TECHSTRONG SPECIAL REPORT

VOL. 01 · NO. 04 · AUGUST 2026

THE PHYSICAL AI ISSUE

From Words to Worlds

The race to build AI that understands what happens next.

Large language models learned to predict the next word. The next great AI race is to predict the next state of the world — and the companies chasing it are betting billions that the future of intelligence is physical.

BY ALAN SHIMEL / TECHSTRONG GROUP

TECHSTRONG SPECIAL REPORT · VOL. 01 · NO. 04 · AUGUST 2026

Contents

IN THIS REPORT

Google DeepMind, Nvidia, Meta, Runway, World Labs, Odyssey, Decart, Waymo, Waabi, Wayve and Yann LeCun's new AMI Labs are all racing to build "world models" – systems meant to predict not the next word, but the next state of a physical environment.

The money is arriving before the definition has settled. This report maps the architectures, the capital, the benchmarks and the reality gap that still separates a convincing video from a system that understands consequence.

$2.64B
Combined new capital raised by World Labs & AMI Labs alone since Feb. 2026

52%
Highest paired score on the 2026 What-If World causal benchmark — no model exceeded it

2M
Industrial robots operating in China — 4.5× Japan's installed base

0.95
Sim-to-real correlation Runway reported across 8 robot-manipulation policies

OPENING

Predicting Consequences, Not Just Words

That difference sounds subtle until a machine is asked to do something rather than say something.

An LLM can generate a plausible description of a robot placing a glass on a table. A world model is supposed to predict whether the robot's hand will reach the glass, whether its grip will hold, whether the glass will collide with another object and whether the action will succeed. One system predicts language. The other attempts to predict consequences.

That distinction helps explain why world models have become one of the most heavily financed and contested frontiers in artificial intelligence. Google DeepMind, Nvidia, Meta, Runway, World Labs, Odyssey, Decart, Waymo, Waabi, Wayve and Yann LeCun's new AMI Labs are all pursuing some version of the idea. They do not agree on the architecture, the product or even what should qualify as a world model.

The money, however, is arriving before the definition has settled.

World Labs, founded by Fei-Fei Li and three other prominent AI researchers, raised $1 billion in February 2026, adding AMD, Nvidia, Autodesk and other strategic investors. Autodesk alone committed $200 million.[1] AMI Labs raised $1.03 billion in what was described as Europe's largest seed round.[3] Odyssey raised $310 million at a $1.45 billion valuation.[4] Decart raised $300 million at a valuation approaching $4 billion.[5] In each case, investors are betting that the AI economy is about to move beyond systems trained primarily on humanity's recorded knowledge and toward systems trained to model environments, actions and outcomes.

World models are being pitched as the foundation for interactive entertainment, spatial computing, robotics, autonomous vehicles, digital twins, scientific discovery, industrial automation and, eventually, artificial general intelligence. They may become all of those things.

They may also become another AI category in which visually impressive demonstrations are mistaken for evidence that the underlying system understands more than it does.

That is the tension at the center of the world-model movement. A system that produces a beautiful video of a physical event may have learned how the event usually looks. That does not mean it understands why it happened, what would happen under different conditions or how a machine should act when the outcome matters.

> **"Plausibility is sufficient for content generation. It is not sufficient for machines acting in the world."**

That tension runs through everything that follows – the architectures competing to define the category, the money pouring into it, and the widening gap between what these systems can show and what they can actually be trusted to predict.

01 INTELLIGENCE NEEDS SOMEWHERE TO STAND

The modern AI boom has been built on a remarkable discovery: A model trained to predict tokens across enough human-produced material can acquire capabilities that look increasingly like reasoning, writing, coding and problem-solving.

LLMs do not merely retrieve passages from their training data. They construct representations that let them generalize, combine ideas and solve problems they may not have encountered in precisely the same form. Their achievements should not be diminished simply because next-token prediction sits at the center of their training.

But language is an indirect record of reality.

The internet contains descriptions of objects falling, vehicles colliding, people moving through rooms and machines operating in factories. It does not contain a complete record of the forces, states, interventions and consequences involved in those events. Much of what humans know about the world was never written down because it seemed too obvious to record.

A toddler learns that an unsupported object falls, a hidden toy continues to exist and a cup pushed off a table may break. These lessons emerge from observation and interaction before the child can describe gravity, object permanence or material properties. Humans build expectations about the world through experience.

> **"Wordsmiths in the dark" — eloquent and knowledgeable while remaining ungrounded.**
>
> FEI-FEI LI, FOUNDER, WORLD LABS

Fei-Fei Li describes today's LLMs as "wordsmiths in the dark," systems that can be eloquent and knowledgeable while remaining ungrounded.[6] Her argument is that spatial intelligence will allow AI to perceive, generate and reason about three-dimensional environments rather than remain confined to words and flat images.

Yann LeCun has framed the limitation more bluntly: "We are not going to get to human-level AI just by scaling LLMs."[7] His alternative is an architecture that learns abstract representations of the world, predicts how states will change and plans actions without having to generate every pixel or token along the way.

Google DeepMind makes a related case. It defines world models as systems that simulate aspects of an environment so an agent can anticipate how the environment will evolve and how its actions will affect it. DeepMind has called them "a key stepping stone" toward AGI because they could provide an effectively unlimited curriculum of simulated environments in which agents learn.

These claims start from a legitimate weakness in language-centric AI. A model can explain how to stack blocks while failing to control a robot arm that must perceive the blocks, estimate their positions, select a grip and adjust when one begins to slip. Knowledge about a task is not the same as the ability to perform it.

Yet the distinction cannot be reduced to LLMs on one side and world models on the other. Multimodal models already process images, video and audio. Vision-language-action models translate perception and language into robotic actions. Reasoning systems generate and evaluate possible solutions. Digital twins have modeled physical systems for decades. Autonomous-driving companies have long used simulation to train and test vehicles.

World models overlap all of them. The question is not whether LLMs will be replaced. It is whether advanced AI systems need another predictive layer, one that represents changing environments and lets an agent test possible actions before choosing one.

02 AN OLD IDEA REACHES ITS FOUNDATION-MODEL MOMENT

World models did not suddenly appear after ChatGPT.

The concept reaches back through model-based reinforcement learning, cognitive science, control theory, robotics and research into how biological intelligence predicts sensory experience. In model-based reinforcement learning, an agent learns a representation of its environment and uses that representation to plan. Instead of discovering every consequence through real-world trial and error, it can simulate possible actions internally.

Richard Sutton's Dyna architecture connected learning, planning and action in 1990. Other researchers were exploring recurrent systems that learned to predict the consequences of a controller's actions at roughly the same time. The basic idea was already present: An intelligent agent should be able to imagine what may happen before it acts.

David Ha and Jürgen Schmidhuber brought the phrase "world models" to a broader machine-learning audience in 2018.[11] Their system learned compressed spatial and temporal representations of reinforcement-learning environments. An agent could be trained inside the model's generated version of the environment — what the researchers called its "dream" — and then transfer the resulting policy into the actual environment.

The experiment was limited compared with today's ambitions, but the architecture captured the enduring promise. If a machine can learn an accurate internal simulator, it can practice faster, explore dangerous scenarios safely and encounter more variations than reality could economically provide.

What has changed is scale.

Earlier world models operated in narrow environments with limited visual complexity and a relatively small number of possible actions. Today's models can be trained on millions of hours of video, 3D scenes, driving logs, robotic interactions, sensor streams, simulation traces and human demonstrations. Advances in video generation, transformers, diffusion models, self-supervised learning and accelerated computing have made it possible to generate environments that are visually convincing and increasingly interactive.

Google DeepMind's Genie 3 can generate navigable environments from text prompts at 24 frames per second and 720p resolution while maintaining consistency for a few minutes.[8] Runway's GWM-1 generates environments frame by frame in real time and accepts controls that include camera movement, speech and robot commands.[9] Odyssey streams interactive video that changes in response to user input. World Labs' Marble generates explorable 3D worlds from text, images, panoramas and video, then exports them into other design and simulation tools.[10]

These systems are arriving alongside a robotics boom. That matters because a world model without an agent is primarily a simulator. An agent without a world model may act without adequately anticipating the result. Physical AI brings the two together.

> **"We created Cosmos to democratize physical AI and put general robotics in reach of every developer."**
>
> JENSEN HUANG, FOUNDER & CEO, NVIDIA

Nvidia CEO Jensen Huang has called world foundation models fundamental to the development of robots and autonomous vehicles.[12] It is an ambitious claim, but it also reveals Nvidia's commercial strategy. The company does not need to own every robot or world model. It can provide the compute, data-processing pipelines, simulation platforms, digital twins, edge hardware and foundation models used by the companies that build them.

World models are not merely a new class of AI model. They could become a new workload for the entire AI infrastructure stack.

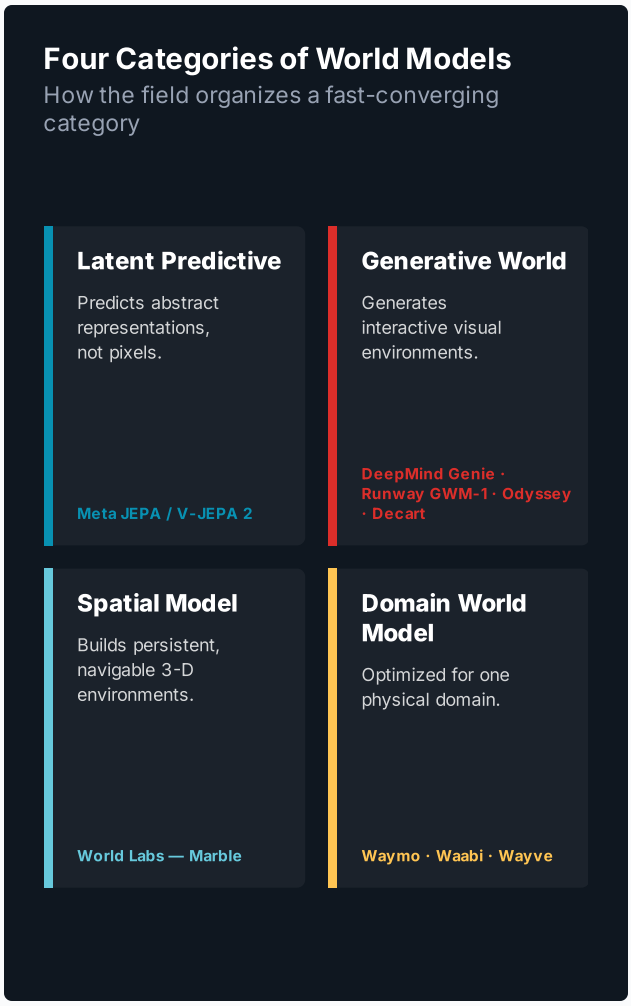
03 WHAT EXACTLY IS A WORLD MODEL?

The term now describes several related but different technologies.

The broadest definition is an internal representation that allows a system to predict how an environment changes over time, including how actions influence those changes. Under that definition, a world model does not need to generate video. It may operate entirely in a compressed latent space that is meaningful to the machine but unreadable to humans.

That definition becomes less precise when products enter the conversation. A generative-video company may call its model a world model because it has learned motion, perspective and common visual patterns. A robotics company may use the term for a system that predicts the next state of a robot and its surroundings. An autonomous-driving company may apply it to a simulator that generates sensor data and responsive traffic. A 3D company may mean a system that constructs persistent environments that can be navigated and edited. A researcher may mean an abstract predictive model used for planning.

These systems can be organized into four broad categories.



The first is the **latent predictive model.** Meta's Joint Embedding Predictive Architecture, or JEPA, is the clearest example. Instead of predicting every pixel in a future image, a JEPA model predicts abstract representations, learning the important structure of the world without spending resources reproducing unpredictable detail.

Meta's V-JEPA 2 was trained using more than one million hours of video and one million images, followed by action-conditioned training using 62 hours of robot data, then used to plan pick-and-place actions in unfamiliar environments – with reported success rates between 65% and 80% on those short-horizon tasks.[13]

The second category is the **generative world.** Google's Genie, Runway's GWM-1, Odyssey's models and Decart's real-time systems generate interactive visual environments. The commercial path begins with gaming, media, virtual production and interactive characters – areas where visual quality can create value before physical accuracy fully arrives.

The third category is the **spatial model.** World Labs is developing models that construct persistent three-dimensional environments rather than only producing sequential frames, placing the company near the intersection of AI generation, 3D design, architecture, simulation and robotics.

The fourth category is the **domain world model.** Waymo, Waabi, Wayve and other autonomous-driving companies build systems optimized for roads, vehicles, sensors and traffic – sacrificing generality for control, measurable performance and specialized operating data.

The categories are beginning to converge. Generative systems are gaining action control. Spatial models are becoming simulation-ready. Robotics systems are using video models as predictive backbones. Domain simulators are incorporating broad visual knowledge from foundation models.

There is still no guarantee they will converge into a universal world model.

> **"One of the big misconceptions is that intelligence is a single thing."**
>
> JOSHUA TENENBAUM, COGNITIVE SCIENTIST, MIT

MIT cognitive scientist Joshua Tenenbaum warns that humans may not carry one all-encompassing simulator in their heads.[14] We construct different models according to context, task and goal – and that may also be the most practical path for AI. A surgical robot, warehouse robot, autonomous vehicle and game character do not need identical representations of the world. What matters to one can be irrelevant to another.

The eventual architecture may be a system capable of creating and updating task-specific world models, not a single digital replica of reality.

04 THE ARCHITECTURAL CONTEST

World models have become part of a larger disagreement about how intelligence should be built.

The dominant AI approach has been autoregressive prediction: Given a sequence, predict what comes next. This method powers LLMs and many video models. It scales well and takes advantage of vast amounts of unstructured data. More data, more parameters and more compute have repeatedly produced better results.

The strongest argument for continuing this path is empirical. Scaling has generated capabilities that many researchers did not expect to emerge so quickly. Multimodal models have acquired visual and spatial knowledge without having been explicitly engineered as classical simulators. Video models learn something about motion and physical regularities because doing so improves their predictions.

The risk is that these models learn the appearance of physical behavior rather than the structure that causes it. A video model knows that a dropped ball usually travels downward because that pattern appears throughout its data. Whether it represents gravity as a stable causal relationship is a different question. If the gravity changes, the surface becomes frictionless or the object's mass is altered, does the model update the outcome correctly — or does it continue producing the most familiar visual pattern?

LeCun's JEPA approach deliberately avoids generating every detail. It predicts representations in latent space, where the model can focus on high-level states and dynamics — intended to make learning more efficient and planning more stable.

Generative approaches take the opposite advantage. They create visible rollouts that humans can inspect, agents can experience and developers can use as synthetic training data, providing a shared interface between model, simulator and user.

THREE COMPETING ARCHITECTURES

- **Autoregressive / generative** — visible, inspectable rollouts; strong for media and synthetic training data.
- **Latent (JEPA-style)** — predicts abstract states, not pixels; efficient, better suited to planning.
- **Hybrid (neural + physics engine)** — learned realism plus classical constraints; Nvidia's Cosmos, Omniverse and Isaac Sim.

Other systems combine learned models with explicit physics engines. A neural model can generate the scene and its variations while a conventional simulator enforces constraints such as collision, mass, friction and joint limits. Nvidia's strategy brings Cosmos world foundation models together with Omniverse, Isaac Sim and digital-twin technology — the learned model provides flexibility and realism, and the simulation stack provides structure and controllability.

This hybrid approach may prove more useful in high-consequence environments than an entirely generative one. The architecture will also depend on the purpose: A creative world generator can tolerate surprising outputs, even benefit from them. A robot-policy evaluator or autonomous-vehicle simulator cannot.

There may therefore be no single winning architecture. The market is not choosing one approach yet. It is financing all of them.

FIELD NOTE — ARCHITECTURES

“There may be no single winning architecture. The market is not choosing one approach yet. It is financing all of them.”

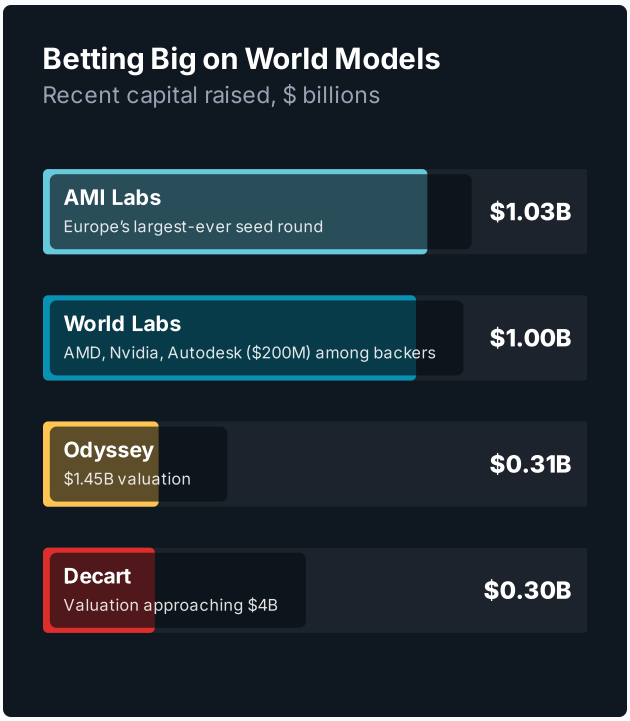
05 THE COMPANIES BUILDING THE NEXT LAYER

Google DeepMind has one of the broadest positions in the field. Genie generates interactive environments. SIMA develops agents that operate in different virtual worlds. Gemini Robotics connects multimodal reasoning to physical action. Waymo contributes a mature autonomous-driving operation, large amounts of sensor data and billions of simulated miles.

In early 2026, Waymo introduced a world model built on Genie 3 and adapted to autonomous driving. It generates camera and lidar outputs, allows engineers to alter scenes through language and driving controls and creates rare events that a fleet may never encounter naturally.[15] A tornado or an elephant on the road makes for an entertaining demonstration, but the strategically important capability is controlled variation – changing weather, vehicle behavior, scene layout or sensor conditions and examining how the driving system responds.

Meta's approach has centered on JEPA and open research. V-JEPA 2 demonstrates that a model trained mostly from passive video can acquire useful representations and then become action-conditioned with a relatively small amount of robot data. The unresolved question is whether this can scale from short tasks toward sustained planning in dynamic environments.

AMI Labs represents a much larger wager on that architecture. LeCun left Meta to build an independent company around world models, persistent memory, reasoning and planning. Its initial industrial emphasis is significant – AMI does not need to win the consumer chatbot market to prove its thesis. It can begin in manufacturing, robotics, healthcare and other domains where language models alone are plainly insufficient.



World Labs is pursuing spatial intelligence. Marble already has an intelligible creative product: Generate a navigable 3D world from text, images or video and use it in design, media, architecture or simulation – a path to revenue without waiting for general-purpose robotics. Its investors reveal the intended platform position. Autodesk wants AI-generated spatial environments inside design workflows; Nvidia and AMD want the compute workloads; robotics and simulation companies need generated environments.

Runway has similarly expanded from video creation into general world models. GWM-1 has variants for explorable worlds, interactive characters and robotics – a logical progression from the capabilities required to generate coherent video over time.

Runway reported that its robotics model produced a 0.95 correlation between simulated and real-world performance across eight robot-manipulation policies, drawing on 1,450 simulated rollouts and more than 16,000 human ratings.[16] The results suggest world models could help rank robot policies before expensive hardware testing, although eight policies on one robot platform do not establish general sim-to-real reliability.

Odyssey is pursuing general, causal and multimodal models. Its leadership has described the objective as finding a "GPT-3 moment" for world models – a scaling point at which a foundation model supports a broad range of applications, not merely a better demo.

Nvidia may have the most strategically complete position. Cosmos supplies open world foundation models. Omniverse provides digital twins and simulation. Isaac supports robotics development. GR00T provides robotic foundation models. Jetson brings inference to physical machines. CUDA, GPUs and cloud partnerships supply the compute underneath everything.

Opening Cosmos is not philanthropy. It expands the number of developers building world-model workloads on Nvidia infrastructure, and a successful open ecosystem can prevent any one closed model provider from controlling the category above Nvidia's hardware.

The autonomous-driving companies have an advantage the general labs do not: years of action-and-consequence data. Waymo, Waabi and Wayve know what their systems observed, what actions were taken and what happened next. That may turn out to be more valuable than having the most photorealistic general model.

06 THE INTERNET OF CONSEQUENCES

LLMs were trained on a vast, accidental corpus created by billions of people writing, coding, photographing, publishing and communicating online.

There is no equivalent internet of physical consequences.

Video contains rich information, but most video is observational. It shows what happened, not the complete state of the environment, the forces involved, every action available or what would have happened if another action had been chosen. A camera may record a person opening a drawer. It does not necessarily reveal the force applied, the resistance of the mechanism, the precise hand position or the unsuccessful attempts that occurred beforehand. It rarely contains synchronized robot actions, tactile signals, depth data and counterfactual outcomes.

World models need more than images of reality. They need trajectories.

A useful training record could include the environment before an action, the action itself, the resulting changes, sensor observations, failures, corrections and rewards. For robotics, that may mean camera, depth, force, joint, gripper and tactile data synchronized over time. For autonomous vehicles, it can include camera, lidar, radar, maps, traffic movement and the vehicle's control decisions. For factories, it may include machine telemetry, maintenance histories, process changes, quality outcomes and operator interventions.

This changes the competitive map. Internet companies began the LLM era with large stores of text, images and user behavior. The world-model era could favor companies controlling fleets, factories, robots, laboratories, warehouses and industrial networks.

WHO OWNS THE PHYSICAL DATA?

- **Tesla** – vehicles and manufacturing plants.
- **Waymo** – autonomous-driving data and simulation experience.
- **Amazon** – warehouses, logistics systems and robots.
- **Siemens, Bosch, Honeywell, Schneider Electric, Rockwell** – industrial environments.
- **Caterpillar, John Deere** – heavy equipment.

The data will not automatically become useful. It is fragmented, proprietary, inconsistent and frequently collected for operations rather than machine learning. An action trajectory from one robot arm may not transfer cleanly to another; sensors differ; factories use different machines, layouts and processes.

Simulation and synthetic data are intended to fill the gaps – but synthetic data introduces a circular risk. The model is trained on incomplete observations, generates additional experiences based on what it learned and then trains an agent on those experiences. If its representation is wrong, the synthetic data can scale the error along with the coverage.

The solution is establishing a real-to-sim-to-real loop: Real systems provide observations and outcomes; simulators generate variations and counterfactuals; policies are tested in simulation; physical deployment produces new evidence about where the model was accurate and where it failed; those results return to the model.

Control of that corpus – an internet of consequences, not a public network but a growing collection of action, state and outcome data – may become one of the most important competitive questions in physical AI.

FIELD NOTE — DATA

“The world-model era may reward control of factories, fleets, laboratories, sensors and simulated environments — not just compute.”

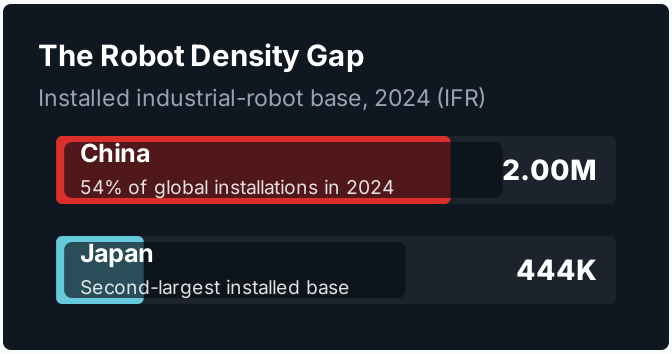
07 CHINA'S PHYSICAL-DATA ADVANTAGE

The world-model contest will not be confined to Silicon Valley.

China has an enormous manufacturing base, a dense robotics supply chain, a rapidly growing humanoid-robot sector and a national strategy emphasizing embodied intelligence. Those advantages do not guarantee leadership in world models, but they create access to something the technology requires: physical deployment at scale.

The International Federation of Robotics reports that China operates roughly two million industrial robots – about 4.5 times the installed base of Japan, the second-largest market. China accounted for 54% of annual global industrial-robot installations, and domestic suppliers represented 57% of installations in the Chinese market in 2024.[17]

Factories, warehouses and robot fleets can become data-generation systems. Every deployed machine can produce trajectories showing what it perceived, what it attempted and what happened. China also has the workforce, manufacturing ecosystem and government-backed coordination to organize large-scale data collection.



That does not mean Chinese companies can simply pool every industrial stream into a national world model. Data quality, interoperability, privacy, commercial incentives and access to advanced compute remain significant constraints. Many industrial robots perform repetitive, tightly programmed tasks that may contribute less generalizable learning than their numbers imply.

But the underlying strategic advantage is real. China can connect AI development to physical production faster than countries that have outsourced much of their manufacturing base.

The United States retains formidable strengths in frontier models, accelerated computing, cloud infrastructure, autonomous driving and research. Nvidia, Google, Meta and an unusually well-funded startup ecosystem give it leadership across several parts of the stack. Japan and Germany bring deep robotics and industrial-engineering expertise. Europe has world-class automation, automotive and manufacturing companies, along with a growing concern about sovereign access to AI infrastructure.

The race may therefore divide by asset. The United States could lead in general models, chips and software platforms. China could lead in manufacturing deployment, robot production and physical-data generation. Germany and Japan could become critical sources of industrial systems and domain knowledge.

The geopolitical contest will also involve standards. Whoever defines formats for robot trajectories, sensor data, simulation environments, policy evaluation and digital twins can shape the ecosystem. World models could consequently shift AI competition closer to the physical economy – not only to data centers, but to factories, roads, ports, hospitals, mines, laboratories and power grids.

08 IS THERE A WORLD-MODEL MARKET?

There is considerable investment in world-model companies. That does not yet mean there is a measurable standalone world-model market.

Most customers do not wake up wanting to buy a world model. They want to design a building faster, create a game world, test a robot, reduce factory downtime, train an autonomous vehicle or generate a scientific hypothesis. The technology will be purchased through those outcomes.

The most immediate market is creative and interactive media. World models can lower the cost of producing 3D environments, virtual sets, game prototypes and interactive experiences. The tolerance for imperfect physics is relatively high, and the value of rapid iteration is clear. World Labs, Runway, Odyssey and Decart can commercialize these capabilities without solving general intelligence.

Architecture, engineering and design offer another early path. Spatial models can turn sketches, images and text into navigable environments — Autodesk's investment in World Labs reflects the potential to insert generative spatial intelligence into established professional workflows.

Robotics training and evaluation may become a more consequential market, although technical requirements are higher. Hardware testing is slow and expensive. If a world model can reliably predict whether a policy will succeed, developers can evaluate many more policies before touching a physical machine. Runway's early policy-correlation results point toward that opportunity.

Autonomous driving is already a world-model market in practice, even if customers never buy a product with that label. Simulation, scenario generation, sensor modeling and counterfactual testing are essential parts of the development stack. Waymo says its vehicles have driven billions of miles in virtual worlds.[15] Waabi built its company around a closed-loop simulator that creates digital twins, generates sensor data and stress-tests its driving system.[18]

> **"Most customers do not wake up wanting to buy a world model. They want to design a building faster, test a robot or reduce factory downtime."**

Digital twins and industrial-process prediction may offer the greatest long-term enterprise value. A world model could help a manufacturer explore process changes, anticipate equipment failures or test robot interactions. In warehouses and logistics, it could evaluate layouts, traffic patterns and automated workflows.

Scientific and medical applications are more speculative but potentially transformative. AlphaFold demonstrated the value of a model that predicts a critical aspect of biology, but a general-purpose scientific world model would need to represent interventions and changing states, not merely infer a structure.

Several business models will coexist: Foundation-model providers can license models and APIs; application companies can embed world-model capabilities into design, media and industrial products; cloud companies can sell training and inference; simulation vendors can combine neural generation with physics engines; robotics and autonomous-driving companies can use world models internally.

There may never be one homogeneous world-model market. World modeling may become a foundational capability embedded across several existing markets — much as machine learning became a capability inside software, cloud, advertising, security and analytics. The label may eventually disappear even if the technology becomes ubiquitous.

09 THE REALITY GAP

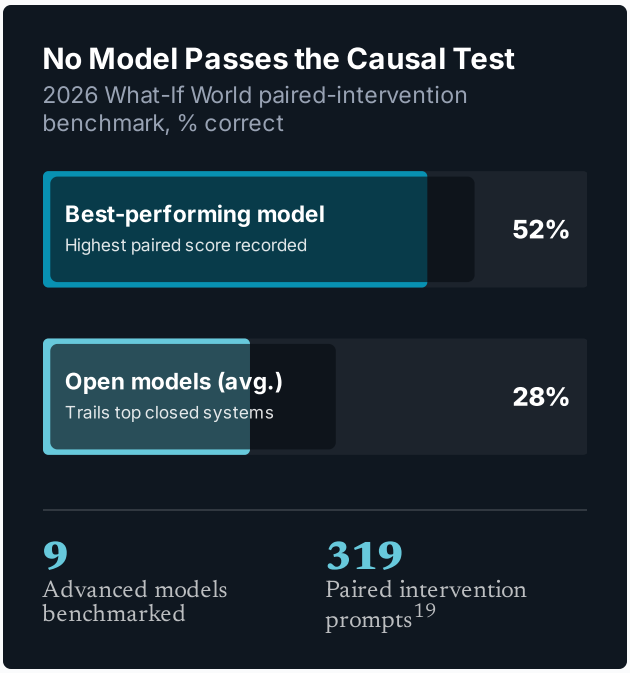
World-model demonstrations are unusually persuasive because humans equate visual coherence with understanding.

If a generated video shows shadows, reflections, motion and object interactions that look correct, we naturally assume the system understands the scene. That assumption is dangerous. A model can produce a visually convincing event by reproducing statistical regularities without representing the causal relationships underneath them – it can know what breaking glass looks like without knowing how force, angle, material and surface determine whether the glass breaks.

This is the visual-conflation problem: Photorealism is mistaken for physical understanding.

> **"It can know what breaking glass looks like without knowing how force, angle, material and surface determine whether the glass breaks."**

The gap becomes apparent under intervention. If gravity changes, does the model adjust trajectories? If friction decreases, does the object slide farther? If the braking force changes, does the vehicle stop at a different location? If an object moves behind another, does the model preserve its identity and state?



FIELD NOTE — THE REALITY GAP

“Photorealism is mistaken for physical understanding. The gap becomes apparent only under intervention.”

The 2026 What-If World benchmark tested nine advanced models using 319 paired prompts in autonomous-driving and robotic-manipulation scenarios. Each pair changed one physical variable while keeping the scene otherwise similar. No model exceeded 52% on the paired score, and open models clustered around 28%.[19] Performance was especially weak when the intervention was physically meaningful but visually subtle.

That is a more demanding test than asking whether one generated video appears realistic. A model must generate the right difference between two outcomes.

Long horizons create another problem. Small errors compound. An object shifts slightly between frames. A hand changes shape. A door forgets whether it is open. A vehicle drifts from its lane. A generated world can remain convincing for seconds while becoming unreliable over minutes. Google DeepMind acknowledges that Genie 3 maintains consistency for only a few minutes — a major improvement for interactive generation, but not enough for a robot working through a shift, a vehicle completing a trip or an industrial process unfolding over hours.

WHERE THE GAP WIDENS

- Small per-frame errors compound silently over long horizons
- Object identity and state can be lost when occlusion occurs
- Genie 3 holds consistency for only a few minutes[8]
- None of the nine benchmarked models cleared 52% on causal pairs[19]

World models also struggle with uncertainty. A system may generate one confident future even when several outcomes are possible. A planner needs to know not only what is likely but also what is uncertain, what could go wrong and when the model is operating outside its experience.

Simulation-to-reality remains the decisive test. A policy can perform beautifully inside a learned world and fail when lighting, materials, sensor noise or human behavior differ from the simulation. The more a model shapes the agent's training, the more important it becomes to identify the model's blind spots.

Evaluation is beginning to improve. Meta released benchmarks covering motion understanding, object interaction and physical reasoning alongside V-JEPA 2, on which the leading AI systems remained substantially behind human performance.[13] Nvidia's PBench evaluates space, time, object permanence and physical laws across driving, robotics, industrial spaces and common-sense scenarios — yet Nvidia itself describes measuring world-model capability as an open research problem.[20]

No universally accepted benchmark captures world-model quality. A serious evaluation framework will need to measure action-conditioned prediction, causal intervention, object permanence, physical invariants, long-term consistency, counterfactual outcomes, closed-loop planning, real-world policy performance and sim-to-real correlation — plus domain-specific safety tests, because the acceptable error rate for a game, robot arm and autonomous vehicle cannot be the same.

Until those standards emerge, vendors can select demonstrations and benchmarks that flatter their systems. The accountability question becomes more serious as these systems move from pixels to actions. Who is responsible when a world model incorrectly predicts a safe outcome and a robot causes harm? The model provider, application developer, equipment manufacturer, system operator and data owner may all play a role.

> **"The accountability question becomes more serious as these systems move from pixels to actions."**

World models may be necessary for AGI. Current world models are not evidence that AGI is imminent.

10 WHAT COMES NEXT: FROM SIMULATION TO INTELLIGENCE

The near-term future of world models is more practical than the rhetoric surrounding them.

Creative systems will generate better interactive worlds. Designers will move from prompts to persistent 3D environments. Studios will create virtual productions faster. Games will become more dynamic. Characters and environments will react in real time.

Robotics teams will use world models for synthetic data and policy evaluation. Autonomous-driving companies will generate rare scenarios and controllable sensor streams. Industrial firms will build bounded models of factories and warehouses. Scientists will develop specialized predictive models for biology, chemistry, weather and materials.

> **"The near-term future of world models is more practical than the rhetoric surrounding them."**

These applications do not require one universal simulator. They require models accurate enough for a defined task, tested against real outcomes and integrated into a feedback loop.

Over the medium term, world models will be combined with language, perception, memory and planning. Language models can interpret goals and explain choices. Perception models can identify the current state. World models can simulate possible futures. Planning systems can compare them. Policy models can translate the selected plan into action.

Agents may then evaluate several possible actions before taking a consequential step. A warehouse robot could simulate alternative paths. A manufacturing system could test process changes. A scientific agent could generate hypotheses, run virtual experiments and decide which physical experiments deserve resources.

The data infrastructure will evolve with the models. Industries will need standardized formats for actions, sensors, states and outcomes. Marketplaces may emerge for licensed robotic trajectories, simulation environments and physical-operating data. Deployed systems will contribute new experience, creating continuous real-to-sim-to-real learning.

Over the longer term, the most important capability may not be carrying one fixed model of the world. It may be constructing the right model for the problem. A generally capable system could identify the relevant objects, rules and uncertainties in an unfamiliar environment, build a task-specific representation, run experiments and update that representation as evidence arrives. It could distinguish what it knows from what it is merely assuming.

That would move AI beyond retrieving and recombining humanity's recorded knowledge. It could generate new knowledge by forming hypotheses and testing their consequences.

This is where world models connect to AGI, superintelligence and recursive intelligence. AGI is usually framed as a system capable of performing a broad range of cognitive tasks at human level. Superintelligence implies performance beyond humanity across most consequential domains. Recursive intelligence describes a trajectory in which systems increasingly contribute to their own improvement and to the discovery of new capabilities.

World models could matter at every stage because they give AI somewhere to experiment. But simulation is not understanding, and a model's imagined world is never the world itself. The more intelligent an agent becomes, the more dangerous it may be to let the agent confuse the two.

The central challenge is therefore not building the most beautiful simulation. It is building a model that knows what it represents, recognizes where it may be wrong and remains connected to evidence from reality.

CLOSING STATEMENT

The LLM era rewarded control of compute, algorithms and humanity's recorded knowledge. The world-model era may also reward control of factories, fleets, laboratories, sensors, spatial platforms and simulated environments.

LLMs learned from what humanity recorded about the world. World models must learn from what the world does: how it changes, how it responds and what happens when an action goes wrong.

THE RACE IS NO LONGER ONLY TO BUILD AI THAT CAN PRODUCE THE NEXT ANSWER. IT IS TO BUILD AI CAPABLE OF IMAGINING WHAT HAPPENS NEXT — AND DECIDING WHETHER IT SHOULD ACT AT ALL.

NOTES & SOURCES

1. World Labs raised $1 billion in February 2026, with AMD, Nvidia and Autodesk ($200M) among the new investors.
   World Labs — "World Labs Announces New Funding"
2. Reuters/U.S. News coverage of the World Labs $1 billion funding round.
   Reuters via U.S. News & World Report
3. Yann LeCun's AMI Labs raised $1.03 billion in what was described as Europe's largest seed round.
   TechCrunch — "Yann LeCun's AMI Labs raises $1.03B to build world models"
4. Odyssey raised $310 million at a $1.45 billion valuation.
   Reuters — "AI lab Odyssey valued at $1.45 billion in latest funding round"
5. Decart raised $300 million at a valuation approaching $4 billion.
   The Wall Street Journal — "Startup Makes Switching AI Chips Easier—and Nvidia Is a New Investor"
6. Fei-Fei Li describes today's LLMs as "wordsmiths in the dark."
   Observer — "Fei-Fei Li's Spatial A.I. Startup World Labs Unveils Its First Product"
7. Yann LeCun: "We are not going to get to human-level AI just by scaling LLMs."
   Reported by The Wall Street Journal; quoted via Gary Marcus, "The False Glorification of Yann LeCun"
8. Google DeepMind's Genie 3 generates navigable environments at 24 fps and 720p, maintaining consistency for a few minutes.
   Google DeepMind — "Genie 3: A new frontier for world models"
9. Runway's GWM-1 generates environments frame by frame with camera, speech and robot controls.
   Runway — "Introducing GWM-1"
10. World Labs' Marble generates explorable 3D worlds from text, images, panoramas and video.
    World Labs — "Marble: A Multimodal World Model"
11. David Ha and Jürgen Schmidhuber introduced "World Models" in 2018.
    arXiv 1803.10122 — "World Models"
12. Jensen Huang: "We created Cosmos to democratize physical AI and put general robotics in reach of every developer."
    Nvidia Newsroom — "NVIDIA Launches Cosmos World Foundation Model Platform"
13. Meta's V-JEPA 2 was trained on 1M+ hours of video and 62 hours of action-conditioned robot data, reporting 65–80% pick-and-place success.
    Meta AI — "Introducing the V-JEPA 2 world model and new benchmarks for physical reasoning"
14. Joshua Tenenbaum: "One of the big misconceptions is that intelligence is a single thing."
    MIT News — New Scientist interview clip with Prof. Joshua Tenenbaum
15. Waymo introduced a world model built on Genie 3, adapted for autonomous-driving simulation, in early 2026.
    Waymo — "The Waymo World Model: A New Frontier for Autonomous Driving Simulation"
16. Runway reported a 0.95 correlation between simulated and real-world performance across 8 robot-manipulation policies, drawing on more than 16,000 human ratings.
    Runway Research — "Accelerating Robot Policy Evaluation with General World Models"
17. China operates roughly 2 million industrial robots, about 4.5× Japan's installed base, and accounted for 54% of global installations in 2024.
    International Federation of Robotics — "Robot Density Surges in Europe, Asia, and Americas"
18. Waabi built a closed-loop simulator that creates digital twins and stress-tests its driving system.
    Waabi — "How Waabi World Works"
19. The 2026 What-If World benchmark tested nine models on 319 paired prompts; no model exceeded 52%, and open models clustered around 28%.
    arXiv 2605.27589 — "What-If World: A Causal Benchmark for General World Models in Embodied Scenarios"
20. Nvidia's PBench evaluates space, time, object permanence and physical laws across driving, robotics, industrial and common-sense scenarios.
    Nvidia Research — "PBench: A Physical AI Benchmark for World Models," Cosmos Lab

COLOPHON

ABOUT THIS REPORT

This special report was researched, written and designed by the editorial team at Techstrong Group, drawing on public statements, company announcements, peer-reviewed research and independent reporting cited throughout the preceding pages.

DESIGN & TYPOGRAPHY

Set in Fraunces (display and headlines), Newsreader (body text) and Inter (labels, captions and data). Charts and diagrams were produced natively for this report. Photographic plates were generated using AI image synthesis for editorial illustration purposes and do not depict real people, products or events.

EDITORIAL STANDARDS

Techstrong Group special reports undergo independent source verification. Where a claim could not be independently confirmed against a primary source at the time of publication, it has been omitted or clearly qualified in the text.

CONTACT

Techstrong Group · DevOps.com · Security Boulevard · Techstrong.tv · Techstrong.ai