic control with dynamics randomization. In *IEEE International Conference on Robotics and Automation*, 2018.
- [138] Tobin J, Fong R, Ray A, Schneider J, Zaremba W, Abbeel P. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE International Conference on Intelligent Robots and Systems*, 2017.
- [139] Mehta B, Diaz M, Golemo F, Pal CJ, Paull L. Active domain randomization. In *Conference on Robot Learning*, 2020.
- [140] Ramos F, Possas R, Fox D. BayesSim: Adaptive Domain Randomization Via Probabilistic Inference for Robotics Simulators. In *Proceedings of Robotics: Science and Systems*, 2019, doi:10.15607/RSS.2019.XV.029.
- [141] Memmel M, Wagenmaker A, Zhu C, Fox D, Gupta A. ASID: Active Exploration for System Identification in Robotic Manipulation. In *International Conference on Learning Representations*, 2024.
- [142] Rubinstein RY, Kroese DP. The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning, 2004.
- [143] Murthy JK, Macklin M, Golemo F, Voleti V, Petrini L, Weiss M, Considine B, Parent-Lévesque J, Xie K, Erleben K, Paull L, Shkurti F, Nowrouzezahrai D, Fidler S. gradSim: Differentiable simulation for system identification and visuomotor control. In *International Conference on Learning Representations*, 2021.
- [144] Ruan Q, Lei J, Yuan W, Zhang Y, Lu D, Liu G, Jia K. Prof. Robot: Differentiable Robot Rendering Without Static and Self-Collisions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025.
- [145] Hu Y, Liu J, Spielberg A, Tenenbaum JB, Freeman WT, Wu J, Rus D, Matusik W. ChainQueen: A Real-Time Differentiable Physical Simulator for Soft Robotics. In *IEEE International Conference on Robotics and Automation*, 2019, 6265–6271.
- [146] Billard A, Albu-Schaeffer A, Beetz M, Burgard W, Corke P, Ciocarlie M, Dahiya R, Kragic D, Goldberg K, Nagai Y, et al.. A roadmap for AI in robotics. *Nature Machine Intelligence*, 2025.
- [147] Hu Y, Liu J, Yang X, Xu M, Kuang Y, Xu W, Dai Q, Freeman WT, Durand F. Quantaichi: a compiler for quantized simulations. *ACM Transactions on Graphics*, 2021, 40(4): 1–16.
- [148] Liu J, Shi H, Zhang S, Yang Y, Ma C, Xu W. Automatic quantization for physics-based simulation. *ACM Transactions on Graphics*, 2022, 41(4): 1–16.
- [149] Du T, Wu K, Ma P, Wah S, Spielberg A, Rus D, Matusik W. Diffpd: Differentiable projective dynamics. *ACM Transactions on Graphics*, 2021, 41(2): 1–21.
- [150] Li Z, Xu Q, Ye X, Ren B, Liu L. Diffrr: Differentiable sph-based fluid-rigid coupling for rigid body control. *ACM Transactions on Graphics*, 2023, 42(6): 1–17.
- [151] Stuyck T, Chen Hy. Diffxpbd: Differentiable position-based simulation of compliant constraint dynamics. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 2023, 6(3): 1–14.
- [152] Taylor AH, Le Cleac’h S, Kolter Z, Schwager M, Manchester



- Z. Dojo: A Differentiable Simulator for Robotics. *arXiv preprint arXiv:2203.00806*, 2022.
- [153] Macklin M. Warp: A High-performance Python Framework for GPU Simulation and Graphics. <https://github.com/nvidia/warp>, 2022, nVIDIA GPU Technology Conference (GTC).
- [154] DeepMind G. MuJoCo XLA. <https://mujoco.readthedocs.io/en/stable/mjx.html>, 2025.
- [155] Barcellona L, Zadaianchuk A, Allegro D, Papa S, Ghidoni S, Gavves E. Dream to Manipulate: Compositional World Models Empowering Robot Imitation Learning with Imagination. In *International Conference on Learning Representations*, 2025.
- [156] Yang X, Ji Z, Lai YK. Differentiable physics-based system identification for robotic manipulation of elastoplastic materials. *The International Journal of Robotics Research*, 2025: 02783649251334661.
- [157] Audley D, Lee D. Ill-posed and well-posed problems in systems identification. *IEEE Transactions on Automatic Control*, 1974.
- [158] Lou H, Zhang M, Geng H, Zhou H, He S, Gao Z, Zhao S, Mao J, Abbeel P, Malik J, Seita D, Wang Y. DREAM: Differentiable Real-to-Sim-to-Real Engine for Learning Robotic Manipulation. In *RSS Workshop on Dexterous Manipulation: Learning and Control with Diverse Data*, 2025.
- [159] Shi L, Xu Y, Wang S, Huang J, Zhao W, Jia Y, Yan Z, Gu W, Zhou G. An Real-Sim-Real (RSR) Loop Framework for Generalizable Robotic Policy Transfer with Differentiable Simulation. *arXiv preprint arXiv:2503.10118*, 2025.
- [160] Kuroki S, Guo J, Matsushima T, Okubo T, Kobayashi M, Ikeda Y, Takanami R, Yoo P, Matsuo Y, Iwasawa Y. Gendom: Generalizable one-shot deformable object manipulation with parameter-aware policy. In *IEEE International Conference on Robotics and Automation*, 2024.
- [161] Lutter M, Silberbauer J, Watson J, Peters J. Differentiable physics models for real-world offline model-based reinforcement learning. In *IEEE International Conference on Robotics and Automation*, 2021, 4163–4170.
- [162] Bjelonic F, Tischhauser F, Hutter M. Towards bridging the gap: Systematic sim-to-real transfer for diverse legged robots. *arXiv preprint arXiv:2509.06342*, 2025.
- [163] Degraeve J, Hermans M, Dambre J, Wyffels F. A differentiable physics engine for deep learning in robotics. *Frontiers in Neurorobotics*, 2019, 13: 6.
- [164] Ravi N, Reizenstein J, Novotny D, Gordon T, Lo WY, Johnson J, Gkioxari G. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020.
- [165] Abou-Chakra J, Rana K, Dayoub F, Suenderhauf N. Physically Embodied Gaussian Splatting: A Visually Learnt and Physically Grounded 3D Representation for Robotics. In *Conference on Robot Learning*, 2024.
- [166] Zhu S, Mou L, Li D, Ye B, Huang R, Zhao H. Vr-robo: A real-to-sim-to-real framework for visual robot navigation and locomotion. *IEEE Robotics and Automation Letters*, 2025.
- [167] NVIDIA Isaac Sim: Robotics Simulation and Synthetic Data Generation. <https://developer.nvidia.com/isaac/sim>, 2025.
- [168] Moran B, Comi M, Bohez S, Erez T, Li Z, Hasenclever L. Splatting Physical Scenes: End-to-End Real-to-Sim from Imperfect Robot Data. *arXiv preprint arXiv:2506.04120*, 2025.
- [169] Wu Y, Pan L, Wu W, Wang G, Miao Y, Xu F, Wang H. RL-gsbridge: 3d gaussian splatting based real2sim2real method for robotic manipulation learning. In *IEEE International Conference on Robotics and Automation*, 2025, 192–198.
- [170] Coumans E, Bai Y. Pybullet, a python module for physics simulation for games, robotics and machine learning. *URL http://pybullet.org*, 2016.
- [171] Le Cleac'h S, Yu HX, Guo M, Howell T, Gao R, Wu J, Manchester Z, Schwager M. Differentiable physics simulation of dynamics-augmented neural objects. *IEEE Robotics and Automation Letters*, 2023.
- [172] Howell TA, Le Cleac'h S, Kolter JZ, Schwager M, Manchester Z. Dojo: A differentiable simulator for robotics. *arXiv preprint arXiv:2203.00806*, 2022, 9(2): 4.
- [173] Jin Y, Lv J, Jiang S, Lu C. Diffgen: Robot demonstration generation via differentiable physics simulation, differentiable rendering, and vision-language model. *arXiv preprint arXiv:2405.07309*, 2024.
- [174] Li TM, Aittala M, Durand F, Lehtinen J. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics*, 2018, 37(6): 1–11.
- [175] Werling K, Omens D, Lee J, Exarchos I, Liu CK. Fast and Feature-Complete Differentiable Physics Engine for Articulated Rigid Bodies with Contact Constraints. In *Robotics: Science and Systems*, 2021.
- [176] Qureshi MN, Garg S, Yandun F, Held D, Kantor G, Silwal A. SplatSim: Zero-shot sim2real transfer of rgb manipulation policies using gaussian splatting. In *IEEE International Conference on Robotics and Automation*, 2025.
- [177] Li X, Li J, Zhang Z, Zhang R, Jia F, Wang T, Fan H, Tseng KK, Wang R. RoboGsim: A real2sim2real robotic gaussian splatting simulator. *arXiv preprint arXiv:2411.11839*, 2024.
- [178] Zhai AJ, Shen Y, Chen EY, Wang GX, Wang X, Wang S, Guan K, Wang S. Physical property understanding from language-embedded feature fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024, 28296–28305.
- [179] Wang M, Tian S, Wu J, Schwager M. Phys2Real: Physically-Informed Gaussian Splatting for Adaptive Sim-to-Real Transfer in Robotic Manipulation. In *Structured World Models for Robotic Manipulation*, 2025.
- [180] Xu X, Ge W, Qiu D, Chen Z, Yan D, Liu Z, Zhao H, Zhao H, Zhang S, Liang J, et al.. Gaussianproperty: Integrating physical properties to 3d gaussians with lmms. In *International Conference on Computer Vision*, 2025, 7231–7240.
- [181] Zhou H, Guo Y, Wang X, Xu K. Monomobility: Zero-shot 3d mobility analysis from monocular videos. In *International Conference on Computer Vision*, 2025, 8800–8809.

- [182] Jiang T, Guan Y, Ma L, Xu J, Meng J, Chen W, Zeng Z, Li L, Wu D, Chen R. DexSim2Real²: Building Explicit World Model for Precise Articulated Object Dexterous Manipulation. *IEEE Transactions on Robotics*, 2025, 41: 4360–4379.
- [183] Laine S, Hellsten J, Karras T, Seol Y, Lehtinen J, Aila T. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 2020.
- [184] Yu Q, Yuan X, Chen J, Zheng D, Hao C, You Y, Chen Y, Mu Y, Liu L, Lu C, et al.. ArtGS: 3D Gaussian Splatting for Interactive Visual-Physical Modeling and Manipulation of Articulated Objects. In *IEEE International Conference on Intelligent Robots and Systems*, 2025.
- [185] Kim S, Ha J, Kim YH, Lee Y, Park FC. Screwsplat: An end-to-end method for articulated object recognition. In *Conference on Robot Learning*, 2025.
- [186] Guo J, Xin Y, Liu G, Xu K, Liu L, Hu R. Articulatedgds: Self-supervised digital twin modeling of articulated objects using 3d gaussian splatting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025, 27144–27153.
- [187] Hu K, Yu P, Tan N. Learning high-fidelity robot self-model with articulated 3D Gaussian splatting. *The International Journal of Robotics Research*, 2024: 02783649251396980.
- [188] Xie T, Zong Z, Qiu Y, Li X, Feng Y, Yang Y, Jiang C. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.
- [189] Li X, Qiao Y, Chen PY, Jatavallabhula KM, Lin MC, Jiang C, Gan C. PAC-NeRF: Physics Augmented Continuum Neural Radiance Fields for Geometry-Agnostic System Identification. In *International Conference on Learning Representations*, 2023.
- [190] Jiang H, Hsu HY, Zhang K, Yu HN, Wang S, Li Y. Phys-Twin: Physics-Informed Reconstruction and Simulation of Deformable Objects from Videos. *ICCV*, 2025.
- [191] Huang Z, Hu Y, Du T, Zhou S, Su H, Tenenbaum JB, Gan C. PlasticineLab: A Soft-Body Manipulation Benchmark with Differentiable Physics. In *International Conference on Learning Representations*, 2021.
- [192] Liu F, Wang H, Yao S, Zhang S, Zhou J, Duan Y. Physics3d: Learning physical properties of 3d gaussians via video diffusion. *arXiv preprint arXiv:2406.04338*, 2024.
- [193] Zhang T, Yu HX, Wu R, Feng BY, Zheng C, Snavely N, Wu J, Freeman WT. Physdreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, 2024.
- [194] Li SL, Zhang A, Chen B, Matusik H, Liu C, Rus D, Sitzmann V. Controlling diverse robots by inferring Jacobian fields with deep networks. *Nature*, 2025: 1–7.
- [195] Irshad MZ, Comi M, Lin YC, Heppert N, Valada A, Ambrus R, Kira Z, Tremblay J. Neural Fields in Robotics: A Survey. *arXiv preprint arXiv:2410.20220*, 2024.
- [196] Wu T, Yuan YJ, Zhang LX, Yang J, Cao YP, Yan LQ, Gao L. Recent advances in 3d gaussian splatting. *Computational Visual Media*, 2024.
- [197] Valentin JP, Sengupta S, Warrell J, Shahrokni A, Torr PH. Mesh based semantic modelling for indoor and outdoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
- [198] Nie Y, Hou J, Han X, Nießner M. Rfd-net: Point scene understanding by semantic instance reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.
- [199] Park JJ, Florence P, Straub J, Newcombe R, Lovegrove S. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [200] Loper MM, Black MJ. OpenDR: An approximate differentiable renderer. In *European Conference on Computer Vision*, 2014, 154–169.
- [201] Kato H, Ushiku Y, Harada T. Neural 3d mesh renderer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 3907–3916.
- [202] Liu S, Li T, Chen W, Li H. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *International Conference on Computer Vision*, 2019, 7708–7717.
- [203] Boss M, Braun R, Jampani V, Barron JT, Liu C, Lensch H. Nerd: Neural reflectance decomposition from image collections. In *International Conference on Computer Vision*, 2021.
- [204] Bi S, Xu Z, Srinivasan P, Mildenhall B, Sunkavalli K, Hašan M, Hold-Geoffroy Y, Kriegman D, Ramamoorthi R. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020.
- [205] Srinivasan PP, Deng B, Zhang X, Tancik M, Mildenhall B, Barron JT. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.
- [206] Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021.
- [207] Kerbl B, Kopanas G, Leimkühler T, Drettakis G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 2023.
- [208] Kandukuri RK, Strecke M, Stueckler J. Physics-Based Rigid Body Object Tracking and Friction Filtering From RGB-D Videos. In *International Conference on 3D Vision*, 2024.
- [209] Lynch KM, Park FC. Modern robotics, 2017.
- [210] Ma P, Du T, Tenenbaum J, Matusik W, Gan C. RISP: Rendering-Invariant State Predictor with Differentiable Simulation and Rendering for Cross-Domain Parameter Estimation. In *International Conference on Learning Representations*, 2022.
- [211] Liu R, Canberk A, Song S, Vondrick C. Differentiable Robot Rendering, 2024.

- [212] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, et al.. On the Opportunities and Risks of Foundation Models, 2022.
- [213] Xiang J, Lv Z, Xu S, Deng Y, Wang R, Zhang B, Chen D, Tong X, Yang J. Structured 3d latents for scalable and versatile 3d generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025.
- [214] Yao K, Zhang L, Yan X, Zeng Y, Zhang Q, Xu L, Yang W, Gu J, Yu J. Cast: Component-aligned 3d scene reconstruction from an rgb image. *ACM Transactions on Graphics*, 2025, 44(4): 1–19.
- [215] Lai Z, Zhao Y, Liu H, Zhao Z, Lin Q, Shi H, Yang X, Yang M, Yang S, Feng Y, et al.. Hunyuan3D 2.5: Towards High-Fidelity 3D Assets Generation with Ultimate Details. *arXiv preprint arXiv:2506.16504*, 2025.
- [216] Dai T, Wong J, Jiang Y, Wang C, Gokmen C, Zhang R, Wu J, Fei-Fei L. Automated Creation of Digital Cousins for Robust Policy Learning. In *Conference on Robot Learning*, 2024.
- [217] Grieves M, Vickers J. Digital Twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. *Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches*, 2017.
- [218] Cherian A, Corcodel R, Jain S, Romeres D. Llmphy: Complex physical reasoning using large language models and world models. *arXiv preprint arXiv:2411.08027*, 2024.
- [219] Blattmann A, Dockhorn T, Kulal S, Mendelevitch D, Kilian M, Lorenz D, Levi Y, English Z, Voleti V, Letts A, et al.. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- [220] Huang T, Zeng Y, Li H, Zuo W, Lau RW. Dreamphysics: Learning physical properties of dynamic 3d gaussians with video diffusion priors. *arXiv e-prints*, 2024: arXiv–2406.
- [221] Authors G. Genesis: A Universal and Generative Physics Engine for Robotics and Beyond, 2024.
- [222] Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan KR, Cao Y. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.
- [223] SpeedBot Robotics. Kunwu Industrial Intelligence Platform. <https://kunwucloud.speedbot.net/>, 2025.
- [224] Pan J, Xing J, Reiter R, Zhai Y, Aljalbout E, Scaramuzza D. Learning on the Fly: Rapid Policy Adaptation via Differentiable Simulation. *arXiv preprint arXiv:2508.21065*, 2025.
- [225] He T, Gao J, Xiao W, Zhang Y, Wang Z, Wang J, Luo Z, He G, Sobanbabu N, Pan C, Yi Z, Qu G, Kitani K, Hodgins JK, Fan L, Zhu Y, Liu C, Shi G. ASAP: Aligning Simulation and Real-World Physics for Learning Agile Humanoid Whole-Body Skills. In *Robotics: Science and Systems*, 2025, doi: 10.15607/RSS.2025.XXI.066.
- [226] Lutter M, Ritter C, Peters J. Deep Lagrangian Networks: Using Physics as Model Prior for Deep Learning. In *International Conference on Learning Representations*, 2019.
- [227] Greydanus S, Dzamba M, Yosinski J. Hamiltonian neural networks. *Advances in Neural Information Processing Systems*, 2019, 32.
- [228] Saemundsson S, Terenin A, Hofmann K, Deisenroth M. Variational integrator networks for physically structured embeddings. In *International Conference on Artificial Intelligence and Statistics*, 2020, 3078–3087.
- [229] Havens A, Chowdhary G. Forced variational integrator networks for prediction and control of mechanical systems. In *Learning for Dynamics and Control*, 2021, 1142–1153.
- [230] Duruisseaux V, Duong TP, Leok M, Atanasov N. Lie group forced variational integrator networks for learning and control of robot systems. In *Learning for Dynamics and Control*, 2023, 731–744.
- [231] Ramesh A, Ravindran B. Physics-informed model-based reinforcement learning. In *Learning for Dynamics and Control*, 2023, 26–37.
- [232] Zhang J, Geng H, You Y, Deng C, Abbeel P, Malik J, Guibas L. Rodrigues Network for Learning Robot Actions. *arXiv preprint arXiv:2506.02618*, 2025.
- [233] Sholokhov A, Liu Y, Mansour H, Nabi S. Physics-informed neural ODE (PINODE): embedding physics into models using collocation points. *Scientific Reports*, 2023, 13(1): 10166.
- [234] Lutter M, Ritter C, Peters J. Deep Lagrangian Networks: Using Physics as Model Prior for Deep Learning. In *International Conference on Learning Representations*, 2019.
- [235] Sanyal S, Roy K. RAMP-Net: A Robust Adaptive MPC for Quadrotors via Physics-informed Neural Network. In *IEEE International Conference on Robotics and Automation*, 2023, 1019–1025, doi:10.1109/ICRA48891.2023.10161410.
- [236] Nicodemus J, Kneifl J, Fehr J, Unger B. Physics-informed neural networks-based model predictive control for multi-link manipulators. *IFAC*, 2022, 55(20): 331–336.
- [237] Hao C, Lu W, Xu Y, Chen Y. Neural Motion Simulator Pushing the Limit of World Models in Reinforcement Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025, 27608–27617.
- [238] Zhao Z, Li B, Du Y, Fu T, Wang C. PhysORD: a neuro-symbolic approach for physics-infused motion prediction in off-road driving. In *IEEE International Conference on Intelligent Robots and Systems*, 2024, 11670–11677.
- [239] Li Z, Mei W, Yu K, Bai Y, Li S. ICODE: Modeling dynamical systems with extrinsic input information. *IEEE Transactions on Automation Science and Engineering*, 2025.
- [240] Ishihara Y, Takahashi M. Empirical study of future image prediction for image-based mobile robot navigation. *Robotics and Autonomous Systems*, 2022, 150: 104018.
- [241] Chen L, Lu K, Rajeswaran A, Lee K, Grover A, Laskin M, Abbeel P, Srinivas A, Mordatch I. Decision transformer: Reinforcement learning via sequence modeling. *Advances in Neural Information Processing Systems*, 2021, 34: 15084–15097.

- [242] Shang J, Kahatapitiya K, Li X, Ryoo MS. Starformer: Transformer with state-action-reward representations for visual reinforcement learning. In *European Conference on Computer Vision*, 2022, 462–479.
- [243] Zhang L, Xiong Y, Yang Z, Casas S, Hu R, Urtasun R. Copilot4D: Learning Unsupervised World Models for Autonomous Driving via Discrete Diffusion. In *International Conference on Learning Representations*, 2024.
- [244] Yang J, Chitta K, Gao S, Chen L, Shao Y, Jia X, Li H, Geiger A, Yue X, Chen L. ReSim: Reliable World Simulation for Autonomous Driving. In *Advances in Neural Information Processing Systems*, 2025.
- [245] Zhou G, Pan H, LeCun Y, Pinto L. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.
- [246] Oquab M, Darcet T, Moutakanni T, Vo HV, Szafraniec M, Khalidov V, Fernandez P, HAZIZA D, Massa F, El-Nouby A, Assran M, Ballas N, Galuba W, Howes R, Huang PY, Li SW, Misra I, Rabbat M, Sharma V, Synnaeve G, Xu H, Jegou H, Mairal J, Labatut P, Joulin A, Bojanowski P. DI-NOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024, featured Certification.
- [247] Huang Y, Zhang J, Zou S, Liu X, Hu R, Xu K. LaDi-WM: A Latent Diffusion-based World Model for Predictive Manipulation. In *Conference on Robot Learning*, 2025.
- [248] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al.. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021, 8748–8763.
- [249] Locatello F, Weissenborn D, Unterthiner T, Mahendran A, Heigold G, Uszkoreit J, Dosovitskiy A, Kipf T. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 2020, 33: 11525–11538.
- [250] Wu Z, Dvornik N, Greff K, Kipf T, Garg A. SlotFormer: Unsupervised Visual Dynamics Simulation with Object-Centric Models. In *International Conference on Learning Representations*, 2023.
- [251] Kapl F, Mamaghan AMK, Horn M, Marr C, Bauer S, Dittadi A. Object-centric representations generalize better compositionally with less compute. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*, 2025.
- [252] Li Y, Wu J, Tedrake R, Tenenbaum JB, Torralba A. Learning Particle Dynamics for Manipulating Rigid Bodies, Deformable Objects, and Fluids. In *ICLR*, 2019.
- [253] Sanchez-Gonzalez A, Godwin J, Pfaff T, Ying R, Leskovec J, Battaglia P. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning*, 2020, 8459–8468.
- [254] Tuomainen N, Blanco-Mulero D, Kyrki V. Manipulation of granular materials by learning particle interactions. *IEEE Robotics and Automation Letters*, 2022, 7(2): 5663–5670.
- [255] Kuaishou. Kling AI. <https://klingai.kuaishou.com/>, 2024.
- [256] Seaweed T, Yang C, Lin Z, Zhao Y, Lin S, Ma Z, Guo H, Chen H, Qi L, Wang S, et al.. Seaweed-7b: Cost-effective training of video generation foundation model. *arXiv preprint arXiv:2504.08685*, 2025.
- [257] Wan T, Wang A, Ai B, Wen B, Mao C, Xie CW, Chen D, Yu F, Zhao H, Yang J, et al.. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
- [258] Kong W, Tian Q, Zhang Z, Min R, Dai Z, Zhou J, Xiong J, Li X, Wu B, Zhang J, et al.. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- [259] Teng H, Jia H, Sun L, Li L, Li M, Tang M, Han S, Zhang T, Zhang W, Luo W, et al.. MAGI-1: Autoregressive Video Generation at Scale. *arXiv preprint arXiv:2505.13211*, 2025.
- [260] Yang Z, Teng J, Zheng W, Ding M, Huang S, Xu J, Yang Y, Hong W, Zhang X, Feng G, et al.. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *International Conference on Learning Representations*, 2024.
- [261] Peng X, Zheng Z, Shen C, Young T, Guo X, Wang B, Xu H, Liu H, Jiang M, Li W, et al.. Open-sora 2.0: Training a commercial-level video generation model in \$200 k. *arXiv preprint arXiv:2503.09642*, 2025.
- [262] HaCohen Y, Chiprut N, Brazowski B, Shalem D, Moshe D, Richardson E, Levin E, Shiran G, Zabari N, Gordon O, et al.. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.
- [263] Team G. Mochi 1. <https://github.com/genmoai/models>, 2024.
- [264] Yang S, Du Y, Ghasemipour SKS, Tompson J, Schuurmans D, Abbeel P. Learning Interactive Real-World Simulators. In *NeurIPS 2023 Workshop on Generalization in Planning*.
- [265] Team A, Zhu H, Wang Y, Zhou J, Chang W, Zhou Y, Li Z, Chen J, Shen C, Pang J, et al.. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025.
- [266] Xiang J, Liu G, Gu Y, Gao Q, Ning Y, Zha Y, Feng Z, Tao T, Hao S, Shi Y, et al.. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024.
- [267] Gao S, Yang J, Chen L, Chitta K, Qiu Y, Geiger A, Zhang J, Li H. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 2024, 37: 91560–91596.
- [268] Russell L, Hu A, Bertoni L, Fedoseev G, Shotton J, Arani E, Corrado G. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.
- [269] Ren X, Shen T, Huang J, Ling H, Lu Y, Nimier-David M, Müller T, Keller A, Fidler S, Gao J. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025, 6121–6132.

- [270] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, et al.. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [271] Hao S, Gu Y, Ma H, Hong J, Wang Z, Wang D, Hu Z. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, 8154–8173.
- [272] Xie K, Yang I, Gunerli J, Riedl M. Making large language models into world models with precondition and effect knowledge. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, 7532–7545.
- [273] Gu Y, Zhang K, Ning Y, Zheng B, Gou B, Xue T, Chang C, Srivastava S, Xie Y, Qi P, Sun H, Su Y. Is Your LLM Secretly a World Model of the Internet? Model-Based Planning for Web Agents. *Transactions on Machine Learning Research*, 2025.
- [274] Yang D, Hu L, Tian Y, Li Z, Kelly C, Yang B, Yang C, Zou Y. WorldGPT: a Sora-inspired video AI agent as Rich world models from text and image inputs. *arXiv preprint arXiv:2403.07944*, 2024.
- [275] Ge Z, Huang H, Zhou M, Li J, Wang G, Tang S, Zhuang Y. Worldgpt: Empowering llm as multimodal world model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, 7346–7355.
- [276] Xiang J, Gu Y, Liu Z, Feng Z, Gao Q, Hu Y, Huang B, Liu G, Yang Y, Zhou K, et al.. PAN: A World Model for General, Interactable, and Long-Horizon World Simulation. *arXiv preprint arXiv:2511.09057*, 2025.
- [277] He H, Patrikar J, Kim DK, Smith M, McGann D, Aghamohammadi Aa, Omidshafiei S, Scherer S. GrndCtrl: Grounding World Models via Self-Supervised Reward Alignment. *arXiv preprint arXiv:2512.01952*, 2025.
- [278] World Labs. Marble: A Multimodal World Model. <https://www.worldlabs.ai/blog/marble-world-model>, 2025.
- [279] Huang Y, Chen W, Zheng W, Tao X, Wan P, Zhou J, Lu J. Terra: Explorable Native 3D World Model with Point Latents. *arXiv preprint arXiv:2510.14977*, 2025.
- [280] Team H, Wang Z, Liu Y, Wu J, Gu Z, Wang H, Zuo X, Huang T, Li W, Zhang S, et al.. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025.
- [281] Ball PJ, Bauer J, Belletti F, Brownfield B, Ephrat A, Fruchter S, Gupta A, Holsheimer K, Holynski A, Hron J, Kaplanis C, et al.. Genie 3: A New Frontier for World Models. <https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/>, 2025, google DeepMind Blog.
- [282] Huang S, Wu J, Zhou Q, Miao S, Long M. Vid2World: Crafting Video Diffusion Models to Interactive World Models. *arXiv preprint arXiv:2505.14357*, 2025.
- [283] Sutton RS. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 1991, 2(4): 160–163.
- [284] Janner M, Fu J, Zhang M, Levine S. When to trust your model: Model-based policy optimization. *Advances in Neural Information Processing Systems*, 2019, 32.
- [285] Yu T, Kumar A, Rafailov R, Rajeswaran A, Levine S, Finn C. Combo: Conservative offline model-based policy optimization. *Advances in Neural Information Processing Systems*, 2021, 34: 28954–28967.
- [286] Li C, Krause A, Hutter M. Offline Robotic World Model: Learning Robotic Policies without a Physics Simulator. *arXiv preprint arXiv:2504.16680*, 2025.
- [287] Kurutach T, Clavera I, Duan Y, Tamar A, Abbeel P. Model-Ensemble Trust-Region Policy Optimization. In *International Conference on Learning Representations*, 2018.
- [288] Zhou ZH. Ensemble methods: foundations and algorithms, 2025.
- [289] Dedieu A, Ortiz J, Lou X, Wendelken C, Guntupalli JS, Lehrach W, Lazaro-Gredilla M, Murphy KP. Improving Transformer World Models for Data-Efficient RL. In *International Conference on Machine Learning*, 2025.
- [290] Wang L, Zhang X, Wang Y, Zhan G, Wang W, Gao H, Duan J, Li SE. Off-policy Reinforcement Learning with Model-based Exploration Augmentation. In *Advances in Neural Information Processing Systems*, 2025.
- [291] Barkley B, Fridovich-Keil D. Stealing That Free Lunch: Exposing the Limits of Dyna-Style Reinforcement Learning. In *International Conference on Machine Learning*, 2025.
- [292] Kidambi R, Rajeswaran A, Netrapalli P, Joachims T. Morel: Model-based offline reinforcement learning. *Advances in Neural Information Processing Systems*, 2020, 33: 21810–21823.
- [293] Rafailov R, Yu T, Rajeswaran A, Finn C. Offline reinforcement learning from images with latent space models. In *Learning for Dynamics and Control*, 2021, 1154–1168.
- [294] Liu Z, Li S, Lee WS, YAN S, Xu Z. Efficient Offline Policy Optimization with a Learned Model. In *International Conference on Learning Representations*, 2023.
- [295] Chen XH, Yu Y, Li Q, Luo FM, Qin Z, Shang W, Ye J. Offline model-based adaptable policy learning. *Advances in Neural Information Processing Systems*, 2021, 34: 8432–8443.
- [296] Lee D, Kwon M. Temporal Distance-aware Transition Augmentation for Offline Model-based Reinforcement Learning. In *International Conference on Machine Learning*, 2025.
- [297] Qin RJ, Zhang X, Gao S, Chen XH, Li Z, Zhang W, Yu Y. NeoRL: A near real-world benchmark for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 2022, 35: 24753–24765.
- [298] As Y, Qu C, Unger B, Kang D, van der Hart M, Shi L, Coros S, Wierman A, Krause A. SPiDR: A Simple Approach for Zero-Shot Safety in Sim-to-Real Transfer. In *Advances in Neural Information Processing Systems*, 2025.
- [299] Yin S, Wu J, Huang S, Su X, He X, HAO J, Long M. Trajectory World Models for Heterogeneous Environments. In *ICLR*

- 2025 Workshop on World Models: Understanding, Modelling and Scaling.
- [300] Ye X, Gao X, Wu K, Pan Z, Komura T. SDRS: Shape-Differentiable Robot Simulator. *IEEE Transactions on Robotics*, 2025: 1–20, doi:10.1109/TRO.2025.3636344.
- [301] Todorov E, Li W. A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems. In *American Control Conference*, 2005, 300–306.
- [302] Tassa Y, Erez T, Todorov E. Synthesis and stabilization of complex behaviors through online trajectory optimization. In *IEEE International Conference on Intelligent Robots and Systems*, 2012, 4906–4913.
- [303] Wan W, Wang Y, Erickson Z, Held D. Differentiable Trajectory Optimization as a Policy Class for Reinforcement and Imitation Learning, 2024.
- [304] Zhu J, Wang Y, Wu L, Qin T, Zhou W, Liu TY, Li H. Making better decision by directly planning in continuous control. In *International Conference on Learning Representations*, 2021.
- [305] Heess N, Wayne G, Silver D, Lillicrap T, Erez T, Tassa Y. Learning continuous control policies by stochastic value gradients. *Advances in Neural Information Processing Systems*, 2015, 28.
- [306] Byravan A, Springenberg JT, Abdolmaleki A, Hafner R, Neunert M, Lampe T, Siegel N, Heess N, Riedmiller M. Imagined value gradients: Model-based policy optimization with transferable latent dynamics models. In *Conference on Robot Learning*, 2020, 566–589.
- [307] Clavera I, Fu Y, Abbeel P. Model-Augmented Actor-Critic: Backpropagating through Paths. In *International Conference on Learning Representations*, 2020.
- [308] Amos B, Stanton S, Yarats D, Wilson AG. On the model-based stochastic value gradient for continuous reinforcement learning. In *Learning for Dynamics and Control*, 2021, 6–20.
- [309] Bharadhwaj H, Xie K, Shkurti F. Model-predictive control via cross-entropy and gradient-based optimization. In *Learning for Dynamics and Control*, 2020, 277–286.
- [310] Jyothir S, Jalagam S, LeCun Y, Sobal V. Gradient-based planning with world models. *arXiv preprint arXiv:2312.17227*, 2023.
- [311] Pascanu R, Li Y, Vinyals O, Heess N, Buesing L, Racanière S, Reichert D, Weber T, Wierstra D, Battaglia P. Learning model-based planning from scratch. *arXiv preprint arXiv:1707.06170*, 2017.
- [312] Chua K, Calandra R, McAllister R, Levine S. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in Neural Information Processing Systems*, 2018, 31.
- [313] Schrittwieser J, Antonoglou I, Hubert T, Simonyan K, Sifre L, Schmitt S, Guez A, Lockhart E, Hassabis D, Graepel T, et al.. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 2020, 588(7839): 604–609.
- [314] Wang T, Ba J. Exploring Model-based Planning with Policy Networks. In *International Conference on Learning Representations*, 2020.
- [315] Feng Y, Hansen N, Xiong Z, Rajagopalan C, Wang X. Fine-tuning Offline World Models in the Real World. In *Conference on Robot Learning*, 2023.
- [316] Liu Z, Fu G, Du C, Lee WS, Lin M. Continual Reinforcement Learning by Planning with Online World Models. In *International Conference on Machine Learning*, 2025.
- [317] Williams G, Aldrich A, Theodorou EA. Model predictive path integral control: From theory to parallel computation. *Journal of Guidance, Control, and Dynamics*, 2017, 40(2): 344–357.
- [318] Sobal V, Zhang W, Cho K, Balestriero R, Rudner TG, LeCun Y. Learning from reward-free offline data: A case for planning with latent dynamics models. *arXiv preprint arXiv:2502.14819*, 2025.
- [319] Gao C, Zhang H, Xu Z, Cai Z, Shao L. Flip: Flow-centric generative planning as general-purpose manipulation world model. *arXiv preprint arXiv:2412.08261*, 2024.
- [320] Pu Y, Niu Y, Tang J, Xiong J, Hu S, Li H. One Model for All Tasks: Leveraging Efficient World Models in Multi-Task Planning. *arXiv preprint arXiv:2509.07945*, 2025.
- [321] Quevedo J, Sharma AK, Sun Y, Suryavanshi V, Liang P, Yang S. WorldGym: World Model as An Environment for Policy Evaluation, 2025.
- [322] Tseng WC, Gu J, Zhang Q, Mao H, Liu MY, Shkurti F, Yen-Chen L. Scalable Policy Evaluation with Video World Models. *arXiv preprint arXiv:2511.11520*, 2025.
- [323] Racanière S, Weber T, Reichert D, Buesing L, Guez A, Jimenez Rezende D, Puigdomènech Badia A, Vinyals O, Heess N, Li Y, et al.. Imagination-augmented agents for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 2017, 30.
- [324] Mu Y, Zhuang Y, Wang B, Zhu G, Liu W, Chen J, Luo P, Li SE, Zhang C, HAO J. Model-Based Reinforcement Learning via Imagination with Derived Memory. In A Beygelzimer, Y Dauphin, P Liang, JW Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.
- [325] Mazzaglia P, Verbelen T, Dhoedt B, Lacoste A, Rajeswar S. Choreographer: Learning and Adapting Skills in Imagination. In *International Conference on Learning Representations*, 2023.
- [326] Mattes P, Schlosser R, Herbrich R. Hieros: Hierarchical Imagination on Structured State Space Sequence World Models. In *International Conference on Machine Learning*, 2024.
- [327] Samsami MR, Zholus A, Rajendran J, Chandar S. Mastering Memory Tasks with World Models. In *International Conference on Learning Representations*, 2024.
- [328] Hu Y, Guo Y, Wang P, Chen X, Wang YJ, Zhang J, Sreenath K, Lu C, Chen J. Video Prediction Policy: A Generalist Robot Policy with Predictive Visual Representations. In *International Conference on Machine Learning*, 2025.



- [329] Xu H, Ding J, Xu J, Wang R, Chen J, Mai J, Fu Y, Ghanem B, Xu F, Elhoseiny M. Diffusion-based imaginative coordination for bimanual manipulation. In *International Conference on Computer Vision*, 2025, 11469–11479.
- [330] Nematollahi I, DeMoss B, Chandra AL, Hawes N, Burgard W, Posner I. LUMOS: Language-Conditioned Imitation Learning with World Models. In *IEEE International Conference on Robotics and Automation*, 2025.
- [331] Du Y, Yang M, Florence P, Xia F, Wahid A, Ichter B, Sermanet P, Yu T, Abbeel P, Tenenbaum JB, et al.. Video Language Planning. *arXiv preprint arXiv:2310.10625*, 2023.
- [332] Zhou S, Du Y, Chen J, Li Y, Yeung DY, Gan C. RoboDreamer: learning compositional world models for robot imagination. In *International Conference on Machine Learning*, 2024, 61885–61896.
- [333] Li J, Wang Q, Wang Y, Jin X, Li Y, Zeng W, Yang X. Open-World Reinforcement Learning over Long Short-Term Imagination. In *International Conference on Learning Representations*, 2025.
- [334] Fan L, Wang G, Jiang Y, Mandlekar A, Yang Y, Zhu H, Tang A, Huang DA, Zhu Y, Anandkumar A. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 2022, 35: 18343–18362.
- [335] Bharadhwaj H, Dwivedi D, Gupta A, Tulsiani S, Doersch C, Xiao T, Shah D, Xia F, Sadigh D, Kirmani S. Gen2Act: Human Video Generation in Novel Scenarios enables Generalizable Robot Manipulation. In *1st Workshop on X-Embodiment Robot Learning*, 2024.
- [336] Villar-Corrales A, Behnke S. PlaySlot: Learning Inverse Latent Dynamics for Controllable Object-Centric Video Prediction and Planning. In *International Conference on Machine Learning*, 2025.
- [337] Wen C, Lin X, So JIR, Chen K, Dou Q, Gao Y, Abbeel P. Any-point Trajectory Modeling for Policy Learning. In *Robotics: Science and Systems*, 2024, doi:10.15607/RSS.2024.XX.092.
- [338] Routray S, Pan H, Jain U, Bahl S, Pathak D. ViPRA: Video Prediction for Robot Actions. In *NeurIPS 2025 Workshop on Space in Vision, Language, and Embodied AI*, 2025.
- [339] Tan H, Feng Y, Mao X, Huang S, Liu G, Hao Z, Su H, Zhu J. AnyPos: Automated Task-Agnostic Actions for Bimanual Manipulation. *arXiv preprint arXiv:2507.12768*, 2025.
- [340] Xiao H, Liu X, Zhao H, Liu J, Xu K. Designing Pin-pression Gripper and Learning its Dexterous Grasping with Online In-hand Adjustment. *ACM Transactions on Graphics*, 2025, 44(4): 1–17.
- [341] She Q, Zhang S, Ye Y, Hu R, Xu K. Learning Cross-Hand Policies of High-DOF Reaching and Grasping. In *European Conference on Computer Vision*, 2024, 269–285.
- [342] Wu J, Antonova R, Kan A, Lepert M, Zeng A, Song S, Bohg J, Rusinkiewicz S, Funkhouser T. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 2023, 47(8): 1087–1102.
- [343] Liu J, Jiang F, Tashiro S, Chen S, Tanaka M. A physics-informed and data-driven framework for robotic welding in manufacturing. *Nature Communications*, 2025, 16(1): 4807.
- [344] Zamiela C, Stokes R, Tian W, Doude H, Priddy MW, Bian L. Physics-informed approximation of internal thermal history for surface deformation predictions in wire arc directed energy deposition. *Journal of Manufacturing Science and Engineering*, 2024, 146(8): 081007.
- [345] Narang YS, Storey K, Akinola I, Macklin M, Reist P, Wawrzyniak L, Guo Y, Moravánszky Á, Lu M, Handa A, et al.. Factory: Fast Contact for Robotic Assembly. In *Robotics: Science and Systems*, 2022.
- [346] Tang B, Lin MA, Akinola I, Handa A, Sukhatme GS, Ramos F, Fox D, Narang Y. IndustReal: Transferring contact-rich assembly tasks from simulation to reality. In *Robotics: Science and Systems*, 2023.
- [347] Zhao H, Yu Y, Xu K. Learning efficient online 3D bin packing on packing configuration trees. In *International Conference on Learning Representations*, 2021.
- [348] Zhao H, She Q, Zhu C, Yang Y, Xu K. Online 3D bin packing with constrained deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 35, 2021, 741–749.
- [349] Zhao H, Zhu C, Xu X, Huang H, Xu K. Learning practically feasible policies for online 3D bin packing. *Science China Information Sciences*, 2022, 65(1): 112105.
- [350] Zhao H, Pan Z, Yu Y, Xu K. Learning physically realizable skills for online packing of general 3D shapes. *ACM Transactions on Graphics*, 2023, 42(5): 1–21.
- [351] Xu J, Gong M, Zhang H, Huang H, Hu R. Neural packing: from visual sensing to reinforcement learning. *ACM Transactions on Graphics*, 2023, 42(6): 1–11.
- [352] Ai B, Tian S, Shi H, Wang Y, Tan C, Li Y, Wu J. RoboPack: Learning Tactile-Informed Dynamics Models for Dense Packing. In *Robotics: Science and Systems*, 2024.
- [353] Xu K, Wu F, Zhao J. Model-based deep reinforcement learning with heuristic search for satellite attitude control. *Industrial Robot: the international journal of robotics research and application*, 2019, 46(3): 415–420.
- [354] Aborizk A, Fitz-Coy N. Multiphase autonomous docking via model-based and hierarchical reinforcement learning. *Journal of Spacecraft and Rockets*, 2024, 61(4): 993–1005.
- [355] El-Hariry M, Orsula A, Geist M, Olivares-Mendez M. RL-AVIST: Reinforcement Learning for Autonomous Visual Inspection of Space Targets. *arXiv preprint arXiv:2510.22699*, 2025.
- [356] Reiner M. Modelica FMI based hybrid reinforcement learning enhanced trajectory planning for an ADR scenario for combined control of a satellite with a 7-axis robotic arm using Modelica/FMI. In *Modelica Conferences*, 2025, 489–496.
- [357] Watter M, Springenberg J, Boedecker J, Riedmiller M. Embed to control: A locally linear latent dynamics model for control



- from raw images. *Advances in Neural Information Processing Systems*, 2015, 28.
- [358] Hansen N, Lin Y, Su H, Wang X, Kumar V, Rajeswaran A. Mo-dem: Accelerating visual model-based reinforcement learning with demonstrations. *arXiv preprint arXiv:2212.05698*, 2022.
- [359] Yamada J, Rigter M, Collins J, Posner I. Twist: Teacher-student world model distillation for efficient sim-to-real transfer. In *IEEE International Conference on Robotics and Automation*, 2024, 9190–9196.
- [360] Cen J, Yu C, Yuan H, Jiang Y, Huang S, Guo J, Li X, Song Y, Luo H, Wang F, et al.. WorldVLA: Towards Autoregressive Action World Model. *arXiv preprint arXiv:2506.21539*, 2025.
- [361] Zhang K, Ren P, Lin B, Lin J, Ma S, Xu H, Liang X. Pivot-r: Primitive-driven waypoint-aware world model for robotic manipulation. *Advances in Neural Information Processing Systems*, 2024, 37: 54105–54136.
- [362] Lyu J, Li Z, Shi X, Xu C, Wang Y, Wang H. DyWA: Dynamics-adaptive World Action Model for Generalizable Non-prehensile Manipulation. In *ICRA 2025 Workshop: Beyond Pick and Place*.
- [363] Li Y, Zhang Y, Xiao W, Pan C, Weng H, He G, He T, Shi G. Hold My Beer: Learning Gentle Humanoid Locomotion and End-Effector Stabilization Control. In *RSS 2025 Workshop on Whole-body Control and Bimanual Manipulation: Applications in Humanoids and Beyond*.
- [364] Sun W, Chen L, Su Y, Cao B, Liu Y, Xie Z. Learning humanoid locomotion with world model reconstruction. *arXiv preprint arXiv:2502.16230*, 2025.
- [365] Hansen N, Jyothir S, Sobal V, LeCun Y, Wang X, Su H. Hierarchical World Models as Visual Whole-Body Humanoid Controllers. In *CoRL 2024 Workshop on Whole-body Control and Bimanual Manipulation: Applications in Humanoids and Beyond*.
- [366] Lai H, Cao J, Xu J, Wu H, Lin Y, Kong T, Yu Y, Zhang W. World model-based perception for visual legged locomotion. In *IEEE International Conference on Robotics and Automation*, 2025.
- [367] Gu X, Wang YJ, Zhu X, Shi C, Guo Y, Liu Y, Chen J. Advancing Humanoid Locomotion: Mastering Challenging Terrains with Denoising World Model Learning. In *Robotics: Science and Systems*, 2024.
- [368] Zheng H, Cheng Y, Liu H, Ye L, Liu H. HuWo: Building Physical Interaction World Models for Humanoid Robot Locomotion. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.
- [369] Lei Y, Luo Z, He T, Cao J, Shi G, Kitani K. Scalable Humanoid Whole-Body Control via Differentiable Neural Network Dynamics. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.
- [370] Sun R, Zang H, Li X, Islam R. Learning latent dynamic robust representations for world models. *arXiv preprint arXiv:2405.06263*, 2024.
- [371] Liu H, Gao Y, Teng S, Chi Y, Shao YS, Li Z, Ghaffari M, Sreenath K. Ego-Vision World Model for Humanoid Contact Planning. *arXiv preprint arXiv:2510.11682*, 2025.
- [372] Wang S, Fei Z, Cheng Q, Zhang S, Cai P, Fu J, Qiu X. World Modeling Makes a Better Planner: Dual Preference Optimization for Embodied Task Planning. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.
- [373] Werby A, Huang C, Büchner M, Valada A, Burgard W. Hierarchical Open-Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation. In *Robotics: Science and Systems*, 2024.
- [374] Hu A, Corrado G, Griffiths N, Murez Z, Gurau C, Yeo H, Kendall A, Cipolla R, Shotton J. Model-based imitation learning for urban driving. *Advances in Neural Information Processing Systems*, 2022, 35: 20703–20716.
- [375] Xia F, Zamir AR, He Z, Sax A, Malik J, Savarese S. Gibson env: Real-world perception for embodied agents. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 9068–9079.
- [376] Deitke M, Han W, Herrasti A, Kembhavi A, Kolve E, Mottaghi R, Salvador J, Schwenk D, VanderBilt E, Wallingford M, et al.. Robothor: An open simulation-to-real embodied ai platform. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, 3164–3174.
- [377] Savva M, Kadian A, Maksymets O, Zhao Y, Wijmans E, Jain B, Straub J, Liu J, Koltun V, Malik J, et al.. Habitat: A platform for embodied ai research. In *International Conference on Computer Vision*, 2019, 9339–9347.
- [378] Dosovitskiy A, Ros G, Codevilla F, Lopez A, Koltun V. CARLA: An open urban driving simulator, 2017: 1–16.
- [379] Li Q, Peng Z, Feng L, Zhang Q, Xue Z, Zhou B. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 45(3): 3461–3475.
- [380] Li Q, Peng ZM, Feng L, Liu Z, Duan C, Mo W, Zhou B. Scenarionet: Open-source platform for large-scale traffic scenario simulation and modeling. *Advances in Neural Information Processing Systems*, 2023, 36: 3894–3920.
- [381] Zhou H, Lin L, Wang J, Lu Y, Bai D, Liu B, Wang Y, Geiger A, Liao Y. Hugsim: A real-time, photo-realistic and closed-loop simulator for autonomous driving. *arXiv preprint arXiv:2412.01718*, 2024.
- [382] Sun J, Xie Y, Chen L, Zhou X, Bao H. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, 15598–15607.
- [383] Guo H, Peng S, Lin H, Wang Q, Zhang G, Bao H, Zhou X. Neural 3d scene reconstruction with the manhattan-world assumption. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, 5511–5520.
- [384] Wu R, Mildenhall B, Henzler P, Park K, Gao R, Watson D, Srinivasan PP, Verbin D, Barron JT, Poole B, et al.. Reconfusion: 3d reconstruction with diffusion priors. In *IEEE*



- Conference on Computer Vision and Pattern Recognition*, 2024, 21551–21561.
- [385] Dong S, Xu K, Zhou Q, Tagliasacchi A, Xin S, Nießner M, Chen B. Multi-robot collaborative dense scene reconstruction. *ACM Transactions on Graphics*, 2019, 38(4): 1–16.
- [386] Tang Y, Zhang J, Yu Z, Wang H, Xu K. Mips-fusion: Multi-implicit-submaps for scalable and robust online neural rgb-d reconstruction. *ACM Transactions on Graphics*, 2023, 42(6): 1–16.
- [387] Xi Y, Zhu C, Duan Y, Yi R, Zheng L, He H, Xu K. THP: Tensor-field-driven hierarchical path planning for autonomous scene exploration with depth sensors. *Computational Visual Media*, 2024, 10(6): 1121–1135.
- [388] Zheng L, Zhu C, Zhang J, Zhao H, Huang H, Niessner M, Xu K. Active scene understanding via online semantic reconstruction. In *Computer Graphics Forum*, volume 38, 2019, 103–114.
- [389] Zhao G, Wu S, Ma J, Chen X, et al.. DriveDreamer: Towards Real-World-Driven World Models for Driving Video Generation. In *European Conference on Computer Vision*, 2024.
- [390] Hu A, Gafton P, Kosiorek A, Posner I, Kendall A, et al.. GAIA-1: A Generative World Model for Autonomous Driving. *arXiv preprint arXiv:2309.17080*, 2023.
- [391] Chen Y, Wang Y, Zhang Z. Drivingsgpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers. In *International Conference on Computer Vision*, 2025, 26890–26900.
- [392] Bar A, Zhou G, Tran D, Darrell T, LeCun Y. Navigation World Models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025, 15791–15801.
- [393] Mahjourian R, Kim J, Chai Y, Tan M, Sapp B, Angelov D. Occupancy flow fields for motion forecasting in autonomous driving. *IEEE Robotics and Automation Letters*, 2022, 7(2): 5639–5646.
- [394] Zyrianov V, Zhu X, Wang S. Learning to generate realistic lidar point clouds, 2022: 17–35.
- [395] Guo J, Ye Y, He T, Wu H, Jiang Y, Pearce T, Bian J. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025.
- [396] Valevski D, Leviathan Y, Arar M, Fruchter S. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.
- [397] Zhang Y, Peng C, Wang B, Wang P, Zhu Q, Kang F, Jiang B, Gao Z, Li E, Liu Y, et al.. Matrix-Game: Interactive World Foundation Model. *arXiv preprint arXiv:2506.18701*, 2025.
- [398] He X, Peng C, Liu Z, Wang B, Zhang Y, Cui Q, Kang F, Jiang B, An M, Ren Y, et al.. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.
- [399] Peebles W, Xie S. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, 2023, 4195–4205.
- [400] Magne L, Awadalla A, Wang G, Xu Y, Belofsky J, Hu F, Kim J, Schmidt L, Gkioxari G, Kautz J, Yue Y, Choi Y, Zhu Y, Fan L. NitroGen: A Foundation Model for Generalist Gaming Agents. <https://nitrogen.minedojo.org/>.
- [401] Yu J, Qin Y, Wang X, Wan P, Zhang D, Liu X. GameFactory: Creating New Games with Generative Interactive Videos. In *International Conference on Computer Vision*, 2025, 11590–11599.
- [402] Asadi K, Misra D, Kim S, Littman ML. Combating the compounding-error problem with a multi-step model. *arXiv preprint arXiv:1905.13320*, 2019.
- [403] Lai H, Shen J, Zhang W, Yu Y. Bidirectional model-based policy optimization. In *International Conference on Machine Learning*, 2020, 5618–5627.
- [404] Xu T, Li Z, Yu Y. Error bounds of imitating policies and environments for reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 44(10): 6968–6980.
- [405] Plaata A, Kusters W, Preuss M. High-accuracy model-based reinforcement learning, a survey. *Artificial Intelligence Review*, 2023, 56(9): 9541–9573.
- [406] Buckman J, Hafner D, Tucker G, Brevdo E, Lee H. Sample-efficient reinforcement learning with stochastic ensemble value expansion. *Advances in Neural Information Processing Systems*, 2018, 31.
- [407] Pan F, He J, Tu D, He Q. Trust the model when it is confident: Masked model-based actor-critic. *Advances in Neural Information Processing Systems*, 2020, 33: 10537–10546.
- [408] Kim K, Sano M, De Freitas J, Haber N, Yamins D. Active world model learning with progress curiosity. In *International Conference on Machine Learning*, 2020, 5306–5315.
- [409] Sankarar C, Blaes S, Martius G. Curious exploration via structured world models yields zero-shot object manipulation. *Advances in Neural Information Processing Systems*, 2022, 35: 24170–24183.
- [410] Mavor-Parker A, Young K, Barry C, Griffin L. How to stay curious while avoiding noisy tvs using aleatoric uncertainty estimation. In *International Conference on Machine Learning*, 2022, 15220–15240.
- [411] Hou Z, An Z, Du W. Beyond Noisy-TVs: Noise-Robust Exploration Via Learning Progress Monitoring. *arXiv preprint arXiv:2509.25438*, 2025.
- [412] Mazzaglia P, Catal O, Verbelen T, Dhoedt B. Curiosity-driven exploration via latent bayesian surprise. In *AAAI Conference on Artificial Intelligence*, volume 36, 2022, 7752–7760.
- [413] Iten K, Treven L, Sukhija B, Dörfler F, Krause A. Sample-efficient and Scalable Exploration in Continuous-Time RL. *arXiv preprint arXiv:2510.24482*, 2025.
- [414] Schmitt S, Shawe-Taylor J, van Hasselt H. Exploration via epistemic value estimation. In *AAAI Conference on Artificial Intelligence*, volume 37, 2023, 9742–9751.
- [415] Fan Y, Ming Y. Model-based reinforcement learning for

- continuous control with posterior sampling. In *International Conference on Machine Learning*, 2021, 3078–3087.
- [416] Cai S, Wang Z, Wang S, Perdikaris P, Karniadakis GE. Physics-informed neural networks for heat transfer problems. *Journal of Heat Transfer*, 2021, 143(6): 060801.
- [417] Zhao C, Zhang F, Lou W, Wang X, Yang J. A comprehensive review of advances in physics-informed neural networks and their applications in complex fluid dynamics. *Physics of Fluids*, 2024, 36(10).
- [418] Vahab M, Haghighat E, Khaleghi M, Khalili N. A physics-informed neural network approach to solution and identification of biharmonic equations of elasticity. *Journal of Engineering Mechanics*, 2022, 148(2): 04021154.
- [419] Feng H, Hu P, Wang Y, Fan D, Wu T, Zhang Y. Physics-informed super-resolution and forecasting method based on inaccurate partial differential equations and partial observation. *Physics of Fluids*, 2025, 37(6).
- [420] Dalal K, Kocejka D, Xu J, Zhao Y, Han S, Cheung KC, Kautz J, Choi Y, Sun Y, Wang X. One-minute video generation with test-time training. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025, 17702–17711.
- [421] Zhang L, Zhang Q, Jiang H, Bai Y, Yang W, Xu L, Yu J. BANG: Dividing 3D Assets via Generative Exploded Dynamics. *ACM Transactions on Graphics*, 2025, 44(4): 1–21.
- [422] Niu C, Li J, Xu K. Im2struct: Recovering 3d shape structure from a single rgb image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 4521–4529.
- [423] Li J, Xu K, Chaudhuri S, Yumer E, Zhang H, Guibas L. Grass: Generative recursive autoencoders for shape structures. *ACM Transactions on Graphics*, 2017, 36(4): 1–14.
- [424] Cao Z, Chen Z, Pan L, Liu Z. PhysX-3D: Physical-Grounded 3D Asset Generation. In *Advances in Neural Information Processing Systems*, 2025.
- [425] Dai Q, Ni X, Shen Q, Chen W, Chen B, Chu M. RainyGS: Efficient Rain Synthesis with Physically-Based Gaussian Splatting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025, 16153–16162.
- [426] Anonymous. FieryGS: In-the-Wild Fire Synthesis with Physics-Integrated Gaussian Splatting. In *International Conference on Learning Representations*, 2025, under review.
- [427] Xu L, Mohaddes D, Wang Y. LLM Agent for Fire Dynamics Simulations. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*, 2024.
- [428] Gosmar D, Dahl DA. Hallucination mitigation using agentic ai natural language-based frameworks. *arXiv preprint arXiv:2501.13946*, 2025.
- [429] Ravi N, Gabeur V, Hu YT, Hu R, Ryali C, Ma T, Khedr H, Rädle R, Rolland C, Gustafson L, et al.. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- [430] Yang J, Gao M, Li Z, Gao S, Wang F, Zheng F. Track Anything: Segment Anything Meets Videos, 2023.
- [431] Wen B, Yang W, Kautz J, Birchfield S. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.
- [432] Yang Y, Wu X, He T, Zhao H, Liu X. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023.
- [433] Xing E, Deng M, Hou J, Hu Z. Critiques of world models. *arXiv preprint arXiv:2507.05169*, 2025.
- [434] Sahoo P, Meharia P, Ghosh A, Saha S, Jain V, Chadha A. A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models. In *EMNLP (Findings)*, 2024, 11709–11724.
- [435] Liu Y, Zhang K, Li Y, Yan Z, Gao C, Chen R, Yuan Z, Huang Y, Sun H, Gao J, et al.. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.
- [436] Lin H, Chen S, Liew J, Chen DY, Li Z, Shi G, Feng J, Kang B. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.
- [437] Wang J, Chen M, Karaev N, Vedaldi A, Rupprecht C, Novotny D. Vggt: Visual geometry grounded transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025.
- [438] Wang C, Chen C, Huang Y, Dou Z, Liu Y, Gu J, Liu L. PhysCtrl: Generative Physics for Controllable and Physics-Grounded Video Generation. In *Advances in Neural Information Processing Systems*, 2025.

Author Biography



Kai Xu is a professor in the School of Computing, NUDT, where he received his Ph.D. degree in 2011. He serves on the editorial boards of ACM Transactions on Graphics, Computer Graphics Forum, Computers & Graphics, etc.

