Preprint · October 30, 2026

Simulate, Then Act: World Models, Evaluation Gaps and Enterprise Agents

Qiang Wang[1], Great Unicorn Intelligence Research Team[2*]

Abstract

A world model predicts how an environment evolves given its state and a candidate action. Drawing on recent work, this review proposes a taxonomy by state representation and primary use; compares pixel-reconstruction, latent-prediction and autoregressive-token objectives on fidelity, plannability and compute; organizes evaluation from perceptual quality to closed-loop decision value, showing that coverage thins near decision-making and that few benchmarks contrast world models with direct policies; and examines how to couple world models with LLM agents to "simulate, then act". We argue that enterprises need decision-relevant, calibratable and auditable structured world models whose outputs are reasoning to verify, not fact.

Keywords: world models; latent dynamics; joint-embedding predictive architecture (JEPA); video generation; LLM agents; evaluation benchmarks; agent governance

Introduction

Before crossing a street, a person runs a quick rehearsal in their head: how fast is the car, how many steps do I need, will I make it? This ability to simulate in the mind before acting in the world is what AI researchers call a world model. In 2018, Ha and Schmidhuber turned the intuition into a trainable system: learn a compressed spatio-temporal representation of the environment without supervision, train a policy entirely inside the "dream" the model generates, and then transfer the policy back to the real environment[1].

Eight years later, the world model has become a concept shared across several research communities. Video generation models are described as "world simulators"[2]; interactive environment models can turn a line of text into a three-dimensional world one can walk through in real time[3]; self-supervised video models let a robot arm pick and place objects toward image goals with almost no new data collection[4]; and LLM agents have started to ask "what will this page look like if I click here?" before they click[5]. The expansion brings two problems. One is inconsistent terminology[6,7]. The other is misaligned evaluation: many papers report whether the frames look right, while enterprises care whether decisions made with the model turn out better.

Author information: [1] Qiang Wang, Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd.; research interests: artificial general intelligence, computer vision; Shenzhen, Guangdong, China.

[2] Great Unicorn Intelligence Research Team, Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd.; research interests: AI agents, AGI training; Shenzhen, Guangdong, China. Corresponding author. E-mail: research@hxallai.com.

This review is written for practicing engineers and enterprise decision-makers. The “Background and Related Work” section gives definitions and the taxonomy; the “Core Methods and Technical Analysis” section compares training objectives, action conditioning and planning; the “Comparison and Evaluation” section reviews evaluation benchmarks; the “LLM agents: simulate, then act” section discusses coupling with LLM agents; the “Enterprise applications and case types” section discusses deployable types in enterprise settings; the “Challenges and Limitations” section summarizes limitations; the “Viewpoint and outlook” section presents our view and testable outlook; the “Conclusion” section concludes.

Method. We take the Dreamer, JEPA, Genie and Cosmos lines as the backbone and screened, by hand, arXiv preprints from November 2025 to October 2026 whose titles contain "world model" or "JEPA". Every cited source was opened to check its title and authors. All figures quoted are as reported by the original authors; preprints have not been peer-reviewed and should be read with caution.

Background and Related Work

Definition: starting from the POMDP

In an introductory paper from October 2026, Zhu notes that the notion of a world model remains loosely defined, spanning predictive dynamics models, learned simulators and models that jointly generate future observations and actions; he proposes the partially observable Markov decision process (POMDP) as a unifying frame[6]. Following that line, a world model has at least three parts:

- **An encoder** `z_t = f_θ(o_≤t)` that compresses the observation history into a state representation;
- **A transition model** `ẑ_(t+1) ~ p_θ(z_(t+1) | z_t, a_t)` that predicts the next state given an action;
- **A readout head** (optional) that decodes back to an observation `ô_(t+1)`, or predicts reward, value or termination.

The approaches differ mainly in what the state z looks like—pixels, continuous latents, discrete tokens, text or a structured table—and in which signal constrains the transition model during training. The survey by Ding et al. divides world models by function: those that build internal representations to understand the world, and those that predict future states to simulate and guide decisions[7]. We focus on the latter, because an enterprise agent ultimately has to answer "what happens if we do this?"

Lineage: four converging lines

Latent dynamics and reinforcement learning. Dreamer learns behaviors by imagining trajectories in a learned compact state space and backpropagating value gradients through them[8]. DreamerV2 switched to discrete latents and reached human-level performance on Atari[9]. DreamerV3 outperformed specialized methods on more than 150 tasks with a single configuration and, according to its authors, collected diamonds in Minecraft from scratch without human data or curricula; the work appeared in Nature in 2025[10]. Dreamer 4 (September 2025) pushes the idea offline: the world model learns general knowledge from large amounts of unlabeled video, learns action conditioning

from only a small amount of action-labeled data, and runs real-time inference on a single GPU thanks to a "shortcut forcing" objective and an efficient Transformer. The agent learns entirely in imagination; the authors report it is the first to obtain diamonds in Minecraft purely from offline data, a task that requires choosing more than 20,000 mouse-and-keyboard actions from raw pixels[11]. In parallel, MuZero learns a model that predicts only reward, value and policy and searches a tree over it[12], while TD-MPC2 performs local trajectory optimization in the latent space of an implicit, decoder-free world model and covers 104 online RL tasks with a single set of hyperparameters[13].

Tokens and diffusion. IRIS builds a world model from a discrete autoencoder and an autoregressive Transformer and reaches a mean human-normalized score of 1.046 on Atari 100k with the equivalent of about two hours of gameplay[14]. DIAMOND argues that compressing frames into discrete tokens discards visual details that matter for decisions; using a diffusion model of the environment instead, it reaches 1.46 on the same benchmark[15].

Representation prediction (JEPA). I-JEPA does not reconstruct pixels; from one context block it predicts the representations of other target blocks in the same image[16]. V-JEPA makes feature prediction the sole objective of video self-supervision[17]. V-JEPA 2 is pre-trained on more than one million hours of internet video and then post-trained, with less than 62 hours of unlabeled robot video, into an action-conditioned world model, V-JEPA 2-AC, which performs image-goal pick-and-place zero-shot on Franka arms in two labs—without collecting data in those environments and without task-specific training or reward[4]. LeJEPA (November 2025) offers a theory of JEPAs: the ideal embedding distribution is an isotropic Gaussian, and a regularizer called SIGReg removes the need for heuristics such as stop-gradient and teacher–student networks, with a core implementation of about 50 lines of code[18]. DINO-WM learns dynamics directly on pre-trained DINOv2 patch features for zero-shot planning[19].

Video generation and interactive environments. OpenAI's Sora report calls video generation models "world simulators" and argues that scaling diffusion Transformers over spacetime latent patches is a promising path toward general-purpose simulators of the physical world, while conceding that Sora does not accurately model basic interactions such as glass shattering[2]. Genie learns without supervision from unlabeled internet video and turns text, images or even sketches into virtual worlds that can be controlled frame by frame; it has 11 billion parameters[20]. Genie 3 generates dynamic worlds from a text prompt that can be navigated in real time at 24 frames per second and 720p, staying consistent for a few minutes, with visual memory reaching back about one minute[3]. GameNGen simulates the classic game DOOM in real time at 20 frames per second on a single TPU[21]; UniSim attempts a universal simulator of real-world interaction[22]. NVIDIA's Cosmos platform releases, with open weights, a video curation pipeline, pre-trained world foundation models and tokenizers[23]; its successor Cosmos-Predict2.5 is trained on 200 million curated video clips, ships at 2B and 14B parameters, and targets synthetic data generation, policy evaluation and closed-loop simulation[24]. In autonomous driving, GAIA-1 and GAIA-2 control multi-camera video generation with structured inputs[25,26]. Wiedemer et al. further find that Veo 3 solves, zero-shot, tasks it was not trained for, such as segmentation, understanding physical properties and simple maze solving[27].

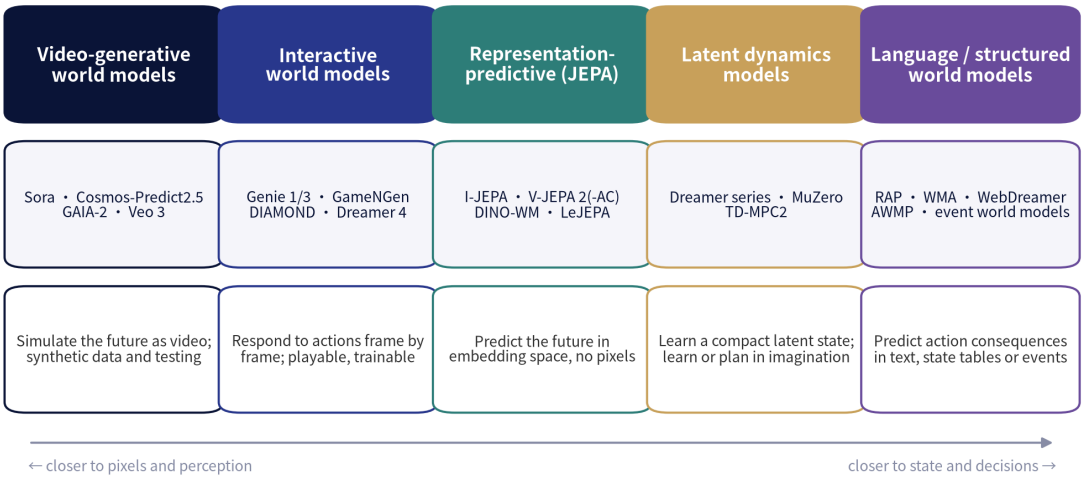
Language and structured world models. RAP uses an LLM as both world model and reasoning agent

and explores with Monte Carlo tree search, framing reasoning as planning[28]. WMA and WebDreamer let web agents predict page changes before they click[5,29]. Work from the past year extends world models to dynamic scheduling[30], modular business systems[31], multi-actor event streams[32] and economic simulation[33].

Five families

Combining these lines, we group world models into five families along two axes: state representation and primary use (Fig. 1, Table 1). The grouping is an approximation for discussion, and families overlap: Dreamer 4 is a latent dynamics model that also runs interactively; V-JEPA 2-AC is a representation-predictive model that is also used for model predictive control.

Figure 1 Five families of world models (state representation × primary use)



Source: the authors' synthesis of [1–27][41–54]. The grouping is an approximation for discussion; families overlap.

Fig. 1 | Five families of world models. Source: the authors' synthesis of the literature.

Table 1 | The five families compared.

Family	State representation	Representative work	Training signal	Action-conditioned?	Primary use
Video-generative	Video latent patches / pixels	Sora[2], Cosmos-Predict2.5[24], GAIA-2[26], Veo 3[27]	Denoising reconstruction	Partly (text, trajectory or structured conditions)	Synthetic data, scenario testing, content
Interactive	Frame sequence + latent actions	Genie[20], Genie 3[3], GameNGen[21], DIAMOND[15]	Denoising or autoregressive prediction	Yes (per frame)	Training environments for agents, interactive simulation
Representation-predictive (JEPA)	Encoder embeddings	I-JEPA[16], V-JEPA 2[4], DINO-WM[19], LeJEPA[18]	Prediction in embedding space + anti-collapse	Yes, after post-training	Understanding, planning, robot control
Latent dynamics	Continuous / discrete latents	Dreamer series[8–11], MuZero[12], TD-	Reconstruction, reward, value,	Yes	Learning in imagination, MPC

Family	State representation	Representative work	Training signal	Action-conditioned?	Primary use
		MPC2[13]	temporal consistency		
Language / structured	Text, state tables, event logs	RAP[28], WMA[29], WebDreamer[5], AWMP[30]	Next-observation or state-delta prediction	Yes	LLM agent planning, business-process simulation

Positioning

Existing surveys mostly approach world models from robotics, autonomous driving or generative modeling[7,34]. This review differs in three ways: it places LLM agents and structured world models in the same taxonomy; it layers evaluation by distance from the decision; and it discusses deployment and governance from the standpoint of enterprise decision-making.

Core Methods and Technical Analysis

A unified formulation

Writing the three parts of the “Definition: starting from the POMDP” section as a training objective, most world models can be expressed as

```
L = L_pred(ẑ_(t+k), target_(t+k)) + λ1 · L_reg + λ2 · L_task
```

Here L_pred is the prediction term, L_reg the anti-collapse/regularization term, and L_task the task term (reward, value, etc.).

The three training objectives differ fundamentally in what the target is: the raw observation (or its compressed latent) for pixel reconstruction; an embedding produced by another encoder for latent prediction; and the next discrete token for autoregressive token models.

Three training objectives compared

Pixel reconstruction (including diffusion denoising). The model must reproduce every detail of future frames. Supervision is dense and the results are visible; the cost is that much capacity goes into details irrelevant to the decision, such as the texture of leaves. DIAMOND shows that in some games those details do matter for decisions[15]; Physics-IQ shows that visual realism and physical understanding are unrelated[35].

Latent prediction (JEPA). The model predicts the future only in representation space, so it need not paint every pixel and can concentrate capacity on predictable structure. V-JEPA 2, pre-trained on over a million hours of video and post-trained on less than 62 hours of robot video, supports planning[4]—a sign of the approach's data efficiency. Its main risk is representation collapse: if the encoder maps every input to the same point, the prediction loss drops to zero while nothing is learned. Earlier methods relied on tricks such as an exponential-moving-average (EMA) target network and stop-gradient; LeJEPA's SIGReg offers a better-founded alternative[18]. A second risk is invisible error: the

loss is measured in the model's own embedding units, which are hard to interpret. Cheung and He find, for one end-to-end JEPA world model, that after the first epoch a further 15% fidelity gain across the next fifteen checkpoints coincides with a correlation of only r = −0.010 between one-step error and planning success[36].

Autoregressive tokens. Frames are discretized into tokens, and the model predicts the next token as a language model would. IRIS and Genie both use this structure[14,20]. It reuses LLM training and serving infrastructure and naturally supports step-by-step interaction; the costs are detail lost to discretization[15] and errors that accumulate step by step. The Genie 3 post states plainly that generating an environment autoregressively is generally harder than generating a whole video, because inaccuracies accumulate over time[3].

Table 2 | Trade-offs among the three training objectives.

Dimension	Pixel reconstruction / diffusion	Latent prediction (JEPA)	Autoregressive tokens
Prediction target	Future observation or its compressed latent	Embedding from a target encoder	Next discrete token
Supervision density	High	Medium	High
Capacity spent on decision-irrelevant detail	Likely	Less	Depends on the tokenizer
Main failure mode	Realistic but physically wrong[35]	Collapse; hard-to-read error[18,36]	Discretization loss; error accumulation[3,15]
Interpretability and human inspection	Intuitive (frames can be viewed)	Needs a decoder or probes	Moderate
Interactivity and real-time use	Needs distillation or few-step sampling[11,21]	Light inference, suits inner planning loops[4,19]	Step-by-step by construction[20]
Representative work	Sora, Cosmos, DIAMOND, GameNGen	I-JEPA, V-JEPA 2, DINO-WM, LeJEPA	IRIS, Genie

For enterprises the table has a direct corollary: if the goal is a decision rather than content to be watched, the training objective should target decision-relevant state as closely as possible. Lambert et al. named this "objective mismatch" in 2020: the dynamics model is trained on one-step likelihood while the real goal is downstream control performance, and the two are often uncorrelated; a model that works for a specific task need not be globally accurate, and a globally accurate model may not be accurate enough locally[37]. Several 2026 papers reconfirm the point from different angles[36,38].

Action conditioning and latent actions

To serve decisions, a world model must answer "what happens if I do A?" That requires conditioning on actions. The difficulty is that most internet video carries no action labels. Three representative approaches:

- **Latent action models.** Genie infers a discrete latent action from adjacent frames without supervision; at inference time users steer the world by choosing latent actions[20].
- **Alignment with a little labeled data.** Dreamer 4 learns world knowledge from large amounts

of unlabeled video and needs only a small amount of action-labeled data to learn general action conditioning[11]; V-JEPA 2-AC does the same with less than 62 hours of robot video[4].

- **Structured conditions.** GAIA-2 conditions on structured variables such as ego-vehicle dynamics, agent configurations, weather and road semantics[26]. Genie 3 lets text trigger "world events", such as changing the weather or introducing new objects; DeepMind argues that this broadens the range of counterfactual, "what if" scenarios available to learning agents[3].

For enterprise settings the third approach is the most realistic: business actions—order quantities, promotion depth, process parameters—are structured variables already and need not be inferred from video.

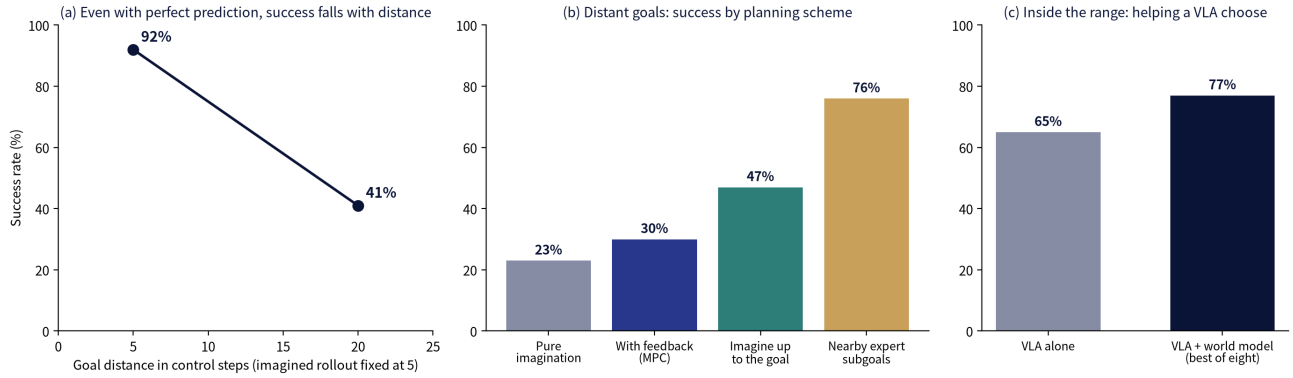
Planning: from predictions to decisions

With an action-conditioned world model in hand, four planning schemes are common:

1. **Learning in imagination**: train a policy network on model-generated trajectories, as in the Dreamer series[8–11].
2. **Model predictive control (MPC)**: at each step search for an action sequence in the model, execute only its first action, and re-plan from the new observation, as in TD-MPC2, DINO-WM and V-JEPA 2-AC[4,13,19].
3. **Tree search**: expand a search tree over the model, as in MuZero and RAP[12,28].
4. **Candidate ranking**: an external policy or LLM proposes candidates that the world model simulates and scores, as in Navigation World Models and WebDreamer[5,39].

Work from the past year exposes a key boundary on planning. Alrasheed et al. build action-conditioned predictors on five frozen visual backbones (including V-JEPA 2 and DINOv2) and, on Meta-World and BridgeData V2, find that a world model guides action selection reliably only when the goal lies within, or slightly beyond, the trajectory it imagines. With five-step rollouts, the model ranks actions reliably only for targets five to ten control steps ahead, whereas task goals lie 16 to 53 steps away, and an 81-fold larger predictor does not extend that range. More fundamentally, even with perfect prediction—using the real simulator in place of the learned model—success falls from 92% to 41% as the target moves from five to twenty steps ahead[40] (Fig. 2). Their conclusion is that planning needs either longer imagined trajectories or closer subgoals: for distant goals, pure imagination succeeds in 23% of episodes, MPC with feedback in 30%, imagining as far as the goal in 47%, and nearby expert subgoals in 76%. Within its plannable range, letting the world model choose among eight actions proposed by a vision-language-action (VLA) policy raises success from 65% to 77%[40].

Figure 2 The plannable range of latent world models (Alrasheed et al., 2026)



Data: figures reported in the abstract of [55] (Meta-World and BridgeData V2; five-step imagined rollouts). In (a) only the 5- and 20-step points are from the paper; the line only indicates the trend.

Fig. 2 | The plannable range of latent world models. Data:[40].

The lesson for enterprise agents is direct: a world model is better at near-range comparison than at long-range prophecy. Splitting a long-horizon goal into verifiable nearby subgoals, and letting the world model compare candidates within each, is more reliable than asking it to roll all the way to the end.

Real-time and long-horizon engineering

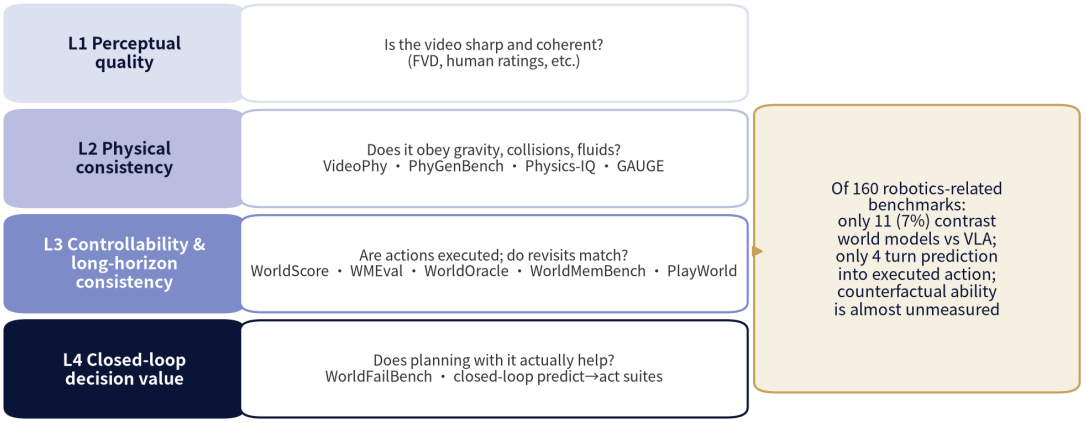
Interactive world models face two engineering constraints. The first is latency: Genie 3 must respond to inputs several times per second while looking back over an ever-longer history[3]; GameNGen's next-frame prediction reaches a peak signal-to-noise ratio (PSNR) of 29.4, comparable to lossy JPEG compression[21]; and Dreamer 4 achieves real-time interactive inference on a single GPU through shortcut forcing[11]. The second is long-horizon consistency: Genie 3 environments stay largely consistent for several minutes, but DeepMind acknowledges that the model currently supports a few minutes of continuous interaction rather than hours[3]. For enterprise simulation, minutes of visual consistency are not the key metric; what matters is whether business state remains coherent after tens or hundreds of decision cycles—a question we return to in the “Long-horizon consistency” and “Challenges and Limitations” sections.

Comparison and Evaluation

Four levels of evaluation

We sort current evaluations by their distance from the decision into four levels (Fig. 3). L1, perceptual quality: are the frames sharp and coherent? L2, physical consistency: do they obey gravity, collisions and fluid behavior? L3, controllability and long-horizon consistency: are actions executed correctly, and does the world stay the same when revisited? L4, closed-loop decision value: when the world model is used for planning or evaluation, do task outcomes actually improve?

Figure 3 Four levels of world-model evaluation and representative benchmarks



Sources: benchmarks [28–40]; statistics on the right from the survey by Jena et al. [40]. The lower the level, the closer to what enterprises care about, and the thinner current coverage.

Fig. 3 | Four levels of world-model evaluation and representative benchmarks. Sources:[34,35,41–51].

Physics: realism is not understanding

VideoPhy uses human evaluation to check whether interactions among various materials follow physical common sense[43]. PhyGenBench covers 27 physical laws across four fundamental domains with 160 prompts[42]. Physics-IQ spans fluid dynamics, optics, solid mechanics, magnetism and thermodynamics; after testing models including Sora, Runway and Pika, its authors conclude that physical understanding is "severely limited" and unrelated to visual realism[35].

Kang et al. use a deterministic 2D physics simulator to generate training data and test systematically whether diffusion video models can "discover" physical laws. Generalization is perfect within distribution and improves with scale for combinatorial cases, but fails out of distribution; further analysis shows that the models do not abstract general rules but generalize "case by case", mimicking similar training examples[41]. GAUGE (August 2026) anchors evaluation in measured real-world trajectories across 22 task families, including rigid bodies, flexible cables, textiles and volumetric deformables. No physics engine is uniformly faithful; the largest gaps appear in impulsive contact, rapid textile motion and volumetric deformation. Video world models, for their part, can produce trajectories with the expected equation form while recovering the wrong accelerations, momentum transfer and oscillation timing[46].

For enterprises, "looks right" is an especially dangerous kind of wrong: plausible in form, off in the numbers, and hard to catch by eye.

Controllability: did the action happen?

WorldScore decomposes world generation into a sequence of next-scene tasks with explicit camera-trajectory layouts and evaluates 19 models on controllability, quality and dynamics over 3,000 test examples[44]. WorldModelBench adds a physics-adherence dimension that catches subtle violations, such as irregular changes in object size that breach mass conservation, and evaluates 14 models using

67,000 human labels[45].

Benchmarks from 2026 have started to replace human impressions with computable state as the reference. WMEval uses Unity physics simulation as ground truth: in nine scenes, the simulator and the world model execute the same action sequences and are compared on action controllability, scene and character consistency, kinematics, fluid mechanics, physical causality and lighting; many weaknesses invisible to standard perceptual metrics leave large room for improvement[47]. WorldOracle compares directly against real executions: given the same initial observation and low-level action sequence, does the predicted rollout reproduce the physical consequences and task-state transitions that actually occurred? It evaluates seven action-conditioned world models[48].

Long-horizon consistency

WorldMemBench evaluates the "world memory" of video world models over 260 test cases, covering static memory of people, objects and environments and dynamic memory of event progression and state persistence. Across nine representative models, strong static retention does not imply reliable dynamic evolution, and repeated revisits reveal memory degradation beyond the first return[49]. PlayWorld uses multimodal agents as "players" that pursue long-horizon objectives in 171 scenarios, scoring geometric consistency, interaction fidelity, out-of-sight evolution and in-sight evolution; the nine models tested remain unreliable on long-horizon interactive objectives, especially spatial consistency and persistent state evolution[50].

For language world models, the World Agent benchmark asks whether a model can keep a world running—grounding events in correct transitions of an explicit state. Across eight models, causal-relation checking is the weak spot for every model[52].

Closed-loop decision value: key, scarce

Jena et al. catalogue 160 web-verified robotics-related benchmarks from 2017 to 2026. Of these, 138 are model-agnostic, only 11 (7%) explicitly contrast world models with direct VLA policies, only four turn prediction into executed action, and counterfactual ability is almost entirely unmeasured[34]. The authors argue that the field cannot yet say whether world models deliver a measurable closed-loop advantage, and that the obstacle lies in how things are measured[34].

WorldFailBench fills part of the gap. It asks not only whether a world model can judge whether a policy succeeds, but whether it faithfully simulates how a policy fails, at which stage, and under which conditions success turns into failure. Across seven world models it finds substantial gaps; the authors warn that misrepresenting failures can lead to wrong diagnoses and can make a policy look ready for deployment even when a small scene change breaks it[51].

Table 3 | Representative benchmarks by evaluation level.

Level	Benchmark	Ground truth	Scale (as reported)	Main finding
L2	VideoPhy[43]	Human evaluation	Prompts on material interactions	Physical common sense lacking
L2	PhyGenBench[42]	Hierarchical model-based judging	160 prompts / 27 laws	Physical correctness needs work

Level	Benchmark	Ground truth	Scale (as reported)	Main finding
L2	Physics-IQ[35]	Real filmed video	Several physics domains	Physical understanding unrelated to realism
L2	GAUGE[46]	Measured real trajectories	22 task families	Right equation form, wrong numbers
L3	WorldScore[44]	Camera-trajectory layouts	3,000 cases / 19 models	Large spread across controllability, quality, dynamics
L3	WMEval[47]	Unity simulation state	9 scenes	Perceptual metrics hide causal errors
L3	WorldOracle[48]	Paired real executions	7 models	Distinct strengths and failure modes
L3	WorldMemBench[49]	Fixed initial references	260 cases / 9 models	Memory degrades with revisits
L3	PlayWorld[50]	Agent-player objectives	171 scenarios / 9 models	Long-horizon spatial consistency unreliable
L4	WorldFailBench[51]	Simulation and real robot	7 models	Failure simulation clearly lacking
Survey	Jena et al.[34]	—	160 benchmarks	7% contrast; counterfactuals almost untested

Three systematic biases in evaluation

Taken together, these studies point to three systematic biases.

First, one-step error stands in for decision value. Cheung and He find almost no correlation between one-step fidelity and planning success[36], and Lambert et al.'s objective mismatch explains why[37]. Across environments the exchange rate is unstable too: a 93% reduction in fidelity error on TwoRoom yields −0.7 percentage points of planning, while a 40% reduction on PushT yields +4.9 points[36].

Second, replay stands in for real branching. Gonuguntla forks SWE-bench agent trajectories at controlled points and continues them with a different model. Swaps rewrite 61% to 94% of post-fork actions, only 3% of replayed states remain valid, and a log-stitching replay evaluator mispredicts every success-relevant outcome[53]. The warning for counterfactual simulation is clear: once an action changes the downstream state, static replay of historical logs scores the wrong world.

Third, side effects of training are ignored. In two interactive text environments, Che et al. replace true next-observation targets with in-distribution but mismatched ones. Prediction accuracy drops by 15.3% to 61.6%, yet most of the task gains remain; even training on random rewards unrelated to the environment raises pass@64 on VisualWebArena by 14.3%[54]. A gain after adding a world-model objective therefore cannot simply be credited to the model having learned to predict the world.

LLM agents: simulate, then act

Why LLM agents need world models

The web, enterprise systems and the physical world are full of irreversible actions. The WMA authors give an example: today's web agents may repeatedly buy a non-refundable flight ticket, a mistake

humans avoid because they foresee the consequence[29]. The WebDreamer authors likewise note that many actions on real websites cannot be undone, so backtracking-based tree search is hard to apply[5]. A world model offers another route: rehearse the consequences of candidate actions in the model, then commit to one.

Four modes of coupling

(1) The LLM itself as world model. RAP predicts reasoning states with an LLM and searches a tree[28]. WebDreamer uses an LLM as both world model and value function; on VisualWebArena it is competitive with tree search while being four to five times more efficient, and its specially trained Dreamer-7B world model performs comparably to GPT-4o[5].

(2) A dedicated, trained world model. WMA finds that predicting the next page directly runs into repeated elements and very long HTML, so it predicts only natural-language descriptions of the important state differences between time steps[29]. DGS-WM goes further, using as world-modeling targets only the observation units that carry decision value for subsequent actions, weighting them adaptively by confidence, and optimizing jointly with the RL loss[38]. Together they suggest that a language agent's world model should predict what changed, not what is seen.

(3) On-demand invocation and risk anticipation. PAWM's pilot study finds that world-model guidance is most useful when it anticipates pitfalls within a bounded multi-step future. It intervenes only when a pitfall may arise, predicting the impending failure to steer the agent away; it improves Claude Sonnet 4.5 by 24.3% on ScienceWorld and 7.7% on WebArena-Lite[55]. DreamGuard uses a risk-aware world model as a runtime guardrail: it maintains a compact recurrent latent state over the trajectory, predicts future states to derive immediate-hazard and prefix-risk evidence, and decides before execution whether to intervene, with an average end-to-end latency of 25 ms per call[56].

(4) Synthesizing trainable environments. WildAWM uses open-source models to synthesize executable environments that support graphical interfaces, MCP services and command lines together, producing 341 long-horizon tasks; models trained in these worlds improve markedly on eight out-of-distribution benchmarks[57].

Table 4 | Representative work coupling LLM agents with world models.

Mode	Representative work	What the world model predicts	When it is called	Reported effect (as reported)
LLM as world model	RAP[28], WebDreamer[5]	Next reasoning state / page	Every step	On par with tree search, 4–5× more efficient[5]
Dedicated world model	WMA[29], DGS-WM[38]	State deltas / decision-relevant observations	Every step	Better than full-observation modeling[38]
On-demand risk anticipation	PAWM[55], DreamGuard[56]	Failures / risk within a bounded future	When risk is detected	+24.3% (ScienceWorld)[55]; 25 ms[56]
Environment synthesis	WildAWM[57]	Full executable environments	At training time	Gains on 8 OOD benchmarks[57]
Structured exogenous	AWMP[30]	Future job arrivals	When observations deviate from forecast	Weighted tardiness down 6–46%[30]

Mode	Representative work	What the world model predicts	When it is called	Reported effect (as reported)
modeling				

When can a world model be trusted?

"Simulate, then act" presumes the simulation is trustworthy. Dual-Frontier formalizes the question: when a world-model-guided decision fails, the trajectory alone may not reveal whether the decision rule or the world model was at fault, and the authors prove that the two components are not identifiable from passive interaction. The resulting principle admits a decision only when its predicted advantage exceeds a certified bound on decision-relevant world-model error; otherwise evidence goes first to verifying the world model[58]. In our view this verify-then-promote idea lines up closely with the enterprise governance principle that advice is not a decision (see “Viewpoint and outlook”).

Structured decomposition is another route to trust. AWMP splits the transitions of dynamic scheduling into the known, deterministic transitions of the shop floor and the uncertain, exogenous arrival of future jobs, and lets the LLM forecast only the latter. A surprise-triggered mechanism re-queries the LLM only when new arrivals deviate sufficiently from the previous forecast, and the LLM is calibrated online on its recent forecast errors. Against the best of eight classical priority-dispatching rules, weighted tardiness falls by 6% to 21% on dynamic job shops and by 22% to 46% on flexible job shops[30]. Song and Cai study modular services such as order, payment, inventory and shipment, and show that a world model fitted only to observational traces incurs an irreducible interventional error when back-door paths are unblocked: a trace may show that payment precedes shipment without revealing whether payment authorizes shipment or a hidden trigger explains both. Causal interfaces help most where API signatures expose preconditions and downstream effects[31].

Multi-actor settings are harder still. Ni et al. build "event world" benchmarks and find that when only the agent's own action is known and other actors strongly affect the outcome, models struggle to predict; larger models help little, and models overestimate plan success[32]. Genie 3's official limitations likewise list the modeling of complex interactions among multiple independent agents as an open problem[3].

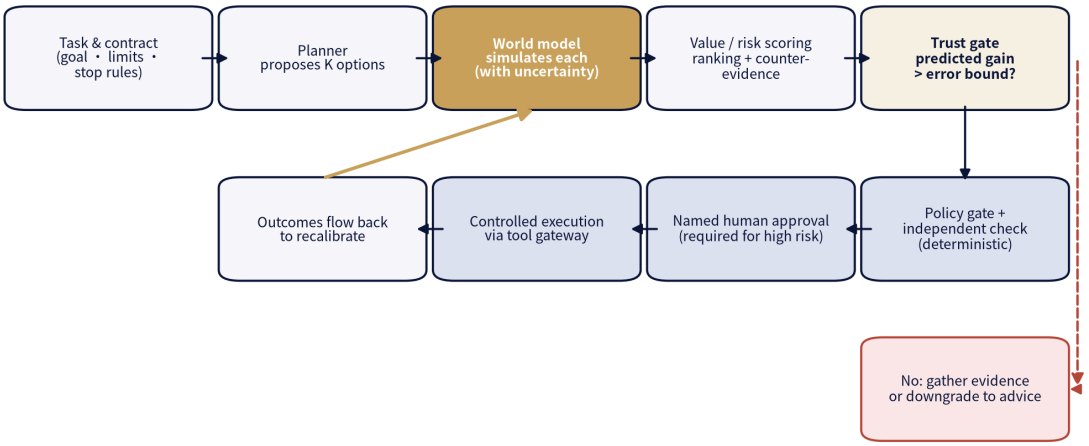
Cost: simulation is not free

Every "simulate first" means extra inference calls. TAO, a recent preprint, targets agentic serving in which GPU inference alternates with CPU tool calls. It jointly considers near-term tool readiness, node-level cache credit with time-aware decay, and token footprint when admitting and placing sessions; the authors report that across 12 settings, relative to the strongest evaluated baseline for each metric, it cuts latency by up to 26.8% and mean time to first token by up to 36.3%, and raises decode throughput by up to 29.5%[59]. Our reading is that when a world model becomes a tool the agent calls repeatedly, this kind of joint scheduling directly affects whether simulating first pays for itself. PAWM's "call only when something may go wrong"[55] and AWMP's surprise-triggered refresh[30] give the algorithmic answer to the same question: on-demand simulation is more economical than simulating every step.

An essay published by the DeepMind Institute on October 29 adds that what changes the scale of

organizations is the ability of agents to coordinate, which current benchmarks barely measure[60]. When several agents share one world model to coordinate their plans, the model becomes the common draft of their collaboration, and its consistency and traceability matter all the more. CoWorld Reasoner wraps a video world model as a set of tools, uses skills to guide a vision-language model in deciding whether to "imagine", and lets tools and skills evolve with feedback without changing model weights, reporting relative gains of 11.8% on MMSI and 13.0% on ERQA[61]. This "harness-orchestrated world model" is quite close to how an enterprise agent platform is organized.

Figure 4 "Simulate, then act": where a world model sits inside a governed agent (our view, schematic)



Note: the trust gate follows the idea of Dual-Frontier [48]; on-demand invocation follows PAWM [44]; governance components correspond to Harness modules listed on the Great Unicorn Intelligence website. Schematic of our view, not a description of any deployment.

Fig. 4 | "Simulate, then act": where a world model sits inside a governed agent. Our view, schematic.

Enterprise applications and case types

This section discusses types only. All examples and charts use illustrative data and do not represent the actual results of any company or product.

Decision framework: five questions first

Before building a world model for a business problem, we suggest answering five questions:

1. **Can the state be observed in structured form?** Tabular state—inventory, goods in transit, prices, process parameters—makes a better world-model state than "what the store looks like".
2. **Can the actions be enumerated or parameterized?** Order quantities, discount depth and temperature set-points can; "improve customer relationships" cannot.
3. **Which parts are known mechanisms and which are exogenous uncertainty?** Following AWMP's decomposition[30], write deterministic business rules (such as inventory conservation) as code, and let the learned model predict only the exogenous part (such as demand).
4. **Is there interventional data?** With only historical observations and no active experiments or natural experiments, the risk of causal confusion is high[31,62].

5. **What does an error cost, and can it be undone?** The higher and less reversible the cost, the more the simulation should stay at the level of advice and go through human approval.

Type one: FMCG counterfactual simulation

A typical question: "If next month we switch these stores' replenishment rule from 'last week's sales' to 'rolling forecast', and add extra stock in promotion weeks, what happens to stock-outs and near-expiry inventory?" The state is inventory, goods in transit and price by store × SKU × week; the actions are the replenishment rule and the promotion calendar; the exogenous variables are demand, weather and competitor activity.

Fig. 5 uses illustrative data to make two points. One: intervals matter more than point estimates. Differences in the median between the three policies must be read together with the simulated intervals; where the intervals overlap, no policy should be declared better. Two: long rollouts must be recalibrated regularly. Rolled forward step by step without actual data, the error compounds with the horizon; recalibrated with actual sales every few weeks, it stays low. This matches MBPO's lesson of branching short rollouts from real data[63].

Figure 5 Illustration: counterfactual simulation of store replenishment policies (illustrative data only)

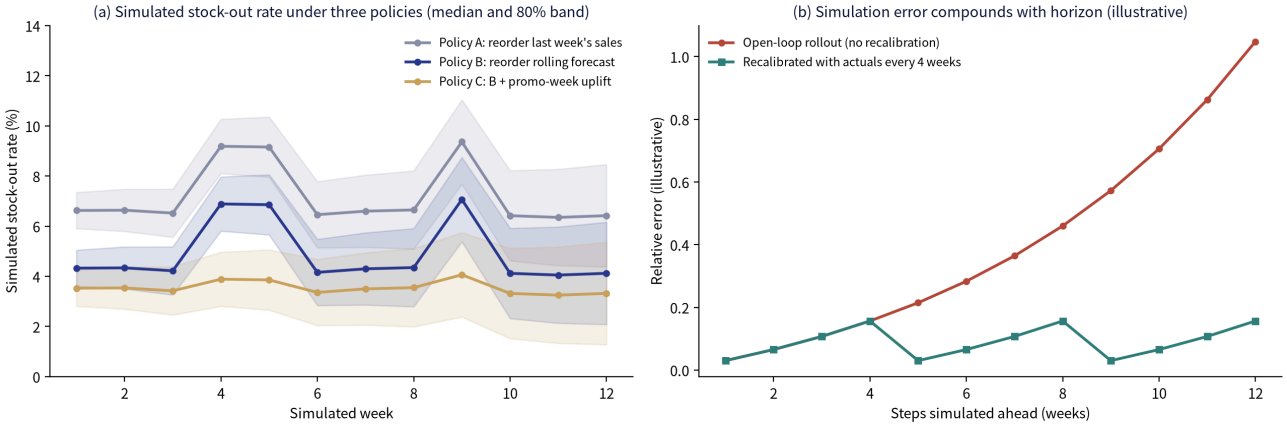


Fig. 5 | Illustration: counterfactual simulation of store replenishment policies. All data illustrative.

Type two: pharma process simulation

In pharmaceutical manufacturing, the relation between process parameters and critical quality attributes can partly be described by mechanistic models and partly has to be learned from batch data, which suits a hybrid world model of "mechanistic model + data-driven residual". GAUGE reminds us that even mature physics engines show systematic gaps on contact, flexible and deformable processes[46]; simulation results can therefore narrow the range of experiments and flag risk directions, but cannot replace process validation. The data, model versions and parameters used in a simulation should be logged and replayable, in line with requirements for computerized systems and data integrity.

Type three: risk stress scenarios

In risk management, a world model can serve as a scenario generator: under a given external shock—

say a sudden drop in demand or a supply disruption—it generates multiple possible paths, helping analysts find weak links invisible under a single assumption. Research on economic world models proposes a six-level capability ladder and observes that existing work clusters at the lower levels[33]. In public health, EpiMind uses a graph-factored recurrent state-space model to generate joint, policy-conditioned rollouts across regions while guaranteeing per-period feasibility under shared-resource constraints, reducing the RMSE of hospital admissions by 29% relative to graph-free dynamics modeling[64]. Such work shows that structured, constrained world models can support scenario analysis; but in any regulated setting, scenario simulation can only be supporting material for analysts, and methods and conclusions should be approved by named professionals.

Table 5 | Deployable types of world models in enterprise settings. Discussion of types.

Scenario type	State	Actions	Exogenous variables	Recommended form	Main risk	Required gates
FMCG replenishment and promotion	Inventory, in-transit, price	Replenishment rules, promotion calendar	Demand, weather, competitors	Deterministic inventory rules + learned demand model	Causal confusion (promotions coincide with peak season)	Show intervals, recalibrate, human sign-off
Pharmaceutical process	Process parameters, intermediate metrics	Parameter settings	Raw-material batch variation	Mechanistic model + data residual	Extrapolation failure	Pinned model versions; used for decisions only after validation
Risk stress scenarios	Aggregate portfolio-level metrics	Scenario assumptions	External shocks	Structured scenario generator	Thin scenario coverage, overconfidence	Approval by named professionals; no automatic execution
Business-process agents	System and document state	API calls	Other actors' behavior	Causal interfaces + risk anticipation[31,56]	Irreversible writes	Policy gate, idempotent writes, approval

Challenges and Limitations

Compounding error. Errors in autoregressive rollouts accumulate with the horizon, as the Genie 3 post[3], MBPO's short-branch design[63] and the plannable-range experiments of Alrasheed et al.[40] all show. Even with perfect prediction, success on distant goals drops markedly[40], so the problem lies not only in model accuracy but in the planning scheme itself.

Causal confusion. De Haan et al. find that ignoring causal structure leads to "causal misidentification": more information can yield worse performance[62]. A world model's training objective cannot tell what an action caused from what merely co-occurred with it[65]. Do-JEPA starts from one saved simulator state, runs the target action and a reference action, and trains the model to predict the difference between the two futures, cutting latent effect error by 28.4% on an end-to-end model[65]. But this relies on a resettable simulator; in real enterprise data, interventional evidence is often scarce.

Missing evaluation. As the "Comparison and Evaluation" section showed, closed-loop and counterfactual evaluation is very thin[34], one-step error is not a reliable proxy[36], and static replay scores the wrong world[53].

Compute and data. V-JEPA 2 draws on more than one million hours of video[4] and Cosmos-Predict2.5 on 200 million video clips[24]. For enterprises the more realistic constraint is tabular data. Jeon et al. compare, on 147 real tabular datasets, a tabular foundation model with a JEPA latent term against one trained on value prediction alone: the JEPA arm trails 32:70 in wins to losses on classification and 8:24 on regression, and needs 1.42 times the steps and 1.66 times the wall-clock time to reach its plateau[66]. Transplanting a recipe that works for vision onto enterprise tables is no guarantee of gains.

Multiple actors and open worlds. Other actors' decisions change outcomes, yet models overestimate plan success[32]; video world models remain immature at modeling interactions among multiple independent agents[3].

Safety and governance. Because its open-ended, real-time capabilities raise new safety challenges, Genie 3 was released as a limited research preview[3]. In an enterprise, if a world model's output is written into systems as fact, model error turns into business data.

Table 6 | Main limitations and mitigations.

Limitation	Symptom	References	Mitigation
Compounding error	Long rollouts drift	3,40,63	Short branches, nearby subgoals, regular recalibration with actuals
Objective mismatch	One-step error unrelated to decisions	36,37	Select models on decision metrics; decision-relevant targets[38]
Causal confusion	Correlation taken as causation	31,62,65	Interventional data, causal interfaces, counterfactual-difference objectives
Missing evaluation	Closed-loop and counterfactual untested	34,53	Real-branch evaluation, closed-loop metrics
Compute and data	Video-scale data unavailable	4,24,66	Structured state, mechanism + residual
Multiple actors	Plan success overestimated	3,32	Model other actors explicitly; conservative decisions

Viewpoint and outlook

What follows is the view and judgment of Great Unicorn Intelligence, not established fact.

View 1: Enterprise decision agents need decision-relevant, structured world models, not pixel worlds. Three arguments support this: enterprise actions are structured variables to begin with (see "Action conditioning and latent actions"); one-step fidelity correlates weakly with decision value (see "Three systematic biases in evaluation"); and in operational problems such as scheduling, the decomposition into known deterministic transitions plus learned exogenous variables has already shown results[30]. In enterprises, video world models are more likely to serve synthetic data and

testing.

View 2: A world model's output is reasoning, not fact. The website of our Enterprise Agent Platform summarizes its governance framework as five inequalities, two of which are "reasoning ≠ fact" and "advice ≠ decision"[67]. We believe world-model simulations belong in the same framework: they can enter the approval package as evidence for candidate options, but they may not bypass the policy gate and independent verifier in Harness, and still less trigger write operations directly. Dual-Frontier's rule of admitting a decision only when its predicted advantage exceeds the error bound[58] can be read as an algorithmic statement of this principle.

View 3: World models need an online calibration loop. Fig. 5 shows that long rollouts must be recalibrated with actual data. Deep Hope S1, as described on our website, uses a "train while serving" architecture: an online path handles real-time inference and decisions, an offline path handles batch modeling, and samples and outcomes flow back to drive offline iteration; in FMCG it carries rolling store and regional demand forecasts[68]. We see this kind of dual online/offline path as a suitable foundation for calibratable world models, while the Harness model router—which pins model versions and forbids automatic upgrades mid-run—keeps every simulation reproducible.

View 4: Use simulation at low autonomy first. The website divides agent autonomy into levels L0 to L5, with the main field of play at L2 (advise), L3 (call tools) and conditional L4 (execute low-risk tasks)[67]. We recommend using world models first at L1–L2: in scenario agents such as FMCG campaign reviews and initial review of pharmaceutical expense documents, simulations help people compare options faster, while named staff still make the decisions.

View 5: Evaluate on decision regret and calibration. We suggest that enterprises accepting a world model report at least three metrics: improvement over a baseline when used for decisions (a closed-loop metric); the coverage of actual outcomes by simulated intervals (a calibration metric); and the accuracy of predictions for actions not taken, checked against later natural experiments (a counterfactual metric).

Table 7 | Trend predictions testable within two to three years. Our view.

No.	Prediction	Time frame	How to test
P1	More than half of newly released world-model benchmarks will report both open-loop prediction metrics and closed-loop execution metrics	By end of 2028	Count new benchmarks from 2027–2028 against the 7% baseline in[34]
P2	Mainstream LLM agent frameworks will ship an on-demand world model / risk anticipator as a standard component	By 2028	Survey the official component documentation of major open-source agent frameworks
P3	Most published enterprise-operations world-model cases will use a hybrid of deterministic rules plus learned exogenous variables rather than end-to-end generation	By 2028	Review operations cases in peer-reviewed papers and official technical reports
P4	In enterprises, video and interactive world models will be used mainly for synthetic data and testing; published cases of direct use in business decisions will remain a minority	By end of 2027	Tally the uses of world models in published cases
P5	Regulated industries will require that simulations entering decision materials carry uncertainty intervals and calibration records	By 2029	Track wording in regulatory documents and internal-control guidance

Conclusion

World models are moving from generating futures that look real to supporting better decisions. Our main conclusions are four. World models fall into five families, and enterprise decisions most need the structured, decision-relevant one. The training objective determines where capacity goes, and one-step fidelity is not a reliable proxy for decision value. The closer an evaluation is to the decision, the thinner its coverage. "Simulate, then act" already shows gains in LLM agents, provided the model is called on demand, compares near-range options, is recalibrated regularly, and hands its simulations to a governance system for the final call.

Key takeaways

- A world model's value lies in near-range comparison, not long-range prophecy[40].
- Realistic frames do not mean physical understanding[35,46]; low one-step error does not mean good decisions[36].
- Counterfactual evaluation cannot rely on log replay[53].
- In enterprise settings, start with structured state and a hybrid of deterministic rules and learned exogenous variables[30].
- Simulations are reasoning, not fact; before entering a decision they should pass a policy gate, independent verification and named approval.

Acknowledgments and disclaimer: the views expressed are those of the Great Unicorn Intelligence research team. All third-party figures are as reported by the original authors; product capabilities are as published on the Great Unicorn Intelligence website, https://www.hxallai.com/en/.

References

1. Ha, D. & Schmidhuber, J. World Models. Preprint at https://arxiv.org/abs/1803.10122 (2018).

2. Brooks, T. et al. Video generation models as world simulators. OpenAI Technical Report (2024-02-15) (2024). https://openai.com/index/video-generation-models-as-world-simulators/

3. Parker-Holder, J. & Fruchter, S. Genie 3: A new frontier for world models. Google DeepMind Blog (2025-08-05) (2025). https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/

4. Assran, M. et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. Preprint at https://arxiv.org/abs/2506.09985 (2025).

5. Gu, Y. et al. Is Your LLM Secretly a World Model of the Internet? Model-Based Planning for Web Agents. arXiv preprint arXiv:2411.06559 (2024). https://arxiv.org/abs/2411.06559

6. Zhu, Z. An Introduction to World Models for Generalist Agents. Preprint at https://arxiv.org/abs/2610.30789 (2026).

7. Ding, J. et al. Understanding World or Predicting Future? A Comprehensive Survey of World Models. ACM Computing Surveys (extended version on arXiv). arXiv:2411.14499 (2024). https://arxiv.org/abs/2411.14499

8. Hafner, D., Lillicrap, T., Ba, J. & Norouzi, M. Dream to Control: Learning Behaviors by Latent Imagination. Preprint at https://arxiv.org/abs/1912.01603 (2019).

9. Hafner, D., Lillicrap, T., Norouzi, M. & Ba, J. Mastering Atari with Discrete World Models. In International Conference on Learning Representations (ICLR 2021). arXiv:2010.02193 (2020). https://arxiv.org/abs/2010.02193

10. Hafner, D., Pasukonis, J., Ba, J. & Lillicrap, T. Mastering diverse control tasks through world models. Nature 640, 647–653 (2025). https://doi.org/10.1038/s41586-025-08744-2

11. Hafner, D., Yan, W. & Lillicrap, T. Training Agents Inside of Scalable World Models. Preprint at https://arxiv.org/abs/2509.24527 (2025).

12. Schrittwieser, J. et al. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. Preprint at https://arxiv.org/abs/1911.08265 (2019).

13. Hansen, N., Su, H. & Wang, X. TD-MPC2: Scalable, Robust World Models for Continuous Control. In International Conference on Learning Representations (ICLR 2024). arXiv:2310.16828 (2023). https://arxiv.org/abs/2310.16828

14. Micheli, V., Alonso, E. & Fleuret, F. Transformers are Sample-Efficient World Models. In International Conference on Learning Representations (ICLR 2023). arXiv:2209.00588 (2022). https://arxiv.org/abs/2209.00588

15. Alonso, E. et al. Diffusion for World Modeling: Visual Details Matter in Atari. In Advances in Neural Information Processing Systems (NeurIPS 2024). arXiv:2405.12399 (2024). https://arxiv.org/abs/2405.12399

16. Assran, M. et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023). arXiv:2301.08243 (2023). https://arxiv.org/abs/2301.08243

17. Bardes, A. et al. Revisiting Feature Prediction for Learning Visual Representations from Video. Preprint at https://arxiv.org/abs/2404.08471 (2024).

18. Balestriero, R. & LeCun, Y. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics. Preprint at https://arxiv.org/abs/2511.08544 (2025).

19. Zhou, G., Pan, H., LeCun, Y. & Pinto, L. DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning. Preprint at https://arxiv.org/abs/2411.04983 (2024).

20. Bruce, J. et al. Genie: Generative Interactive Environments. Preprint at https://arxiv.org/abs/2402.15391 (2024).

21. Valevski, D., Leviathan, Y., Arar, M. & Fruchter, S. Diffusion Models Are Real-Time Game Engines. In International Conference on Learning Representations (ICLR 2025). arXiv:2408.14837 (2024). https://arxiv.org/abs/2408.14837

22. Yang, S. et al. Learning Interactive Real-World Simulators. Preprint at https://arxiv.org/abs/2310.06114 (2023).

23. Agarwal, N. et al. (NVIDIA). Cosmos World Foundation Model Platform for Physical AI. Preprint at https://arxiv.org/abs/2501.03575 (2025).

24. Ali, A. et al. (NVIDIA). World Simulation with Video Foundation Models for Physical AI. Preprint at https://arxiv.org/abs/2511.00062 (2025).

25. Hu, A. et al. GAIA-1: A Generative World Model for Autonomous Driving. Preprint at https://arxiv.org/abs/2309.17080 (2023).

26. Russell, L. et al. GAIA-2: A Controllable Multi-View Generative World Model for Autonomous Driving. Preprint at https://arxiv.org/abs/2503.20523 (2025).

27. Wiedemer, T. et al. Video models are zero-shot learners and reasoners. Preprint at https://arxiv.org/abs/2509.20328 (2025).

28. Hao, S. et al. Reasoning with Language Model is Planning with World Model. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023). arXiv:2305.14992 (2023). https://arxiv.org/abs/2305.14992

29. Chae, H. et al. Web Agents with World Models: Learning and Leveraging Environment Dynamics in Web Navigation. In International Conference on Learning Representations (ICLR 2025). arXiv:2410.13232 (2024). https://arxiv.org/abs/2410.13232

30. Zhang, N. et al. LLM-Based Exogenous World Modeling for Dynamic Scheduling. Preprint at https://arxiv.org/abs/2610.29308 (2026).

31. Song, X. & Cai, Z. When Do Causal World Models Help Modular LLM Agents. Preprint at https://arxiv.org/abs/2610.00012 (2026).

32. Ni, J. et al. From Single-Actor Environments to Multi-Actor Event Worlds: Can LLMs Be Event World Models? Preprint at https://arxiv.org/abs/2610.30488 (2026).

33. Han, J. et al. From Economic Agents to Agentic Economies: A Systems Blueprint for Economic World Models. Preprint at https://arxiv.org/abs/2608.06020 (2026).

34. Jena, G. et al. Do World Models Make Better Robots? A Survey of Evaluation Benchmarks for Predictive Embodied Intelligence. arXiv preprint arXiv:2609.29669 (2026). https://arxiv.org/abs/2609.29669

35. Motamed, S. et al. Do generative video models understand physical principles? Preprint at https://arxiv.org/abs/2501.09038 (2025).

36. Cheung, J. & He, Z. One-Step Prediction Fidelity Is a Poor Proxy for Planning Performance in JEPA World Models. Preprint at https://arxiv.org/abs/2610.19375 (2026).

37. Lambert, N., Amos, B., Yadan, O. & Calandra, R. Objective Mismatch in Model-based Reinforcement Learning. In Proceedings of the 2nd Conference on Learning for Dynamics and Control (L4DC), PMLR 120, 761–770. arXiv:2002.04523 (2020). https://arxiv.org/abs/2002.04523

38. Ran, M. et al. Model What Matters: Decision-Guided Selective World Modeling for Language Agents. Preprint at https://arxiv.org/abs/2610.19642 (2026).

39. Bar, A. et al. Navigation World Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025). arXiv:2412.03572 (2024). https://arxiv.org/abs/2412.03572

40. Alrasheed, A., Azam, B. & Akhtar, N. The Planning Limits of Latent World Models. Preprint at https://arxiv.org/abs/2609.39235 (2026).

41. Kang, B. et al. How Far is Video Generation from World Model: A Physical Law Perspective. In International Conference on Machine Learning (ICML 2025). arXiv:2411.02385 (2024). https://arxiv.org/abs/2411.02385

42. Meng, F. et al. Towards World Simulator: Crafting Physical Commonsense-Based Benchmark for Video Generation. Preprint at https://arxiv.org/abs/2410.05363 (2024).

43. Bansal, H. et al. VideoPhy: Evaluating Physical Commonsense for Video Generation. Preprint at https://arxiv.org/abs/2406.03520 (2024).

44. Duan, H. et al. WorldScore: A Unified Evaluation Benchmark for World Generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2025). arXiv:2504.00983 (2025). https://arxiv.org/abs/2504.00983

45. Li, D. et al. WorldModelBench: Judging Video Generation Models As World Models. Preprint at https://arxiv.org/abs/2502.20694 (2025).

46. Wang, S. et al. GAUGE: A Measurement-Grounded Benchmark for Physical Fidelity in Simulation Engines and Video World Models. Preprint at https://arxiv.org/abs/2608.05948 (2026).

47. Hu, L. et al. WMEval: Evaluating Interactive World Models Grounded on Physical Simulations. Preprint at https://arxiv.org/abs/2610.33894 (2026).

48. Duan, H. et al. WorldOracle: Benchmarking Action-Conditioned World Models against Real-World Executions. Preprint at https://arxiv.org/abs/2610.18826 (2026).

49. Zhou, N. et al. WorldMemBench: A Comprehensive Benchmark for Memory of Video World Models. Preprint at https://arxiv.org/abs/2610.19063 (2026).

50. Ding, K. et al. PlayWorld: Benchmarking World Models with Agent Players over Long-Horizon Objectives. Preprint at https://arxiv.org/abs/2608.13552 (2026).

51. Zhang, Z. et al. WorldFailBench: Do World Models Understand Policy Failures? Preprint at https://arxiv.org/abs/2610.33565 (2026).

52. Chen, W. et al. World Agent: Can Language Models Keep a World Running? Preprint at https://arxiv.org/abs/2609.32692 (2026).

53. Gonuguntla, A. The Replay Gap: Static Evaluation of Model Switching in LLM Agents Scores the Wrong World. Preprint at https://arxiv.org/abs/2608.08239 (2026).

54. Che, X. et al. Does Learning to Predict the World Help Agents Act? Auditing World-Model Post-Training. arXiv preprint arXiv:2609.33335 (2026). https://arxiv.org/abs/2609.33335

55. Wang, J. et al. Steering Black-Box Large Language Model Agents with Pitfall Anticipation World

Model. Preprint at https://arxiv.org/abs/2610.29139 (2026).

56. Lin, W. et al. DreamGuard: Efficient Runtime Guardrail for LLM Agents via Risk-Aware World Model. Preprint at https://arxiv.org/abs/2608.05695 (2026).

57. Wang, Z. et al. Agent World Model in the Wild: Computer-Use Agent Environment Synthesis. Preprint at https://arxiv.org/abs/2610.26828 (2026).

58. Zhu, H. et al. Dual-Frontier: When Can an Agent Trust Its World Model? Preprint at https://arxiv.org/abs/2609.26293 (2026).

59. Li, S. et al. TAO: Tool-Aware Joint Scheduling for LLM-Based Agentic Serving. Preprint at https://arxiv.org/abs/2610.37230 (2026).

60. Imas, A., Koh, A. & Legg, S. The Economics of Intelligence: How and when does AGI lead to economic transformation? DeepMind Institute Essay (2026-10-29) (2026). https://institute.deepmind.com/essays/the-economics-of-intelligence/

61. Cheng, D. et al. CoWorld Reasoner: Agentic Multimodal Reasoning with the Evolving Video World Model-Driven Harness. Preprint at https://arxiv.org/abs/2610.21910 (2026).

62. de Haan, P., Jayaraman, D. & Levine, S. Causal Confusion in Imitation Learning. In Advances in Neural Information Processing Systems (NeurIPS 2019). arXiv:1905.11979 (2019). https://arxiv.org/abs/1905.11979

63. Janner, M., Fu, J., Zhang, M. & Levine, S. When to Trust Your Model: Model-Based Policy Optimization. In Advances in Neural Information Processing Systems (NeurIPS 2019). arXiv:1906.08253 (2019). https://arxiv.org/abs/1906.08253

64. Su, Y. et al. Graph World Models for Constrained Epidemic Policy Planning. Preprint at https://arxiv.org/abs/2609.35545 (2026).

65. Resani, H. & Shi, J. Q. Do-JEPA: From Masking to Intervention in Latent World Models. Preprint at https://arxiv.org/abs/2609.37378 (2026).

66. Jeon, M., Cho, S. & Suh, J. Y. A JEPA Recipe for Tabular Foundation Models. Preprint at https://arxiv.org/abs/2609.25541 (2026).

67. 弘犀智能科技（深圳）有限公司 (Great Unicorn Intelligence). 企业智能体平台 · Harness 治理引擎 / Enterprise Agent Platform · Harness governance engine. Official website (2026). https://www.hxallai.com/platform/agent/

68. 弘犀智能科技（深圳）有限公司 (Great Unicorn Intelligence). 推训一体大模型平台 Deep Hope S1 / Deep Hope S1 unified inference–training LLM platform. Official website (2026). https://www.hxallai.com/platform/deep-hope/

Acknowledgements We thank the Great Unicorn Intelligence Research Team, Tsinghua University, and Professor Qiang Wang's research team at Xi'an Jiaotong University.

Competing interests Qiang Wang and the Great Unicorn Intelligence Research Team are affiliated with Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd., whose products and views are

discussed in this article.

Additional information Correspondence and requests for materials should be addressed to the Great Unicorn Intelligence Research Team (research@hxallai.com).

Copyright Notice: © 2026 Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd. All rights reserved. Any reposting, quotation in commentary, or adaptation must credit the authors (Qiang Wang and Great Unicorn Intelligence Research Team) and cite the source (Great Unicorn Intelligence, https://www.hxallai.com).