Preprints.org

Article

Not peer-reviewed version

Learning the Physical World from Videos: A Prospective Study on World Models

Jiawei Li *, Jiarui Yang *, Peidong Liu, Shu-Tao Xia, Liang Lin *

Posted Date: 8 April 2026

doi: 10.20944/preprints202604.0503.v1

Keywords: physical consistency; world models; video generator; embodied intelligence

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Learning the Physical World from Videos: A Prospective Study on World Models

Jiawei Li [1,*], Jiarui Yang [1,2,*], Peidong Liu [1], Shu-Tao Xia [1,2] and Liang Lin [1,*]

[1] JD Explore Academy, Beijing, China
[2] Tsinghua Shenzhen International Graduate School, Shenzhen, China
* Correspondence: li-jw15@tsinghua.org.cn (J.L.); yangjiarui55@163.com (J.Y.); linliang@jd.com (L.L.)

Abstract

World models aim to enable agents to perceive states, predict future outcomes, and reason for decision-making by simulating real-world environments, and are widely regarded as a crucial pathway toward artificial general intelligence (AGI). **Video, as one of the most accessible and intuitively representative media of dynamic environments, naturally contains rich implicit representations of the physical world**. Consequently, learning world models from videos has become a prominent research direction. However, a significant gap remains between **video data and the real physical world**: videos capture only superficial visual phenomena and lack explicit representations of three-dimensional structure, physical properties, and causal mechanisms. This limitation severely constrains the physical consistency and practical applicability of world models. Motivated by this, the present work provides a prospective study of recent research in this domain, encompassing: (1) key challenges arising from the video–physical world gap and representative solutions; (2) three major construction paradigms of physical world models; (3) a thorough summary of existing evaluation benchmarks; and (4) future research directions and discussions. It is noteworthy that this study is the first to systematically examine video-driven world model research from the perspective of physical world. In contrast to prior study that primarily focus on generative modeling or provide broad overviews, this work emphasizes world models with tangible physical grounding, explicitly excluding generative tasks such as video synthesis or 3D/4D modeling that diverge conceptually from the goal of modeling the physical world. Adopting a problem-oriented perspective, this study aims to provide subsequent researchers with a systematic framework and decision-making guidance for understanding existing work, designing innovative methods, and facilitating the deployment of world models in real-world applications.

Keywords: physical consistency; world models; video generator; embodied intelligence

1. Introduction

World models refer to computational frameworks in which agents understand and predict external dynamics through internal environment simulations, thereby supporting perception, reasoning, and decision-making [1]. As a key pathway toward artificial general intelligence (AGI), world models aim to capture the physical laws, social interaction mechanisms, and environmental uncertainties inherent in the real world, enabling agents to perform efficient planning and adaptive behaviors in complex and dynamic scenarios. In practical applications, world models demonstrate broad potential: in robotics [2], they facilitate agents' understanding and simulation of interactions with the environment; in autonomous driving [3–5], they enable the prediction of future traffic scenarios and potential risks; and in game AI [6,7], they support long-horizon strategy reasoning and multi-step planning.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

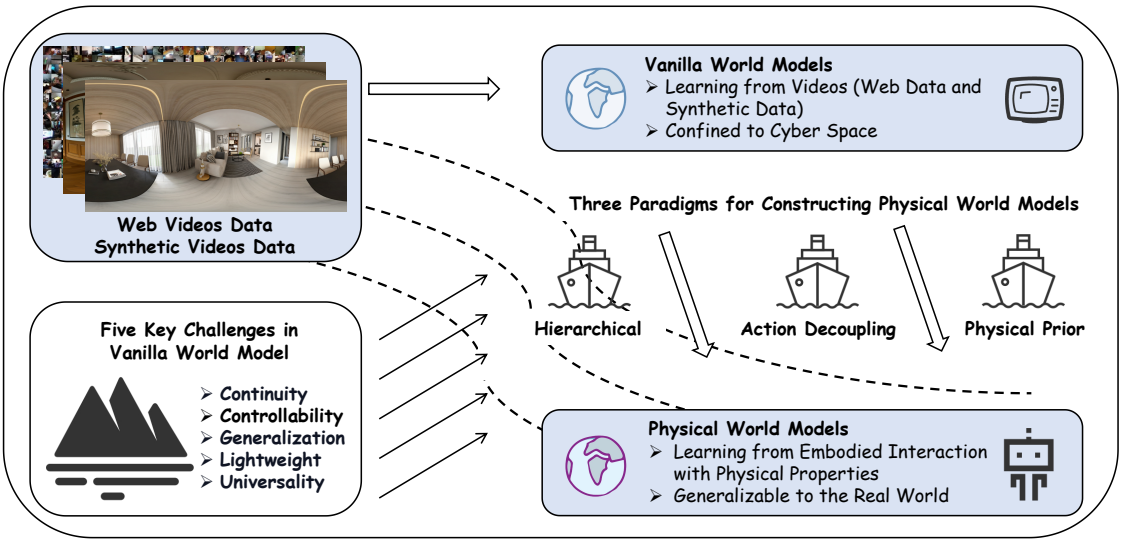

Figure 1. Conventional world models primarily rely on learning from videos (e.g., web-scale or synthetic data), which constrains them to operate within cyber space. As a result, they face five fundamental challenges, including limitations in continuity, controllability, generalization, efficiency, and universality. To address these issues, we identify three key paradigms—hierarchical modeling, action decoupling, and physical prior integration—that provide complementary solutions. By incorporating these paradigms, world models can be extended beyond purely observational learning toward physically grounded modeling, enabling learning from embodied interaction with explicit physical properties. This transition ultimately leads to physical world models that generalize more effectively and perform robustly in real-world environments.

Over the past decade, with the rapid advancement of computational power, data scale, and model performance, research on world models has achieved significant progress. In 2018, Ha and Schmidhuber [1] proposed a world model framework that integrates a Variational Autoencoder (VAE) with a Recurrent Neural Network (RNN) [8], marking a paradigm shift from traditional symbolic representations to data-driven approaches. Subsequently, the Dreamer series of methods [9–11] and Joint Embedding Predictive Architectures (JEPA) [12] introduced latent state modeling and predictive learning, demonstrating the feasibility and advantages of latent dynamics for planning and decision-making tasks. Benefiting from the rise of large-scale models, world models have entered a phase of rapid iteration since 2024, accompanied by a surge in related research publications. Representative systems such as Sora2 [13] and Genie3 [14] have significantly advanced the practical realization of learning world models directly from video data, further highlighting the potential of this research direction.

Video, as the most intuitive and readily accessible medium for capturing the dynamics of the physical world, has increasingly emerged as a natural substrate for constructing world models. It inherently encodes temporal continuity, motion patterns, and rich contextual information, thereby providing an indispensable foundation for learning structured representations of the environment. However, video is fundamentally a 2D projection of reality—it captures only superficial visual phenomena and lacks explicit representations of three-dimensional structure, physical properties (e.g., mass, friction, elasticity), contact interactions, and underlying causal mechanisms. This intrinsic limitation causes video-driven world models to remain largely confined to the level of visual pattern fitting, giving rise to the so-called "pixel-to-physics" gap. Nevertheless, compared to real-world data acquisition, video data is safer, more cost-effective, and can support virtually unlimited experimentation within simulated environments. Bridging this gap is therefore a critical research direction toward developing physically grounded and practically deployable world models.

Existing surveys have provided valuable insights into the field of world models, yet they are often limited to specific application domains, generative tasks, or high-level framework overviews, and generally lack a systematic analysis from the perspective of physical consistency [15]. For instance, the surveys by Tu et al. [3], Feng et al. [4], and Guan et al. [5] on autonomous driving world models emphasize applications in perception, prediction, and planning, but treat video merely as an auxiliary

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

input and do not specifically address biases arising from missing causal information. Ding et al. [16] and Zhu et al. [17] explore, at a broader level, the potential of extending generative models toward generalized world models, yet their focus lies primarily on delineating the boundaries of generation and simulation, overlooking in-depth analyses of physical modeling mechanisms. Some studies concentrate on physics simulation in embodied intelligence [18,19], but these are constrained to specific simulated environments or emphasize functional outcomes rather than the underlying gaps. With the rapid evolution of world model research, there is an urgent need for a survey that systematically reviews video-driven world model paradigms and key challenges, with physical consistency as the guiding principle.

Motivated by the above, this work presents a systematic survey of video-driven physical world models from a problem-oriented perspective, explicitly excluding directions that deviate from the goal of modeling the physical world (e.g., pure video generation or 3D/4D reconstruction [20]). Specifically, our discussion is organized along four key dimensions:

- **Pixel–Physics Challenges**: We distill five core challenges—continuity, controllability, generalization, lightweight, and universality—and systematically summarize the sub-problems and representative solutions for each challenge through the lens of physical consistency.
- **Three Paradigms of Physical World Model:** Existing approaches toward physically grounded modeling can be broadly categorized into three classes: prior injection, dynamic–static decoupling, and hierarchical abstraction.
- **Benchmarks**: Existing evaluation benchmarks are systematically reviewed, with an emphasis on assessment frameworks related to physical perception and dynamics prediction.
- **Future Directions**: We identify five major open problems for future research and provide a systematic discussion on industrial deployment and current safety issues.

Through this systematic organization, the survey aims to illuminate the key challenges and opportunities in bridging the gap from pixels to physics, providing a clear research framework and guidance for developing the next generation of world models with true physical understanding.

2. Challenges of Learning from Video

Videos naturally encode spatiotemporal continuity, object interactions, event evolution, and latent causal chains, making them highly valuable for capturing physical and dynamic processes. Large-scale video generation models even exhibit an implicit adherence to real-world constraints to some extent. However, as 2D projections, videos discard critical 3D geometry, object properties, and causal mechanisms, causing world models trained on them to overfit superficial visual patterns rather than underlying physical principles. This, in turn, gives rise to a series of core challenges in consistency, controllability, generalization, physical grounding, efficiency, and universality, as illustrated in Figure 2 and Table 1.

2.1. Physical Continuity

2.1.1. Temporal Continuity

The gap in temporal continuity arises because videos capture only discrete snapshots and lack explicit causal signals, causing autoregressive world models to accumulate prediction errors and break continuous state evolution. Short temporal horizons and stepwise error compounding further prevent generating long-range physically consistent sequences. Existing solutions can be summarized as follows:

Autoregressive improvements. Conditioning on preceding frames (single or multiple) provides a simple yet efficient means of ensuring continuity and has been widely adopted in recent works [21–24]. Several improvements have been proposed: EnerVerse [25] introduces a sparse keyframe memory mechanism that predicts future video blocks based on block-level context, effectively avoiding the redundant storage and error propagation associated with continuous memory while preserving long-term dependencies. EVA [26], on the other hand, incorporates a reflection-based conditioning

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

mechanism into a block-level autoregressive framework, enabling the model to adaptively extend video length. Yume [27] combines block-level autoregression with FramePack-based historical context compression and hierarchical sampling, achieving theoretically unlimited interactive generation and effectively mitigating temporal discontinuities in long sequences.

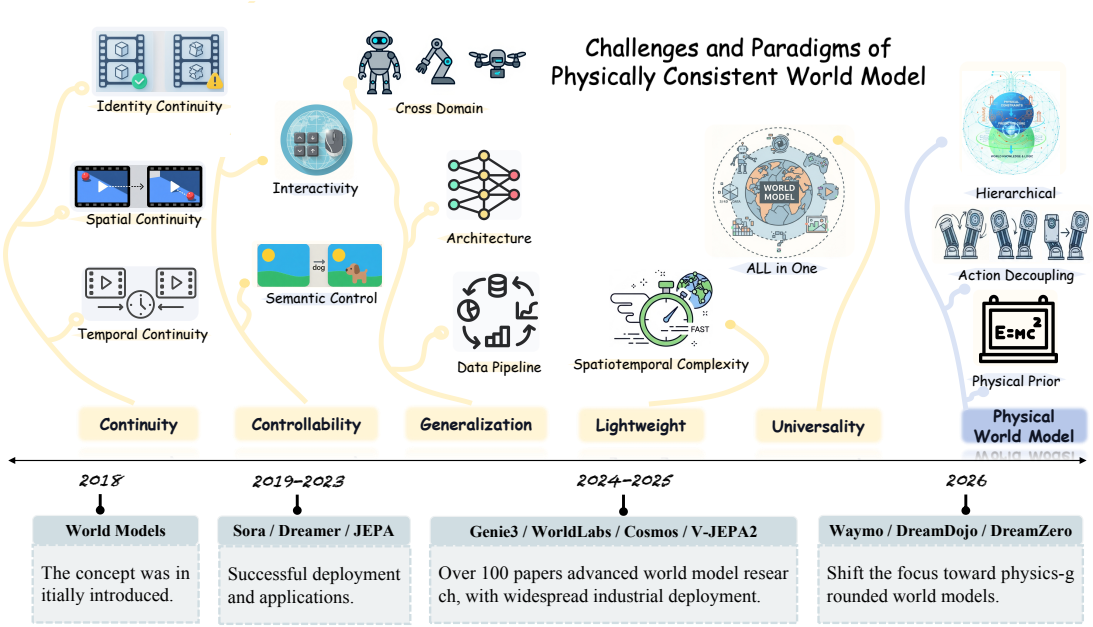

Figure 2. Since the concept of World Models was first introduced in 2018, advancements in data availability, computational power, and model scalability have significantly accelerated progress in this field. By 2025, world model research has reached a historical milestone, with nearly a hundred academic papers published within a single year. Nevertheless, learning world models from videos continues to face five key challenges that remain unresolved.

Diffusion schedule optimization. Most diffusion-based approaches typically assume that all frames share the same noise level [28,29]. Recent studies, however, demonstrate that optimizing noise schedules—such as diffusion forcing [30,31]—facilitates autoregressive long-horizon generation [32]. Similarly, Yan [33] progressively injects noise within temporal windows and applies a sliding-window sampling strategy to ensure local continuity while preserving long-term coherence. GEM [34] maintains temporal continuity in long video generation by denoising in multiple stages, conditioning on previous frames, and enforcing progressively increasing noise levels over time during training.

Conditional constraints. Imposing physical structural constraints on world models substantially enhances temporal physical continuity. Pathdreamer [35] employs a hierarchical two-stage architecture, first generating semantic and depth maps as structural contexts before predicting future frames conditioned on them. This layered strategy improves long-term fidelity since predicting structural representations grounded in physical geometry is inherently more continuous than directly generating RGB pixels. PlayerOne [36] adopts a joint reconstruction framework that simultaneously models 4D scenes and video frames, thereby ensuring physically consistent scene continuity. VRAG [22] incorporates global state information (e.g., character coordinates and poses) and retrieves relevant frames from a historical buffer, allowing the model to leverage past contexts and spatial awareness for improved temporal continuity.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026
doi:10.20944/preprints202604.0503.v1

Table 1. A detailed summary of the physical challenges encountered when learning world models from videos.

Challenges	Importance	Classes	The Pixel–Physics Gap	Solutions
Continuity	⭐ *Physical continuity is fundamental to stable world models. Maintaining continuity across time, space, and object identity is crucial for reliable long-term prediction and decision-making.*	Temporal (Sec. 2.1.1)	⚠ *Videos sample continuous processes as discrete frames and lack explicit causality, causing autoregressive models to accumulate temporal errors and break physical continuity.*	Autoregression improvements Optimization of schedules Conditional constraints Optimization-level
		Spatial (Sec. 2.1.2)	⚠ *Videos are 2D projections that lose key 3D geometry (e.g., depth and structure)*	Implicit alignment Explicit alignment Memory mechanisms
		Identity (Sec. 2.1.3)	⚠ *Videos lack explicit object properties (e.g., mass, material, shape).*	Identity perception
Controllability	⭐ *Videos passively record events without action–outcome causality, limiting world models in controllability, causal reasoning, and goal-directed behavior.*	Semantic (Sec. 2.2.1)	⚠ *Videos lack semantic–physical alignment, hindering grounded instruction.*	Semantic alignment mechanism
		Interactivity (Sec. 2.2.2)	⚠ *Videos record past events and cannot model counterfactuals or real-time responses to interventions.*	Low-level control signals Global control
Generalization	⭐ *Videos capture appearance rather than underlying physics, leading to overfitting. Generalization requires learning physics-grounded representations transferable across scenes and tasks.*	Data Augmentation (Sec. 2.3.1)	⚠ *Physically diverse, well-annotated video data are scarce, limiting coverage of rare dynamics.*	Automated data pipeline Simulation-to-real engine Foundation models
		Architecture (Sec. 2.3.2)	⚠ *Naive architectures capture visual correlations rather than physical invariances.*	Advanced modeling
		Behavioral & Environmental (Sec. 2.3.3)	⚠ *Videos from different embodiments and environments exhibit large distribution shifts in physical dynamics.*	Regularization & disentangling Context adaptation
Lightweight	⭐ *Lightweight models save resources, enable real-time interaction, and learn compact, generalizable dynamics, balancing efficiency and performance for constrained settings.*	Representation & Efficiency (Sec. 2.4)	⚠ *Videos are high-dimensional and redundant, with much irrelevant pixel data, hindering efficient extraction of physically meaningful representations.*	Latent modeling Sequence optimization Parameter-efficient / transfer Training and sample efficiency
Universality	⭐ *Videos capture only a limited slice of reality and lack unified cross-modal representations. Universal world models need shared physical abstractions that generalize broadly.*	Modeling Multi-task Architecture (Sec. 2.5)	⚠ *Videos' passive, single-view, and modality-limited nature prevents learning unified physical representations.*	Shared representations Multi-task learning Knowledge generalization and transfer

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

Optimization-level. Regularization strategies and loss-based enforcement provide direct optimization routes to penalize physical discontinuities in learned representations. SSD [37] employs a state-space modeling framework to efficiently handle long-horizon sequential information, enabling the extraction of rich long-term contextual dependencies. Vid2World [38] modifies both architecture and training objectives—particularly through adjustments to temporal attention layers and temporal convolution weight sharing—to enable causal generation and autoregressive capability in video diffusion models, ensuring predictions depend only on past physical states. SGF [39] enforces temporal continuity by minimizing the mean squared error between consecutive observations while introducing variance–covariance regularization to prevent representation collapse.

Table 2. Summary of Works on Physical Continuity Challenges.

Modalities: V indicates video or multi-frame; I denotes single-frame or image; L refers to latent representation; S to spatial representation; T to text modality; A to action representation; C to camera (pose) information; D to depth feature; N to neural signal; O to object-level feature; P to physical information; and M to historical memory feature. ***Evaluations:*** **V** assesses visual quality, generation, prediction, and control in downstream tasks, encompassing qualitative evaluations. **R** evaluates robotic tasks in simulation, while **R̂** includes real-robot experiments. **P** measures physical understanding and perception. **C** evaluates planning and decision-making in games, tasks, and navigation.

Work	Venue	Main Solution	Method	Input / Output	Conditions	Evals
Temporal Consistency						
EnerVerse [25]	NeurIPS'25	Autoregression Improvements	Sparse Chunks	V / V	T S A C D	VR̂
EVA [26]	arXiv'25		Reflection	V / V	T A	VR̂
Yume [27]	arXiv'25		Framepack	V / V	T A O N	V
SAMPO [23]	NeurIPS'25	Diffusion Schedules	Scale-Wise	V A T / V	None	VR̂
Yan [33]	arXiv'25		Progressive Noise	V / V	T A	V
GEM [34]	CVPR'25		Increasing Noise	V I O / V O	O P	V
Diamond [32]	NeurIPS'24		Adaptive Noise	V A M / V	None	V
Epona [31]	ICCV'25	Condition Constraints	Diffusion Forcing	V / V	A	VC
Pathdreamer [35]	ICCV'21		HR. Modeling	I / I	S D	VC
PlayerOne [36]	NeurIPS'25		Rec. Constraints	I S A C / V S	T	V
VRAG [22]	NeurIPS'25	Optimization Level	Global Constraints	V A / V	None	CVR̂
Vid2World [38]	ICLR'26		Loss-based	V A / V A	T	V
SSD [37]	NeurIPS'25		State-space	V A / V	None	C
SGF [39]	ICLR'25		Regularization	V A / L	None	C
Emu3.5 [40]	arXiv'25		Pre-training	V T / V	None	V
Spatial Consistency						
RoboScape [41]	NeurIPS'25	Implicit Representation Alignment	HR. Modeling	V A D / I D P	None	VR
ManipDreamer [42]	arXiv'25		Action Tree	I / V	T D O	VR
WorldGrow [43]	AAAI'26		Block Inpainting	S / S	T I O	V
DeepVerse [21]	arXiv'25		Structural Alignment	I D C / I D C	T M A	V
GAIA-2 [44]	arXiv'25	Explicit Representation Alignment	Semantic Alignment	V / V	P S T	VC
WVD [45]	CVPR'25		Spatial Joint Modeling	V S / V S	None	V
FlashWorld [46]	ICLR'26		Dual-mode Pre-training	I / S	T C N	V
Geom. Forcing [47]	ICLR'26		Rep. Alignment	V / V	None	V
InfiniCube [48]	ICCV'25		HR. Constraints	A / S	S I O	V
MindJourney [49]	NeurIPS'25		Language Guidance	I T / I T	C T	VCP
UniFuture [50]	ICRA'26		Multi-modal	V D / V D	S	V
Edeline [51]	NeurIPS'25		Mem. Enhancement	I A / L	None	C
Ctrl-World [52]	ICLR'26		Space Constraints	I D / V D	S O C	V
Spatial-Mem [53]	NeurIPS'25	Memory Mechanism	Semantic Alignment	V S / I	S M	V
WorldMEM [54]	NeurIPS'25		Memory Bank	V / V	C N M	V
Voyager(LLM) [55]	TMLR'24		Skill Library	T / A	P M	C
SSM-World [56]	ICCV'25		State-Space Models	V / V	A	V
Mem. Forcing [57]	arXiv'25		Memory Replay	V M / I	S C	V
Identity Consistency						
SSWM [58]	arXiv'24	Attention	Semantic Alignment	I / L	P A	P
Loci-v1 [59]	ICLR'23	Occlusion	Imagination Tracking	V / L	O	C
SAVi++ [60]	NeurIPS'22	Tracking	Identity Tracking	V / D O	None	P
ForeDiff [61]	arXiv'25	Anchors	Arch. Decoupling	V / V	A T	VR̂

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

2.1.2. Spatial Continuity

Videos are 2D projections that discard crucial 3D geometry, causing world models to miss physically grounded spatial relationships. Human spatial perception uses multi-view associations and long-term memory [47], inspiring recent methods to recover 3D geometry and preserve long-range spatial dependencies through memory mechanisms.

Implicit representation alignment. These methods do not directly parameterize 3D structures but instead encode physical spatial correlations implicitly within the model through cross-modal fusion or attention mechanisms. For instance, EnerVerse [25] introduces cross-view spatial attention, leveraging camera intrinsics/extrinsics and ray direction maps to model view correspondences and enhance holistic multi-view generation. RoboScape [41] jointly learns from RGB and depth, using depth features as geometric constraints to implicitly acquire 3D physical scene priors rather than merely fitting 2D images. Similarly, ManipDreamer [42], GEM [34] and DeepVerse [21] integrate depth, semantics, RGB, and dynamic masks to improve spatial physical continuity. GAIA-2 [44], as a surround-view world model, ensures multi-camera spatial continuity by aligning streams via structured conditions such as environmental factors and road semantics, enabling high-resolution spatiotemporally coherent generation.

Explicit representation alignment. These approaches provide explicit physical spatial information to supervise joint distributions, spatial guidance, and alignment. WVD [45] encodes global 3D coordinates into spatial pixels, learning the joint distribution of 2D space and 3D coordinates from 6D (RGB+XYZ) video with explicit geometric supervision. Geometry Forcing [47] aligns intermediate video model representations with features from pretrained geometric foundation models, guiding the model to internalize physical geometric alignment over latent perspectives and scales. Pathdreamer [35] accumulates past observations into a 3D point cloud and reprojects it into 2D space as context, explicitly reasoning about the 3D physical geometry of the next frame. InfiniCube [48] constructs an external voxel-based “3D ground buffer" from videos, providing explicit physical grounding to mitigate spatial drift in long sequences. DSG-World [62] explicitly builds a 3D Gaussian world model from dual-state video observations, introducing dual-segmentation-aware Gaussian fields, pseudo-intermediate symmetric alignment, and cooperative pruning/pasting to maintain geometric integrity under occlusions. Voyager-DM [63] aligns RGB with depth and employs efficient world caching for long-range geometrically consistent 3D reconstruction. Spatial-Mem [53] leverages external geometry-grounded 3D world representations (e.g., point clouds) to filter dynamic regions while explicitly memorizing static spatial structures.

Memory mechanisms. Many approaches incorporate external memory banks or retrieval modules [55] to store historical physical observations, mitigating error accumulation and preserving spatial continuity. WorldMem [54] leverages a memory bank to store past visual and state features, achieving long-term 3D physical continuity through a memory attention mechanism. Spatial-Mem [53], in contrast, maintains sparse keyframes as episodic memory and dynamically expands when new regions appear, thereby enhancing the preservation of spatial relationships. Unlike these, Memory Forcing [57] emphasizes backward retrieval of matched source frames from the current viewpoint, centering on geometric continuity to reduce drift and improve structural accuracy when revisiting scenes. Furthermore, PEWM-3D [64] continuously integrates observations into a shared 3D feature map (e.g., Plücker coordinate embeddings), which is referenced throughout the generation phase to ensure globally continuous spatial coherence. In addition, SSM-World [56] introduces structured memory along the spatial dimension via a block-wise scanning State Space Model (SSM) mechanism, partitioning the global space into controllable units. This design balances temporal memory with spatial physical continuity while incorporating frame-level local attention to enhance generation quality.

2.1.3. Identity Continuity

World models that operate directly in pixel space often fail to capture the physically invariant properties of objects across frames. Since pixel-level representations are highly sensitive to noise,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

lighting changes, and viewpoint variation, they cannot maintain the stable attributes — such as shape, material, and appearance — that define object identity in the physical world, leading to identity drift in long sequences. Fine-grained object-level modeling is therefore crucial for maintaining physically grounded identity continuity across frames [65]. In this regard, the GEM model [34] introduces "identity embeddings" for objects to eliminate operation ambiguities and adopts customized integration mechanisms for different types of control signals (such as self-motion, object manipulation, and human posture), ensuring strict physical identity continuity between generated content and diverse control signals. Similarly, FOCUS [66] allocates a unique one-hot identity vector s^t_{obj} to each object, and the object latent extractor generates stable latent representations s^t_{obj} based on this identity, thereby ensuring identity continuity. SSWM [58] employs a differentiable, iterative competitive attention process to automatically and end-to-end map visual inputs to a set of stable, semantically aligned latent representations, ensuring that the identity of objects remains coherent both spatially and temporally. Loci-v1 [59] and SAVi++ [60] approach identity continuity by decomposing the scene into multiple object slots. Similarly, ForeDiff's [61] prediction flow learns during pretraining how to precisely identify and preserve physically grounded identity-related features of objects, such as spatial location, shape, and appearance, from video frames, providing an "identity anchor." Recently, the unprecedented success of Sora2 [13] in cross-scene multi-view identity continuity demonstrates the immense potential of learning physically consistent representations from videos.

2.2. Controllability

The controllability gap exists because videos passively record events and cannot represent action-conditioned responses, causal interventions, or counterfactuals. Without this, models cannot predict "what if" scenarios, limiting policy exploration and intervention. Current research addresses this limitation by introducing explicit modeling of action–outcome mappings to ensure that environmental states respond to interventions in accordance with physical laws, while also striving to construct interactive world models.

2.2.1. Semantic Control

Since videos lack explicit representations of the semantics underlying scene changes — such as object material properties, force interactions, and spatial affordances — world models struggle to ground high-level semantic instructions in physically consistent outputs. This approach aims to bridge semantic intent and physical scene evolution by integrating structured semantic inputs, text-conditioned guidance, and multimodal semantic alignment mechanisms [67], effectively preventing semantic drift and physically inconsistent generation. Models such as GAIA-2 [44], InfiniCube [48], and GEM [34] achieve this by directly injecting structured physical and semantic conditions—such as road semantics, high-definition maps, vehicle bounding boxes, textual prompts, or DINOv2 [68] features—into the generative model. This enables the predictable manifestation of advanced semantic instructions, including multi-agent interaction, multi-camera consistency, dynamic object insertion, weather condition adjustments, and object movement or insertion, while maintaining physical plausibility in the generated scenes.

For finer semantic interaction, Yume [27] utilizes novel sampling strategies and stochastic differential equations to enhance the controllability of text conditions, and achieves physically grounded camera control through quantized camera trajectories. LaDi-WM [69] innovatively designs an interactive diffusion process that dynamically adjusts DINO geometric latent codes and SigLip semantic latent codes, aligning geometry with semantics to accurately model the physically consistent evolution of scene semantics within world dynamics. FlowDreamer [70] adopts 3D scene flow as a universal motion representation, capturing physically grounded semantic deformations of non-rigid objects, and supports fine-grained semantic motion control that respects plausibility. Pathdreamer [35] avoids semantic ambiguity in camera modeling by directly specifying future viewpoint trajectory sequences grounded in spatial geometry. Additionally, NWM [71] and ManipDreamer [42] implement real-time semantic planning and semantic command compliance in robotic operations using CDiT [72] architectures with

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

semantic constraints, energy function-encoded semantic rules, and action tree-structured instructions, ensuring that semantic control signals translate into physically executable robot behaviors.

2.2.2. Interactivity

Interactivity enables world models not only to passively predict but also to actively respond to interventions and instructions, thereby supporting causal reasoning and goal-directed policy exploration. Existing studies mainly focus on two directions: (I) interaction video generation driven by local trajectories or action signals (emphasizing sequence-level action conditioning, suitable for stepwise prediction in robotics or games); and (II) real-time explorable or 3D world models driven by global signals (emphasizing holistic prompts or multimodal inputs, suitable for immersive simulation).

The first line of research emphasizes conditioning video generation models on low-level control signals (e.g., robot joint actions, game inputs, or reward feedback) to enable stepwise interactive prediction. Representative works include Vid2World [38], AVID [73], iVideoGPT [74], DWS [75], AirScape [76], and UnifoLM-WMA-0 [77], which employ action-injection mechanisms to map user-provided control signals into sequences of future video frames, thereby supporting embodied interactive exploration such as robotic trajectory evolution. These approaches essentially constitute controllable video-driven world models, establishing a tight coupling between low-level actions and environmental dynamics. In addition, some works further focus on multi-task trajectory integration and generalization. For instance, WLA [78] aggregates cross-environment trajectories to achieve continuous dynamic prediction, while the IAFM [79] extracts action signals from large-scale robotic video trajectories to support cross-task and cross-domain interactive simulation.

The second line of research is driven by global conditions (e.g., text, natural language instructions, or multimodal prompts), aiming to construct navigable and explorable spatial world representations (3D or 4D) and extend video generation into embodied, immersive, or panoramic scene simulation. For example, TesserAct [80] employs instruction-driven grid evolution mechanisms for controllable environment generation, whereas PlayerOne [36] models first-person-view (POV) videos to support navigation. LatticeWorld [81] addresses static 3D world modeling by seamlessly integrating with an industrial-grade computer graphics rendering engine. In virtual environments, several studies investigate open-world simulation based on game scenarios: MineWorld [82] and Hunyuan-Game [6,83] are specifically designed for game trajectories, supporting agent-controlled interactive exploration; Yan [33] introduces a multi-granularity editing framework that combines peripheral control devices for interactive video generation; while Voyager [63], Mirage 2 [84], and Matrix Game [7] further develop real-time explorable game-world modeling. Moreover, some works explicitly incorporate 3D spatial structures into generative models: for instance, YUME [27], Marble [85], and Matrix-3D [86] combine omnidirectional video generation with interactive control, enabling real-time 360° scene exploration driven by external devices.

2.3. Generalization

Although large-scale video data exist, annotated datasets for embodied intelligence and 3D environments remain scarce. Videos capture appearances rather than underlying physical laws, causing overfitting to visual patterns and poor coverage of rare dynamics. Recent work uses data augmentation, improved architectures, and domain generalization to enhance generalization and cross-scene adaptability.

2.3.1. Data

Data Synthesis and Utilization. A typical approach involves fine-tuning generative models to synthesize diverse new data. For example, DREAMGEN [87] employs a video-based world model to generate both familiar and novel tasks across diverse environments, while integrating a latent action model or inverse dynamics model to infer pseudo-action sequences, forming “neural trajectories." A more advanced framework, Dream to Manipulate [67], introduces a learnable digital twin that combines Gaussian splatting [88] with simulators to generate new environments and action configurations with physically consistent dynamics, thereby expanding the diversity of the training distribution.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

As shown in Figure 3, the EnerVerse-D data engine [25] integrates world models with 4D Gaussian Splatting (4DGS) to establish a self-reinforcing data generation loop capable of producing high-quality, multi-view video data with geometrically consistent scene reconstruction.

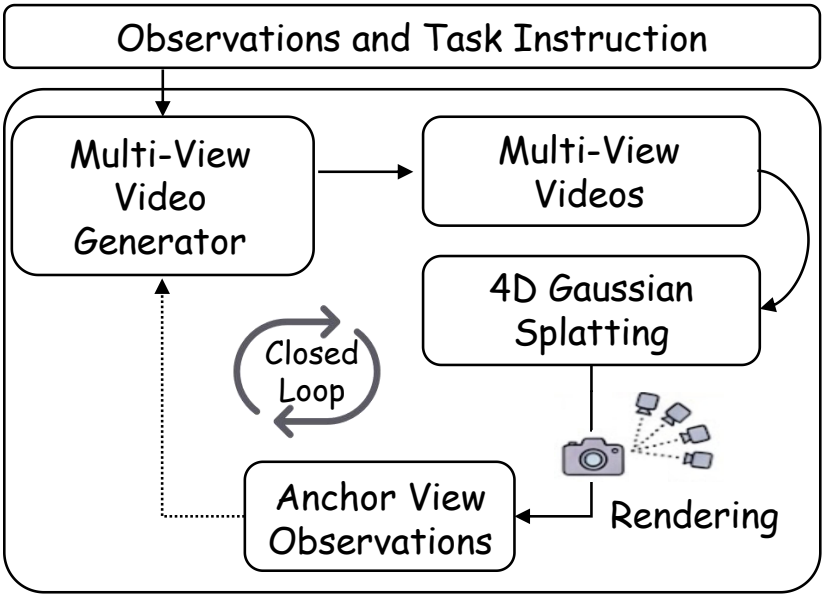

Figure 3. The data engine pipeline of EnerVerse [25] operates as follows: multi-camera observation images and anchor views are processed by a multi-view video generator to produce denoised multi-view videos. Combined with camera pose inputs, these videos are fed into 4DGS for four-dimensional scene reconstruction. The reconstructed results are then rendered into high-fidelity anchor images and iteratively refined through feedback, improving motion consistency and reconstruction accuracy, ultimately achieving geometrically consistent and high-definition outputs.

In addition, many studies explore unsupervised or weakly supervised paradigms. GEM [34] leverages large-scale cross-domain data and pseudo-labels to generate depth maps, trajectories, and poses, improving instance-level perception and generalization. FLARE [89] enhances temporal representation learning by leveraging structured representations and action-aware future embeddings derived from unlabeled data. VLMWM [90] employs a dynamic modeling mechanism to automatically generate pseudo-labels, thereby strengthening the model's supervisory signals for dynamics.

Sim2Real Transfer. Since simulators can explicitly model physical properties such as friction, mass, and contact dynamics, simulation-to-reality transfer offers a principled way to inject physical knowledge into world models. SRCC [91] introduces the Sim2Real Correlation Coefficient to quantify a simulator's predictive fidelity, optimizing simulation parameters to better align dynamics with real-world performance. SimWorld [92], a simulator-conditioned scene generation engine, integrates scene construction and simulation modules to produce synthetic data and labels with physically accurate properties for world model training. Other methods exploit general semantic alignment to bridge the appearance gap. For instance, Lang4Sim2Real [**?**] employs natural language descriptions as a unified signal to bridge the visual gap between simulated and real-world imagery. In addition, generative editability-based approaches, such as Cosmos-transfer [93] and Dreamland [94], leverage multi-conditional multimodal pretraining to achieve photorealistic domain transfer while preserving physical plausibility.

Foundation Model Priors. Cross-scene generalization can be enhanced by leveraging the prior knowledge encoded in pretrained video or multimodal foundation models. For example, UWM [28] and Vidar [95] achieve broad semantic understanding and efficient domain transfer by fine-tuning on only a small amount of domain-specific data. Similarly, Founder [96] maps foundation model representations into the world model latent state space, enabling world models to handle cross-domain scenarios and multimodal tasks. Notably, DINO-WM [97] trains a world model on offline behavioral

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

data using DINOv2-based pretrained visual features, thereby enabling task-agnostic zero-shot planning grounded in physically meaningful visual representations. In addition, Geometry Forcing [47] exploits pretrained 3D foundation models for physically grounded representation learning, alleviating the modeling bottlenecks caused by limited 3D supervision.

2.3.2. Architecture Generalization

One class of approaches introduces structural modeling of visual or geometric inputs using discrete tokens, virtual intermediate states, or hierarchical structures to better capture physically invariant patterns. WorldDreamer [98] learns world dynamics through masked-token prediction, while DSG-World [62] reconciles observational discrepancies via geometric constraints and pseudo-states that encode structure. HMBRL [99] adopts hierarchical models with low-dimensional abstract actions to generalize across complex tasks. Another class emphasizes regularization and robust training to address uncertainty and drift in latent physical representations. MoSim [100] introduces residual flow penalties to avoid physically uncertain regions, while SDE-based [101] and Jacobian-regularized frameworks analyze and correct latent drift errors in physical state representations, thereby improving robustness. A third line of work focuses on multimodal conditioning and policy structures: GenRL [102] achieves task generalization with vision/language prompts, and BPN [103] separates state and task representations with bilinear fusion to enhance transferability and robustness under environmental shifts.

2.3.3. Behavioral and Environmental Generalization

Behavioral Generalization. aims to enhance the adaptability of world models in cross-task and cross-platform action modeling, where distribution shifts in dynamics across embodiments pose a central challenge. RUWM [104] enhances behavioral robustness by analyzing latent representation errors and introducing regularization over physical state transitions. 3DFlowAction [29] and FlowDreamer [70] address challenges across heterogeneous robotic platforms, complex manipulation tasks, and novel objects by leveraging 3D optical flow as a physically grounded motion representation and large-scale pretraining. AdaWorld [105] employs latent action self-supervision to extract and recombine actions, thereby enabling flexible control and behavioral generalization across heterogeneous environments.

Environmental Generalization. focuses on platform differences and adaptation to new environments. Vidar unifies observation spaces across different robotic platforms by integrating multi-view video data, and introduces a masked inverse dynamics mechanism to extract action features from generated videos, achieving few-shot generalization in dual-arm manipulation under varied configurations. DreamerV3 [11] leverages robustness techniques to handle multimodal returns and sparse rewards, improving adaptability in novel environments; its extension, cRSSM [106], systematically incorporates contextual information into DreamerV3, further strengthening generalization to unseen settings. AirSpace [76] adopts a two-stage training scheme—intent-controllable modeling and spatiotemporal constraint learning—to address distribution mismatches in physical dynamics and limited diversity in embodied intelligence for aerial domains. MoSim [100] decouples physics modeling from policy learning via a neural motion simulator, effectively mitigating challenges posed by unseen scenarios and distribution shifts in dynamics, thus enhancing cross-task generalization.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

Table 3. Hardware and configuration comparison for lightweight world-model approaches (for reference only).

Work	Venue	GPUs	Batch Size	Training Steps
DINO-world [107]	arXiv'25	H100*16	1024	350K Iter.
HWM [108]	arXiv'25	A6000*2	128	–
MinD [109]	arXiv'25	A40*4	–	9 Hours
Sparse Imagin. [110]	ICLR'26	3090*4	32	100 Epochs
Simulus [111]	arXiv'25	4090*1	8	100 Epochs
EMERALD [112]	ICML'25	3090*1	16	–
D^2-World [113]	arXiv'24	V100*8	24	24 Epochs
AVID [73]	RLC'25	A100*4	64	7 Days
ScaleZero [114]	arXiv'25	A100*8	512	–
KeyWorld [115]	arXiv'25	A800*8	1	100 Epochs
TWIST [116]	ICRA'24	3090*1	–	500K Iter.
IRIS [117]	ICLR'23	A100*8	256	3.5 Days
Δ-IRIS [118]	ICML'24	A100*1	32	1K Epochs
HERO [119]	arXiv'25	A100*1	–	–
PosePilot [120]	IROS'25	A100*8	-	–
OCWM [121]	ICLR'25	H100*4	32	40 Epochs

2.4. Lightweight

Lightweighting world models is crucial due to high spatial-temporal overheads that limit real-time tasks like robotics and autonomous driving. Beyond efficiency, reducing model size helps filter out visually redundant information, forcing models to capture physically meaningful dynamics rather than irrelevant details. Compact architectures act as inductive biases for learning concise, generalizable, and physics-grounded representations. Existing methods focus on maximizing computational and sample efficiency while preserving essential physical dynamics. Methodologies can be grouped into several complementary strategies:

Learning in structured latent spaces. Since physical dynamics are low-dimensional relative to raw pixel space, learning in structured latent spaces provides a natural way to separate physically meaningful state representations from visual redundancy. Akbulut et al. [122] provided early experimental evidence that structured latent representations significantly improve both efficiency and performance, achieving over a 50% gain compared with modeling directly in observation space. Recent works such as DINO-World [107], MinD [109], and HWM [108] similarly compress high-dimensional pixel streams into compact latent spaces and perform temporal modeling therein, leveraging VQ-VAE [123], denoised latent representations, or low-resolution latent features to fundamentally reduce the computational and memory overhead of sequence modeling while retaining physically relevant state information.

Shortening sequence length via tokenization, sparsification, and parallelism. Physical dynamics are often locally sparse — most video frames contain physically uninformative redundancy between keystate transitions. Methods such as Sparse Imagination (random token dropping and grouped sparse attention) [110], Simulus (modular tokenization) [111], and Δ-IRIS (reduced key-frame encoding with autoregressive optimization) [118] lessen redundant temporal dependencies through token sparsification, concentrating model capacity on significant state changes. In parallel, approaches like EMERALD (MaskGIT-style parallel prediction) [112] and D²-World (non-autoregressive single-stage occupancy prediction) [113] further reduce the cumulative cost of autoregressive training by enabling parallel or masked generation across time.

Parameter-efficient strategies for deployment and transfer. Models including AVID [73], PosePilot [120], and LAMP [124] adapt tasks through lightweight adapters, black-box masking, or plug-and-play modules, avoiding modification — or requiring only minimal fine-tuning — of pretrained backbones that already encode broad physical priors. ScaleZero [114] balances capacity and efficiency via dynamic parameter scaling and staged LoRA expansion. KEYWORLD [115] concentrates computation on a sparse set of semantically and critical frames and employs lightweight convolutional models to synthesize intermediate frames, while TWIST [116] distills a privileged-state teacher world model

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

into a student trained solely on domain-randomized visual observations, effectively reducing both training time and data requirements while preserving generalization.

Table 4. Summary of Works on Physical Grounding Challenges.

Note: Abbreviations used in the table — HR. denotes hierarchical representation. ***Modalities:*** V indicates video or multi-frame; I denotes single-frame or image; L refers to latent representation; S to spatial representation; T to text modality; A to action representation; C to camera (pose) information; D to depth feature; N to neural signal; O to object-level feature; P to physical information; and M to historical memory feature. ***Evals and Downstream Apps***: Physical generation, question answering, interaction, understanding, attributes are abbreviated as $P_{G,Q,I,U,A}$, A refers to action prediction. M stands for motion planning. F stands for fluid dynamics. In downstream applications: W is real world, R is robotics, D is autonomous driving, and O is objects.

Work	Venue	Method	Input / Output	Conditions	Evals and Apps
Explicit Priors & Feedback Integration					
Pandora [126]	arXiv'24	Physical Prompts	I / V	T	$P_G - W$
WorldGPT [127]	MM'24	Modality Alignment	L A / L	V T	$P_G - W$
LLMPhy [128]	arXiv'24	Engine Integration	I / V	V A T	$P_Q - O$
DrivePhysica [129]	arXiv'24	Positional Constraints	V / V	T C S P	$P_G - D$
PhysTwin [130]	ICCV'25	Attribute Fusion	V D / S	None	A – RO
SlotPi [131]	SIGKDD'25	Physical Constraints	V / V	None	$P_{G,Q}F - D$
S2-SSM [132]	arXiv'25	Sparse Regularization	V / I	None	$P_I - O$
RenderWorld [133]	ICRA'25	Pretraining	I D O / A L	S P	M – D
DINO-WM [97]	ICML'25	Pretraining Priors	I A / L	None	M – OR
HERMES [134]	ICCV'25	Multi-view Modeling	I T / L T	A P	A – D
Cosmos [93]	arXiv'25	Multimodal Constraints	V / V	T O D	$P_Q - W$
Disentangling Static and Dynamic Factors					
AdaWorld [105]	ICML'25	Action Decoupling	V / I A	T M	$P_{I,M} - O$
Dyn-O [65]	NeurIPS'25	Dynamic Decoupling	I A / L	O	A – O
ContextWM [135]	NeurIPS'23	Dynamic Decoupling	V A / L	T O S	P_IAM – DR
DisWM [136]	ICCV'25	Dynamic Decoupling	I / L	None	$P_{A,U} - O$
DreamDojo [70]	RAL'26	Explicit Action Modeling	I D A / I D	S	AM – R
DreamZero [29]	arXiv'26	Action Decoupling	I T C / S	A D S	A – R
OC-STORM [137]	arXiv'25	Object Extraction	I O A / L O	None	P_UM – O
AD3 [138]	ICML'24	Action Decoupling	V A / L	L	P_UM – O
LongDWM [139]	arXiv'25	Action Decoupling	V A / V	T S C	$P_G - D$
Vidar [95]	arXiv'25	Action Decoupling	V / V	T N I C	$P_G - R$
DREAMGEN [87]	arXiv'25	Pseudo Action Estimation	V / V A	V A N	$P_G - R$
VLMWM [90]	arXiv'25	Fine-tuning	I A / I A	None	P_GA – W
WorldDreamer [98]	arXiv'24	Disentangled Modeling	V / V	A T	$P_G - W$
Simulus [111]	arXiv'25	Dynamic Decoupling	I A / L	N S	M – O
SCALOR [140]	ICLR'20	Background Modeling	V / L S	O	P_UA – O
AETHER [141]	ICCV'25	Unified Modeling	V A C / V A D	None	P_GA – W
UWM [28]	RSS'25	Action Decoupling	V A / V A	None	P_GA – R
FLARE [89]	CoRL'25	Unified Modeling	I A / A	N T	A – W
Progressive Constraints & Hierarchical Abstraction					
DWS [75]	AAAI'26	Regularization	I / A / V	T	$P_{G,I} - RO$
Dreamland [94]	arXiv'25	Engine Simulation	L O S / I	T	$P_{U,I} - D$
GWM [142]	ICCV'25	Hierarchical Abstraction	I / S / A	A S	A – R
PIWM [143]	arXiv'24	Interpretability	V A P / L P	None	$P_{U,I} - O$
Ross et al. [144]	ICLR'25	Theoretical Framework	L M P / L P	None	$P_A - O$
SimWorld [92]	arXiv'25	Simulation-based Modeling	I T / I	P O N D	$P_{G,U} - D$
MoSim [100]	CVPR'25	Multi-constraint	L A / P	None	A – R
WALL-E [145]	NeurIPS'25	Rule Learning	L A M / L A T	I T	M – O
FOLIAGE [146]	arXiv'25	Hierarchical Abstraction	L A / L	I S O P	$P_A - O$
LLMPHY [128]	arXiv'24	Hierarchical Abstraction	V T / T	A I P	$P_{A,I} - O$
PILWM [147]	arXiv'25	Soft Mask	V / I	None	$P_I - O$
VLWM [148]	arXiv'25	Hierarchical Abstraction	V / T A	M T P	P_QM – W
V-JEPA 2 [12]	arXiv'25	Hierarchical pretraining	V / L	T A N	AM – R

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

These approaches have demonstrated substantial improvements in training/inference speed, model size, and downstream control sample efficiency. Nevertheless, they share several trade-offs rooted in the tension between fidelity and computational economy: the invertibility and fidelity of latent representations constrain the upper bound of variable recovery and fine-grained dynamics modeling; sparsification and non-autoregressive strategies face challenges in maintaining long-horizon continuity and handling multimodal uncertainty in state transitions; and black-box adaptation methods exhibit limited generalization when faced with cross-domain transfer involving substantially different environments or dynamics distributions beyond pretraining.

2.5. Universality

The lack of unification arises because videos capture only limited slices of reality, lacking a shared cross-modal physical representation. Truly universal world models rely on shared physical abstractions that generalize across domains. The Plato hypothesis [149] suggests vision, language, and action are projections of the same underlying reality, implying a unified semantic space. Contemporary models leverage this by mapping heterogeneous modalities into shared physical representations, enabling cross-modal consistency, knowledge transfer, and multi-task robustness.

As shown in Figure 4, Yue et al.'s unified framework [125] decomposes the system into five components: interaction and perception; unified reasoning (dynamics, causality, explicit and latent reasoning); memory; environment (learnable and generative); and multimodal generation (video, image, audio, 3D, prediction). This framework exhibits important complementarities with Eq. (1)–(3) in this paper: while Eq. (1)–(3) define the formal backbone of world models, the above framework explicitly incorporates memory modules and multimodal generative capabilities, thereby providing broader coverage for emerging variants of world models

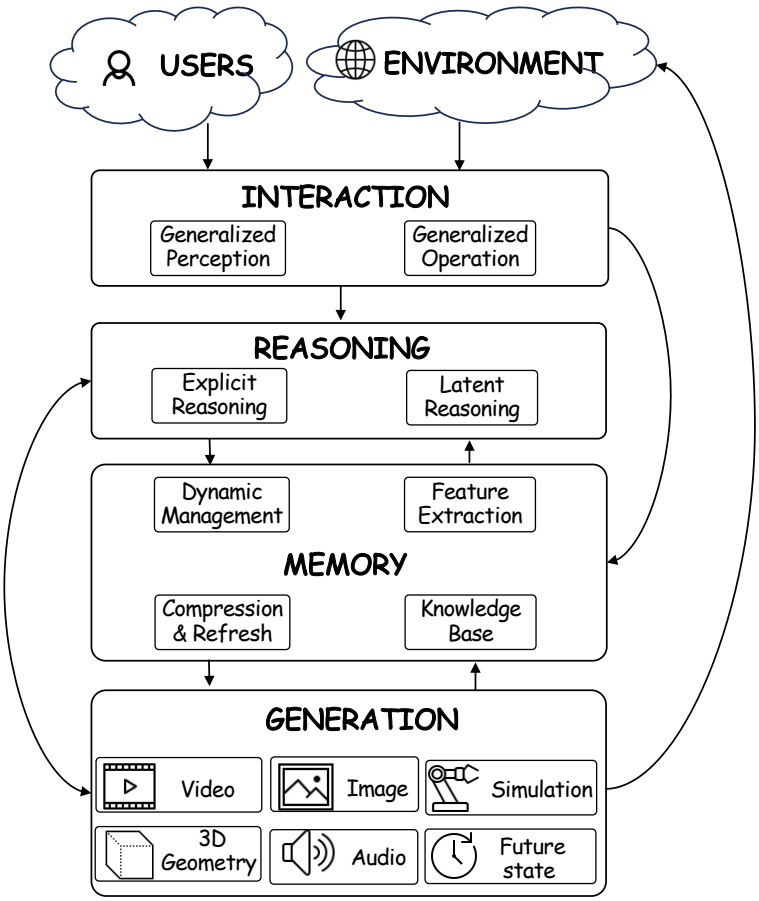

Figure 4. The unified world model framework advocated by Yue et al. [125], featuring interaction, reasoning, memory, and multimodal generation.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

Data pretraining and shared latent-space modeling form the core technical foundation for generality. DINO-world [107] leverages large-scale, uncurated web videos for pretraining, combined with the DINOv2 [68] visual encoder to perform latent-space predictions. Its compact Transformer autoregressive architecture supports variable frame rates and context lengths, extracting physically generalizable knowledge from diverse data and demonstrating seamless cross-task adaptability. WorldDreamer [98] frames world modeling as an unsupervised visual sequence problem, integrating text, image, and action modalities through masked token prediction, with parallel masked-token prediction enhancing efficiency and exhibiting high generalization and flexibility across tasks. V-JEPA 2 [12] performs self-supervised pretraining on Internet videos combined with limited robotic data to learn visually predictable representations of physical dynamics, enabling zero-shot planning and future prediction across downstream tasks. WPT [150] systematically leverages reward-free, non-expert, multi-robot, uncurated offline data, filling in action dimensions to pretrain a single world model across embodied agents. Cosmos [93], as a foundational world model, is pretrained on diverse video and physical scenarios to acquire universal structure and causal patterns, and after finetuning, can generate, predict, and simulate future states, handle cross-modal inputs, and produce controllable outputs, providing a unified framework for AI applications.

Modular design and multi-task shared representations provide architectural guarantees for generality. DreamerV3 [11] jointly learns environment models to predict action outcomes and plan ahead, using normalization and balancing techniques to handle diverse reward distributions and fixed hyperparameters for stable multi-task learning, demonstrating general adaptability without task-specific tuning. AETHER [141] integrates 4D dynamic reconstruction, action prediction, and visual planning within a single physical backbone, emphasizing geometric reasoning and achieving zero-shot generalization to real environments from synthetic data, while its global action representation further unifies and enhances generalization in navigation and robotic tasks. HERMES [134] unifies 3D scene understanding and future evolution via bird's-eye representations, preserving geometric relations while integrating multi-view spatial information, and introduces a "world query" mechanism to fuse physical knowledge within a unified architecture. UniFuture [50] employs dual latent-variable sharing and multi-scale interaction to integrate future generation and depth perception, refining cross-modal features and generating temporally consistent future images and depth maps from single frames, exhibiting zero-shot generalization in unseen environments.

Flexible architectural extensions and instruction-based control further expand the applicability of generality. Simulus [111] uses a modular unified design, independently handling tokenizers, embedding tables, and prediction heads for each modality, decoupling representation learning from physical dynamics modeling and supporting various modality combinations, achieving sample-efficient generality across benchmarks. 1X World Model [151] predicts environment changes from specific action trajectories, capturing cross-task patterns and promoting generalization through multi-task training, enabling unified planning and decision-making across physical tasks. Pandora [126] integrates pretrained language and video models in a two-stage unified training, with instruction finetuning enabling real-time natural language control over video generation, demonstrating cross-domain unification and controllability in indoor, outdoor, and gaming environments. UWM [28] unifies action and video diffusion within a Transformer architecture, decoupling diffusion step execution, policy learning, and world modeling, learning causal and physical understanding from heterogeneous data and supporting flexible unified reasoning.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

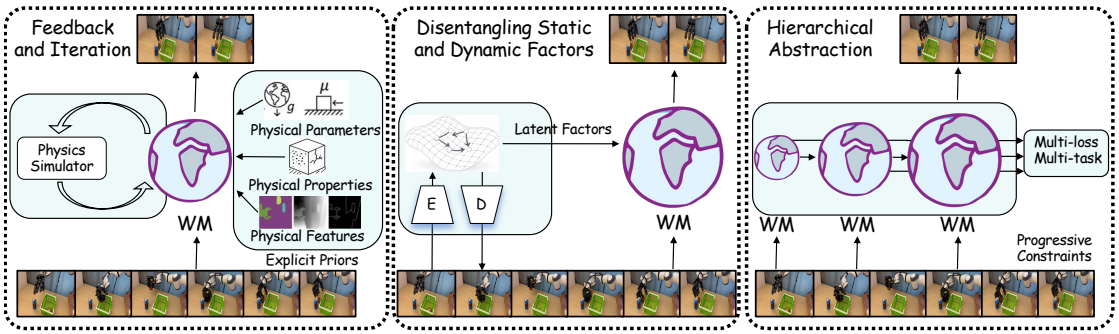

Figure 5. Three paradigms for explicit physical knowledge injection in world models. **Left:** integrating domain knowledge via explicit physical priors and simulation feedback to mitigate hallucinations in dynamic prediction. **Middle:** disentangling physically invariant (static) and physically variant (dynamic) latent factors for scene representation. **Right:** progressive constraints with hierarchical abstractions, from pixel-level fidelity to abstract causal structures.

3. Three Paradigm of Physical World Model

Physical knowledge in videos is often implicitly present and challenging for models to learn effectively. Explicit physical embedding endows world models with physical perception—the ability to understand physical laws, object properties, and causal dynamics—and is crucial for bridging the pixel–physics gap. Existing works can be broadly categorized into three paradigms of explicit injection, as follows:

3.1. Learning from Physical Priors

A prominent strategy is to incorporate physical priors and simulation feedback, anchoring the world model within classical mechanics frameworks to mitigate hallucinations in dynamic prediction. By explicitly injecting domain knowledge—such as object properties, conservation laws, or interaction constraints—models are guided toward physically plausible trajectories even under partial observability or long-horizon rollout. This paradigm is particularly important in safety-critical applications, where purely data-driven models often suffer from compounding errors and unrealistic predictions.

Representative methods include: Pandora [126] and WorldGPT [127], which leverage LLMs to parse textual descriptions of physical properties (e.g., stiffness, friction) and translate them into perceptual priors within simulated environments; LLMPhy [128], which iteratively optimizes scene parameters via LLM-based program synthesis and physics engine feedback to support complex reasoning tasks; and DrivePhysica [129], which explicitly models motion understanding and spatial relationships in autonomous driving scenarios using coordinate alignment, instance flow guidance, and bounding-box conditioning. PhysTwin [130] integrates spring–mass models, sparse-to-dense optimization, and Gaussian splatting to construct digital twins of deformable objects. Other approaches, such as SlotPi [131], introduce Hamiltonian constraints into spatiotemporal reasoning modules to enforce energy-consistent predictions, while S2-SSM [132] infers causal graphs from object interactions using sparsity-regularized state-space models. Renderworld [133] and TRANSDREAMER [152] further enhance physical consistency via 2D–3D occupancy mapping and transformer-based long-range modeling, respectively. These methods exemplify the broader shift toward hybrid neuro-symbolic systems that combine differentiable physics priors with learned representations. Notably, GeoPT [153] proposes a synthetic-dynamics pretraining paradigm that lifts static geometry into a dynamic space, enabling models to acquire physical intuition from unlabeled particle trajectory evolution.

3.2. Learning from Action Decoupling

Another strategy for enhancing physical perception is to explicitly decompose scene representations into physically invariant (static) and physically variant (dynamic) components. This decomposition reflects a physics-inspired inductive bias: background structures and scene geometry typically follow stable constraints over time, while foreground objects evolve under dynamic forces and interac-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

tions. By separating these factors, world models can reduce interference between unrelated signals, leading to improved robustness, generalization, and interpretability in complex environments.

Several recent works operationalize this idea through structured latent representations and disentangled dynamics modeling. Dyn-O [65] employs a Mamba-based state-space model [154] to disentangle object-centric features, enabling stable prediction in cluttered scenes. ContextWM [135] separates perception into a context encoder that extracts temporally invariant cues and a dynamic module that models action-dependent evolution. 3DFlowAction [29] leverages 3D optical flow to unify motion representations across different agents, providing object-centric motion signals independent of the actor identity. OC-STORM [137] extracts precise object states and fuses them with raw visual input to mitigate background noise and repeated recognition errors, while AD3 [138] explicitly isolates task-relevant dynamics from distractors in control scenarios. More recently, latent action disentanglement and embedding approaches [105,155], exemplified by DreamDojo [156], learn compact and transferable action representations directly from human videos. These methods significantly improve embodied physical reasoning by aligning latent action spaces with underlying physical causality.

3.3. Hierarchical Progressive Learning

Hierarchical modeling with progressively enforced supervision has emerged as a scalable framework for physical world modeling, enabling representations to evolve from low-level sensory fidelity to high-level causal abstraction. This paradigm reflects the multi-scale nature of physical systems, where local interactions (e.g., pixel motion or contact dynamics) aggregate into global structures and long-term dependencies. By organizing learning across multiple levels of abstraction, hierarchical models can better capture both fine-grained physical details and coarse-grained semantic regularities.

Representative approaches include RoboScape [41], which dynamically samples salient keypoints (e.g., robots or manipulable objects) to implicitly encode deformation and material properties through temporal consistency; and DWS [75], which emphasizes motion-aware supervision to prioritize dynamic over static features. Dreamland [94] introduces hierarchical and controllable world abstractions that enable precise manipulation at both pixel and object levels, while PSI [157] employs a three-phase loop—predictor training, structure extraction (e.g., optical flow and depth), and reintegration—to achieve structured and controllable representations. From a theoretical perspective, the system-theoretic framework of Ross et al. [144] formulates world-model learning as low-dimensional temporal projection and tokenization, spanning models from classical least-squares regression to generative approaches such as GANs [158].

Recent methods further integrate geometry and physics into hierarchical pipelines. GWM [142] and PIWM [143] combine Gaussian splatting propagation with physics-equation mappings to achieve interpretable 3D state prediction and reasoning. Notably, hierarchical modeling has been widely recognized for its advantages in physical interpretability, modularity, and adaptability across tasks. V-JEPA2 [12], for instance, encodes videos into high-level semantic representations, filters out stochastic or texture-level noise, and learns stable temporal dynamics via masked prediction and multi-scale modeling. Through this progressive abstraction process, such models gradually approximate real-world state transitions and underlying physical laws within the latent space.

4. Benchmarks

In robotics and embodied AI, datasets are essential for training, evaluation, and generalization. They span simulated and real-world scenarios, include multimodal data (images, videos, sensors, language), and support challenges like sim-to-real transfer, multi-task learning, and sparse rewards. The datasets fall into four areas—robot manipulation/planning, physical perception, navigation, and observational quality—and increasingly incorporate extensions for video prediction, lifelong learning, and multimodal fusion.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

Figure 6. Visualization of representative datasets for downstream tasks. The red, orange, blue, and green boxes represent datasets for visual evaluation, robotic manipulation, decision-making and planning, and physical evaluation, respectively

4.1. Datasets

4.1.1. Robot Manipulation

Robotic manipulation research relies on diverse datasets to evaluate world models. Open X-Embodiment [159], contributed by 21 institutions, includes over one million trajectories from 22 robot embodiments and unifies 60 prior datasets under a standard format, supporting cross-platform dynamics learning. Its RT-1 component [160] provides 130k demonstrations across 13 robots and 700 tasks, enabling vision-and-language–conditioned actions and zero-shot multi-task control.

For lifelong and continual learning, the LIBERO dataset [161] introduced by UT Austin in 2023 provides 130 tasks organized into four subsets—spatial, object-centric, goal-directed, and composite—designed to evaluate physical knowledge transfer and policy generalization across varied configurations. The DROID dataset [162] contains 76k trajectories and 350 hours of manipulation data collected with unified hardware (e.g., Franka Panda arm and Oculus controllers), offering RGB, depth, joint states, and language annotations across 564 scenes and 86 tasks. The lightweight PushT benchmark [163] focuses on 2D planar pushing with contact dynamics, suitable for rapid prototyping of diffusion policies and sparse-reward actor–critic algorithms in Gym environments [164].

Simulation-based datasets further improve accessibility to physically diverse training scenarios. The RLBench dataset [165] uses CoppeliaSim [166] to simulate a Franka Panda robot performing 100 tabletop tasks with varied object properties, supporting imitation learning and few-shot learning. Bridge [167] aggregates 60k kitchen manipulation demonstrations across 24 scenes and 13 skills, enriched with language annotations for end-to-end policy learning.

Advanced benchmarks are tailored to assess specialized physical capabilities. VP2 [168] provides 310 task instances with scripted trajectories for MPC, revealing how model capacity affects uncertainty-aware planning. RoboCasa [169] is a kitchen simulation platform featuring 2,500 objects and 100 tasks across various robot embodiments, combining human and auto-generated demonstrations for home-robotics research. DexArt [170] targets fine-grained manipulation of articulated objects with complex joints, while ARMBench [171] focuses on warehouse logistics, comprising 235k pick-and-place operations involving 190k objects.

The DMC [172] and its Remastered variant introduce visual perturbations across physically diverse tasks ranging from double pendulum to humanoid control, targeting robustness to visual overfitting while preserving dynamics evaluation. Franka Kitchen [173] simulates a 9-DoF arm interacting with multiple household appliances; Meta-World [174] provides 50 meta-RL tasks; DrawerWorld [175] evaluates Sawyer arm performance under visual invariance; MuJoCo Pusher [176] centers on precise

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

7-DoF object relocation under continuous dynamics. ManiSkill2 [177], built on SAPIEN, supports 20 task families with 4M demonstration frames spanning 2D/3D inputs. MyoSuite [178] incorporates biomechanical models with muscle fatigue, providing realistic musculoskeletal simulation for intricate dexterous tasks. AgiBot-World-Beta [179] covers 5 scenes, 217 tasks, and 100 robot types; Robomimic [180] provides a modular imitation-learning framework; RoboNet [181] aggregates 15M frames from 7 platforms for vision-based dynamics pretraining.

Further innovations include RH20T [182] with 110k multimodal sequences for evaluating contact-rich dexterous generalization; FurnitureBench [183], offering over 5,000 real teleoperated assembly demonstrations with complex physical constraints; RoboVerse [184], a unified multi-simulator platform with 10M physically diverse transitions; RoboMM [185], enabling cross-dataset transfer via 3D augmentation fusion; RoboAgent [186], integrating semantic enhancement for efficient multi-task learning; EmbodiedBench [187], featuring 1,128 multimodal LLM evaluations in physically grounded environments; and BEHAVIOR-1K [188], which simulates human-like household activities with realistic interactions for high-level planning research.

4.1.2. Planning and Decision-Making

Planning and decision-making research relies on diverse datasets covering embodied navigation, autonomous driving, and procedurally generated scenarios. They capture a range of dynamics—from rigid-body and multi-agent interactions to long-horizon state prediction—providing a foundation for evaluating world-model-driven planning. In embodied navigation, PointMaze [189] serves as a classical benchmark for offline reinforcement learning, examining long-horizon planning and sample efficiency through sparse rewards and multi-scale maze configurations. Habitat [190] leverages high-fidelity reconstructions from Matterport3D [191], HM3D [192], and Gibson [193] to create realistic indoor physical environments that support point-goal navigation and interactive policy learning. JRDB [194] complements these resources with 360° panoramic annotations of pedestrian-rich scenes, facilitating research on social dynamics and socially aware navigation. In autonomous driving, KITTI [195] advances image–LiDAR perception, odometry, and 3D reconstruction through synchronized stereo, LiDAR, and GPS recordings; WorldArena dataset [196] is a dual-arm robot manipulation dataset based on RoboTwin 2.0 [197], containing task video frames, corresponding action control sequences, and task instructions.; and the Waymo Open Dataset [198] spans the full autonomous driving stack—from scene perception and multi-agent trajectory forecasting to end-to-end driving supervision. OpenDV [199] aggregates multi-city driving videos for large-scale prediction and planning research. RealEstate10K [200] contributes large-scale camera pose annotations for 3D scene understanding, while DrivingDojo [201] and DriveWorld [202] focus on world-model training by emphasizing action-conditioned physical dynamics and memory-based state-space modeling for 4D dynamic scene pretraining.

Complementing real-world datasets, procedurally generated benchmarks provide controlled environments for evaluating generalization under varied dynamics. DeepMind Lab [203] and the OpenAI Procgen Benchmark [204] construct diverse task spaces through procedural generation, enabling systematic assessment of out-of-distribution generalization. Atari [205] and BSuite [206] remain standard diagnostic suites for core reinforcement learning capabilities, with Atari100k imposing strict sample-efficiency constraints. The λ Benchmark [207] targets mobile manipulation in multi-room environments, using high-quality human demonstrations to evaluate cross-scene generalization. Collectively, these datasets span the spectrum from real to simulated environments and from low-level perception to high-level decision-making, forming a comprehensive foundation for developing physically grounded world-model-driven intelligent agents.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

Table 5. Summary of benchmarks for robotic manipulation in the context of physical world model evaluation.

Abbreviations: Sr – single-arm, **Dr** – dual-arm, **F** – finger, **G** – general-purpose, **Sk** – skeletal. Suffix **–R/S** indicates real/simulation. **Tr.** – trajectory, **Ts** – task. **Modalities: S** – State, **J** – Joint information, **R** – Reward, **A** – Action, **I** – Image, **M** – Multi-view, **D** – Depth map, **C** – Camera pose, **T** – Text, **N** – Neural signals, **L** – LiDAR, **O** – Object-level.

Name	Category	Modalities	Composition	Size	Brief Summary
AgiBot-World [179]	G-R	JTIDC	1M Tr./217 Ts	43.8T	Humanoid robot training in manipulation, tools, and collaboration
EmbodiedBench [187]	G-RS	JNDT	None/1k Ts	–	Evaluating embodied agents in planning and control
Open X [159]	G-R	–	1M+ Tr./160k Ts	8.9T	Multi-robot learning for cross-embodiment manipulation transfer
BEHAVIOR-1K [188]	G-S	JDTOI	None/1k Ts	165G	Everyday household activities in simulation for AI agents
DMC [172]	G-S	JN	Gen. Tr./50+ Ts	–	DeepMind continuous control tasks for RL in locomotion and manipulation
RoboCasa [169]	G-S	JITD	100k Tr./100 Ts	–	Simulation for generalist robots in kitchens with diverse assets
RoboVerse [184]	G-S	MYJI	500k+ Tr./1k+ Ts	23G	Unified simulation for robot learning across tasks
DexArt [170]	F-S	SJ	None/4 Ts	–	Dexterous manipulation of articulated objects with robotic hands
MyoSuite [178]	Sk-S	JN	10k Tr./204 Ts	–	Musculoskeletal models for dexterous human-like control
ARMBench [171]	Sr-R	JTO	240k Tr./3 Ts	–	Amazon warehouse pick-and-place perception and manipulation
Bridge [167]	Sr-R	MT	60k Tr./13 Ts	387G	Multi-task manipulation from demonstration data
DROID [162]	Sr-R	DCT	76k Tr./86 Ts	1.7T	In-the-wild manipulation from mobile robots in offices
FurnitureBench [183]	Sr-R	J	5k Tr./8 Ts	55G	Long-horizon manipulation such as furniture assembly
RH20T [182]	Sr-R	JDNT	110k Tr./150+ Ts	5T	Multimodal contact-rich robotic skills for one-shot learning
RoboAgent [186]	Sr-R	M	100k Tr./38 Ts	425G	Manipulation demonstrations for task-specific learning
RoboNet [181]	Sr-R	J	162k Tr./None	0.8T	Multi-robot transfer learning in tabletop manipulation
RT-1 [160]	Sr-R	T	130k+ Tr./700+ Ts	111G	Visuomotor policies from large-scale robot data
Franka Kitchen [173]	Sr-S	J	513 Tr./22 Ts	–	Kitchen interaction tasks with Franka robot
LIBERO [161]	Sr-S	TJ	1693 Tr./130 Ts	–	Long-horizon task learning and generalization
ManiSkill2 [177]	Sr-S	JDSON	4M Tr./20 Ts	151G	Generalizable manipulation across robots and environments
Meta-World [174]	Sr-S	J	2M Tr./50 Ts	46G	Meta-RL manipulation for fast adaptation
MuJoCo Pusher [176]	Sr-S	JN	5k Tr./1 Ts	–	Continuous control pushing task in MuJoCo
PushT [163]	Sr-S	J	122 Tr./1 Ts	2.8G	Tabletop pushing interaction tasks
RLBench [165]	Sr-S	DOJ	Gen. Tr./100 Ts	–	Simulation for RL and imitation learning
RoboMM [185]	Sr-S	MCDTJ	70k Tr./100+ Ts	–	Multimodal generalist manipulation model
VP2 [168]	Sr-S	J	310 Tr./11 Ts	182G	Visual planning for object manipulation
Robomimic [180]	S/Dr-S	J	5.9k Tr./5 Ts	19G	Offline imitation and RL for manipulation

Table 6. Summary of benchmarks for decision-making & planning in the context of physical world model evaluation.

Modalities: S – State, **J** – Joint information, **R** – Reward, **A** – Action, **I** – Image, **M** – Multi-view, **D** – Depth map, **C** – Camera pose, **T** – Text, **N** – Neural signals, **L** – LiDAR, **O** – Object-level.

Name	Modalities	Composition	Brief Summary
RealEstate10K [200]	–	80k video-extracted trajectories	Camera trajectories, intrinsics, and poses
Procgen Benchmark [204]	–	Programmatically generated	16 diverse games with varying difficulty
DeepMind Lab [109]	AR	80k trajectories	3D first-person navigation and control
DrivingDojo [201]	ASTO	18k videos	Ego-vehicle actions, multi-agent interaction, open-world driving
Atari [205]	IAR	Programmatically generated	50+ classic Atari games
Habitat [190]	IDOSR	–	Indoor navigation and task interaction simulation
World-in-World [208]	IDMATC	4 platforms and 4 tasks	Closed-loop interactive embodied tasks
KITTI [195]	LTDOS	180G videos, 100k trajectories	Mobile robotics: perception, SLAM, planning, and detection
Waymo Open [198]	MLSO	3 video subsets	Multi-city driving perception and behavior prediction
JRDB [194]	MLO	54 scenes	Indoor/outdoor navigation and human-robot interaction
PointMaze [189]	SARI	10M states	Point-mass maze navigation
Bsuite [206]	SAT	Programmatically generated	T-maze, umbrella task, and path exploration
λ Benchmark [207]	TSA	571 demonstrations	Language-guided indoor navigation and interaction
OpenDV [199]	TALM	3T video	~2059 hours of real-world driving videos
WorldArena [196]	AIT	500 videos, 100k trajectories	Open-world embodied evaluation with planning and interaction

4.1.3. Physical Perception

Physical perception datasets help models learn physical laws, action recognition, and dynamics prediction using synthetic or real videos. Meta FAIR IntPhys (and IntPhys2) [209] evaluates persistence, invariance, spatiotemporal continuity, and solidity via a violation-of-expectation paradigm, with models achieving 50% accuracy. InfLevel [210] emphasizes memory and reasoning under variable conditions, while SSv2 [211] contains 220,847 short videos of hand–object interactions across 174 action categories, serving as a challenging benchmark for robotics and video pretraining.

Egocentric video datasets like Epic-Kitchens-100 [212] provide 100 hours of head-mounted kitchen recordings with 90,000 annotated actions for recognition, anticipation, and adaptation, including hand–object boxes. EVA-Bench [26] evaluates embodied video prediction using multimodal vision–language and diffusion models across 125 human and robotic activities. NVIDIA PhysX [213], integrated into Omniverse by 2025, enables large-scale AI physical simulation for tasks like destruction, fluids, and particle dynamics.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

Regarding simulation platforms, CARLA [214] provides a customizable urban driving environment with variable weather conditions to assess the robustness of autonomous driving perception. Kubric [215], integrating PyBullet[216] with Blender [217], enables large-scale synthetic data generation (terabyte-scale) for tasks such as optical flow and object detection. HeterNS [218] extends the Navier–Stokes equations for heterogeneous fluid simulation, supporting PDE solvers like Unisolver for vortex prediction. TraySim [128] generates multi-view collision and impact scenarios in MuJoCo, supporting research that combines LLMs with physics engines for stability prediction. WorldNet [127] offers multimodal state-transition data, including outdoor and real-world subsets, specifically designed for world model training.

DeepMind's Perception Test [219], a multimodal dataset containing 11,620 video clips, evaluates models' memory, physical reasoning, and semantic understanding through tracking and question-answering tasks. Physion [220] includes 1,200 object-interaction videos (e.g., rolling, deformation) for visual prediction research. PHYBench [221] assesses LLMs' physical reasoning capabilities through 500 physics problems ranging from classical mechanics to quantum phenomena, using expression edit distance as the evaluation metric. PhysBench [222] contains over 10,000 data samples combining images, videos, and text, covering object properties, object relations, scene understanding, and physical dynamics across multiple tasks, aiming to assess models' observation and reasoning of real-world physical phenomena. SCOPE [223] provides 17,600 frames of multi-agent urban scene data across varying weather conditions, enabling research on collaborative perception.

4.1.4. Quality and Understanding of Visual Physics

Observation quality and understanding datasets assess models' abilities in semantic parsing, spatial reasoning, and visual fidelity, crucial for accurate world-model interpretation [32,224]. VSPW [225] offers 3,536 high-res videos with 251,632 semantic frames across 200 scenes; Cityscapes [226] provides 5,000 urban images with instance, disparity, and sequence annotations. ShotBench [227] and ShotVL [228] evaluate cinematic QA and multimodal alignment, ManipBench [229] tests manipulation reasoning, and Matterport3D [191] supplies 194,400 RGB-D images for 3D reconstruction and semantic understanding.

Spatial understanding datasets evaluate whether world models maintain geometrically consistent representations across viewpoints and time, essential for physically grounded prediction. R2R [230] provides 22,000 natural language navigation instructions in Matterport3D; GameWorld Score [7] measures Minecraft scene quality and controllability. SAT [231] offers 218,000 QA pairs for dynamic spatial reasoning, WhatsUp [232] tests basic spatial relations, NuInteract [233] includes 150,000 driving-related image-QA pairs with 2D/3D localization, OmniDrive-nuScenes [234] evaluates 3D reasoning and counterfactuals, and Yume-Bench [27] assesses video generation quality.

Multimodal and causal reasoning datasets push evaluation toward the higher-order scene understanding that physically grounded world models ultimately require. OmniWorld [235] aggregates four-dimensional multimodal data from games and public sources for scene reconstruction and prediction. WorldPrediction [236] uses discriminative tasks to evaluate models' causal reasoning in dynamic activity scenarios. Cosmos-Reason1 [93] provides 400 million video-text pairs specifically curated for training world models with physical commonsense reasoning in robotics and autonomous driving. PhyWorldBench [237] targets the physical fidelity of text-to-video generation models and contains approximately 1,050 physical phenomenon prompts to evaluate whether models adhere to real-world physical laws (including basic motion, collisions, energy conservation, etc.). It even includes "Anti-Physics" scenarios to test whether models can deliberately violate conventional physical logic according to instructions. WorldModelBench [238] contains 350 prompts across 7 domains and 56 subdomains, with human annotations to test whether generated videos follow real-world physics, causal consistency, and task instructions.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

Table 7. A summary of commonly used visual evaluation benchmarks in world-model research.

Name	Semantic Category	Annotation Category	Composition
VSPW [225]	Indoor/Outdoor	–	3.5k videos
Cityscapes [226]	Urban	Semantic, depth & camera parameters	25k labeled videos
ShotBench [227]	Photography	Shot size, motion, lighting, layout	3.5k question answering pairs.
ManipBench [229]	Robotics	Task QA, deformation understanding	13k question answering pairs.
Matterport3D [191]	Indoor	Asset labels, depth, normals	90 buildings, 11k rooms
R2R [230]	Indoor	Room, asset, and action descriptions	7k paths
GameWorld Score [7]	Games	Quality evaluation, spatio-temporal consistency	1k hours labeled
SAT [231]	Indoor/Outdoor	Spatial QA, motion annotation	218k question answering pairs
WhatsUp [232]	Indoor/Outdoor	Spatial positions	820 question answering pairs
NuInteract [233]	Autonomous Driving	Description, spatial information, task category	850 scenes
OmniDrive [234]	Autonomous Driving	Lane-object and counterfactual reasoning	–
Sekai [239]	Urban, Games	Position, scene, weather, time, trajectories	5000+ hours videos
OmniWorld [235]	Real/Sim. World	Games, embodied, navigation, planning	600k videos
Cosmos-Reason1 [240]	Real/Sim. World	Spatio-temporal and physical annotations	1.7M question answering pairs
WorldPrediction [236]	Third-person	Healthcare, assembly, repair	810 instructional videos
PhyWorldBench [237]	Real/Sim. World	Physics categories (10 × 5 subcategories)	1,050 prompts
PhysBench [222]	Real/Sim. World	Object properties, relations, scene dynamics	10k question answering pairs
WorldModelBench [238]	Real/Sim. World	Comprehensive domains and disciplines	350 instances

4.2. Metrics

World model development remains fragmented, with diverse designs and evaluation metrics across tasks. Some works prioritize generative quality, others physical plausibility, often neglecting holistic performance. To address this, we systematically categorize existing evaluation methods by focus and application domain as follows:

4.2.1. Visual Physics Evaluation Metrics

Pixel and Structural Consistency: Metrics such as PSNR, SSIM [241] and Chamfer L1 Distance [242] measure the similarity between generated and ground-truth images at the pixel and structural levels.

Distributional and Perceptual Distance: Indicators including FID [243], FVD [244], Fréchet DINO (FDD) [245], LPIPS [246], DreamSim [247], and Fréchet Video Motion Distance (FVMD) [248] assess the statistical similarity between the generated and real data distributions.

Semantic and Aesthetic Metrics: CLIP Score [249], CLIP Aesthetic [250], and DINO-Score [251] reflect semantic alignment and aesthetic quality of generated results. Additionally, benchmark suites such as VBench [252] and WorldScore [253] evaluate video generation performance in terms of subject-background consistency, frame coherence, motion smoothness, image quality, and style stability.

Semantic Matching Metrics: BLEU [254], METEOR [255], ROUGE-L [256], CIDEr [257], Commonsense [238], Scene Revisit Consistency (SRC) [64], DOVER [258], and Embedding Cosine Similarity measure text-to-vision correspondence and scene reconstruction capability.

4.2.2. Control, Planning, and Decision-Making Metrics

While specific implementations vary across robotics, navigation, and autonomous driving, most downstream evaluation logic shares common principles and dimensions. These tasks typically emphasize the overall performance across the perception–prediction–decision–control pipeline, capturing correctness, robustness, dynamic stability, and generalization capacity.

Performance and Correlation Metrics: Return [259], Rank Correlation [260], ELO Rating [261], and Human-Normalized Scores (HNS) [262] quantify policy performance and model consistency across tasks. **Statistical and Error Metrics:** Precision, Recall, F1-score [263], MSE, Inter-Quartile Mean (IQM), and Normalized Hilbert-Schmidt Independence Criterion (NHSIC) [264] measure the deviation between model predictions and real-world distributions.

Dynamic and Model Error Metrics: Model Error Prediction (MEP) [265], Value Gap [266], Return Correlation [267], One-step Error (OSE) [268], Global Consistency Index (GCI) [267], and Global Prediction Error (GPE) [267] assess error propagation and stability in long-term prediction within dynamic environments.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

Efficiency and Action Metrics: Sampling Rate, FPS, Time Costs, Action Accuracy [78], State Coverage [269], Discrete Action Classification (DAC) [82] and Probability of Improvement [270] evaluate computational efficiency, action generation precision, and exploration coverage.

In robotic manipulation scenarios, world models perform end-to-end mappings from high-dimensional sensory inputs to continuous control outputs. Metrics thus emphasize operational precision and task success, such as Success Rate, L1 Error, Instruction Following (IF) [87], End-effector Position Error (EEPE) [12], Mean Per-Joint Position Error (mPJPE) [271], Mean Relative-Root Position Error (mRRPE) [271] and Achieved Horizon [100], reflecting the model's ability to perform fine-grained control and respond accurately to environmental feedback.

In control and navigation tasks (e.g., trajectory generation, autonomous driving), the evaluation focus shifts toward spatial accuracy and temporal coherence. Common metrics include Trajectory Tracking Success Rate (TTSR) [272], Absolute Trajectory Error (ATE) [273], Relative Pose/Translation/Rotation Error [273], Average Displacement Error (ADE) [274], Navigation Error [275], Collision Rate [258], Optimality Gap [276], and Time-to-Collision [277], which collectively measure geometric precision, motion smoothness, and planning optimality and safety. Furthermore, Normalized Dynamic Time Warping (NDTW) [278] and Success weighted by NDTW quantify temporal alignment and dynamic path consistency.

4.2.3. Physical Grounded Metrics

Physical grounded perception metrics assess a world model's ability to capture real-world physical laws, three-dimensional spatial relations, and dynamic consistency—core indicators of its depth of "world understanding."

Geometric and Spatial Consistency: Metrics such as mIoU, Foreground mIoU (FmIoU) [279], and Foreground Adjusted Rand Index (FARI) [280] quantify structural reconstruction accuracy in segmentation and foreground recovery tasks. SPICE [281] and EVA-Score [26] evaluate semantic-level spatial expressiveness in physical reasoning and alignment tasks. Depth and surface metrics such as AbsRel [282], normal error (Median/Mean), and Accuracy Thresholds [282] further quantify geometric deviations in depth and normal estimation, providing evidence for 3D reconstruction fidelity.

Temporal and Dynamic Consistency: Flow Error [283] and Reprojection Error [253] measure projection deviations and optical flow stability in motion estimation and viewpoint reconstruction, while Revisit Error [54] tests physical consistency across repeated scene observations, verifying whether object representations remain temporally coherent.

Physical Constraint and Interaction Rationality: Physics Alignment [87] measures conformity to physical laws (e.g., gravity, collision, friction), while Control of Object Manipulation (COM) [34] and Implicit Risk Assessment (IRA) [34] capture safety and physical rationality in interactive tasks. GameWorld Score [7] offers a holistic evaluation of object consistency, scene coherence, and input control precision (e.g., keyboard, mouse) in simulation environments, indicating cross-scenario generalizability.

Representation and Feature Consistency: DINOv2 L2 Distance [68] and Structural Hamming Distance (SHD) [284] quantify the preservation of structural and causal relationships in learned feature space. Key Points Matching (KPM) [285] and Hits@k [286] assess performance in structural retrieval and correspondence, indicating adherence to physical constraints. Finally, WorldScore (3D Consistency, Object Content Control) [253] integrates 3D coherence and object-level controllability, serving as a comprehensive indicator of multimodal physical consistency.

5. Future Directions and Discussion

5.1. Open Challenges

Deep Integration of Physics Engines and Neural Networks. Current video-driven world models primarily rely on large-scale data pretraining to implicitly learn physical laws. However, they still exhibit significant limitations in out-of-distribution generalization and in modeling complex phe-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

nomena such as contact dynamics and non-rigid deformations. Hybrid paradigms, exemplified by ContactGaussian-WM [287], integrate differentiable physics engines (e.g., MuJoCo [176], PhysX [213], Isaac [288]) with 3DGS representations, enabling more physically consistent modeling through the synergy of physical priors and neural networks. Nevertheless, several key challenges remain, including differentiability bottlenecks (e.g., discontinuities in contact events that hinder gradient propagation), computational efficiency (e.g., mismatches between simulation time scales and video frame rates), and the difficulty of accurately inferring material properties from visual observations.

Interpretability and Verifiability of World Models. Most existing world models operate as end-to-end black-box systems. While effective for perception and generation tasks, their lack of interpretability hinders deployment in safety-critical scenarios such as surgical robotics and nuclear facility inspection. Approaches such as PIWM [143] attempt to align prediction processes with explicit physical equations (e.g., Newtonian mechanics and energy conservation), incorporating physically meaningful latent variables, symbolic regression, and formal verification to enhance interpretability and verifiability. Recently, the divergence between JEPA [12] and generative paradigms has raised a fundamental question: is pixel-level prediction necessary for learning physical laws, or should models instead operate in abstract representation spaces to improve efficiency and generalization?

Cross-Domain Transfer. Section 2.5 discussed recent progress toward unified paradigms across multiple tasks, environments, and frameworks. However, achieving seamless generalization across domains—such as gaming, robotic manipulation, autonomous driving, and the real world—remains an open problem. First, general-to-specialized adaptation remains challenging. For instance, although the "pretraining–adaptation" paradigm (e.g., Waymo's driving model [198]) has shown promise, heterogeneity in physical dynamics across domains (e.g., vehicle dynamics vs. robotic manipulators vs. human motion) limits the effectiveness of straightforward fine-tuning. Second, unified action spaces are still lacking. Cross-embodiment transfer is constrained by differences in action representations. Promising directions include the development of universal intermediate action representations, embodiment-agnostic motion primitives, and hierarchical action abstractions [156]. Third, commonsense-driven transfer remains underexplored. Works such as DisWM [136] and DINO-WM [97] have taken initial steps, but systematic transfer of physical commonsense still lacks robust evaluation protocols and unified modeling frameworks.

Standardization of Evaluation Metrics. Section 4 provided a detailed overview of existing benchmarks. However, visual realism does not guarantee physical correctness, and strong task performance does not necessarily imply genuine understanding of physical laws. Although efforts such as WorldScore [253] and WorldModelBench [238] aim to establish comprehensive evaluation frameworks, no community-wide consensus has yet been reached. Recent initiatives, including the CVPR 2025 WorldModelBench Workshop and the ICLR 2026 World Models Workshop, indicate that the community is actively moving toward standardized evaluation methodologies.

Data Bottlenecks and Synthetic Data Closed Loops. High-quality datasets must exhibit rich physical diversity and include multimodal signals such as depth, force/torque, and joint states, along with high-precision annotations. However, existing datasets remain insufficient in these aspects. Approaches such as NVIDIA Cosmos [93] and EnerVerse [25] propose a closed-loop data paradigm—"real → simulation → generation → augmentation of real data"—which improves data efficiency. Nevertheless, this paradigm introduces the risk of amplifying physical biases through generative models, highlighting the need for independent physical validation mechanisms.

5.2. Industrialization and Deployment

World models are rapidly transitioning from academic research to industrial deployment, forming a synergistic ecosystem across foundational platforms, vertical applications, and data infrastructures. According to a Frost & Sullivan white paper, over 80% of autonomous driving algorithm development pipelines have incorporated world models or simulation for training and validation.

On the platform level, NVIDIA's Cosmos [93] has evolved into a comprehensive toolchain encompassing multimodal data management, video tokenization, pretrained models, and inference services.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

Google DeepMind's Genie 3 [14] pushes world models toward real-time interaction, demonstrating general-purpose capabilities in autonomous driving and game generation. Meanwhile, World Labs [85] has productized 3D scene generation through Marble as a SaaS platform for creators, validating a viable B2C commercialization pathway. In vertical domains, Waymo's world model [198] built upon Genie 3 demonstrates multi-sensor fusion, hierarchical control, and long-tail scenario generation, representing a benchmark implementation of the "general pretraining + domain adaptation" paradigm. Similarly, Huawei's WEWA architecture and NIO's Neural World Model (NWM) system emphasize embedding world models into closed-loop perception–prediction–planning pipelines, enabling online simulation and decision optimization. On the data side, large-scale robotic datasets such as AgiBot-World are lowering entry barriers for the industry.

5.3. Safety and Ethical Challenges

The *International AI Safety Report 2026* [289] highlights that current AI systems exhibit "striking limitations" in physical world reasoning: while performing well within training distributions, they may fail catastrophically in edge-case physical scenarios. Representative examples include multi-vehicle collisions on icy roads in autonomous driving and millimeter-scale operational errors in surgical robotics.

Moreover, although synthetic data closed-loop training (see Section 2.3.1) improves data diversity, most existing studies overlook the risk of recursive bias amplification. Specifically, a generative model G, trained on a real dataset $\mathcal{D}$, produces synthetic data $\tilde{\mathcal{D}} = G(\mathcal{D})$ that inevitably inherits statistical biases from $\mathcal{D}$, while also introducing new biases arising from the model's inductive priors. These biases accumulate through nonlinear feedback loops, leading to systematic under representation or overestimation in specific physical scenarios. For instance, in rare robotic manipulation settings—such as handling irregular soft objects, left-handed operations, or cluttered environments—synthetic data may reinforce physical inconsistencies rather than correct them if diversity is not explicitly enforced.

Finally, unlike conventional deepfake content that often lacks strong physical consistency, the ultimate objective of world models is to construct highly realistic world simulators. At that stage, the indistinguishability of fine-grained physical details may enable malicious applications, including public opinion manipulation, scientific fraud, and data poisoning attacks targeting physical AI systems.

6. Conclusion

This paper provides a systematic overview of the cutting-edge field of video-driven physical world models. The in-depth analysis was conducted from four different perspectives: the modeling paradigms, the pixel–physics gap, benchmarks and outlooks.

Regarding the pixel–physics gap, we identified and analyzed six core challenges: continuity, controllability, generalization, physical grounding, lightweight, and universality. For each challenge, we summarized representative solutions and research progress. These analyses show that while the videos provide rich learning resources for world models, they also lead to a series of gaps that need to be bridged through modeling paradigms and training strategies.

Physical consistency is not only a technological challenge, but also the most critical bottleneck for world models to evolve from a "video predictor" to a "trustworthy physical simulator." In general, this article reveals the trends and several potential pathways for the evolution towards modeling the physical world.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ha, D.; Schmidhuber, J. World models. *arXiv preprint arXiv:1803.10122* **2018**, *2*.
2. Zhang, P.F.; Cheng, Y.; Sun, X.; Wang, S.; Zhu, L.; Shen, H.T. A Step Toward World Models: A Survey on Robotic Manipulation. *arXiv preprint arXiv:2511.02097* **2025**.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

3. Tu, S.; Zhou, X.; Liang, D.; Jiang, X.; Zhang, Y.; Li, X.; Bai, X. The role of world models in shaping autonomous driving: A comprehensive survey. *arXiv preprint arXiv:2502.10498* **2025**.
4. Feng, T.; Wang, W.; Yang, Y. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260* **2025**.
5. Guan, Y.; Liao, H.; Li, Z.; Hu, J.; Yuan, R.; Zhang, G.; Xu, C. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles* **2024**.
6. Li, J.; Tang, J.; Xu, Z.; Wu, L.; Zhou, Y.; Shao, S.; Yu, T.; Cao, Z.; Lu, Q. Hunyuan-GameCraft: High-dynamic Interactive Game Video Generation with Hybrid History Condition. *arXiv preprint arXiv:2506.17201* **2025**.
7. 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**.
8. Medsker, L.R.; Jain, L.; et al. Recurrent neural networks. *Design and applications* **2001**, *5*, 2.
9. Hafner, D.; Lillicrap, T.; Ba, J.; Norouzi, M. Dream to Control: Learning Behaviors by Latent Imagination. In Proceedings of the International Conference on Learning Representations, 2020.
10. Hafner, D.; Lillicrap, T.; Norouzi, M.; Ba, J. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193* **2020**.
11. Hafner, D.; Pasukonis, J.; Ba, J.; Lillicrap, T. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104* **2023**.
12. 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**.
13. OpenAI. Sora 2 is here. https://openai.com/index/sora-2/, 2025. Accessed: 2025-06-05.
14. DeepMind. Genie 3: A new frontier for world models. https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/, 2025. Accessed: 2025-06-05.
15. Yue, J.; Huang, Z.; Chen, Z.; Wang, X.; Wan, P.; Liu, Z. Simulating the Visual World with Artificial Intelligence: A Roadmap. *arXiv preprint arXiv:2511.08585* **2025**.
16. 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**.
17. 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**.
18. 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**.
19. Liu, D.; Zhang, J.; Dinh, A.D.; Park, E.; Zhang, S.; Mian, A.; Shah, M.; Xu, C. Generative physical ai in vision: A survey. *arXiv preprint arXiv:2501.10928* **2025**.
20. Xie, N.; Tian, Z.; Yang, L.; Zhang, X.P.; Guo, M.; Li, J. From 2D to 3D Cognition: A Brief Survey of General World Models. *arXiv preprint arXiv:2506.20134* **2025**.
21. Chen, J.; Zhu, H.; He, X.; Wang, Y.; Zhou, J.; Chang, W.; Zhou, Y.; Li, Z.; Fu, Z.; Pang, J.; et al. DeepVerse: 4D Autoregressive Video Generation as a World Model. *arXiv preprint arXiv:2506.01103* **2025**.
22. Chen, T.; Hu, X.; Ding, Z.; Jin, C. Learning World Models for Interactive Video Generation. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
23. Wang, S.; Tian, J.; Wang, L.; Liao, Z.; lijiayi.; Dong, H.; Xia, K.; Zhou, S.; Tang, W.; Hua, G. SAMPO: Scale-wise Autoregression with Motion Prompt for Generative World Models. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
24. 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**.
25. Huang, S.; Chen, L.; Zhou, P.; Chen, S.; Liao, Y.; Jiang, Z.; Hu, Y.; Gao, P.; Li, H.; Yao, M.; et al. EnerVerse: Envisioning Embodied Future Space for Robotics Manipulation. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
26. Chi, X.; Fan, C.K.; Zhang, H.; Qi, X.; Zhang, R.; Chen, A.; Chan, C.m.; Xue, W.; Liu, Q.; Zhang, S.; et al. Eva: An embodied world model for future video anticipation. *arXiv preprint arXiv:2410.15461* **2024**.
27. Mao, X.; Lin, S.; Li, Z.; Li, C.; Peng, W.; He, T.; Pang, J.; Chi, M.; Qiao, Y.; Zhang, K. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744* **2025**.
28. Zhu, C.; Yu, R.; Feng, S.; Burchfiel, B.; Shah, P.; Gupta, A. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792* **2025**.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

29. 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**.
30. Chen, B.; Martí Monsó, D.; Du, Y.; Simchowitz, M.; Tedrake, R.; Sitzmann, V. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems* **2024**, *37*, 24081–24125.
31. Zhang, K.; Tang, Z.; Hu, X.; Pan, X.; Guo, X.; Liu, Y.; Huang, J.; Yuan, L.; Zhang, Q.; Long, X.X.; et al. Epona: Autoregressive diffusion world model for autonomous driving. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 27220–27230.
32. Alonso, E.; Jelley, A.; Micheli, V.; Kanervisto, A.; Storkey, A.J.; Pearce, T.; Fleuret, F. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems* **2024**, *37*, 58757–58791.
33. Ye, D.; Zhou, F.; Lv, J.; Ma, J.; Zhang, J.; Lv, J.; Li, J.; Deng, M.; Yang, M.; Fu, Q.; et al. Yan: Foundational interactive video generation. *arXiv preprint arXiv:2508.08601* **2025**.
34. Hassan, M.; Stapf, S.; Rahimi, A.; Rezende, P.; Haghighi, Y.; Brüggemann, D.; Katircioglu, I.; Zhang, L.; Chen, X.; Saha, S.; et al. Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 22404–22415.
35. Koh, J.Y.; Lee, H.; Yang, Y.; Baldridge, J.; Anderson, P. Pathdreamer: A world model for indoor navigation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14738–14748.
36. Tu, Y.; Luo, H.; Chen, X.; Bai, X.; Wang, F.; Zhao, H. PlayerOne: Egocentric World Simulator. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
37. Savov, N.; Kazemi, N.; Zhang, D.; Paudel, D.P.; Wang, X.; Gool, L.V. StateSpaceDiffuser: Bringing Long Context to Diffusion World Models. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
38. 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**.
39. Robine, J.; Höftmann, M.; Harmeling, S. Simple, Good, Fast: Self-Supervised World Models Free of Baggage. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
40. Cui, Y.; Chen, H.; Deng, H.; Huang, X.; Li, X.; Liu, J.; Liu, Y.; Luo, Z.; Wang, J.; Wang, W.; et al. Emu3. 5: Native Multimodal Models are World Learners. *arXiv preprint arXiv:2510.26583* **2025**.
41. Shang, Y.; Zhang, X.; Tang, Y.; Jin, L.; Gao, C.; Wu, W.; Li, Y. RoboScape: Physics-informed Embodied World Model. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
42. Li, Y.; Wei, X.; Chi, X.; Li, Y.; Zhao, Z.; Wang, H.; Ma, N.; Lu, M.; Zhang, S. ManipDreamer: Boosting Robotic Manipulation World Model with Action Tree and Visual Guidance. *arXiv preprint arXiv:2504.16464* **2025**.
43. Li, S.; Yang, C.; Fang, J.; Yi, T.; Lu, J.; Cen, J.; Xie, L.; Shen, W.; Tian, Q. Worldgrow: Generating infinite 3d world. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 6433–6441.
44. 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**.
45. Zhang, Q.; Zhai, S.; Martin, M.A.B.; Miao, K.; Toshev, A.; Susskind, J.; Gu, J. World-consistent video diffusion with explicit 3d modeling. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 21685–21695.
46. Li, X.; Wang, T.; Gu, Z.; Zhang, S.; Guo, C.; Cao, L. FlashWorld: High-quality 3D Scene Generation within Seconds. In Proceedings of the The Fourteenth International Conference on Learning Representations, 2026.
47. Wu, H.; Wu, D.; He, T.; Guo, J.; Ye, Y.; Duan, Y.; Bian, J. Geometry Forcing: Marrying Video Diffusion and 3D Representation for Consistent World Modeling. In Proceedings of the The Fourteenth International Conference on Learning Representations, 2026.
48. Lu, Y.; Ren, X.; Yang, J.; Shen, T.; Wu, Z.; Gao, J.; Wang, Y.; Chen, S.; Chen, M.; Fidler, S.; et al. Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 27272–27283.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

49. Yang, Y.; Liu, J.; Zhang, Z.; Zhou, S.; Tan, R.; Yang, J.; Du, Y.; Gan, C. MindJourney: Test-Time Scaling with World Models for Spatial Reasoning. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
50. Liang, D.; Zhang, D.; Zhou, X.; Tu, S.; Feng, T.; Li, X.; Zhang, Y.; Du, M.; Tan, X.; Bai, X. Seeing the Future, Perceiving the Future: A unified driving world model for future generation and perception. *arXiv preprint arXiv:2503.13587* **2025**.
51. Lee, J.H.; Lin, B.J.; Sun, W.F.; Lee, C.Y. EDELINE: Enhancing Memory in Diffusion-based World Models via Linear-Time Sequence Modeling. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
52. Guo, Y.; Shi, L.X.; Chen, J.; Finn, C. Ctrl-World: A Controllable Generative World Model for Robot Manipulation. In Proceedings of the International Conference on Learning Representations (ICLR), 2026.
53. Wu, T.; Yang, S.; Po, R.; Xu, Y.; Liu, Z.; Lin, D.; Wetzstein, G. Video World Models with Long-term Spatial Memory. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
54. Xiao, Z.; LAN, Y.; Zhou, Y.; Ouyang, W.; Yang, S.; Zeng, Y.; Pan, X. WorldMem: Long-term Consistent World Simulation with Memory. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
55. Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; Anandkumar, A. Voyager: An Open-Ended Embodied Agent with Large Language Models. *Transactions on Machine Learning Research* **2024**.
56. Po, R.; Nitzan, Y.; Zhang, R.; Chen, B.; Dao, T.; Shechtman, E.; Wetzstein, G.; Huang, X. Long-context state-space video world models. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 8733–8744.
57. Huang, J.; Hu, X.; Han, B.; Shi, S.; Tian, Z.; He, T.; Jiang, L. Memory Forcing: Spatio-Temporal Memory for Consistent Scene Generation on Minecraft. *arXiv preprint arXiv:2510.03198* **2025**.
58. Collu, J.; Majellaro, R.; Plaat, A.; Moerland, T.M. Slot Structured World Models. *arXiv preprint arXiv:2402.03326* **2024**.
59. Traub, M.; Otte, S.; Menge, T.; Karlbauer, M.; Thuemmel, J.; Butz, M.V. Learning What and Where: Disentangling Location and Identity Tracking Without Supervision. In Proceedings of the The Eleventh International Conference on Learning Representations, 2023.
60. Elsayed, G.; Mahendran, A.; Van Steenkiste, S.; Greff, K.; Mozer, M.C.; Kipf, T. Savi++: Towards end-to-end object-centric learning from real-world videos. *Advances in Neural Information Processing Systems* **2022**, *35*, 28940–28954.
61. Zhang, Y.; Guo, X.; Xu, H.; Long, M. Consistent World Models via Foresight Diffusion. *arXiv preprint arXiv:2505.16474* **2025**.
62. Hu, W.; Wen, X.; Li, X.; Wang, G. DSG-World: Learning a 3D Gaussian World Model from Dual State Videos. *arXiv preprint arXiv:2506.05217* **2025**.
63. Huang, T.; Zheng, W.; Wang, T.; Liu, Y.; Wang, Z.; Wu, J.; Jiang, J.; Li, H.; Lau, R.; Zuo, W.; et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)* **2025**, *44*, 1–15.
64. Zhou, S.; Du, Y.; Yang, Y.; Han, L.; Chen, P.; Yeung, D.Y.; Gan, C. Learning 3D Persistent Embodied World Models. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
65. Wang, Z.; Wang, K.; Zhao, L.; Stone, P.; Bian, J. Dyn-O: Building Structured World Models with Object-Centric Representations. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
66. Ferraro, S.; Mazzaglia, P.; Verbelen, T.; Dhoedt, B. FOCUS: object-centric world models for robotic manipulation. *Frontiers in Neurorobotics* **2025**, *19*, 1585386.
67. Barcellona, L.; Zadaianchuk, A.; Allegro, D.; Papa, S.; Ghidoni, S.; Gavves, S. Dream to Manipulate: Compositional World Models Empowering Robot Imitation Learning with Imagination. In Proceedings of the Greeks in AI Symposium 2025, 2025.
68. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research Journal* **2024**.
69. Huang, Y.; Zhang, J.; Zou, S.; Liu, X.; Hu, R.; Xu, K. LaDi-WM: A Latent Diffusion-Based World Model for Predictive Manipulation. In Proceedings of the Proceedings of The 9th Conference on Robot Learning; Lim,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

J.; Song, S.; Park, H.W., Eds. PMLR, 27–30 Sep 2025, Vol. 305, *Proceedings of Machine Learning Research*, pp. 1726–1743.

70. Guo, J.; Ma, X.; Wang, Y.; Yang, M.; Liu, H.; Li, Q. Flowdreamer: A rgb-d world model with flow-based motion representations for robot manipulation. *IEEE Robotics and Automation Letters* **2026**, *11*, 2466–2473.
71. Bar, A.; Zhou, G.; Tran, D.; Darrell, T.; LeCun, Y. Navigation world models. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 15791–15801.
72. Peebles, W.; Xie, S. Scalable diffusion models with transformers. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 4195–4205.
73. Rigter, M.; Gupta, T.; Hilmkil, A.; Ma, C. AVID: Adapting Video Diffusion Models to World Models. In Proceedings of the Reinforcement Learning Conference, 2024.
74. Wu, J.; Yin, S.; Feng, N.; He, X.; Li, D.; Hao, J.; Long, M. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems* **2024**, *37*, 68082–68119.
75. He, H.; Zhang, Y.; Lin, L.; Xu, Z.; Pan, L. Pre-trained video generative models as world simulators. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 4645–4653.
76. Zhao, B.; Tang, R.; Jia, M.; Wang, Z.; Man, F.; Zhang, X.; Shang, Y.; Zhang, W.; Wu, W.; Gao, C.; et al. AirScape: An Aerial Generative World Model with Motion Controllability. In Proceedings of the Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 12519–12528.
77. Robotics, U. UnifoLM-WMA-0: A World-Model-Action (WMA) Framework under UnifoLM Family. https://github.com/unitreerobotics/unifolm-world-model-action, 2025. Open-source world-model–action architecture spanning multiple types of robotic embodiments.
78. Hayashi, K.; Koyama, M.; Guerreiro, J.J.A. Inter-environmental world modeling for continuous and compositional dynamics. *arXiv preprint arXiv:2503.09911* **2025**.
79. Durante, Z.; Gong, R.; Sarkar, B.; Wake, N.; Taori, R.; Tang, P.; Lakshmikanth, S.; Schulman, K.; Milstein, A.; Vo, H.; et al. An interactive agent foundation model. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 3652–3662.
80. Zhen, H.; Sun, Q.; Zhang, H.; Li, J.; Zhou, S.; Du, Y.; Gan, C. TesserAct: learning 4D embodied world models. *arXiv preprint arXiv:2504.20995* **2025**.
81. Duan, Y.; Zou, Z.; Gu, T.; Jia, W.; Zhao, Z.; Xu, L.; Liu, X.; Lin, Y.; Jiang, H.; Chen, K.; et al. LatticeWorld: A Multimodal Large Language Model-Empowered Framework for Interactive Complex World Generation. *arXiv preprint arXiv:2509.05263* **2025**.
82. 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**.
83. Li, J.; Tang, J.; Xu, Z.; Wu, L.; Zhou, Y.; Shao, S.; Yu, T.; Cao, Z.; Lu, Q. Hunyuan-GameCraft: High-dynamic Interactive Game Video Generation with Hybrid History Condition. *arXiv preprint arXiv:2506.17201* **2025**.
84. Lab, D. Mirage 2 — Generative World Engine. https://www.mirage2.org/, 2025. Browser-based system to generate explorable 3D worlds from images/text.
85. Labs, W. World Labs: spatial intelligence for large world models. https://www.worldlabs.ai/, 2025. Accessed: 2025-06-05.
86. Yang, Z.; Ge, W.; Li, Y.; Chen, J.; Li, H.; An, M.; Kang, F.; Xue, H.; Xu, B.; Yin, Y.; et al. Matrix-3d: Omnidirectional explorable 3d world generation. *arXiv preprint arXiv:2508.08086* **2025**.
87. Jang, J.; Ye, S.; Lin, Z.; Xiang, J.; Bjorck, J.; Fang, Y.; Hu, F.; Huang, S.; Kundalia, K.; Lin, Y.C.; et al. DreamGen: Unlocking Generalization in Robot Learning through Video World Models. In Proceedings of the 9th Annual Conference on Robot Learning, 2025.
88. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **2023**, *42*, 139–1.
89. Zheng, R.; Wang, J.; Reed, S.; Bjorck, J.; Fang, Y.; Hu, F.; Jang, J.; Kundalia, K.; Lin, Z.; Magne, L.; et al. FLARE: Robot Learning with Implicit World Modeling. In Proceedings of the Proceedings of The 9th Conference on Robot Learning; Lim, J.; Song, S.; Park, H.W., Eds. PMLR, 27–30 Sep 2025, Vol. 305, *Proceedings of Machine Learning Research*, pp. 3952–3971.
90. Qiu, Y.; Ziser, Y.; Korhonen, A.; Cohen, S.B.; Ponti, E.M. Bootstrapping World Models from Dynamics Models in Multimodal Foundation Models. *arXiv preprint arXiv:2506.06006* **2025**.
91. Kadian, A.; Truong, J.; Gokaslan, A.; Clegg, A.; Wijmans, E.; Lee, S.; Savva, M.; Chernova, S.; Batra, D. Sim2real predictivity: Does evaluation in simulation predict real-world performance? *IEEE Robotics and Automation Letters* **2020**, *5*, 6670–6677.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

92. Li, X.; Song, R.; Xie, Q.; Wu, Y.; Zeng, N.; Ai, Y. Simworld: A unified benchmark for simulator-conditioned scene generation via world model. In Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2025, pp. 927–934.
93. 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. *arXiv preprint arXiv:2501.03575* **2025**.
94. Mo, S.; Leng, Z.; Liu, L.; Wang, W.; He, H.; Zhou, B. Dreamland: Controllable World Creation with Simulator and Generative Models. *arXiv preprint arXiv:2506.08006* **2025**.
95. Feng, Y.; Tan, H.; Mao, X.; Liu, G.; Huang, S.; Xiang, C.; Su, H.; Zhu, J. Vidar: Embodied video diffusion model for generalist bimanual manipulation. *arXiv preprint arXiv:2507.12898* **2025**.
96. Wang, Y.; Yu, R.; Wan, S.; Gan, L.; Zhan, D.C. Founder: Grounding foundation models in world models for open-ended embodied decision making. In Proceedings of the Forty-second International Conference on Machine Learning, 2025.
97. Zhou, G.; Pan, H.; LeCun, Y.; Pinto, L. DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning. In Proceedings of the Forty-second International Conference on Machine Learning, 2025.
98. Wang, X.; Zhu, Z.; Huang, G.; Wang, B.; Chen, X.; Lu, J. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985* **2024**.
99. Schiewer, R.; Subramoney, A.; Wiskott, L. Exploring the limits of hierarchical world models in reinforcement learning. *Scientific Reports* **2024**, *14*, 26856.
100. Hao, C.; Lu, W.; Xu, Y.; Chen, Y. Neural Motion Simulator Pushing the Limit of World Models in Reinforcement Learning. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 27608–27617.
101. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* **2020**.
102. Mazzaglia, P.; Verbelen, T.; Dhoedt, B.; Courville, A.; Rajeswar, S. GenRL: Multimodal-foundation world models for generalization in embodied agents. *Advances in neural information processing systems* **2024**, *37*, 27529–27555.
103. Fang, F.; Liang, W.; Wu, Y.; Xu, Q.; Lim, J.H. Improving generalization of reinforcement learning using a bilinear policy network. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022, pp. 991–995.
104. Fang, Q.; Du, W.; Wang, H.; Zhang, J. Towards Unraveling and Improving Generalization in World Models. *arXiv preprint arXiv:2501.00195* **2024**.
105. Gao, S.; Zhou, S.; Du, Y.; Zhang, J.; Gan, C. AdaWorld: Learning Adaptable World Models with Latent Actions. In Proceedings of the Forty-second International Conference on Machine Learning, 2025.
106. Prasanna, S.; Farid, K.; Rajan, R.; Biedenkapp, A. Dreaming of Many Worlds: Learning Contextual World Models aids Zero-Shot Generalization. In Proceedings of the Seventeenth European Workshop on Reinforcement Learning, 2024.
107. Baldassarre, F.; Szafraniec, M.; Terver, B.; Khalidov, V.; Massa, F.; LeCun, Y.; Labatut, P.; Seitzer, M.; Bojanowski, P. Back to the features: Dino as a foundation for video world models. *arXiv preprint arXiv:2507.19468* **2025**.
108. Ali, M.Q.; Sridhar, A.; Matiana, S.; Wong, A.; Al-Sharman, M. Humanoid World Models: Open World Foundation Models for Humanoid Robotics. *arXiv preprint arXiv:2506.01182* **2025**.
109. Chi, X.; Ge, K.; Liu, J.; Zhou, S.; Jia, P.; He, Z.; Liu, Y.; Li, T.; Han, L.; Han, S.; et al. MinD: Unified Visual Imagination and Control via Hierarchical World Models. *arXiv preprint arXiv:2506.18897* **2025**.
110. Chun, J.; Jeong, Y.; Kim, T. Sparse Imagination for Efficient Visual World Model Planning. In Proceedings of the The Fourteenth International Conference on Learning Representations, 2026.
111. Cohen, L.; Wang, K.; Kang, B.; Gadot, U.; Mannor, S. Uncovering Untapped Potential in Sample-Efficient World Model Agents. *arXiv preprint arXiv:2502.11537* **2025**.
112. Burchi, M.; Timofte, R. Accurate and Efficient World Modeling with Masked Latent Transformers. In Proceedings of the Proceedings of the 42nd International Conference on Machine Learning; Singh, A.; Fazel, M.; Hsu, D.; Lacoste-Julien, S.; Berkenkamp, F.; Maharaj, T.; Wagstaff, K.; Zhu, J., Eds. PMLR, 13–19 Jul 2025, Vol. 267, *Proceedings of Machine Learning Research*, pp. 5894–5912.
113. Zhang, H.; Yan, X.; Xue, Y.; Guo, Z.; Cui, S.; Li, Z.; Liu, B. D^2-world: An Efficient World Model through Decoupled Dynamic Flow. *arXiv preprint arXiv:2411.17027* **2024**.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

114. 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**.
115. Li, S.; Hao, Q.; Shang, Y.; Li, Y. KeyWorld: Key Frame Reasoning Enables Effective and Efficient World Models. *arXiv preprint arXiv:2509.21027* **2025**.
116. Yamada, J.; Rigter, M.; Collins, J.; Posner, I. Twist: Teacher-student world model distillation for efficient sim-to-real transfer. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 9190–9196.
117. Micheli, V.; Alonso, E.; Fleuret, F. Transformers are Sample-Efficient World Models. In Proceedings of the The Eleventh International Conference on Learning Representations, 2023.
118. Micheli, V.; Alonso, E.; Fleuret, F. Efficient World Models with Context-Aware Tokenization. In Proceedings of the International Conference on Machine Learning. PMLR, 2024, pp. 35623–35638.
119. Song, Q.; Wang, X.; Zhou, D.; Lin, J.; Chen, C.; Ma, Y.; Li, X. Hero: Hierarchical extrapolation and refresh for efficient world models. *arXiv preprint arXiv:2508.17588* **2025**.
120. Jin, B.; Li, W.; Yang, B.; Zhu, Z.; Jiang, J.; Gao, H.a.; Sun, H.; Zhan, K.; Hu, H.; Zhang, X.; et al. PosePilot: Steering camera pose for generative world models with self-supervised depth. In Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2025, pp. 8051–8058.
121. Jeong, Y.; Chun, J.; Cha, S.; Kim, T. Object-Centric World Model for Language-Guided Manipulation. In Proceedings of the ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling, 2025.
122. Akbulut, T.; Merlin, M.; Parr, S.; Quartey, B.; Thompson, S. Sample Efficient Robot Learning with Structured World Models. *arXiv preprint arXiv:2210.12278* **2022**.
123. Van Den Oord, A.; Vinyals, O.; et al. Neural discrete representation learning. *Advances in neural information processing systems* **2017**, *30*.
124. Chen, R.; Ko, Y.; Zhang, Z.; Cho, C.; Chung, S.; Giuffré, M.; Shung, D.L.; Stadie, B.C. LAMP: Extracting Locally Linear Decision Surfaces from LLM World Models. *arXiv preprint arXiv:2505.11772* **2025**.
125. Zeng, B.; Zhu, K.; Hua, D.; Li, B.; Tong, C.; Wang, Y.; Huang, X.; Dai, Y.; Zhang, Z.; Yang, Y.; et al. Research on World Models Is Not Merely Injecting World Knowledge into Specific Tasks. *arXiv preprint arXiv:2602.01630* **2026**.
126. 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**.
127. 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 Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 7346–7355.
128. 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**.
129. Yang, Z.; Guo, X.; Ding, C.; Wang, C.; Wu, W. Physical informed driving world model. *arXiv preprint arXiv:2412.08410* **2024**.
130. Jiang, H.; Hsu, H.Y.; Zhang, K.; Yu, H.N.; Wang, S.; Li, Y. Phystwin: Physics-informed reconstruction and simulation of deformable objects from videos. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 7219–7230.
131. Li, J.; Wan, H.; Lin, N.; Zhan, Y.L.; Chengze, R.; Wang, H.; Zhang, Y.; Liu, H.; Wang, Z.; Yu, F.; et al. SlotPi: Physics-informed Object-centric Reasoning Models. In Proceedings of the Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, 2025, pp. 1376–1387.
132. Petri, F.; Asprino, L.; Gangemi, A. Learning Local Causal World Models with State Space Models and Attention. *arXiv preprint arXiv:2505.02074* **2025**.
133. Yan, Z.; Dong, W.; Shao, Y.; Lu, Y.; Liu, H.; Liu, J.; Wang, H.; Wang, Z.; Wang, Y.; Remondino, F.; et al. Renderworld: World model with self-supervised 3d label. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2025, pp. 6063–6070.
134. Zhou, X.; Liang, D.; Tu, S.; Chen, X.; Ding, Y.; Zhang, D.; Tan, F.; Zhao, H.; Bai, X. Hermes: A unified self-driving world model for simultaneous 3d scene understanding and generation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 27817–27827.
135. Wu, J.; Ma, H.; Deng, C.; Long, M. Pre-training contextualized world models with in-the-wild videos for reinforcement learning. *Advances in Neural Information Processing Systems* **2023**, *36*, 39719–39743.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

136. Wang, Q.; Zhang, Z.; Xie, B.; Jin, X.; Wang, Y.; Wang, S.; Zheng, L.; Yang, X.; Zeng, W. Disentangled world models: Learning to transfer semantic knowledge from distracting videos for reinforcement learning. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 2599–2608.
137. Zhang, W.; Jelley, A.; McInroe, T.; Storkey, A. Objects matter: object-centric world models improve reinforcement learning in visually complex environments. *arXiv preprint arXiv:2501.16443* **2025**.
138. Wang, Y.; Wan, S.; Gan, L.; Feng, S.; Zhan, D.C. AD3: implicit action is the key for world models to distinguish the diverse visual distractors. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning, 2024, pp. 51546–51568.
139. Wang, X.; Wu, Z.; Peng, P. LongDWM: Cross-Granularity Distillation for Building a Long-Term Driving World Model. *arXiv preprint arXiv:2506.01546* **2025**.
140. Jiang, J.; Janghorbani, S.; De Melo, G.; Ahn, S. SCALOR: Generative World Models with Scalable Object Representations. In Proceedings of the International Conference on Learning Representations, 2020.
141. Zhu, H.; Wang, Y.; Zhou, J.; Chang, W.; Zhou, Y.; Li, Z.; Chen, J.; Shen, C.; Pang, J.; He, T. Aether: Geometric-aware unified world modeling. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 8535–8546.
142. Lu, G.; Jia, B.; Li, P.; Chen, Y.; Wang, Z.; Tang, Y.; Huang, S. Gwm: Towards scalable gaussian world models for robotic manipulation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 9263–9274.
143. Mao, Z.; Ruchkin, I. Towards Physically Interpretable World Models: Meaningful Weakly Supervised Representations for Visual Trajectory Prediction. *arXiv preprint arXiv:2412.12870* **2024**.
144. Ross, E.; Drygala, C.; Schwarz, L.; Kaiser, S.; di Mare, F.; Breiten, T.; Gottschalk, H. When do World Models Successfully Learn Dynamical Systems? *arXiv preprint arXiv:2507.04898* **2025**.
145. Zhou, S.; Zhou, T.; Yang, Y.; Long, G.; Ye, D.; Jiang, J.; Zhang, C. WALL-E: World Alignment by NeuroSymbolic Learning improves World Model-based LLM Agents. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
146. Liu, X.; Tang, H. FOLIAGE: Towards Physical Intelligence World Models Via Unbounded Surface Evolution. *arXiv preprint arXiv:2506.03173* **2025**.
147. Wang, D.; Sun, Z.; Li, Z.; Wang, C.; Peng, Y.; Ye, H.; Zarrouki, B.; Li, W.; Piccinini, M.; Xie, L.; et al. Enhancing Physical Consistency in Lightweight World Models. *arXiv preprint arXiv:2509.12437* **2025**.
148. Chen, D.; Moutakanni, T.; Chung, W.; Bang, Y.; Ji, Z.; Bolourchi, A.; Fung, P. Planning with reasoning using vision language world model. *arXiv preprint arXiv:2509.02722* **2025**.
149. Huh, M.; Cheung, B.; Wang, T.; Isola, P. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987* **2024**.
150. Zhao, Y.; Scannell, A.; Hou, Y.; Cui, T.; Chen, L.; Büchler, D.; Solin, A.; Kannala, J.; Pajarinen, J. Generalist World Model Pre-Training for Efficient Reinforcement Learning. In Proceedings of the ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling, 2025.
151. Technologies, X. 1X World Model. https://www.1x.tech/discover/1x-world-model, 2024. Accessed: 2025-11-14.
152. Chen, C.; Wu, Y.F.; Yoon, J.; Ahn, S. Transdreamer: Reinforcement learning with transformer world models. *arXiv preprint arXiv:2202.09481* **2022**.
153. Wu, H.; Guo, M.; Li, Z.; Dou, Z.; Long, M.; He, K.; Matusik, W. GeoPT: Scaling Physics Simulation via Lifted Geometric Pre-Training. *arXiv preprint arXiv:2602.20399* **2026**.
154. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In Proceedings of the First conference on language modeling, 2024.
155. Ye, S.; Ge, Y.; Zheng, K.; Gao, S.; Yu, S.; Kurian, G.; Indupuru, S.; Tan, Y.L.; Zhu, C.; Xiang, J.; et al. World Action Models are Zero-shot Policies. *arXiv preprint arXiv:2602.15922* **2026**.
156. Gao, S.; Liang, W.; Zheng, K.; Malik, A.; Ye, S.; Yu, S.; Tseng, W.C.; Dong, Y.; Mo, K.; Lin, C.H.; et al. DreamDojo: A Generalist Robot World Model from Large-Scale Human Videos. *arXiv preprint arXiv:2602.06949* **2026**.
157. Kotar, K.; Lee, W.; Venkatesh, R.; Chen, H.; Bear, D.; Watrous, J.; Kim, S.; Aw, K.L.; Chen, L.N.; Stojanov, S.; et al. World modeling with probabilistic structure integration. *arXiv preprint arXiv:2509.09737* **2025**.
158. Saxena, D.; Cao, J. Generative adversarial networks (GANs) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)* **2021**, *54*, 1–42.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

159. Vuong, Q.; Levine, S.; Walke, H.R.; 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 Proceedings of the Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023, 2023.
160. Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Hsu, J.; et al. RT-1: Robotics Transformer for Real-World Control at Scale. *Robotics: Science and Systems XIX* **2023**.
161. Liu, B.; Zhu, Y.; Gao, C.; Feng, Y.; Liu, Q.; Zhu, Y.; Stone, P. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems* **2023**, *36*, 44776–44791.
162. Khazatsky, A.; Pertsch, K.; Nair, S.; Balakrishna, A.; Dasari, S.; Karamcheti, S.; Nasiriany, S.; Srirama, M.K.; Chen, L.Y.; Ellis, K.; et al. DROID: A large-scale in-the-wild robot manipulation dataset. In Proceedings of the Robotics: Science and Systems, 2024.
163. Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; Song, S. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. *The International Journal of Robotics Research* **2024**.
164. Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540* **2016**.
165. James, S.; Ma, Z.; Arrojo, D.R.; Davison, A.J. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters* **2020**, *5*, 3019–3026.
166. Rohmer, E.; Singh, S.P.; Freese, M. V-REP: A versatile and scalable robot simulation framework. *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems* **2013**, pp. 1321–1326.
167. Walke, H.; Black, K.; Lee, A.; Kim, M.J.; Du, M.; Zheng, C.; Zhao, T.; Hansen-Estruch, P.; Vuong, Q.; He, A.; et al. BridgeData V2: A Dataset for Robot Learning at Scale. In Proceedings of the Conference on Robot Learning (CoRL), 2023.
168. Tian, S.; Finn, C.; Wu, J. A Control-Centric Benchmark for Video Prediction. In Proceedings of the The Eleventh International Conference on Learning Representations, 2024.
169. Nasiriany, S.; Maddukuri, A.; Zhang, L.; Parikh, A.; Lo, A.; Joshi, A.; Mandlekar, A.; Zhu, Y. RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots. In Proceedings of the Robotics: Science and Systems, 2024.
170. Bao, C.; Xu, H.; Qin, Y.; Wang, X. DexArt: Benchmarking Generalizable Dexterous Manipulation with Articulated Objects. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21190–21200.
171. Mitash, C.; Wang, F.; Lu, S.; Terhuja, V.; Garaas, T.; Polido, F.; Nambi, M. ARMBench: An Object-centric Benchmark Dataset for Robotic Manipulation. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023, pp. 9132–9139.
172. Tunyasuvunakool, S.; Muldal, A.; Doron, Y.; Liu, S.; Bohez, S.; Merel, J.; Erez, T.; Lillicrap, T.; Heess, N.; Tassa, Y. dm_control: Software and tasks for continuous control. *Software Impacts* **2020**, *6*, 100022.
173. Gupta, A.; Kumar, V.; Lynch, C.; Levine, S.; Hausman, K. Relay Policy Learning: Solving Long-Horizon Tasks via Imitation and Reinforcement Learning. In Proceedings of the Conference on Robot Learning. PMLR, 2020, pp. 1025–1037.
174. McLean, R.; Chatzaroulas, E.; McCutcheon, L.; Röder, F.; Yu, T.; He, Z.; Zentner, K.; Julian, R.; Terry, J.K.; Woungang, I.; et al. Meta-World+: An Improved, Standardized, RL Benchmark. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2025.
175. Wang, X.; Lian, L.; Yu, S.X. Unsupervised visual attention and invariance for reinforcement learning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 6677–6687.
176. Henderson, P.; Chang, W.D.; Shkurti, F.; Hansen, J.; Meger, D.; Dudek, G. Benchmark environments for multitask learning in continuous domains. *arXiv preprint arXiv:1708.04352* **2017**.
177. Gu, J.; Xiang, F.; Li, X.; Ling, Z.; Liu, X.; Mu, T.; Tang, Y.; Tao, S.; Wei, X.; Yao, Y.; et al. ManiSkill2: A Unified Benchmark for Generalizable Manipulation Skills. In Proceedings of the International Conference on Learning Representations, 2023.
178. Vittorio, C.; Huawei, W.; Guillaume, D.; Massimo, S.; Vikash, K. MyoSuite – A contact-rich simulation suite for musculoskeletal motor control. https://github.com/myohub/myosuite, 2022.
179. contributors, A.W.C. AgiBot World Colosseum. https://github.com/OpenDriveLab/AgiBot-World, 2024.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

180. Mandlekar, A.; Xu, D.; Wong, J.; Nasiriany, S.; Wang, C.; Kulkarni, R.; Fei-Fei, L.; Savarese, S.; Zhu, Y.; Martín-Martín, R. What Matters in Learning from Offline Human Demonstrations for Robot Manipulation. In Proceedings of the Conference on Robot Learning (CoRL), 2021.
181. Dasari, S.; Ebert, F.; Tian, S.; Nair, S.; Bucher, B.; Schmeckpeper, K.; Singh, S.; Levine, S.; Finn, C. RoboNet: Large-Scale Multi-Robot Learning. In Proceedings of the CoRL 2019: Volume 100 Proceedings of Machine Learning Research, 2019, [arXiv:cs.RO/1910.11215].
182. Fang, H.S.; Fang, H.; Tang, Z.; Liu, J.; Wang, J.; Zhu, H.; Lu, C. RH20T: A Robotic Dataset for Learning Diverse Skills in One-Shot. In Proceedings of the RSS 2023 Workshop on Learning for Task and Motion Planning, 2023.
183. Heo, M.; Lee, Y.; Lee, D.; Lim, J.J. FurnitureBench: Reproducible Real-World Benchmark for Long-Horizon Complex Manipulation. In Proceedings of the Robotics: Science and Systems, 2023.
184. Geng, H.; Wang, F.; Wei, S.; Li, Y.; Wang, B.; An, B.; Cheng, C.T.; Lou, H.; Li, P.; Wang, Y.J.; et al. RoboVerse: Towards a Unified Platform, Dataset and Benchmark for Scalable and Generalizable Robot Learning, 2025, [arXiv:cs.RO/2504.18904].
185. Yan, F.; Liu, F.; Huang, Y.; Guan, Z.; Zheng, L.; Zhong, Y.; Feng, C.; Ma, L. RoboTron-Mani: All-in-One Multimodal Large Model for Robotic Manipulation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2025, pp. 13707–13718.
186. Bharadhwaj, H.; Vakil, J.; Sharma, M.; Gupta, A.; Tulsiani, S.; Kumar, V. RoboAgent: Generalization and Efficiency in Robot Manipulation via Semantic Augmentations and Action Chunking, 2023, [arXiv:cs.RO/2309.01918].
187. Yang, R.; Chen, H.; Zhang, J.; Zhao, M.; Qian, C.; Wang, K.; Wang, Q.; Koripella, T.V.; Movahedi, M.; Li, M.; et al. EmbodiedBench: Comprehensive Benchmarking Multi-modal Large Language Models for Vision-Driven Embodied Agents. In Proceedings of the Forty-second International Conference on Machine Learning, 2025.
188. Li, C.; Zhang, R.; Wong, J.; Gokmen, C.; Srivastava, S.; Martín-Martín, R.; Wang, C.; Levine, G.; Lingelbach, M.; Sun, J.; et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In Proceedings of the Conference on Robot Learning. PMLR, 2023, pp. 80–93.
189. Fu, J.; Kumar, A.; Nachum, O.; Tucker, G.; Levine, S. D4RL: Datasets for Deep Data-Driven Reinforcement Learning, 2020, [arXiv:cs.LG/2004.07219].
190. 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 Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9339–9347.
191. Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niebner, M.; Savva, M.; Song, S.; Zeng, A.; Zhang, Y. Matterport3D: Learning from RGB-D Data in Indoor Environments. In Proceedings of the 2017 International Conference on 3D Vision (3DV). IEEE Computer Society, 2017, pp. 667–676.
192. Yadav, K.; Ramrakhya, R.; Ramakrishnan, S.K.; Gervet, T.; Turner, J.; Gokaslan, A.; Maestre, N.; Chang, A.X.; Batra, D.; Savva, M.; et al. Habitat-matterport 3d semantics dataset. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4927–4936.
193. Xia, F.; Zamir, A.R.; He, Z.; Sax, A.; Malik, J.; Savarese, S. Gibson env: Real-world perception for embodied agents. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9068–9079.
194. Martin-Martin, R.; Patel, M.; Rezatofighi, H.; Shenoi, A.; Gwak, J.; Frankel, E.; Sadeghian, A.; Savarese, S. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence* **2021**, *45*, 6748–6765.
195. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
196. Shang, Y.; Li, Z.; Ma, Y.; Su, W.; Jin, X.; Wang, Z.; Jin, L.; Zhang, X.; Tang, Y.; Su, H.; et al. WorldArena: A Unified Benchmark for Evaluating Perception and Functional Utility of Embodied World Models. *arXiv preprint arXiv:2602.08971* **2026**.
197. Chen, T.; Chen, Z.; Chen, B.; Cai, Z.; Liu, Y.; Li, Z.; Liang, Q.; Lin, X.; Ge, Y.; Gu, Z.; et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088* **2025**.
198. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2446–2454.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

199. Yang, J.; Gao, S.; Qiu, Y.; Chen, L.; Li, T.; Dai, B.; Chitta, K.; Wu, P.; Zeng, J.; Luo, P.; et al. Generalized Predictive Model for Autonomous Driving. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.
200. Zhou, T.; Tucker, R.; Flynn, J.; Fyffe, G.; Snavely, N. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)* **2018**, *37*, 1–12.
201. Wang, Y.; Cheng, K.; He, J.; Wang, Q.; Dai, H.; Chen, Y.; Xia, F.; Zhang, Z. Drivingdojo dataset: Advancing interactive and knowledge-enriched driving world model. *Advances in Neural Information Processing Systems* **2024**, *37*, 13020–13034.
202. Min, C.; Zhao, D.; Xiao, L.; Zhao, J.; Xu, X.; Zhu, Z.; Jin, L.; Li, J.; Guo, Y.; Xing, J.; et al. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 15522–15533.
203. Beattie, C.; Leibo, J.Z.; Teplyashin, D.; Ward, T.; Wainwright, M.; Küttler, H.; Lefrancq, A.; Green, S.; Valdés, V.; Sadik, A.; et al. Deepmind lab. *arXiv preprint arXiv:1612.03801* **2016**.
204. Cobbe, K.; Hesse, C.; Hilton, J.; Schulman, J. Leveraging Procedural Generation to Benchmark Reinforcement Learning. *arXiv preprint arXiv:1912.01588* **2019**.
205. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* **2013**.
206. Osband, I.; Doron, Y.; Hessel, M.; Aslanides, J.; Sezener, E.; Saraiva, A.; McKinney, K.; Lattimore, T.; Szepesvári, C.; Singh, S.; et al. Behaviour Suite for Reinforcement Learning. In Proceedings of the International Conference on Learning Representations, 2020.
207. Paperno, D.; Kruszewski, G.; Lazaridou, A.; Pham, N.Q.; Bernardi, R.; Pezzelle, S.; Baroni, M.; Boleda, G.; Fernández, R. The LAMBADA dataset: Word prediction requiring a broad discourse context. In Proceedings of the Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers), 2016, pp. 1525–1534.
208. Zhang, J.; Jiang, M.; Dai, N.; Lu, T.; Uzunoglu, A.; Zhang, S.; Wei, Y.; Wang, J.; Patel, V.M.; Liang, P.P.; et al. World-in-World: World Models in a Closed-Loop World. *arXiv preprint arXiv:2510.18135* **2025**.
209. Bordes, F.; Garrido, Q.; Kao, J.T.; Williams, A.; Rabbat, M.; Dupoux, E. IntPhys 2: Benchmarking Intuitive Physics Understanding In Complex Synthetic Environments. *arXiv preprint arXiv:2506.09849* **2025**.
210. Weihs, L.; Yuile, A.R.; Baillargeon, R.; Fisher, C.; Marcus, G.; Mottaghi, R.; Kembhavi, A. Benchmarking Progress to Infant-Level Physical Reasoning in AI. *TMLR* **2022**.
211. Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The" something something" video database for learning and evaluating visual common sense. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 5842–5850.
212. Damen, D.; Doughty, H.; Farinella, G.M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *43*, 4125–4141.
213. NVIDIA PhysX SDK. https://developer.nvidia.com/physx-sdk, 2025. Accessed: 2025-11-15.
214. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Koltun, V. CARLA: An open urban driving simulator. In Proceedings of the Conference on robot learning. PMLR, 2017, pp. 1–16.
215. Greff, K.; Belletti, F.; Beyer, L.; Doersch, C.; Du, Y.; Duckworth, D.; Fleet, D.J.; Gnanapragasam, D.; Golemo, F.; Herrmann, C.; et al. Kubric: a scalable dataset generator **2022**.
216. PyBullet Physics Simulation. https://pybullet.org, 2025. Accessed: 2025-11-15.
217. Blender – a 3D modelling and rendering package. https://www.blender.org, 2025. Accessed: 2025-11-15.
218. Zhou, H.; Ma, Y.; Wu, H.; Wang, H.; Long, M. Unisolver: PDE-Conditional Transformers Towards Universal Neural PDE Solvers. In Proceedings of the Forty-second International Conference on Machine Learning, 2025.
219. Pătrăucean, V.; Smaira, L.; Gupta, A.; Continente, A.R.; Markeeva, L.; Banarse, D.; Koppula, S.; Heyward, J.; Malinowski, M.; Yang, Y.; et al. Perception Test: A Diagnostic Benchmark for Multimodal Video Models. In Proceedings of the Advances in Neural Information Processing Systems, 2023.
220. Bear, D.; Wang, E.; Mrowca, D.; Binder, F.J.; Tung, H.Y.; Pramod, R.; Holdaway, C.; Tao, S.; Smith, K.A.; Sun, F.Y.; et al. Physion: Evaluating Physical Prediction from Vision in Humans and Machines. In Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

221. Qiu, S.; Guo, S.; Song, Z.Y.; Sun, Y.; Cai, Z.; Wei, J.; Luo, T.; Yin, Y.; Zhang, H.; Hu, Y.; et al. PHYBench: Holistic Evaluation of Physical Perception and Reasoning in Large Language Models, 2025, [arXiv:cs.CL/2504.16074].
222. Chow, W.; Mao, J.; Li, B.; Seita, D.; Guizilini, V.C.; Wang, Y. PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
223. Gamerdinger, J.; Teufel, S.; Schulz, P.; Amann, S.; Kirchner, J.P.; Bringmann, O. Scope: A synthetic multi-modal dataset for collective perception including physical-correct weather conditions. In Proceedings of the 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2024, pp. 2622–2628.
224. Xing, E.; Deng, M.; Hou, J.; Hu, Z. Critiques of world models. *arXiv preprint arXiv:2507.05169* **2025**.
225. Miao, J.; Wei, Y.; Wu, Y.; Liang, C.; Li, G.; Yang, Y. VSPW: A Large-scale Dataset for Video Scene Parsing in the Wild. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021.
226. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213–3223.
227. Liu, H.; He, J.; Jin, Y.; Zheng, D.; Dong, Y.; Zhang, F.; Huang, Z.; He, Y.; Li, Y.; Chen, W.; et al. ShotBench: Expert-Level Cinematic Understanding in Vision-Language Models, 2025, [arXiv:cs.CV/2506.21356].
228. Xue, W.; Qian, C.; Wu, J.; Zhou, Y.; Liu, W.; Ren, J.; Fan, S.; Zhang, Y. ShotVL: Human-Centric Highlight Frame Retrieval via Language Queries. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 9050–9058.
229. Zhao, E.; Raval, V.; Zhang, H.; Mao, J.; Shangguan, Z.; Nikolaidis, S.; Wang, Y.; Seita, D. ManipBench: Benchmarking Vision-Language Models for Low-Level Robot Manipulation. In Proceedings of the Proceedings of The 9th Conference on Robot Learning; Lim, J.; Song, S.; Park, H.W., Eds. PMLR, 27–30 Sep 2025, Vol. 305, *Proceedings of Machine Learning Research*, pp. 3413–3462.
230. Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; van den Hengel, A. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
231. Ray, A.; Duan, J.; Brown, E.; Tan, R.; Bashkirova, D.; Hendrix, R.; Ehsani, K.; Kembhavi, A.; Plummer, B.A.; Krishna, R.; et al. SAT: Dynamic Spatial Aptitude Training for Multimodal Language Models, 2025, [arXiv:cs.CV/2412.07755].
232. Kamath, A.; Hessel, J.; Chang, K.W. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 9161–9175.
233. Zhao, Z.; Fu, H.; Liang, D.; Zhou, X.; Zhang, D.; Xie, H.; Wang, B.; Bai, X. Extending Large Vision-Language Model for Diverse Interactive Tasks in Autonomous Driving. *arXiv preprint arXiv:2505.08725* **2025**.
234. Wang, S.; Yu, Z.; Jiang, X.; Lan, S.; Shi, M.; Chang, N.; Kautz, J.; Li, Y.; Alvarez, J.M. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In Proceedings of the Proceedings of the computer vision and pattern recognition conference, 2025, pp. 22442–22452.
235. 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, 2025, [arXiv:cs.CV/2509.12201].
236. Chen, D.; Chung, W.; Bang, Y.; Ji, Z.; Fung, P. WorldPrediction: A Benchmark for High-level World Modeling and Long-horizon Procedural Planning. In Proceedings of the ICML 2025 Workshop on Assessing World Models, 2025.
237. Gu, J.; Liu, X.; Zeng, Y.; Nagarajan, A.; Zhu, F.; Hong, D.; Fan, Y.; Yan, Q.; Zhou, K.; Liu, M.Y.; et al. " PhyWorldBench": A Comprehensive Evaluation of Physical Realism in Text-to-Video Models. *arXiv preprint arXiv:2507.13428* **2025**.
238. Li, D.; Fang, Y.; Chen, Y.; Yang, S.; Cao, S.; Wong, J.; Luo, M.; Wang, X.; Yin, H.; Gonzalez, J.E.; et al. WorldModelBench: Judging Video Generation Models As World Models. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2025.
239. Li, Z.; Li, C.; Mao, X.; Lin, S.; Li, M.; Zhao, S.; Li, X.; Feng, Y.; Sun, J.; Li, Z.; et al. Sekai: A Video Dataset towards World Exploration. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2025.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

240. Azzolini, A.; Bai, J.; Brandon, H.; Cao, J.; Chattopadhyay, P.; Chen, H.; Chu, J.; Cui, Y.; Diamond, J.; Ding, Y.; et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558* **2025**.
241. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **2004**, *13*, 600–612.
242. Nacken, P.F. Chamfer metrics, the medial axis and mathematical morphology. *Journal of Mathematical Imaging and Vision* **1996**, *6*, 235–248.
243. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **2017**, *30*.
244. Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; Gelly, S. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* **2018**.
245. Stein, G.; Cresswell, J.; Hosseinzadeh, R.; Sui, Y.; Ross, B.; Villecroze, V.; Liu, Z.; Caterini, A.L.; Taylor, E.; Loaiza-Ganem, G. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems* **2023**, *36*, 3732–3784.
246. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.
247. Fu, S.; Tamir, N.Y.; Sundaram, S.; Chai, L.; Zhang, R.; Dekel, T.; Isola, P. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems, 2023.
248. Liu, J.; Qu, Y.; Yan, Q.; Zeng, X.; Wang, L.; Liao, R. Fréchet Video Motion Distance: A Metric for Evaluating Motion Consistency in Videos. In Proceedings of the First Workshop on Controllable Video Generation@ ICML24, 2024.
249. Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. In Proceedings of the Proceedings of the 2021 conference on empirical methods in natural language processing, 2021, pp. 7514–7528.
250. Hentschel, S.; Kobs, K.; Hotho, A. CLIP knows image aesthetics. *Frontiers in Artificial Intelligence* **2022**, *5*, 976235.
251. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9650–9660.
252. Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. Vbench: Comprehensive benchmark suite for video generative models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 21807–21818.
253. Duan, H.; Yu, H.X.; Chen, S.; Fei-Fei, L.; Wu, J. Worldscore: A unified evaluation benchmark for world generation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 27713–27724.
254. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
255. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.
256. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text summarization branches out, 2004, pp. 74–81.
257. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.
258. Stolcke, A.; Yoshioka, T. DOVER: A method for combining diarization outputs. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019, pp. 757–763.
259. Sutton, R.S.; Barto, A.G.; et al. *Reinforcement learning: An introduction*; Vol. 1, MIT press Cambridge, 1998.
260. Kendall, M.G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93.
261. Elo, A.E. *The Rating of Chessplayers, Past and Present*; Arco Publishing: New York, 1978.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

262. Badia, A.P.; Piot, B.; Kapturowski, S.; Sprechmann, P.; Vitvitskyi, A.; Guo, Z.D.; Blundell, C. Agent57: Outperforming the atari human benchmark. In Proceedings of the International conference on machine learning. PMLR, 2020, pp. 507–517.
263. Christen, P.; Hand, D.J.; Kirielle, N. A review of the F-measure: its history, properties, criticism, and alternatives. *ACM Computing Surveys* **2023**, *56*, 1–24.
264. Gretton, A.; Bousquet, O.; Smola, A.; Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In Proceedings of the International conference on algorithmic learning theory. Springer, 2005, pp. 63–77.
265. Botchkarev, A. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv preprint arXiv:1809.03006* **2018**.
266. Stuart Jr, H.W. Value gaps and profitability. *Strategy Science* **2016**, *1*, 56–70.
267. Shi, Z.; Liu, M.; Zhang, S.; Zheng, R.; Dong, S.; Wei, P. GAWM: Global-Aware World Model for Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2501.10116* **2025**.
268. Lambert, N.; Pister, K.; Calandra, R. Investigating compounding prediction errors in learned dynamics models. *arXiv preprint arXiv:2203.09637* **2022**.
269. Xu, Y.; Parker-Holder, J.; Pacchiano, A.; Ball, P.; Rybkin, O.; Roberts, S.; Rocktäschel, T.; Grefenstette, E. Learning general world models in a handful of reward-free deployments. *Advances in Neural Information Processing Systems* **2022**, *35*, 26820–26838.
270. Qin, C.; Klabjan, D.; Russo, D. Improving the expected improvement algorithm. *Advances in Neural Information Processing Systems* **2017**, *30*.
271. Prakash, A.; Tu, R.; Chang, M.; Gupta, S. 3d hand pose estimation in everyday egocentric images. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 183–202.
272. Bento, J.; Zhu, J.J. A metric for sets of trajectories that is practical and mathematically consistent. *arXiv preprint arXiv:1601.03094* **2016**.
273. Sturm, J.; Burgard, W.; Cremers, D. Evaluating egomotion and structure-from-motion approaches using the TUM RGB-D benchmark. In Proceedings of the Proc. of the Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RJS International Conference on Intelligent Robot Systems (IROS), 2012, Vol. 13, p. 6.
274. Mohamed, A.; Zhu, D.; Vu, W.; Elhoseiny, M.; Claudel, C. Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 463–479.
275. Perille, D.; Truong, A.; Xiao, X.; Stone, P. Benchmarking metric ground navigation. In Proceedings of the 2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR). IEEE, 2020, pp. 116–121.
276. Georgiou, T.T.; Smith, M.C. Optimal robustness in the gap metric. In Proceedings of the Proceedings of the 28th IEEE Conference on Decision and Control,. IEEE, 1989, pp. 2331–2336.
277. Ward, J.R.; Agamennoni, G.; Worrall, S.; Bender, A.; Nebot, E. Extending time to collision for probabilistic reasoning in general traffic scenarios. *Transportation Research Part C: Emerging Technologies* **2015**, *51*, 66–82.
278. Senin, P. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA* **2008**, *855*, 40.
279. Zhang, Y.; Mehta, S.; Caspi, A. Rethinking semantic segmentation evaluation for explainability and model selection. *arXiv preprint arXiv:2101.08418* **2021**.
280. Steinley, D.; Brusco, M.J.; Hubert, L. The variance of the adjusted Rand index. *Psychological methods* **2016**, *21*, 261.
281. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Spice: Semantic propositional image caption evaluation. In Proceedings of the European conference on computer vision. Springer, 2016, pp. 382–398.
282. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems* **2014**, *27*.
283. Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; Gelly, S. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* **2018**.
284. Tsamardinos, I.; Brown, L.E.; Aliferis, C.F. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning* **2006**, *65*, 31–78.
285. Belongie, S.; Malik, J.; Puzicha, J. Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence* **2002**, *24*, 509–522.
286. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* **2013**, *26*.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2026 doi:10.20944/preprints202604.0503.v1

287. Wang, M.; Jin, W.; Cao, K.; Xie, L.; Hong, Y. ContactGaussian-WM: Learning Physics-Grounded World Model from Videos. *arXiv preprint arXiv:2602.11021* **2026**.
288. Makoviychuk, V.; Wawrzyniak, L.; Guo, Y.; Lu, M.; Storey, K.; Macklin, M.; Hoeller, D.; Rudin, N.; Allshire, A.; Handa, A. Isaac Gym: High Performance GPU Based Physics Simulation For Robot Learning. *arXiv preprint arXiv:2110.13563* **2021**. NeurIPS 2021 Track: Datasets and Benchmarks.
289. Bengio, Y.; Clare, S.; Prunkl, C.; Andriushchenko, M.; Bucknall, B.; Murray, M.; Bommasani, R.; Casper, S.; Davidson, T.; Douglas, R.; et al. International ai safety report 2026. *arXiv preprint arXiv:2602.21012* **2026**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.