

World Models Should Learn to Perceive Before They Act

Jongseo Lee

Vision and Learning Lab, Kyung Hee University

jong980812@khu.ac.kr <https://jong980812.github.io>

Abstract

World models have become a central idea in AI, and recent work has increasingly judged them by the decisions they enable: a world model is as good as the agent it serves. This position paper argues that the field has been learning to act before it has learned to perceive. The encoders that drive robot planners still fail basic physics, and many systems now called world models are closer to what control theory calls plant models: predictors for one robot in one workspace. Benchmarks have recorded these failures but have rarely explained them, and the papers that build world models have seldom tested what their models understand. A parallel trend toward one model that learns everything at once, language included, has made failures hard to trace. I propose a cognition-first alternative. I define a world model as a system that keeps a belief about the present, including what is out of view, backed by memory; imagines how that state changes; and holds it in a form it can use for new questions. I state five principles, each with a test: perceive first, imagine the unseen, stay at human scale, learn through a borrowed body, and grow rather than unify. In particular, I argue that uncertainty should live in the belief about the present, not in the dynamics: a world model should represent many possible presents, not many possible futures. I then sketch a blueprint and a developmental roadmap, and list research directions that today’s models can pursue.

Prologue

This is not a technical paper. It reports no experiments and proposes no new algorithm. It is a position paper in the spirit of LeCun’s *A Path Towards Autonomous Machine Intelligence* (LeCun, 2022): an attempt to think in the open about what a world model should be and in what order it should acquire its abilities. Many authors have proposed almost every idea here before, in many forms, and I claim priority for none of them. What I offer is a way of putting these ideas together, an argument that the *order* matters, and a short list of research questions that follow. I write for readers in machine learning, cognitive science, and robotics, and I avoid formalism that the argument does not need.

I write from a particular point of view. My research is on trustworthy multimodal video understanding: how models that see, read, and listen are trained, what they actually learn, and whether we can trust their reasoning. I am now turning to physical AI, with video as the window onto the physical world. Hence my two convictions: a machine should first understand the physical world by watching it, and we should check what it has learned directly.

1 What We Have Built

World models have grown from a niche topic in reinforcement learning into one of the central ideas of AI research. The term now covers recurrent networks that let agents learn inside their own “dreams”, predictors that plan in a learned latent space, probabilistic programs that model intuitive physics, video generators that render playable worlds, and language models that researchers probe for hidden maps of the world. Before I survey these lines of work, I state my position.

Position. World-model research has been learning to *act* before it has learned to *perceive*. A world model should first learn to imagine the parts of the physical world it cannot currently see. It should learn this from observation, including the movements of human eyes, bodies, and hands rather than robot motors, and it should rely on memory and on nearly deterministic, everyday rigid-body physics. It should grow as a set of communicating sub-models acquired in stages, not as one model trained on everything at once. Acting on tasks and using language belong on top of this foundation, not beside it.

1.1 The reinforcement-learning line

The idea that an agent can plan by “trying things in [its] head” with an internal model of the world goes back to [Craik \(1943\)](#) (cited in [Sutton, 1991](#), p. 1); it entered machine learning around 1990. [Schmidhuber \(1990\)](#) trained a recurrent network to predict all of its inputs, including reward, and [Sutton \(1990, 1991\)](#) treated planning as learning from experience that a learned model simulates. [Ha and Schmidhuber \(2018\)](#) revived the term *world model* by training a controller inside the “hallucinated dream” of a learned model. The Dreamer family has turned learning in imagination into a general recipe, and its latest version trains agents purely offline inside a world model learned mostly from unlabeled video ([Hafner et al., 2023, 2025](#)). MuZero, by contrast, predicts only “the quantities most directly relevant to planning” ([Schrittwieser et al., 2019](#), p. 1). In this line, a world model has meant an action-conditioned predictor that serves decisions. The tension between predicting everything and predicting what matters is as old as the line itself: in 1990, [Schmidhuber \(1990](#), p. 19) already warned that predicting every future event “leads to the old frame-problem of conventional AI.”

1.2 The JEPA line

LeCun’s position paper ([LeCun, 2022](#)) put the world model at the center of a modular architecture and gave it two roles: to “estimate missing information about the state of the world not provided by perception” and to “predict plausible future states of the world” (p. 7). It bet that a model should predict in a learned representation rather than in pixels, and that it can learn this representation mostly by watching. The JEPA line that followed has grown to include V-JEPA 2, which pretrains on more than a million hours of video and then plans zero-shot after 62 hours of robot data ([Assran et al., 2025](#)); DINO-WM, which plans in frozen self-supervised features ([Zhou et al., 2024](#)); and models that train end to end from pixels on a single GPU ([Balestriero and LeCun, 2025](#); [Maes et al., 2026](#)). Closest to my concerns, [Garrido et al. \(2025\)](#) have shown that V-JEPA learns intuitive physics such as object permanence from natural video alone, although it “predicts the future only as an observer” (p. 8).

1.3 Structured and causal models

A parallel tradition has treated the world model as a structured, causal, and usually probabilistic model of how observations arise. Examples include an “intuitive physics engine” that runs rough simulations of object-based scenes ([Battaglia et al., 2013](#)); the view of human thought as “fundamentally a model-building activity” ([Lake et al., 2016](#), p. 2); causal representations that support intervention ([Schölkopf et al., 2021](#)); built-in biases for deliberate, System-2 reasoning ([Goyal and Bengio, 2020](#)); and intelligence as gathering evidence for a generative model ([Friston et al., 2022](#)). Two models come closest to the cognitive side of this paper. ADEPT keeps a belief about objects, including occluded ones, and reads surprise as how unlikely an observation is under that belief ([Smith et al., 2019](#)); PLATO learns intuitive physics through object-centric representations and passes violation-of-expectation probes taken from developmental psychology ([Piloto et al., 2022](#)). These works have agreed on structure, causality, and data efficiency, and have disagreed on how much structure to build in.

1.4 Generative world simulators

A third line has scaled video generators into simulators, on the bet that “scaling video generation models is a promising path towards building general purpose simulators of the physical world” (Brooks et al., 2024). Genie learns playable environments from unlabeled gameplay video (Bruce et al., 2024), and Genie 3 renders worlds that users explore in real time and that stay consistent for minutes (Parker-Holder and Fruchter, 2025). UniSim and Cosmos have presented such models as training grounds for embodied agents (Yang et al., 2023; NVIDIA et al., 2025), and Fei-Fei Li has defined world models by three abilities: generative, multimodal, and interactive (Li, 2025).

1.5 Language models as world models?

Whether language models contain world models depends on how one defines the term. Probing studies have found board states, maps of space and time, and representations that converge across modalities (Li et al., 2022; Gurnee and Tegmark, 2023; Huh et al., 2024). Others have argued that linguistic form alone cannot yield meaning (Bender and Koller, 2020), or shown that a model can predict the next token almost perfectly while holding an incoherent map and no Newtonian mechanics (Vafa et al., 2024, 2025).

1.6 An emerging consensus: world models for decisions

By 2026 much of this work has begun to converge on two points. First, a world model is not a video generator, and visual fidelity is not the goal (Xing et al., 2025; Motamed et al., 2025). Second, a world model’s value lies in the decisions it enables. “There is no single use-independent world-model score” (Yu et al., 2026, p. 15); world models “live and die by their closed-loop success, not flawless generated visuals” (Zhang et al., 2025, p. 2); the best of them will be “judged less by how vividly they dream and more by how much better they make the agents that query them” (Yang et al., 2026, p. 23). Recent frameworks have ranked world models by levels of decision-relevant ability (Chu et al., 2026; Yao et al., 2026) and placed them between agents and costly real-world interaction (Yang, 2026).

I accept the first point and dispute the priority of the second. Decision utility is a fair test of a finished system but a poor guide to building one. If we judge a world model only by the decisions it enables, research drifts toward tasks, robots, and benchmarks, and we test the foundation of those decisions—perceiving and imagining a world—only indirectly, if at all. A model that perceives the world correctly can serve decisions later. A model tuned for one set of decisions need not perceive the world correctly; it needs only what those decisions require (Schrittwieser et al., 2019). Section 2 examines what this order leaves out.

2 What We Are Missing

These lines of work have produced impressive systems. My concern is what they have left unexamined. Six gaps stand out, and they share one thread: researchers have rarely checked directly what a world model has learned about the world.

2.1 The field is racing to act before it can perceive

Almost every recent world model has been presented as a means to action. V-JEPA 2 moves from understanding to zero-shot robot planning within one paper (Assran et al., 2025); DINO-WM measures itself by planning success (Zhou et al., 2024); DeepMind presents Genie 2 and 3 as unlimited training environments for embodied agents (Google DeepMind, 2024; Parker-Holder and Fruchter, 2025); Cosmos holds that “Physical AI needs to be trained digitally first” (NVIDIA et al., 2025, p. 1); Dreamer 4 trains agents entirely inside a

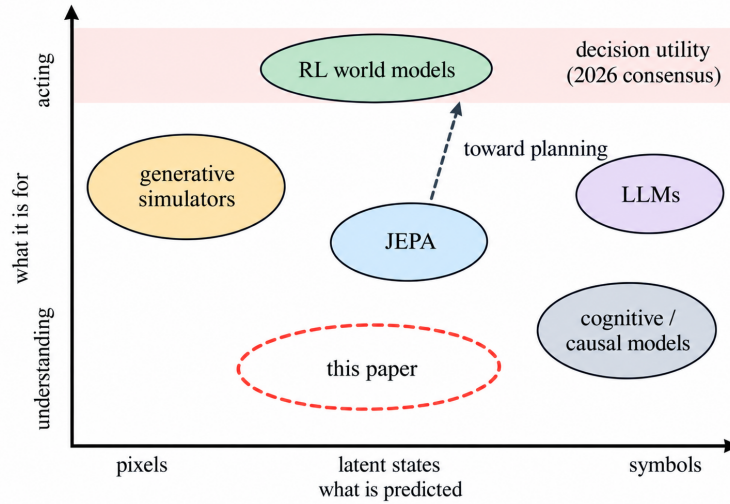


Figure 1: A map of world-model research by what models predict and what they are for. The consensus emerging in 2025–2026 puts the value of a world model in decisions (shaded band), and JEPA work moves from understanding toward planning. This paper argues for the region below: latent, cognition-first world models that aim to understand the present.

learned world model (Hafner et al., 2025); and recent frameworks place world models between agents and costly real-world interaction (Yang, 2026). Much of this work has aimed at useful embodied assistants such as avatars, wearables, and robots (Fung et al., 2025). That is a worthy goal, but it differs from building a model that understands the world.

The rush to act would matter less if perception were solved. It is not, and the evaluation literature says so plainly: “extracting physical representations of scenes is the main bottleneck” (Bear et al., 2021, p. 1); current multimodal models “lack a basic perceptual understanding of simulated scenes” (Jassim et al., 2023, p. 7); perception errors make up 37–45% of the mistakes that vision-language models make on physical questions (Chow et al., 2025, p. 8). A robot that plans better on a benchmark is still a task-specific machine, as robots have always been. What could carry across tasks is the world model beneath it, and that is the part we have examined least.

In fact, many systems that we call world models are closer to what control theory calls *plant models*: a predictor for one robot in one workspace, which an optimizer such as the cross-entropy method searches for actions that bring a predicted embedding close to a goal image, judged mainly by whether the robot gets there (Assran et al., 2025; Zhou et al., 2024; Maes et al., 2026). The predictor needs to be accurate only where the optimizer looks, and the optimizer exploits any error: it will “find a policy that looks good under our dynamics model, but will fail in the actual environment” (Ha and Schmidhuber, 2018, p. 9). Such a system learns how one body’s commands change its view, not how the world works.

2.2 Basic physics is benchmarked, not understood

Intuitive physics has a long evaluation tradition borrowed from developmental psychology: show a model matched pairs of possible and impossible events and measure its surprise, a violation-of-expectation (VoE) test (Riochet et al., 2018; Piloto et al., 2018). A decade of such benchmarks tells a consistent story. Models now pass the simplest tests—V-JEPA, trained only on natural video, reaches 98% on the original IntPhys (Garrido et al., 2025)—but they fail once a test involves interactions between objects, hidden physical properties, or

realistic scenes. On IntPhys 2 the best model scores 57.5%, against 96.4% for humans (Bordes et al., 2025). On Physion++, which requires inferring properties such as mass and elasticity, most models perform “only slightly above chance” (Tung et al., 2023, p. 7). Video-language models score below random on the intuitive-physics category of a benchmark that controls for shortcuts (22.8%, against 97.6% for humans) (Krojer et al., 2025), and the best video generator scores 29.5 out of 100 on Physics-IQ (Motamed et al., 2025). Even V-JEPA does no better than untrained networks on solidity and collision (Garrido et al., 2025).

The measurements are not what is missing; the explanations are. Benchmarks have become careful about *whether* a score is trustworthy—with matched pairs, blind baselines, and minimal-change controls (Krojer et al., 2025; Foss et al., 2025)—and *where* models fail. But papers have usually left *why* models fail as an untested guess: too little memory (Riochet et al., 2018; Bordes et al., 2025), or, for V-JEPA’s failures on collisions, “the model’s framerate constraints” (Garrido et al., 2025, p. 6), a factor the authors never vary. In 2018, Piloto et al. (2018, p. 1) warned that “successful prediction generally relies jointly on a set of physical concepts, making it difficult to crisply analyze the content of acquired knowledge”; in 2026, WorldBench still finds that existing benchmarks “suffer from entanglement” (Upadhyay et al., 2026, p. 1). The few studies that test causes directly are telling. Physion++ finds that models with an explicit property-inference stage do no better than models without one, “indicating that these models are not utilizing physical property inference” (Tung et al., 2023, p. 7); Kang et al. (2024) find that models generalize from similar cases rather than from rules; and a transformer that predicts orbits well does not recover Newtonian mechanics (Vafa et al., 2025).

Interaction alone does not fix the problem either. Buschhoff et al. (2026) fine-tune a vision-language model on physical tasks, either by imitation or by reinforcement learning through interaction. Interaction brings no gain within the task, and neither method yields physical intuitions that carry over to related tasks. Tellingly, quantities such as tower stability are already “highly decodable from model activations”, yet “neither post-training method successfully converts this competence into reliable performance on new tasks” (p. 2). The information is there, but the model cannot use it. That is a problem of perception and representation, and it is the problem this paper asks the field to solve first.

2.3 Action means robot commands

When current world models use actions at all, an action is a command for a particular machine: end-effector commands for a robot arm (Assran et al., 2025; Zhou et al., 2024), keyboard and mouse input for a game (Google DeepMind, 2024; Valevski et al., 2024), or driving controls for a car (Hu et al., 2023). Such data are scarce—“a primary bottleneck in world modeling is the scarcity of action-rich datasets” (Yang, 2026, p. 7)—and they differ from machine to machine, so merging them across datasets becomes a research problem in itself (Yang et al., 2023). Internet video carries no such labels: “such action labels are rarely available in videos from the Internet” (Bruce et al., 2024, p. 3). Genie therefore infers hidden actions from video, but whether this “could scale to more complex and diverse dynamics” remains open (Yang et al., 2024, p. 8). Models trained without actions predict “only as an observer” (Garrido et al., 2025, p. 8) and cannot tell the changes they cause from the changes the world causes.

Between these options lies the most plentiful action-labeled video there is: first-person recordings of people, which capture the wearer’s gaze, head, body, and hand movements along with what they see (Yang et al., 2023; Bai et al., 2025). PEVA shows that a model can predict first-person video from whole-body pose, but treats this as a problem of embodied prediction and control (Bai et al., 2025). A more basic use—letting a model with no body of its own learn which changes in its input it causes—is not the focus of that work.

2.4 Evaluation is tied to tasks or to pixels

Researchers have evaluated world models at one of two extremes. Open-loop reconstruction and final policy success “are the two most common rungs, and a large fraction of works report only this pair as their world-model evidence, leaving the model’s specific contribution entangled with the optimizer, reward model, and data pipeline” (Yu et al., 2026, p. 11). Neither extreme tells us what the model knows. Visual realism does not track physical understanding—across models, the two are not significantly correlated (Motamed et al., 2025)—and visual quality does not guarantee task success (Zhang et al., 2025). Task success, in turn, can rest on shortcuts: on minimally different video pairs, the average accuracy of video-language models falls from 45.4% to 27.3% (Krojer et al., 2025, p. 7).

Meanwhile, the papers that build world models have left physical understanding to other papers. V-JEPA 2 spends about seven pages on robot planning and none on intuitive physics, and a different paper reports its IntPhys 2 score (Assran et al., 2025; Bordes et al., 2025). DINO-WM reports no physics probe, yet concludes from a planning success rate of 0.82 that its model has “effectively learned the general concepts of walls and doors” (Zhou et al., 2024, p. 7). LeWorldModel is a partial exception: it probes physical quantities and measures surprise, though for only two kinds of violation (Maes et al., 2026). The usual recipe has been to build a better encoder and then show that a robot plans better with it. Finally, almost no one has measured whether a model knows where it is wrong: “calibrated uncertainty, not just sharper pixels” remains an open problem (Yang et al., 2026, p. 23), although Schmidhuber (1990, p. 20) noted in 1990 that a model “partly has to model its own ignorance.”

2.5 One model for everything

Another trend is unification: one model that understands, generates, and acts across all modalities. Cosmos 3 argues that separating understanding, generation, and action “is fundamentally limiting” and trains a single model on language, images, video, audio, and actions (NVIDIA et al., 2026, p. 5); the PAN proposal of Xing et al. (2025) likewise puts all modalities on a language-model backbone; and the roadmap of Chen et al. (2026) starts with unified multimodality. The strongest argument for this direction is that representations from different modalities seem to converge as models scale (Huh et al., 2024). But the convergence is partial—alignment between language and vision models “only reaches a score of 0.16”—and its authors note that two models “cannot converge to the same representation if they have access to fundamentally different information” (Huh et al., 2024, pp. 10–11). Seeing, hearing, touching, and remembering make different demands; even an advocate of world models notes that turning the world into 1D or 2D token sequences “makes simple spatial tasks unnecessarily difficult” (Li, 2025). The deeper cost of unification is that it hides failures. When one model learns everything at once, we cannot tell which ability failed, or whether one ability borrows another’s shortcuts. A system that grows in stages can be checked at every stage.

2.6 Language as a shortcut

Language carries much knowledge about the world, and models trained on text alone recover surprising structure, from board states to maps of space and time (Li et al., 2022; Gurnee and Tegmark, 2023). My concern is not what language contains but what it does to a learner that has it from the start: it lets the learner lower its loss without building a perceptual state. The evidence points one way. Multimodal language models perform near chance on intuitive physics (Garrido et al., 2025; Jassim et al., 2023), and below chance on the intuitive-physics category of a benchmark that controls for shortcuts (Krojer et al., 2025). Benchmark designers now have to remove the shortcut that language offers: CausalVQA discarded “approximately 39% of our questions” because a text-only model could answer them without seeing the video (Foss et al., 2025, p. 5). When we test a video model through a language interface, the language model’s abilities mix with the video model’s; the authors of the Veo 3 study warn that “for some tasks, the solution is likely to come from

the LLM instead of the video” (Wiedemer et al., 2025, p. 2). And a language-pretrained model fine-tuned on physical tasks learns “task-specific shortcuts” rather than physical intuitions (Buschhoff et al., 2026, p. 9). A learner that can describe the world before it perceives it may never need to perceive it well.

3 What a World Model Should Be

I now state what I think a world model should be: a definition, five principles, an account of where its uncertainty should live, a blueprint, and a developmental roadmap. None of these requires a particular architecture; a JEPA, a generative model, or an object-centric model could meet them or fail them. To keep the principles concrete, each comes with a test. Each principle also answers gaps from Section 2: Principle 1 the race to act (§2.1), Principles 2 and 3 the physics gap (§2.2), Principle 4 the action gap (§2.3), and Principle 5 unification and language (§2.5, §2.6). Together, the tests answer the evaluation gap (§2.4).

3.1 Definition

A world model is a learned system that, from partial observation, keeps a belief about the present state of the physical world, including what is out of view, backed by memory; imagines how that state changes, on its own and under observed or imagined actions; and holds the state in a form it can use to answer questions it was not trained on.

This definition keeps the two roles that LeCun (2022, p. 7) gives the world model—estimating what perception does not provide, and predicting plausible futures—but puts the first one first. Unlike decision-centric definitions (Yu et al., 2026), it does not require the model to serve a policy; unlike definitions centered on generation (Li, 2025), it does not require the model to render what it imagines. The last clause—a state the model can *use*—is the one most often missed: as Section 2.2 shows, a model can encode a physical quantity that a probe can read and still fail to use it for any new question (Buschhoff et al., 2026). By this definition, an action-conditioned predictor that is trained and tested only to reach goal images in one robot’s workspace is not yet a world model: nothing requires it to hold a belief about what is out of view, and nothing shows that its state can answer any question other than the one its planner asks. Throughout, *the physical world* means objects and events at human scale; modeling other minds is outside the scope of this paper.

3.2 Principles

1. Perceive first. A world model must first perceive and imagine the physical world correctly; acting on tasks and using language build on that ability. By *acting* I mean carrying out actions to reach goals. Observed and imagined actions, discussed under Principle 4, are part of perception, not exceptions to it. The case for this order is the evidence of Section 2: the encoders that current planners rely on fail basic physics, and the V-JEPA 2 authors themselves note that “the capabilities of a representation-space world model . . . are inherently limited by the state information encoded in the learned representation space” (Assran et al., 2025, p. 15). Better planning cannot recover what perception never learned. *Test:* the paper that presents a model measures its physical understanding directly, rather than inferring it from task success.

2. Imagine the unseen. At any moment most of the world is out of view: behind me, inside a box, around a corner. A world model should fill these regions with imagination by default and let observation correct it. This is a claim about knowledge, not about the world: the world behind me is fully definite; only my belief about it is uncertain. So, at human scale, I treat the dynamics as deterministic and place all uncertainty in the belief about the present; Section 3.3 explains why. Memory keeps the uncertainty small. Working memory holds what the model has just seen; long-term memory holds what such scenes usually look like. Together they

narrow the possibilities until imagination is nearly deterministic and objects persist when unseen. Current models are weakest here: JEPA models see clips of “typically 3–4 seconds” (Garrido et al., 2025, p. 8), and Genie 3 stays consistent by remembering about a minute of rendered frames (Parker-Holder and Fruchter, 2025), which is memory as a replay buffer rather than a belief about what is out of view. *Test*: VoE tests after occlusions of increasing length, and viewpoint what-if questions (*what will I see if I turn around?*).

3. Human scale. The target is the physics a person can see: rigid objects that persist, move smoothly, rest on supports, fall, and collide, at everyday scales of space and time. Of the five categories in Physics-IQ—solid mechanics, fluid dynamics, optics, thermodynamics, and magnetism (Motamed et al., 2025)—this principle asks for the first. The aim is not exact Newtonian laws but human-like intuitive physics, which is approximate (Battaglia et al., 2013); what matters is that the model holds it in a state that supports extrapolation. A model that interpolates between training cases (Kang et al., 2024), or that recovers a different law of gravity for each slice of data (Vafa et al., 2025), fails this principle however accurate it is in distribution. *Test*: out-of-distribution (OOD) and compositional extrapolation, tests of how a model generalizes from a few examples of a new task (Vafa et al., 2025), and agreement with human judgments, including human errors.

4. Learn through a borrowed body. The model I have in mind has no body, so in the stages this paper covers it never carries out actions. Actions enter in two other ways. The first is *observed* actions. First-person human video records the gaze, head, body, and hand movements of the person wearing the camera, and these can serve as action signals, as in PEVA, which predicts first-person video from “an action represented by the relative 3D body pose” (Bai et al., 2025, p. 1). Their first use is not to perform tasks but to separate self-motion from world motion: the model must know which changes in its input the borrowed body causes and which the world causes. This needs only a minimal self-model, not a self. Turning one’s head is also the most basic intervention there is, and the aimless looking around of infants produces the kind of data from which a model can recover the world’s hidden variables, while goal-directed data does not (Klindt et al., 2026, p. 9). The second is *imagined* actions: what-if questions, such as what would happen if this were pushed, that the model answers in imagination and never carries out. Observed actions are biased, since people push things when they expect them to move, so causal conclusions from them hold only under strong assumptions (Yu et al., 2026, p. 15). Whether some concepts, such as collisions, need the learner’s own interventions remains open; interaction added after training does not supply them (Buschhoff et al., 2026). *Test*: telling self-motion from object motion, and predicting the effects of human movements.

5. Grow, don’t unify. Seeing, hearing, touching, remembering, and imagining make different demands on a learner. I therefore favor a set of sub-models, each with its own representation, over one model trained on everything at once—the direction of systems that treat the split between understanding, generation, and action as a limitation to remove (NVIDIA et al., 2026). The sub-models talk through a narrow protocol rather than a shared latent space, in the spirit of a global workspace whose content is “analogous to a sentence: it involves only a few variables” (Bengio, 2017, p. 1). The model acquires them in stages (Section 3.5). In the early stages, filling in masks and predicting the future are not enough, because these objectives do not say what matters (Bengio, 2017, p. 4), and pixel reconstruction “biases models toward appearance fidelity at the expense of motion coherence” (Chefer et al., 2025). Infants do not learn alone either: before they understand words, caregivers point and say “look!”. The machine-learning counterpart is extra supervision for structure, such as tracking, segmentation, and motion, which already helps (Jeong et al., 2024; Chefer et al., 2025; Bardes et al., 2023; Locatello et al., 2020). I propose to use it as temporary training signals, or *scaffolds*: added early, removed later, and taken only from teachers trained without language, so that language does not slip in through the back door. Signals of value, such as “careful!” or “well done”, are another form of guidance that I leave for future work. Language comes last. I offer a hypothesis for why: language may be the protocol

between sub-models turned outward, and a protocol cannot be turned outward before it exists. *Test*: staged versus joint training, retention after the scaffolds are removed, and comparisons with and without language.

3.3 Where uncertainty lives

Principle 2 makes a choice that most world models do not make, and the rest of the blueprint depends on it, so I spell it out here.

A model of an uncertain world can put its uncertainty in one of two places. It can make the dynamics stochastic, drawing the next state from a distribution, $s_{t+1} \sim p(\cdot \mid s_t, a_t)$. Or it can make the dynamics deterministic, $s_{t+1} = f(s_t, a_t)$, and keep a distribution over the present state instead: a *belief* $b_t(s_t)$, in the sense of the belief states of partially observable control (Kaelbling et al., 1998). Both can fit the same videos. Most world models have taken the first option or mixed the two. Dreamer’s latent state has a stochastic part (Hafner et al., 2023); video generators sample their futures (Brooks et al., 2024; Bruce et al., 2024); and in LeCun’s architecture a latent variable “represents what cannot be predicted about y (the future) solely from x and from past observations” (LeCun, 2022, p. 19). These models represent many possible futures. I propose that a world model at human scale should represent many possible *presents*, and one future for each.

Three reasons support this choice. First, at human scale the randomness we observe reflects what we do not know, not chance in the world. A ball behind a box, the mass of a closed carton, the friction of a floor: each is a hidden part of the present state, not noise in physics. The intuitive physics engine already works this way. It samples the scene’s geometry, the objects’ hidden masses, and possible outside pushes such as “a breeze, a vibration, or a bump”, and then simulates each sample forward (Battaglia et al., 2013, p. 3); all of its uncertainty enters through what goes in. Second, memory can narrow only the uncertainty that lives in the state. Having seen a ball roll behind a box, the model knows where it is. A stochastic transition draws fresh randomness at every step, so nothing memory holds can remove it, and a sampled future can lose the ball. With deterministic dynamics the ball stays behind the box in every possible present: permanence comes for free, and only the ball’s exact position spreads, slowly, while it stays hidden. Third, the split makes errors traceable. Collisions magnify small differences, so even deterministic predictions fan out after one. In this framing the fan-out is growing uncertainty about the present—where exactly did the ball hit?—that one more look can remove, so it also tells the model where to look. A stochastic predictor would fold the same fan-out into its noise, and we could not tell an unpredictable world from a missing piece of state.

The third reason gives a diagnostic. Give a learned dynamics model a stochastic term, improve its state with more memory, better perception, or scaffolds, and record where the model still needs the noise. In a rigid-body scene the need should shrink as the state improves, and whatever remains marks what the state is missing: contacts, hidden masses, objects out of view. Unlike a benchmark score, such a map points at a cause, which is what Section 2.2 found missing.

The strongest evidence against this choice comes from people. In a task where people predict the path of a ball that bounces while hidden, perceptual uncertainty alone does not explain their judgments, and Smith and Vul (2013, p. 197) conclude that “the internal physics models themselves must be stochastic rather than rely solely on perceptual uncertainty.” ADEPT, too, runs a stochastic physics engine (Smith et al., 2019). I take this seriously, but it does not settle the question. Part of the noise that people need lies in “variability in bounce angles” (Smith and Vul, 2013, p. 186), the very place where the diagnostic predicts that noise should remain: a person rarely sees the details of a contact, such as spin, surface, or exact angle, so a model learned from everyday experience would carry that missing state as noise. Whether such noise belongs to physics or to missing state is exactly what the diagnostic asks. And the principle states what a world model should aim for, not what people achieve.

The choice has limits. It concerns rigid objects at human scale; fluids and other minds lie outside it, and so do the choices of people. LeCun’s own example of a latent variable is another driver’s decision to turn left or right (LeCun, 2022, p. 19). In this paper such choices are not noise in the physics but actions, which enter

the model observed or imagined (Principle 4).

3.4 A blueprint

Figure 2 turns the principles into a blueprint: not a model, but the parts a model needs and the order in which they appear.

At the center is not a predictor of the future but a *belief about the present*: the current state of the world, including what is out of view, with its uncertainty (Principle 2). Sensory sub-models feed it—vision first, hearing and touch later—each with its own representation (Principle 5). Memory is part of the belief, not an add-on: working memory holds what the model has just seen, and long-term memory holds what such scenes usually look like.

An imagination module rolls the belief forward with deterministic dynamics, with no action, with observed actions, or with imagined ones (Principles 2 and 4). Observed actions come from a *borrowed body*: the gaze, head, body, and hand movements of the person wearing the camera. The model itself has no motors. The borrowed body serves two purposes. First, its recorded movements stand in for what neuroscience calls an *effference copy*: a copy of a body’s own movement commands, which nervous systems use, among other things, to tell the input their movements cause from the input the world causes (Crapse and Sommer, 2008). The model has no commands of its own, so the recorded movement plays this role and lets the model discount the changes that the body’s own movement causes. Second, what-if questions let the model query its imagination.

A *surprise monitor* compares what the model imagined with what its senses then report. Its signal does three jobs: it updates the belief and memory; it drives learning and curiosity, much as progress in compression drives curiosity in Schmidhuber (2008); and it shows where the model is wrong, so that the architecture builds in calibration instead of measuring it afterward. VoE tests read out exactly this signal.

The parts talk through a narrow protocol of predictions, prediction errors, and effference copies (Principle 5). The architecture grows: scaffolds join early and leave later, each new stage joins only after the current one passes its tests (Table 2), and language joins last, at the protocol rather than at perception.

Table 1 compares the blueprint with three recent proposals. No single part is new: others take a state-centered view of world models (Kim, 2026), separate self-caused from world-caused change (Zhang et al., 2026), learn action knowledge from human video (Gao et al., 2026; Bai et al., 2025), and keep a belief about hidden objects and read surprise from it (Smith et al., 2019; Piloto et al., 2022). What is new is the arrangement: a belief about the present at the center, no motors, uncertainty only in the belief, memory inside the belief, a built-in monitor of the model’s own errors, and a structure that grows. The two closest systems show the difference. ADEPT borrows a hand-built physics engine and makes it stochastic (Smith et al., 2019). PLATO learns its dynamics, but it learns from synthetic scenes and receives ground-truth object masks in both training and testing (Piloto et al., 2022). The blueprint asks for learned, deterministic dynamics, for object structure that arrives only as a scaffold and is later removed, and for natural first-person video.

3.5 A developmental roadmap

Table 2 orders the abilities of Principles 2–4 by the ages at which infants are reported to show them (Lake et al., 2016, p. 16; Riochet et al., 2018, p. 1). The abilities that infants show earliest—permanence, continuity, support—are also those that V-JEPA learns from passive video, while those that infants show later—gravity, inertia, collision—are among those it fails on (Garrido et al., 2025). The original studies draw the same line: infants respect continuity and solidity by about 2.5 to 4 months, while sensitivity to gravity and inertia develops later in the first year (Spelke et al., 1992, pp. 605, 626), and they then learn, one kind of event at a time, which further variables matter (Baillargeon, 2004). I treat this parallel as a hypothesis to test, not as evidence: it suggests that passive observation can carry a learner through the early stages, and that the later stages need more.

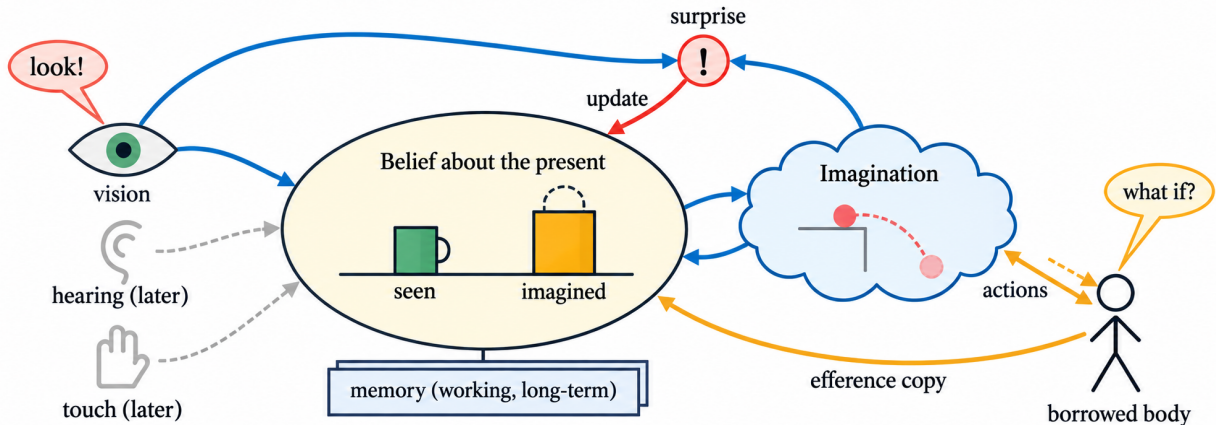


Figure 2: A blueprint for a cognition-first world model. At the center is a belief about the present: what the model sees (the cup) and what it only imagines (the ball behind the box), kept narrow by working and long-term memory. Imagination rolls the belief forward with deterministic dynamics. When its prediction and the observation disagree, surprise updates the belief. The model never carries out actions; they come from a borrowed human body, as observed movements, as what-if questions, and as recorded movements that, like efference copies, discount self-motion. The early “look!” signals are scaffolds that are removed later. Hearing and touch join later, and language last (stages in Table 2). Every arrow is a protocol message, not a shared latent space.

4 Research Directions We Can Pursue Now

The principles lead to research questions that today’s data and models can address. I list six. The first applies to all principles; each of the others follows from one principle.

4.1 Task-free evaluation of what a world model knows

The most urgent need is to evaluate what a model knows rather than what it achieves. A developmental test battery would combine six kinds of test: VoE tests with matched pairs and occlusions of increasing length, which probe permanence over time rather than over a few seconds (Riochet et al., 2018; Bordes et al., 2025; Garrido et al., 2025); controlled OOD and compositional tests, which ask whether a model has rules or only cases (Kang et al., 2024; Upadhyay et al., 2026); viewpoint what-if tests, which ask what becomes visible after a change of gaze and so probe the belief about what is out of view; tests of how a model generalizes from a few examples of a new physical task (Vafa et al., 2025); comparisons with human judgments, including human errors (Battaglia et al., 2013; Bear et al., 2021); and calibration tests, which ask whether the model knows where it is wrong. Most of these tests exist; what is missing is the habit of reporting them next to downstream results and of testing suspected causes of failure.

4.2 Memory-backed imagination

Section 3.3 separates dynamics from uncertainty: deterministic dynamics over latent states, and uncertainty carried by a set of candidate present states that memory narrows—a learned version of the physics engine of Battaglia et al. (2013). Two questions follow. First, do deterministic dynamics, with all uncertainty moved into the state, extrapolate better than stochastic predictors? The diagnostic of Section 3.3 makes this measurable: train both kinds of model on the same rigid-body videos and map where the stochastic one still needs its

	AMI (LeCun, 2022)	PAN (Xing et al., 2025)	Cosmos 3 (NVIDIA et al., 2026)	This paper
Center	world model + cost and actor loop	LLM backbone + generative predictor	one model for all modalities	belief about the present
Action	actor plans and acts	agents trained by RL inside the model	predicted jointly	none carried out ; borrowed body, imagined
Uncertainty	latent variable in the predictor	generative prediction	generative prediction	in the belief only ; deterministic dynamics
Memory	short-term memory module	stateful backbone	—	part of the belief (working + long-term)
Language	—	backbone	core modality	joins last , at the protocol
Over time	fixed (configurable per task)	fixed	fixed	grows in tested stages
Self-check	—	—	—	surprise monitor

Table 1: The blueprint compared with three recent proposals, at the level of architecture rather than implementation.

Stage	Infants	Ability	Scaffold (“look!”)	Test
S0	first days	segmenting objects (what moves together)	motion-based segmentation	segmentation
S1	2–4 mo.	permanence, continuity	tracking through occlusion	VoE with delays
S2	~6 mo.	support, stability, contact	trajectory and motion prediction	VoE, stability
S3	8–10 mo.	gravity, inertia, collision	observed hand actions; imagined interventions; scaffolds removed	VoE, OOD extrapolation
S4	later	hidden properties (mass, elasticity)	—	property inference

Table 2: A developmental roadmap. Ages from Lake et al. (2016) and Riochet et al. (2018); see also Spelke et al. (1992) and Baillargeon (2004). The literature differs in detail. Later stages (more senses, task-directed action, language) follow.

noise. If the map lights up at contacts and occlusions, those are the places where the state falls short. Second, does memory change what the model believes about what is out of view? After an object stays hidden for seconds, for minutes, or until the viewer turns back, is the model surprised if it is gone? Current JEPA models see clips of a few seconds (Garrido et al., 2025), and Genie 3 remembers about a minute of rendered frames (Parker-Holder and Fruchter, 2025); neither tests a belief about what is out of view.

4.3 From decodable to usable physical state

A linear probe asks whether information is present; usability asks whether the model can use it. A vision-language model can encode tower stability in a form a probe can read and still fail to use it on new tasks (Buschhoff et al., 2026). This suggests a simple protocol: freeze the world model, measure how quickly new physical questions can be learned from its state, and check whether the learned answers extrapolate the way Newtonian mechanics would (Vafa et al., 2025). Theory on when a model can recover the world’s hidden variables (Klindt et al., 2026) shows when probing can succeed at all, and which training data should make the state usable: aimless exploration rather than goal-directed behavior.

4.4 Learning through a borrowed body

First-person video already pairs what a person sees with how they move. Existing datasets record gaze, head motion, and body or hand pose (Yang et al., 2023; Bai et al., 2025), and large-scale human video already serves to pretrain robot world models (Gao et al., 2026). The question here is different. Does conditioning a world model on a borrowed body’s movements teach it to separate self-caused from world-caused change, a split studied so far only for an agent’s own actions (Zhang et al., 2026)? And does that split improve the belief about what is out of view? A direct test is to predict the visual effects of the wearer’s head turn and of an object’s motion, and then to swap them. Because observed actions are biased (Yu et al., 2026), simulated scenes with controlled interventions remain a useful check.

4.5 Growing in stages

The developmental roadmap (Table 2) makes predictions that we can test without new hardware. Does training in stages, each gated by its tests, end with better abilities than training on everything at once? Do scaffolds such as tracking, segmentation, and motion prediction, taken from teachers trained without language, help the early stages (Jeong et al., 2024; Chefer et al., 2025; Bardes et al., 2023)? And do the abilities survive when the scaffolds go? Video filmed from infants’ point of view is a natural test bed (Garrido et al., 2025, p. 8). Training on such video alone does not improve intuitive physics (Su et al., 2025), which fits the claim that video alone is not enough and that structure must be added.

4.6 Testing the language shortcut

Finally, we can test the claim that language should come last. Train the same perceptual world model three ways: without language, with language from the start, and with language added after the perceptual stages. Then compare the three on the test battery of Section 4.1, especially on minimal pairs that defeat textual shortcuts (Krojer et al., 2025; Foss et al., 2025). If the language-last model does not perceive better, the claim is wrong. If it does, the hypothesis that language is the protocol turned outward deserves a closer look.

5 Alternative Views and Conclusion

5.1 Alternative views

“Scale and unification will get there.” The most direct alternative is that none of this is needed: scale video generation, unify the modalities, and physics, permanence, and grounding will emerge, as they seemed to emerge in Sora as “purely phenomena of scale” (Brooks et al., 2024). The bitter lesson warns against building in what a model can learn (Sutton, 2019). My position welcomes scale but rejects skipping verification. Controlled studies find that scaling alone does not recover physical laws out of distribution (Kang et al., 2024), and visual realism does not track physical understanding (Motamed et al., 2025). A unified model that fails cannot tell us which ability failed. Ordering and testing are not hand-built knowledge; they are how we would know whether scale has delivered.

“Causality cannot be learned without interaction.” Some argue that agents must learn from their own experience (Silver and Sutton, 2025) and that causal learning requires intervention (Schölkopf et al., 2021); even the authors of the V-JEPA study suspect that collisions require interaction (Garrido et al., 2025, p. 8). I agree that observation has limits, and I do not exclude action. My claim concerns order and form. A model without a body can first learn from borrowed bodies and imagined interventions. Moreover, interaction added by fine-tuning does not produce physical intuitions that generalize when the model’s physical state is not

usable (Buschhoff et al., 2026). Theory shows that general agents must contain world models (Richens et al., 2025); it does not show that they must learn those models by acting first.

“There is no use-independent evaluation.” The strongest objection comes from the decision-centric view: a world model needs to be only as accurate as its policies and tasks require, so “there is no single use-independent world-model score” (Yu et al., 2026, p. 15). I accept that no single score exists. But almost every use assumes some abilities—that objects persist when unseen, move smoothly, and fall when unsupported—and decades of infant research show that we can define and test them without fixing a task. Measuring them does not replace decision-centric evaluation; it tells us whether a decision success rests on understanding or on shortcuts (Krojer et al., 2025).

“Infants are not gradient learners.” Finally, one can question the developmental analogy. Infants have innate structure (Spelke and Kinzler, 2007), bodies, and caregivers that a network trained by gradient descent lacks, and the parallel between V-JEPA’s failures and the infant timeline may be a coincidence. I agree. I use development as a source of hypotheses about order, not as evidence. Section 4 states each hypothesis—stages, scaffolds, and language last—in a form that experiments on today’s models can refute.

5.2 Conclusion

World-model research has been learning to act faster than it has learned to perceive. Planners steer robots with encoders that fail basic physics. Benchmarks have recorded these failures but have rarely explained them. And the field’s emerging consensus judges world models by the decisions they enable. I argue for a different order. A world model should first keep a belief about the present, including what it cannot see; memory should keep that belief narrow, and imagination should roll it forward. Its uncertainty should live in that belief: many possible presents, not many possible futures. It should learn through borrowed bodies before it has one of its own. It should grow in stages, with scaffolds early and language last. And we should check what it has learned directly, rather than infer it from task success. None of this rejects scale, interaction, or decisions; it asks that they rest on a verified foundation. Before world models act, they should learn to perceive.

References

- Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zhohus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL <https://arxiv.org/abs/2506.09985>.
- Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction. *arXiv preprint arXiv:2506.21552*, 2025. URL <https://arxiv.org/abs/2506.21552>.
- Renée Baillargeon. Infants’ physical world. *Current Directions in Psychological Science*, 13(3):89–94, 2004. doi: 10.1111/j.0963-7214.2004.00281.x. URL <https://doi.org/10.1111/j.0963-7214.2004.00281.x>.
- Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. URL <https://arxiv.org/abs/2511.08544>.

- Adrien Bardes, Jean Ponce, and Yann LeCun. Mc-jepa: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv preprint arXiv:2307.12698*, 2023. URL <https://arxiv.org/abs/2307.12698>.
- Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.
- Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021. URL <https://arxiv.org/abs/2106.08261>.
- Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of ACL*, pages 5185–5198, 2020.
- Yoshua Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017. URL <https://arxiv.org/abs/1709.08568>.
- Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025. URL <https://arxiv.org/abs/2506.09849>.
- Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI technical report, <https://openai.com/index/video-generation-models-as-world-simulators/>, 2024.
- Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024. URL <https://arxiv.org/abs/2402.15391>.
- Luca M. Schulze Buschoff, Konstantinos Voudouris, Can Demircan, and Eric Schulz. Can vision language models learn intuitive physics from interaction? *arXiv preprint arXiv:2602.06033*, 2026. URL <https://arxiv.org/abs/2602.06033>.
- Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025. URL <https://arxiv.org/abs/2502.02492>.
- Xinyuan Chen, Haoyu Guo, Shi Guo, Bingqi Jiang, Chunhua Shen, Xing Shen, Tianfan Xue, Yufei Xue, Mulin Yu, Weinan Zhang, Bin Zhao, Bowen Zhou, and Ming Zhou. A definition and roadmap for world models. *arXiv preprint arXiv:2607.06401*, 2026. URL <https://arxiv.org/abs/2607.06401>.
- Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025. URL <https://arxiv.org/abs/2501.16411>.
- Meng Chu, Xuan Billy Zhang, Kevin Qinghong Lin, Lingdong Kong, Jize Zhang, Teng Tu, Weijian Ma, Ziqi Huang, Senqiao Yang, Wei Huang, Yeying Jin, Zhefan Rao, Jinhui Ye, Xinyu Lin, Xichen Zhang, Qisheng Hu, Shuai Yang, Leyang Shen, Wei Chow, Yifei Dong, Fengyi Wu, Quanyu Long, Bin Xia, Shaozuo Yu,

- Mingkang Zhu, Wenhui Zhang, Jiehui Huang, Haokun Gui, Runyi Li, Chenyu Tang, Dong Huang, Xuhang Chen, Rui Liu, Chengzu Li, Shiyi Du, Xu Huang, Haoxuan Che, Long Chen, Qifeng Chen, Wenya Wang, Wenxuan Zhang, Xiaojuan Qi, Yang Deng, Yanwei Li, Mike Zheng Shou, Zhi-Qi Cheng, See-Kiong Ng, Ziwei Liu, Philip Torr, and Jiaya Jia. Agentic world modeling: Foundations, capabilities, laws, and beyond. *arXiv preprint arXiv:2604.22748*, 2026. URL <https://arxiv.org/abs/2604.22748>.
- Kenneth J. W. Craik. *The Nature of Explanation*. Cambridge University Press, Cambridge, UK, 1943.
- Trinity B. Crago and Marc A. Sommer. Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, 9(8):587–600, 2008. doi: 10.1038/nrn2457. URL <https://doi.org/10.1038/nrn2457>.
- Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T. Kao. Causalvqa: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025. URL <https://arxiv.org/abs/2506.09943>.
- Karl J Friston, Maxwell J D Ramstead, Alex B Kiefer, Alexander Tschantz, Christopher L Buckley, Mahault Albarracín, Riddhi J Pitliya, Conor Heins, Brennan Klein, Beren Millidge, Dalton A R Sakthivadivel, Toby St Clere Smithe, Magnus Koudahl, Safae Essafi Tremblay, Capm Petersen, Kaiser Fung, Jason G Fox, Steven Swanson, Dan Mapes, and Gabriel René. Designing ecosystems of intelligence from first principles. *arXiv preprint arXiv:2212.01354*, 2022. URL <https://arxiv.org/abs/2212.01354>.
- Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, Arjun Majumdar, Andrea Madotto, Franziska Meier, Florian Metze, Louis-Philippe Morency, Théo Moutakanni, Juan Pino, Basile Terver, Joseph Tighe, Paden Tomasello, and Jitendra Malik. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*, 2025. URL <https://arxiv.org/abs/2506.22355>.
- Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, Qianli Ma, Seungjun Nah, Loic Magne, Jiannan Xiang, Yuqi Xie, Ruijie Zheng, Dantong Niu, You Liang Tan, K. R. Zentner, George Kurian, Suneel Indupuru, Pooya Jannaty, Jinwei Gu, Jun Zhang, Jitendra Malik, Pieter Abbeel, Ming-Yu Liu, Yuke Zhu, Joel Jang, and Linxi "Jim" Fan. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026. URL <https://arxiv.org/abs/2602.06949>.
- Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025. URL <https://arxiv.org/abs/2502.11831>.
- Google DeepMind. Genie 2: A large-scale foundation world model. <https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/>, 2024.
- Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *arXiv preprint arXiv:2011.15091*, 2020. URL <https://arxiv.org/abs/2011.15091>.
- Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023. URL <https://arxiv.org/abs/2310.02207>.
- David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL <https://arxiv.org/abs/1803.10122>.
- Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. URL <https://arxiv.org/abs/2301.04104>.

- Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025. URL <https://arxiv.org/abs/2509.24527>.
- Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. URL <https://arxiv.org/abs/2309.17080>.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. URL <https://arxiv.org/abs/2405.07987>.
- Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. *arXiv preprint arXiv:2311.09048*, 2023. URL <https://arxiv.org/abs/2311.09048>.
- Hyeonho Jeong, Chun-Hao Paul Huang, Jong Chul Ye, Niloy Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. *arXiv preprint arXiv:2412.06016*, 2024. URL <https://arxiv.org/abs/2412.06016>.
- Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X. URL [https://doi.org/10.1016/S0004-3702\(98\)00023-X](https://doi.org/10.1016/S0004-3702(98)00023-X).
- Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024. URL <https://arxiv.org/abs/2411.02385>.
- Keon Woo Kim. Latent state design for world models under sufficiency constraints. *arXiv preprint arXiv:2605.01694*, 2026. URL <https://arxiv.org/abs/2605.01694>.
- David Klindt, Yann LeCun, and Randall Balestriero. When does lejepa learn a world model? *arXiv preprint arXiv:2605.26379*, 2026. URL <https://arxiv.org/abs/2605.26379>.
- Benno Krojer, Mojtaba Komeili, Candace Ross, Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and Mahmoud Assran. A shortcut-aware video-qa benchmark for physical understanding via minimal video pairs. *arXiv preprint arXiv:2506.09987*, 2025. URL <https://arxiv.org/abs/2506.09987>.
- Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *arXiv preprint arXiv:1604.00289*, 2016. URL <https://arxiv.org/abs/1604.00289>.
- Yann LeCun. A path towards autonomous machine intelligence (version 0.9.2). OpenReview, 2022. URL <https://openreview.net/forum?id=BZ5a1r-kVsF>.
- Fei-Fei Li. From words to worlds: Spatial intelligence is AI’s next frontier. <https://drfeifei.substack.com/p/from-words-to-worlds-spatial-intelligence>, 2025.
- Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022. URL <https://arxiv.org/abs/2210.13382>.
- Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *arXiv preprint arXiv:2006.15055*, 2020. URL <https://arxiv.org/abs/2006.15055>.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026. URL <https://arxiv.org/abs/2603.19312>.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025. URL <https://arxiv.org/abs/2501.09038>.

NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchammi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. URL <https://arxiv.org/abs/2501.03575>.

NVIDIA, :, Aditi, Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, Aarti Basant, Mukesh Beladiya, Mohammad Qazim Bhat, Zaid Pervaiz Bhat, Dan Blick, Vanni Brighella, Han Cai, Tiffany Cai, Eric Cameracci, Jiaxin Cao, Yulong Cao, Mark Carlson, Carlos Casanova, Ting-Yun Chang, Yan Chang, Yu-Wei Chao, Prithvijit Chattopadhyay, Roshan Chaudhari, Chieh-Yun Chen, Junyu Chen, Ke Chen, Qizhi Chen, Wenkai Chen, Xiaotong Chen, Yu Chen, An-Chieh Cheng, Click Cheng, Xiu Chia, Jeana Choi, Chaeyeon Chung, Wenyan Cong, Yin Cui, Magdalena Dadela, Nalin Dadhich, Wenliang Dai, Joyjit Daw, Alperen Degirmenci, Rodrigo Vieira Del Monte, Robert Denomme, Sameer Dharur, Marco Di Lucca, Ke Ding, Wenhao Ding, Yifan Ding, Yuzhu Dong, Nicole Drumheller, Yilun Du, Aigul Dzhumamuratova, Aleksandr Efitorov, Hamid Eghbalzadeh, Naomi Eigbe, Imad El Hanafi, Hassan Eslami, Benedikt Falk, Jiaojiao Fan, Jim Fan, Amol Fasale, Sergiy Fefilatyev, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Vikram Fugro, Prashant Gaikwad, TJ Galda, Katelyn Gao, Yihuai Gao, Wenhao Ge, Sreyan Ghosh, Arushi Goel, Vivek Goel, Akash Gokul, Rama Govindaraju, Jinwei Gu, Miguel Guerrero, Elfie Guo, Aryaman Gupta, Siddharth Gururani, Hugo Hadfield, Song Han, Ankur Handa, Zekun Hao, Mohammad Harrim, Ali Hassani, Nathan Hayes-Roth, Yufan He, Chris Helvig, Cyrus Hogg, Madison Huang, Michael Huang, Sophia Huang, Yufan Huang, Jacob Huffman, DeLesley Hutchins, Suneel Indupuru, Boris Ivanovic, Arihant Jain, Joel Jang, Ryan Ji, Yanan Jian, Dongfu Jiang, Jingyi Jin, Atharva Joshi, Nikhilesh Joshi, Pranjali Joshi, Andy Ju, Jaehun Jung, Weiwei Kang, Scott Kassekert, Jan Kautz, Ashna Khetan, Julia Kiczka, Slawek Kierat, Gwanghyun Kim, Kuno Kim, Sunny Kim, Kezhi Kong, Xin Kong, Zhifeng Kong, Tomasz Kornuta, Egor Krivov, Hui Kuang, Saurav Kumar, Chia-Wen Kuo, George Kurian, Wojciech Kutak, JF Lafleche, Himangshu Lahkar, Omar Laymoun, Jayjun Lee, Sanggil Lee, Gabriele Leone, Boyi Li, Freya Li, Jiajun Li, Jinfeng Li, Ling Li, Pengcheng Li, Shangru Li, Tingle Li, Xiaolong Li, Xuan Li, Zhaoshuo Li, Zhiqi Li, Hao Liang, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Ming-Yu Liu, Sifei Liu, Zihan Liu, Hai Loc Lu, Xiangyu Lu, Alice Luo, Ruipu Luo, Wenjie Luo, Jiangran Lyu, Martin Ding Ma, Nic Ma, Qianli Ma, Dawid Majchrowski, Louis Marcoux, Miguel Martin, Qing Miao, Ashkan Mirzaei, Shreyas Misra, Kaichun Mo, Durra Mohsin, Hyejin Moon, Pawel Morkisz, Saeid Motiian, Kirill Motkov, Seungjun Nah, Yashraj Narang, Deepak Narayanan, Thabang Ngazimbi, Julian Ouyang, Shubham Pachori, David Page, Yatian Pang, Sehwi

- Park, Mahesh Patekar, Mostofa Patwary, Marco Pavone, Trung Pham, Wei Ping, Soha Pouya, Shrimai Prabhumoye, Varun Praveen, Delin Qu, Hesam Rabeti, Morteza Ramezanali, Marilyn Reeb, Xuanchi Ren, Kristen Rumley, Wojciech Rymer, Jun Saito, Yeongho Seol, John Shao, Piyush Shekdar, Tianwei Shen, Humphrey Shi, Min Shi, Stella Shi, Kevin Shih, Mohammad Shoeybi, Mateusz Sienawski, Shuran Song, Alexander Sotelo, Amir Sotoodeh, Sunil Srinivasa, Vignesh Srinivasakumar, Bartosz Stefaniak, Rahul Heinrich Steiger, Shangkun Sun, Jiayang Tang, Shitao Tang, Yangyang Tang, Yue Tang, Tolou Tavakkoli, Kayley Ting, Krzysztof Tomala, Wei-Cheng Tseng, Jibin Varghese, Sergei Vasilev, Thomas Volk, Raju Wagwani, Roger Waleffe, Andrew Z. Wang, Boxiang Wang, Haoxiang Wang, Qiao Wang, Shihao Wang, Shijie Wang, Ting-Chun Wang, Yan Wang, Yu Wang, Rohit Watve, David Wehr, Fangyin Wei, Xinshuo Weng, Jay Zhangjie Wu, Kedi Wu, Hongchi Xia, Summer Xiao, Tianjun Xiao, Kevin Xie, Daguang Xu, Jiashu Xu, Mengyao Xu, Ruqing Xu, Xingqian Xu, Yao Xu, Dinghao Yang, Dong Yang, Hans Yang, Xiaodong Yang, Xuning Yang, Yichu Yang, Yurong You, Zhiding Yu, Hao Yuan, Simon Yuen, Xiaohui Zeng, Pengcuo Zeren, Cindy Zha, Haotian Zhang, Jenny Zhang, Jing Zhang, Liangkai Zhang, Paris Zhang, Shun Zhang, Xuanmeng Zhang, Zhizheng Zhang, Ann Zhao, Yilin Zhao, Yuliya Zhautouskaya, Charles Zhou, Fengzhe Zhou, Shilin Zhu, Yuke Zhu, Dima Zhylko, and Artur Zolkowski. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026. URL <https://arxiv.org/abs/2606.02800>.
- Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models. <https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/>, 2025.
- Luis Piloto, Ari Weinstein, Dhruva TB, Arun Ahuja, Mehdi Mirza, Greg Wayne, David Amos, Chia chun Hung, and Matt Botvinick. Probing physics knowledge using tools from developmental psychology. *arXiv preprint arXiv:1804.01128*, 2018. URL <https://arxiv.org/abs/1804.01128>.
- Luis S. Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6(9):1257–1267, 2022. doi: 10.1038/s41562-022-01394-8. URL <https://doi.org/10.1038/s41562-022-01394-8>.
- Jonathan Richens, David Abel, Alexis Bellot, and Tom Everitt. General agents contain world models. *arXiv preprint arXiv:2506.01622*, 2025. URL <https://arxiv.org/abs/2506.01622>.
- Ronan Riochet, Mario Yncente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018. URL <https://arxiv.org/abs/1803.07616>.
- Juergen Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *arXiv preprint arXiv:0812.4360*, 2008. URL <https://arxiv.org/abs/0812.4360>.
- Jürgen Schmidhuber. Making the world differentiable: On using self-supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments. Technical Report FKI-126-90, Institut für Informatik, Technische Universität München, 1990. URL [https://people.idsia.ch/~juergen/FKI-126-90_\(revised\)bw_ocr.pdf](https://people.idsia.ch/~juergen/FKI-126-90_(revised)bw_ocr.pdf).
- Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019. URL <https://arxiv.org/abs/1911.08265>.

- Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *arXiv preprint arXiv:2102.11107*, 2021. URL <https://arxiv.org/abs/2102.11107>.
- David Silver and Richard S. Sutton. Welcome to the era of experience, 2025. URL <https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf>. Preprint of a chapter in *Designing an Intelligence* (MIT Press).
- Kevin A. Smith and Edward Vul. Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, 5(1):185–199, 2013. doi: 10.1111/tops.12009. URL <https://doi.org/10.1111/tops.12009>.
- Kevin A. Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Joshua B. Tenenbaum, and Tomer D. Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/e88f243bf341ded9b4ced444795c3f17-Abstract.html.
- Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007. doi: 10.1111/j.1467-7687.2007.00569.x. URL <https://doi.org/10.1111/j.1467-7687.2007.00569.x>.
- Elizabeth S. Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowledge. *Psychological Review*, 99(4):605–632, 1992. doi: 10.1037/0033-295X.99.4.605. URL <https://doi.org/10.1037/0033-295X.99.4.605>.
- Ellen Su, Solim Legris, Todd M. Gureckis, and Mengye Ren. Opinion: Learning intuitive physics may require more than visual data. *arXiv preprint arXiv:2512.06232*, 2025. URL <https://arxiv.org/abs/2512.06232>.
- Richard S. Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, pages 216–224, 1990.
- Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991.
- Richard S. Sutton. The bitter lesson. <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>, 2019.
- Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Joshua B. Tenenbaum, Daniel LK Yamins, Judith E Fan, and Kevin A. Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *arXiv preprint arXiv:2306.15668*, 2023. URL <https://arxiv.org/abs/2306.15668>.
- Rishi Upadhyay, Howard Zhang, Jim Solomon, Ayush Agrawal, Yunhao Ba, Alex Wong, Celso M de Melo, and Achuta Kadambi. Worldbench: Benchmarking physical understanding of world models by isolating physics concepts. *arXiv preprint arXiv:2601.21282*, 2026. URL <https://arxiv.org/abs/2601.21282>.
- Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. *arXiv preprint arXiv:2406.03689*, 2024. URL <https://arxiv.org/abs/2406.03689>.

- Keyon Vafa, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. What has a foundation model found? using inductive bias to probe for world models. *arXiv preprint arXiv:2507.06952*, 2025. URL <https://arxiv.org/abs/2507.06952>.
- Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024. URL <https://arxiv.org/abs/2408.14837>.
- Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025. URL <https://arxiv.org/abs/2509.20328>.
- Eric Xing, Mingkai Deng, and Jinyu Hou. Critique of world model. *arXiv preprint arXiv:2507.05169*, 2025. URL <https://arxiv.org/abs/2507.05169>.
- Sherry Yang. World models as an intermediary between agents and the real world. *arXiv preprint arXiv:2602.00785*, 2026. URL <https://arxiv.org/abs/2602.00785>.
- Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023. URL <https://arxiv.org/abs/2310.06114>.
- Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Video as the new language for real-world decision making. *arXiv preprint arXiv:2402.17139*, 2024. URL <https://arxiv.org/abs/2402.17139>.
- Yu Yang, Xuemeng Yang, Licheng Wen, Lingdong Kong, Xiaobin Hu, Dongyue Lu, Wei Chow, Xiyan Huang, Yuxiang Feng, Yue Liao, Jianbiao Mei, Daocheng Fu, Rong Wu, Pinlong Cai, Ran Yi, Ying Tai, Jiangning Zhang, Botian Shi, Yong Liu, and Shuicheng Yan. Quo vadis, world modeling? *arXiv preprint arXiv:2608.02713*, 2026. URL <https://arxiv.org/abs/2608.02713>.
- Nanjie Yao, Hao Wang, Chong Cheng, Zhikang Chen, Wenzhe Li, Jiafei Lyu, Li Shen, Peilin Zhao, Zongqing Lu, Gao Huang, Steven Hoi, Dacheng Tao, and Deheng Ye. World models for embodied intelligence: From plausible to controllable to actionable. *arXiv preprint arXiv:2609.16697*, 2026. URL <https://arxiv.org/abs/2609.16697>.
- Yang Yu, Shiyuan Zhang, Yifei Sheng, Haoxiang Ren, and Haoxin Lin. How should world models be evaluated for embodied decision-making? a decision-making-centric position. *arXiv preprint arXiv:2606.15032*, 2026. URL <https://arxiv.org/abs/2606.15032>.
- Jiahao Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-in-world: World models in a closed-loop world. *arXiv preprint arXiv:2510.18135*, 2025. URL <https://arxiv.org/abs/2510.18135>.
- Yi-Ge Zhang, Tianqi Du, Qi Zhang, and Yisen Wang. Dwm: Separating world effects from actions in latent world models. *arXiv preprint arXiv:2607.18715*, 2026. URL <https://arxiv.org/abs/2607.18715>.
- Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024. URL <https://arxiv.org/abs/2411.04983>.