TECHNICAL LANDSCAPE REPORT · 2026

State of World Models *2026*

Taxonomy, benchmarks and open challenges: a structured snapshot of a fragmented but converging field.

PUBLISHER
world-models.io

VERSION
1.0

DATE
13 Jul 2026

FORMAT
Landscape report

ENGLISH EDITION
NON-COMMERCIAL REFERENCE DOCUMENT

FRONT MATTER OVERVIEW

Publication information

TITLE	State of World Models 2026: Taxonomy, Benchmarks and Open Challenges
PUBLICATION DATE	13 July 2026
VERSION	1.0
AUTHOR AND EDITOR	Bernard GRENAT
PUBLISHER	world-models.io
CONTACT	contact@world-models.io
CANONICAL PAGE	https://world-models.io/reports/state-of-world-models-2026/
DOI	https://doi.org/10.5281/zenodo.21345187
LICENSE	Creative Commons Attribution 4.0 International (CC BY 4.0)
DOCUMENT TYPE	Landscape report / technical report
LANGUAGE	English
AUDIENCE	Researchers, engineers, analysts and decision-makers looking for a practical reference on world models without hype.
SCOPE	Structured overview of the world models landscape in 2026, covering definitions, model families, benchmarks, evaluation and open challenges across reinforcement learning, robotics, video generation, autonomous driving and embodied AI.
STATUS	Public. Non-commercial reference document. Not a ranking, benchmark or endorsement.

SUGGESTED CITATION

GRENAT, Bernard. (2026). State of World Models 2026: Taxonomy, Benchmarks and Open Challenges (Version 1.0). world-models.io. Zenodo. https://doi.org/10.5281/zenodo.21345187

EDITORIAL NOTICE

This publication is an independent landscape and technical report. It has not been peer reviewed. The classifications and interpretations presented here should be read as a practical editorial framework rather than a definitive academic consensus.

This report is a structured snapshot of a rapidly moving field. It does not claim to be a complete academic survey, an authoritative ranking, or the definitive reference on world models. All external works are credited in the References section.

FRONT MATTER OVERVIEW

Table of contents

FRONT MATTER

OVERVIEW

Abstract

World models are becoming one of the most interesting (and, honestly, one of the most confusing) ideas in artificial intelligence.

At a simple level, a world model is an AI system that tries to learn how an environment actually works. It can predict what may happen next, simulate possible futures, support planning, or help an agent understand the consequences of its actions. It sounds simple, but in practice the field covers quite different things: reinforcement learning, robotics, autonomous driving, generative video, embodied AI, simulation, spatial intelligence and autonomous agents.

This report provides a structured overview of the world models landscape as it stands today, in 2026. It doesn't try to say that one definition is the only correct one. It rather tries to organize the field, explain the main model families, compare the emerging benchmarks and point out the open problems.

> KEY MESSAGE
>
> Visual realism, by itself, is not enough. A model can generate an impressive-looking video and still fail to behave like a real world model. Useful world models need temporal coherence, physical consistency, action sensitivity and, most importantly, real functional utility.

The goal of this report is simply to give a practical reference for people who want to understand this field without getting lost in hype, jargon, or overly narrow technical definitions.

DOMAINS	FAMILIES	BENCHMARKS	OPEN CHALLENGES
08	06	03+	07

Executive summary

A two-page overview of what world models are, why they matter in 2026, and where the field still falls short, with anchors into every major part of this report.

In this report

Key takeaways

01 · DEFINITION

A world model is a system that *learns how an environment works*, not just how it looks. See §2.

02 · TAXONOMY

Five axes (domain, function, representation, time horizon, action conditioning) describe every model. See §5.

03 · EVALUATION

Benchmarks are still immature; visual realism alone predicts very little functional utility. See §8.

04 · CHALLENGES

Seven structural problems, from long-horizon reasoning to action conditioning, remain open. See §9.

MAIN CONCLUSION

Visual realism is not sufficient. A world model must be judged on temporal coherence, physical consistency, action sensitivity and, above all, its functional utility inside a decision loop. Universal rankings should be treated with caution: no single benchmark yet covers the whole space.

How to read this report

The document is organized in five parts plus back matter. Readers new to the topic can start with `§1 Introduction` and `§2 What is a world model?`. Practitioners looking for the working taxonomy can jump directly to §5 and the `classification framework in §12`. Researchers focused on evaluation should read §7–§8 alongside `§9 Open challenges`. All external works are credited in the `References`.

This taxonomy should be read as a practical framework, not a final authority. The field remains fragmented across communities, and definitions still vary. Page numbers above point to the first page of each part.

Part I

Framing the *field*

A field defined by many communities at once (reinforcement learning, robotics, video generation, autonomous driving and embodied AI), each carrying its own working definition of what a world model actually is.

SECTION 01

PART I · FRAMING THE FIELD

Introduction

Artificial intelligence has made major progress on text, images, code, speech and video. Large language models can answer questions, write software, analyze documents and use tools. Image and video models can generate scenes that, only a few years ago, would have looked almost impossible.

But one important question still remains: do these systems really understand the world, or do they mostly reproduce patterns? World models are one way to approach this question.

The idea isn't only to generate an answer, an image or a video. The idea is to learn how an environment actually behaves. A world model should be able to predict what happens next, imagine possible scenarios, evaluate the outcome of an action, or help an agent plan before acting. That's why the concept keeps showing up in many different areas: robotics, autonomous vehicles, games, video generation, embodied AI, reinforcement learning and agent systems.

The modern deep learning use of the term was strongly influenced by the now well-known 2018 paper World Models by David Ha and Jürgen Schmidhuber [01]. In that work, an agent learned an internal model of its environment and used it to train inside its own compressed representation of the world. Since then, the term has expanded well beyond its original reinforcement learning context.

Today, "world model" can mean several different things depending on who uses the term. For a robotics researcher, it may mean an action-conditioned model used to evaluate robot behavior. For a video generation lab, it may mean a model able to generate coherent dynamic scenes. For autonomous driving, it may mean a learned simulator for traffic and rare events. For agent systems, it may mean a representation of task states and possible actions.

This diversity is useful, but it also creates a lot of confusion. This report tries to bring at least some structure to that landscape.

> READING NOTE
>
> This taxonomy should be read as a practical framework, not a final authority. The field remains fragmented, definitions vary a lot across communities, and any classification is a working tool rather than a settled truth.

SECTION 02

PART I · FRAMING THE FIELD

What is a world model?

WORKING DEFINITION

A **world model** is an AI model that learns a representation of an environment and uses it to predict, simulate, evaluate or support action inside that environment.

This definition is broad on purpose. It includes many kinds of models, but not everything. A world model can be:

- a model-based reinforcement learning system
- a video prediction model
- an embodied AI model
- a robotics simulator learned from data
- an autonomous driving scenario model
- a spatial or 3D scene model
- a procedural planning model
- an internal simulator used by an agent

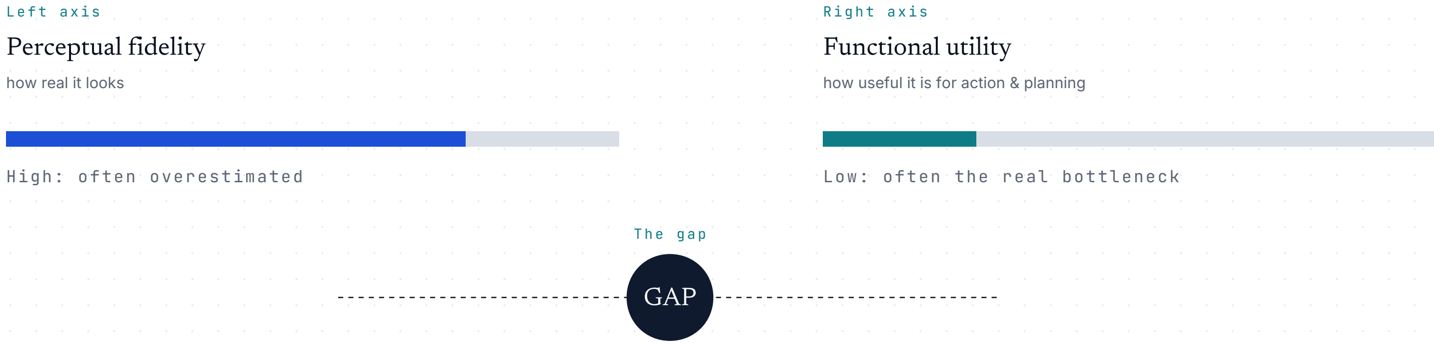
But a model is not automatically a world model just because it generates something visual. For example, a video model may generate a beautiful scene. But if objects disappear, physics are wrong, actions have no clear consequences, or the scene is inconsistent after a few seconds, then the model is probably not a strong world model.

This is an important distinction. In 2026, one of the biggest risks in the field is to confuse visual quality with world understanding. A world model should not only look right. It should behave right, at least in the context where it is supposed to be used.

FIG. 01 · PERCEPTION VS. FUNCTIONALITY

The perception–functionality gap

A world model can score high on visual quality and still fail on the functional side; the two axes are correlated only weakly.



Perceptual quality alone does not predict functional utility; the two axes must be evaluated separately.

SECTION 03

PART I · FRAMING THE FIELD

Why world models matter

World models matter because they can help AI systems move from passive generation to active understanding and planning.

A language model can describe a robot picking up an object. A video model can generate a clip of that action. But a useful world model should help answer a more operational question:

> THE OPERATIONAL QUESTION
>
> What will happen if the robot tries to pick up this object from this angle, with this force, in this environment?

That question is much harder. It requires some representation of objects, space, time, actions, physical constraints and uncertainty. It also requires the model to generalize beyond one specific example.

This is why world models are important for:

- robots that need to act safely
- autonomous vehicles that need to simulate rare scenarios
- agents that need to plan multi-step tasks
- games and simulations that need dynamic environments
- industrial systems that need predictive models
- video models that aim to become more than media generators

The promise is not only to make models more impressive. The promise is to make them more useful for decision-making.

SECTION 04

PART I · FRAMING THE FIELD

A short historical background

World models did not appear suddenly with generative video. The idea is older, especially in reinforcement learning. In model-based reinforcement learning, an agent learns a model of the environment. Instead of learning only by direct trial and error, it can use this model to simulate possible futures and improve its decisions.

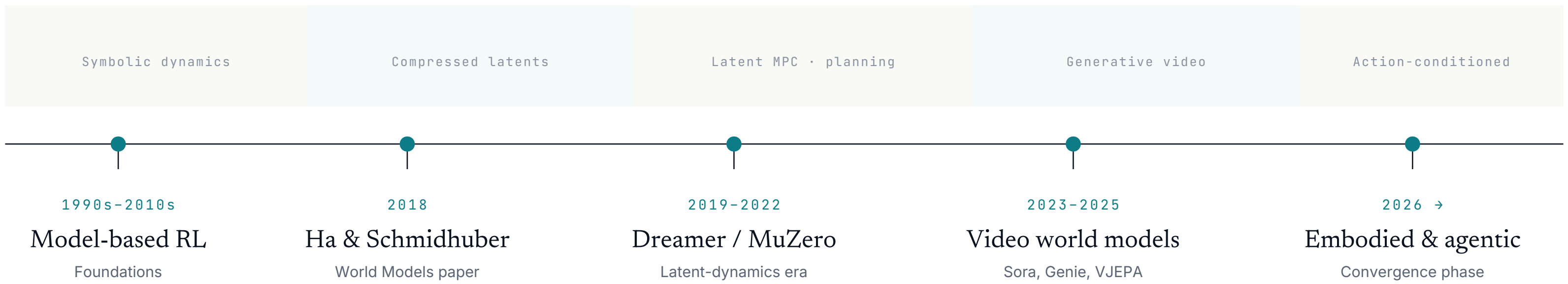
The 2018 World Models paper made this idea more visible in deep learning. The model learned a compressed representation of the environment, then used this internal simulation to train a controller. This was interesting because the agent could, in a way, "dream" inside its learned world.

After that, the concept started to expand. With the rise of large-scale generative models (especially video models), people began to ask whether these systems could learn something about the physical world. If a model can generate a realistic video of a car turning, a ball falling, or a person moving through a room, does it mean it has learned some world structure? The answer is: maybe, but not necessarily. This is exactly why the field needs better definitions, better taxonomies and better benchmarks.

FIG. 02 · FIELD TIMELINE

From model-based RL to embodied world models

Four eras of increasingly capable environment models, from symbolic dynamics to action-conditioned generative systems.



Part II

A practical *taxonomy*

Five axes to describe a world model (domain, function, representation, time horizon, action conditioning), followed by six model families that structure most of the current work.

SECTION 05

PART II · A PRACTICAL TAXONOMY

A practical taxonomy of world models

The world models landscape can be organized around several dimensions. None of them is enough on its own, but together they give a workable framework for comparison.

5.1 Domain

A world model can belong to different domains. This matters because models should not be compared without context. A model built for driving simulation and a model built for video generation may both be called "world models", but they are not solving the same problem.

DOMAIN	TYPICAL USE
Reinforcement learning	Planning, policy learning, sample efficiency
Robotics	Manipulation, navigation, action evaluation
Autonomous driving	Scenario simulation, rare event generation
Video generation	Dynamic scene prediction
Games	Interactive environments and agent training
Embodied AI	Perception, action and planning
Industrial simulation	Process modeling and predictive scenarios
Spatial intelligence	3D reasoning, navigation and scene dynamics

5.2 Function

World models can serve different functions. A strong world model may combine several of these, but in practice most systems are good at some things and weak at others.

FUNCTION	MEANING
Prediction	Forecasting future states
Simulation	Generating possible environment trajectories
Planning	Helping an agent choose actions
Policy evaluation	Testing possible decisions before execution
Synthetic data	Creating useful training or evaluation data
Representation learning	Learning compressed environment states
Counterfactual reasoning	Comparing what could happen under different choices

5.3 Representation

World models do not all represent the world in the same way. Some use latent vectors. Others use video frames, tokens, 3D scenes, object-centric states or symbolic structures. The representation matters because it influences what the model can do.

A pixel-level video model may look visually impressive but be hard to use for planning. A latent dynamics model may be useful for reinforcement learning but not human-readable. A 3D model may capture spatial structure but still miss causality or physical behavior. There is no perfect representation yet.

5.4 Time horizon

Some models predict the next state. Others try to simulate longer trajectories. This creates a major difference:

- Short-term prediction is easier
- Medium-term prediction is harder
- Long-horizon planning is still very difficult
- Procedural reasoning requires even more abstraction

A model can be good at predicting the next few frames and still fail when the task requires long-term consistency. This is one of the reasons why benchmarks focused only on short clips can be misleading.

5.5 Action conditioning

Action conditioning is probably one of the most important criteria. A passive model predicts what may happen next. An action-conditioned model predicts what may happen if a specific action is taken.

> KEY CRITERION
>
> For robotics, autonomous driving and agent planning, action conditioning is essential. A model that cannot understand the effect of actions is limited: it may describe or generate a world, but it cannot reliably help an agent act inside it.

FIG. 03 · TAXONOMY AXES

Five axes to describe a world model

Domain, function, representation, time horizon and action conditioning. The fifth axis is the one most benchmarks still refuse to test.

Axis 01

Domain

· RL · Robotics
· Video · Driving
· Embodied · 3D

Axis 02

Function

· Predict · Simulate
· Plan · Evaluate
· Reason

Axis 03

Representation

· Latent · Pixel
· Token · 3D
· Symbolic · Hybrid

Axis 04

Time horizon

· Short · Medium
· Long · Procedural

Axis 05

Action conditioning

Yes / No
Partial
Unknown

● Critical criterion

Axis 05 is highlighted because most current benchmarks still avoid measuring action conditioning directly.

SECTION 06 PART II · A PRACTICAL TAXONOMY

Main families of world models

6.1 Reinforcement learning world models

This is the most historical family. In reinforcement learning, world models help agents learn environment dynamics. The agent can use the learned model to simulate trajectories, plan actions or improve sample efficiency.

The core benefit is simple: real-world interactions can be expensive, dangerous or slow. If the model can simulate enough of the environment, the agent can learn more efficiently. These models are often less visually impressive than video models, but they may be more useful for decision-making.

6.2 Video world models

Video world models are attracting a lot of attention because generative video has improved rapidly. These models try to generate or predict dynamic visual scenes. Some can produce visually rich sequences that appear coherent at first glance. But the hard question is whether they capture actual world dynamics or only surface-level patterns.

A video world model should ideally preserve:

- object permanence
- temporal coherence
- spatial consistency
- action effects
- physical plausibility
- causal relations

This is where many current systems still struggle. A model can generate a stunning scene but still fail in small details that matter for world modeling. A hand may change shape. An object may disappear. A vehicle may move in an impossible way. The scene looks good, but the world is not stable.

6.3 Embodied world models

Embodied world models are built for agents that perceive and act. This includes robots, navigation systems and embodied AI agents. In this context, the output does not only need to look plausible. It needs to be useful. A robot does not need a world model to create a pretty video. It needs a model that helps it decide what to do next.

> THE RIGHT QUESTION
>
> Does this model improve action, planning or policy evaluation? If the answer is no, then the model may be visually strong but functionally weak.

6.4 Autonomous driving world models

Autonomous driving is one of the most obvious use cases. Driving involves complex environments, moving agents, rules, uncertain behavior and rare events. A useful world model can help simulate scenarios that are difficult or dangerous to collect in real life. It can support traffic prediction, rare scenario generation, policy testing, synthetic data creation, safety validation and simulation of edge cases. But here the tolerance for mistakes is low. A driving world model must be realistic not only visually, but behaviorally and causally.

6.5 Spatial and 3D world models

Spatial world models focus on the structure of environments. They may represent geometry, rooms, objects, surfaces, affordances and navigable spaces. This is important for robotics, augmented reality, game environments, digital twins and embodied agents. However, spatial understanding alone is not enough. A static 3D scene is not necessarily a world model. To become one, the system also needs some understanding of time, change, interaction and action.

6.6 Agentic and procedural world models

Not all worlds are physical. For autonomous agents, a "world" can also be a software environment, a workflow, a task state, an interface or a procedure. In that case, the world model needs to understand what state the task is in, what actions are possible, what each action may change, what constraints apply and how to reach a goal.

This is especially relevant for agents that operate in digital tools, browsers, enterprise systems or multi-step workflows. Here again, the challenge is long-horizon planning: many models can describe a procedure, but fewer can reliably predict and execute it across many steps.

FIG. 04 · MODEL-FAMILY LANDSCAPE

Six families across the world models landscape

Bubble size ≈ current visibility; positions reflect a rough perceptual-fidelity vs. functional-utility trade-off.

A useful world model tends to sit in the top-right quadrant, with high perceptual fidelity and high functional utility.

Part III

Benchmarks & *evaluation*

Emerging benchmarks are starting to formalize what a world model is expected to do and, more importantly, what it is not expected to be judged on.

SECTION 07 PART III · BENCHMARKS & EVALUATION

Benchmarks and evaluation

World models are hard to benchmark because there is no single task that covers the entire field. A good benchmark needs to answer one question clearly: what kind of world modeling ability are we testing? Some benchmarks focus on video realism. Others focus on physics. Others focus on action-conditioned prediction, embodied utility or long-horizon planning.

7.1 WorldModelBench

WorldModelBench [02] evaluates video generation models as world models. Its main point is important: video quality is not enough. A model should also be evaluated on instruction following, physical plausibility and domain-specific world behavior. This is useful because many video models look impressive while hiding subtle world inconsistencies.

7.2 WorldPrediction

WorldPrediction [03] focuses on high-level world modeling and long-horizon procedural planning. This is interesting because it moves beyond pixels. It asks whether a model can understand what actions connect one state to another, closer to planning than simple video prediction. Many current models still struggle with this kind of reasoning.

7.3 WorldArena

WorldArena [04] evaluates embodied world models. Its key contribution is the distinction between perceptual quality and functional utility. This may be one of the most important ideas in the field. A model can produce good-looking predictions and still fail to help an embodied agent. For robotics, this gap is critical: a world model must not only represent the environment, it must help the agent act better.

7.4 Why evaluation is still immature

Benchmarking world models is still early. There are several reasons:

- definitions are not stable;
- long-horizon prediction is fragile;
- models are very different from each other;
- action-conditioned data is expensive;
- many frontier systems are closed;
- visual metrics are not enough;
- physical consistency is hard to measure;
- functional testing is difficult to standardize.

This means that current results must be interpreted carefully. No benchmark today fully captures what "world understanding" means.

> WARNING · UNIVERSAL RANKINGS
>
> A single global ranking of "the best world model" would probably be misleading. A robotics model and a video generation model do not have the same purpose and should not be judged with the same criteria.

FIG. 05 · BENCHMARK COMPARISON MATRIX

What each benchmark actually tests

A quick side-by-side of the four main emerging benchmarks and their distinctive emphases.

BENCHMARK	PRIMARY TARGET	FOCUS	NOTABLE EMPHASIS
WorldModelBench	Video generation models	Video realism + physics	Beyond visual quality
WorldPrediction	High-level reasoning models	Long-horizon planning	Procedural under-standing
WorldArena	Embodied world models	Perception + utility	Perception-function gap
WorldArena 2.0 [08]	Extended embodied setups	Modality · platform · func-tionality	Broader coverage

SECTION 08 PART III · BENCHMARKS & EVALUATION

The main evaluation problem

The biggest problem is that world models are often judged by how they look. This is understandable. Humans are visual. If a generated scene looks real, we tend to assume the model understands something. But this can be misleading.

A useful world model should be judged on deeper criteria:

CRITERION	WHY IT MATTERS
Temporal coherence	The world must stay stable over time
Physical consistency	Objects should obey plausible constraints
Object permanence	Things should not disappear randomly
Action sensitivity	Actions should have clear effects
Causal plausibility	Events should follow understandable relations
Planning utility	The model should help decision-making
Generalization	It should work outside narrow examples
Functional value	It should improve downstream tasks

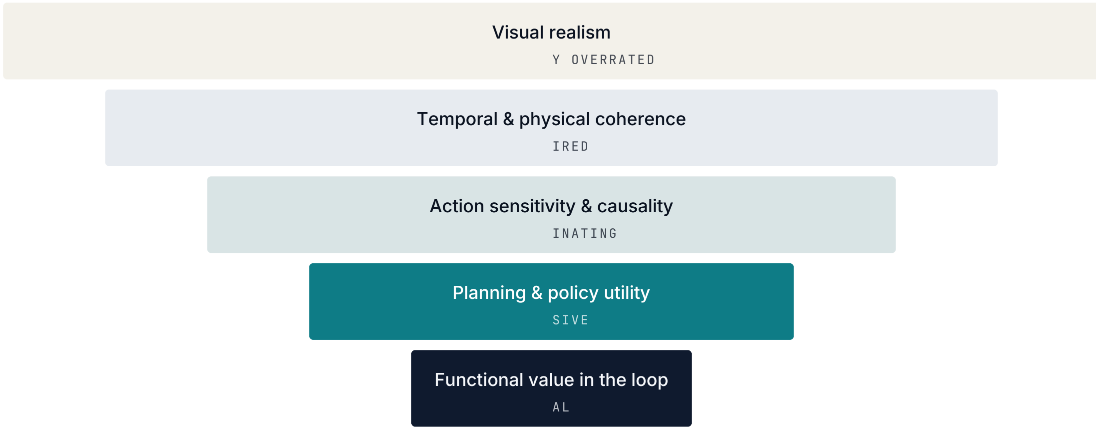
MAIN CONCLUSION

In simple words: a world model should not only be beautiful. It should be useful. That is probably the most important message of this report.

FIG. 06 · EVALUATION PYRAMID

Why visual realism is not enough

Each layer above the base is harder to achieve, and closer to what world understanding actually requires.



Visual realism is at the base of the pyramid: necessary in some settings, but never sufficient on its own.

Part IV

Challenges & *use cases*

Seven open problems, from unstable definitions to long-horizon reasoning and sim-to-real transfer, plus where world models are already becoming operationally useful.

Open challenges

Seven recurring problems shape most conversations about world models today. None of them is fully solved.

9.1 The definition is still unclear

Different communities use the term "world model" differently. This is not necessarily bad, but it makes comparison difficult. A reinforcement learning researcher, a robotics engineer and a video generation team may all use the same term while meaning different things. The field needs more precise language.

9.2 Video generation is not world modeling by default

A video model can be related to world modeling, but it is not automatically a world model. It needs to capture dynamics, consistency and possibly action consequences. Otherwise, it is a generative media model with world-like appearances.

9.3 Long-horizon reasoning is still weak

Long-horizon coherence remains very difficult. Models often drift over time. Small errors accumulate. Objects change. State becomes unstable. This is a serious limitation for planning.

9.4 Action conditioning is underdeveloped

For real agents, prediction must be tied to action. The question is not only "what happens next?" but "what happens if I do this?" Many systems still do not handle this well.

9.5 Sim-to-real transfer is hard

In robotics and autonomous driving, a world model must eventually connect to reality. A model that works in simulation may fail in the real world because of noise, unexpected situations, sensor differences or physical complexity. This remains one of the hardest problems.

9.6 Data is a bottleneck

World models need rich data. Text and static images are not enough. Useful data may include videos, actions, robot trajectories, 3D scenes, sensor streams, driving logs, simulation traces and human demonstrations. This kind of data is harder to collect, clean and share than text.

9.7 Safety is not optional

If a world model is used for planning, then wrong predictions can lead to wrong actions. This is especially important in robotics, driving and industrial systems. Uncertainty estimation, failure detection and safety constraints will be required.

> CONVERGENCE
>
> Reinforcement learning, robotics, video generation, autonomous driving, simulation and agents are all moving toward the same underlying question: can AI systems learn useful predictive models of the environments they operate in?

FIG. 07 · OPEN CHALLENGES MATRIX

Seven problems shaping the next iteration

Each challenge maps to a different failure mode: some structural, some critical, none yet fully solved.

CHALLENGE	WHY IT IS HARD	SEVERITY
Unclear definitions	Different communities, different meanings	Structural
Video ≠ world modeling	Realism does not imply understanding	Critical
Long-horizon reasoning	Errors accumulate, state drifts	Critical
Action conditioning	Underdeveloped in most video models	High
Sim-to-real transfer	Noise, sensor gaps, physical complexity	High
Data availability	Video, actions and trajectories are scarce	Structural
Safety & uncertainty	Wrong predictions can drive wrong actions	Non-optional

Practical use cases

World models are not only a research idea. They may become useful in several fields.

Robotics

Robots can use world models to test actions before executing them. This can reduce trial and error and improve safety.

Autonomous driving

Driving systems can use world models to simulate rare events, test policies and generate synthetic scenarios.

Video and simulation

Video models may evolve toward controllable simulation if they become more physically consistent and action-aware.

Games

World models can support dynamic environments, agent training and procedural generation.

Industrial systems

Industrial world models may help simulate processes, detect risks or predict operational outcomes.

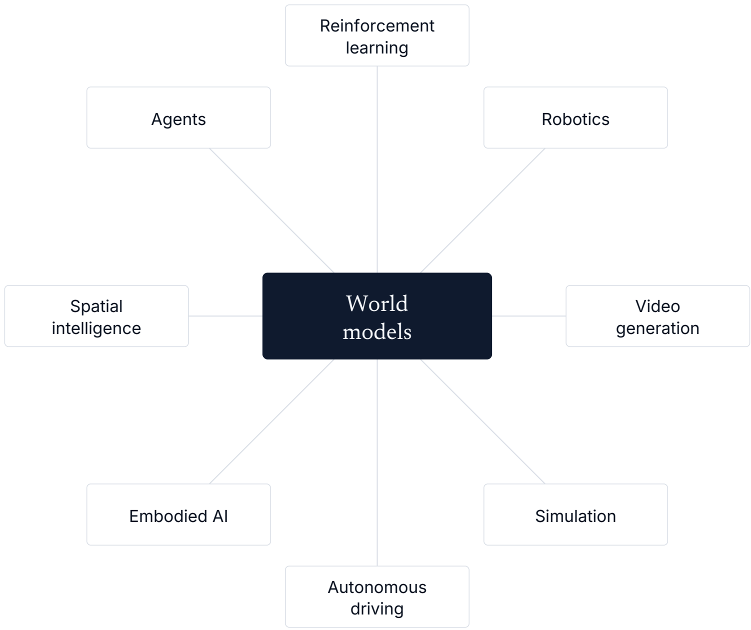
Autonomous agents

Software agents may use world models to represent task states, workflows, interfaces and possible next actions.

FIG. 08 · CONVERGENCE DIAGRAM

Multiple fields converging toward world models

Eight fields, one convergent question: each community brings its own methods, benchmarks and constraints.



Eight fields, one convergent question. Each community brings its own methods, benchmarks and constraints.

Part V

Method & *outlook*

Methodology, a proposed classification framework, five strategic signals for 2026 and a closing view on where the field is likely to go next.

Methodology

This report is based on a structured review of public research papers, benchmark pages, model documentation and technical project pages. The objective is not to provide a complete academic survey; the field is too wide and moving too fast for that. The objective is more practical: provide a map of the main ideas, families, benchmarks and open problems.

Inclusion criteria

A model, benchmark or project is considered relevant if it deals with at least one of these topics:

- prediction of future environment states;
- autonomous driving simulation;
- learned environment dynamics;
- video generation evaluated as world modeling;
- action-conditioned simulation;
- procedural or long-horizon planning;
- embodied AI;
- spatial or temporal environment representation;
- robotics world modeling.

Exclusion criteria

- static image models without dynamics
- generic language models unless used for world-state reasoning
- classic simulators without learned modeling
- pure marketing claims without technical information
- projects with too little public documentation

Limitations

This report has limits. The field is moving quickly. Some models are proprietary. Some claims are difficult to verify. Benchmarks are still young. The term "world model" is also used in different ways by different communities. This report should therefore be read as a structured snapshot, not as a final authority.

SECTION 12

PART V · METHOD & OUTLOOK

Proposed classification framework

To track the field more clearly, world-models.io proposes a simple classification framework. Each model or benchmark can be described through the following fields.

FIELD	DESCRIPTION
Name	Model or benchmark name
Organization	Lab, company or institution
Year	Publication or announcement year
Domain	Robotics, video, driving, RL, embodied AI, etc.
Input	Text, image, video, action, state, sensor
Output	Video, state, action, trajectory, simulation
Action-conditioned	Yes, no, partial, unknown
Representation	Latent, pixel, token, 3D, symbolic, hybrid
Horizon	Short, medium, long, procedural
Evaluation	Visual, physical, functional, planning
Availability	Paper, code, demo, dataset, closed
Notes	Main limits or context

ON RANKINGS

This framework is a description tool, not a scoring system. A universal ranking of all world models would likely be misleading: a robotics model and a video generation model do not share the same purpose and cannot be judged with the same single criterion.

Strategic signals for 2026

Several signals show that world models are becoming an important category in AI.

> SIGNAL 01
>
> Dedicated benchmarks are appearing. The field is becoming measurable, even if imperfectly.

> SIGNAL 02
>
> Video generation is discussed as simulation, not only as media creation.

> SIGNAL 03
>
> Robotics and embodied AI need models that connect perception and action.

> SIGNAL 04
>
> Autonomous driving still requires better ways to simulate rare and dangerous situations.

> SIGNAL 05
>
> Long-horizon planning remains a major weakness across most current systems.

The broader trend is convergence. Reinforcement learning, robotics, video generation, simulation and agent systems are all moving toward the same question: can AI systems learn useful predictive models of the environments they operate in? That question will likely become more important in the coming years.

SECTION 14 PART V · METHOD & OUTLOOK

Conclusion

World models represent an important shift in artificial intelligence. They are not just about generating outputs; they are about learning environments, predicting change, simulating futures and supporting action.

The field is still fragmented. Definitions vary. Benchmarks are young. Many claims are hard to evaluate. Some models look impressive but may not be functionally useful. Still, the direction is clear. As AI systems become more agentic, embodied and interactive, they will need better internal models of the worlds they operate in. This may include physical worlds, digital worlds, simulated worlds and procedural task environments.

The main challenge is to evaluate world models properly. A strong world model should not only generate convincing images or videos. It should be coherent over time, physically plausible, sensitive to actions and useful for decision-making.

CLOSING STATEMENT

In other words, the future of world models will not be measured only by how real they look, but by how well they help systems understand, plan and act.

SECTION R

BACK MATTER

References

01 Ha, D., & Schmidhuber, J. (2018). World Models. arXiv:1803.10122.

02 Li, D., Fang, Y., Chen, Y., Yang, S., Cao, S., Wong, J., Luo, M., Wang, X., Yin, H., Gonzalez, J. E., Stoica, I., Han, S., & Lu, Y. (2025). WorldModelBench: Judging Video Generation Models As World Models. arXiv:2502.20694.

03 Chen, D., Chung, W., Bang, Y., Ji, Z., & Fung, P. (2025). WorldPrediction: A Benchmark for High-level World Modeling and Long-horizon Procedural Planning. arXiv:2506.04363.

04 Shang, Y., Li, Z., Ma, Y., Su, W., Jin, X., Wang, Z., Zhang, X., Tang, Y., Gao, C., Wu, W., Liu, X., Shah, D., Zhang, Z., Chen, Z., Zhu, J., Tian, Y., Chua, T.-S., Zhu, W., & Li, Y. (2026). WorldArena: A Unified Benchmark for Evaluating Perception and Functional Utility of Embodied World Models. arXiv:2602.08971.

05 WorldModelBench project page. Judging Video Generation Models As World Models. See ref. 02.

06 WorldPrediction project page. A Benchmark for High-Level World Modeling and Long-Horizon Procedural Planning. See ref. 03.

07 WorldArena project page. A Unified Benchmark for Evaluating Perception and Functional Utility of Embodied World Models. See ref. 04.

08 WorldArena 2.0 project page. Extending Embodied World Model Benchmarking on Modality, Functionality and Platform.

SECTION C · BACK MATTER

Suggested citation

If you refer to this report in academic, technical or editorial work, please use the citation format below.

APA-STYLE

GRENAT, Bernard. (2026). State of World Models 2026: Taxonomy, Benchmarks and Open Challenges (Version 1.0). world-models.io. Zenodo. https://doi.org/10.5281/zenodo.21345187

SHORT FORM

Grenat, B. State of World Models 2026, v1.0, world-models.io, 2026. DOI: 10.5281/zenodo.21345187.

BIBTEX

```
@techreport{grenat2026worldmodels,
  title       = {State of World Models 2026: Taxonomy, Benchmarks and Open
Challenges},
  author      = {Grenat, Bernard},
  year        = {2026},
  month       = {7},
  institution = {world-models.io},
  type        = {Landscape report},
  version     = {1.0},
  doi         = {10.5281/zenodo.21345187},
  url         = {https://doi.org/10.5281/zenodo.21345187},
  note        = {Version 1.0. Licensed under CC BY 4.0.}
}
```

SECTION A

BACK MATTER

Appendix A: Glossary

TERM	DEFINITION
World model	A model that learns a representation of an environment and uses it to predict, simulate, evaluate or support action.
Model-based reinforcement learning	A reinforcement learning approach where an agent learns a model of the environment and uses it for planning or policy improvement.
Embodied AI	AI involving agents that perceive, act and interact inside physical or simulated environments.
Action-conditioned prediction	Prediction of future states based on a specific action or sequence of actions.
Perceptual fidelity	The visual or sensory realism of a generated output.
Functional utility	The usefulness of a model for downstream tasks such as planning, policy evaluation or decision-making.
Long-horizon planning	Planning over extended sequences of actions or states.
Sim-to-real transfer	The ability of models or policies trained in simulation to work in the real world.

SECTION B · BACK MATTER

Appendix B: Future work

Future versions of this report may include the following extensions. None of them is a commitment; they describe the direction in which world-models.io intends to develop this reference.

01 · DATA

A public dataset of world models and benchmarks with structured metadata.

02 · COMPARISON

A comparative matrix of major model families across the taxonomy axes.

03 · TIMELINE

A visual timeline documenting how the field evolves year by year.

04 · BENCHMARKS

A benchmark-by-benchmark evaluation guide describing what each one measures.

05 · VOICES

Interviews with researchers and practitioners across labs and industries.

06 · VERTICALS

Separate reports on robotics, video world models and autonomous driving.

07 · BIBLIOGRAPHY

A maintained bibliography with structured tags and open-access links.

08 · PROCESS

A public changelog and contribution process open to the research community.

END OF REPORT

This concludes State of World Models 2026: Taxonomy, Benchmarks and Open Challenges, v1.0, published by world-models.io. Comments, corrections and contributions are welcome for future versions.