Atomathic

White paper

Physical AI 2.0: Why Visual World Models and Reasoning Are Not Enough - Why Physical State Recovery Is the Missing Capability

Behrooz Rezvani, Ph.D. | May 2026



Abstract

Artificial intelligence is moving from software workflows into embodied systems that must sense, predict, reason, decide, and act in the real world. The first visible phase of this transition can be described as Physical AI 1.0: a stack centered on large-scale vision, text, and multimodal data, simulation, synthetic-data generation, world models, and increasingly capable reasoning models.

LeCun's 2022 program for autonomous machine intelligence is a useful anchor for this discussion because it assigns world models a dual role: estimate missing information about the current state of the world and predict plausible future states. The later JEPA line sharpened that idea by emphasizing predictive abstraction in latent representation space rather than reconstructing every pixel or token, reinforcing the importance of state-centered representations for planning and control. [1][2][3]

NVIDIA's publicly available platform is one of the clearest current expressions of this stack, spanning Cosmos, Isaac GR00T, Alpamayo, OSMO, NIM, Jetson Thor, DRIVE AGX, and DRIVE Hyperion. [4][6][7][8][9][11][12][13][14] This is a major advance. But it is not the whole story. The current public discussion of Physical AI focuses heavily on world generation, scenario expansion, and action reasoning, especially for long-tail cases in which real-time decisions must be made. It pays less attention to the upstream question of whether the world has been properly observed in the first place.

That gap matters. When the observation basis is sparse, noisy, aliased, cluttered, contested, erased, or physically ambiguous, the bottleneck is not only predicting the next scene or choosing the next action. It is recovering the actual physical state before higher-level reasoning acts on it. This paper argues that Physical AI 2.0 will be defined by a stronger architectural sequence: world models, physical state recovery, reasoning systems, and action. World models encode priors and plausible future world states. **Physical state recovery improves what the system can know about the actual world under incomplete observation**. Reasoning systems compare options and select among actions. Action closes the loop in the physical world and generates new observations. **The central conclusion is clear: vision and language-first world models—the dominant basis of current public world-model systems—are not sufficient on their own to meet the state-estimation goals of Physical AI, especially when critical state is weakly observed or encoded in nonvisual measurements.** In more formal technical language, physical state recovery is trustworthy physical state inference under incomplete observation.

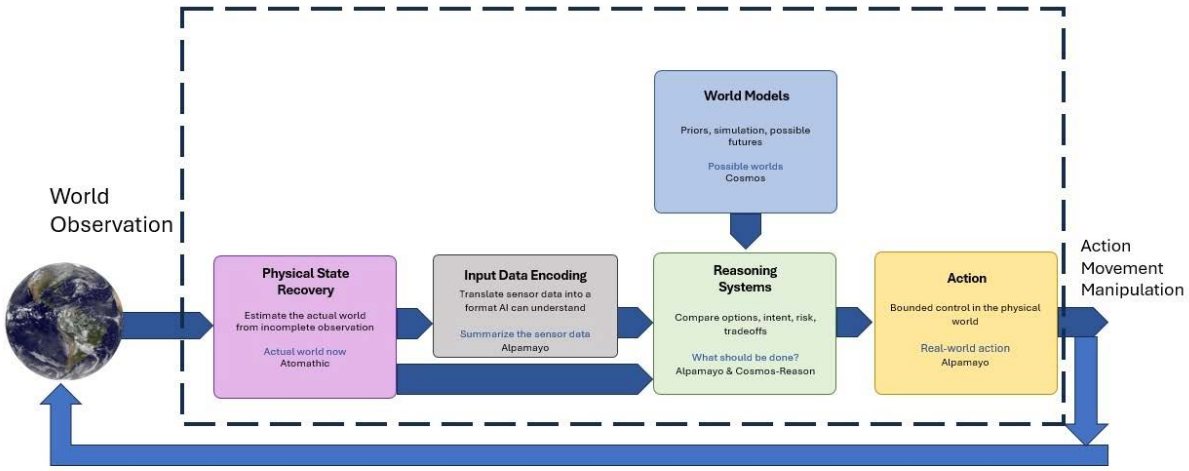

Figure 1. The Physical AI loop. During real-world operation, world observations support physical state recovery, which feeds reasoning and bounded action in the physical world. World models provide priors and plausible future world states, while action changes the world and generates new observations that close the loop.

01. Physical AI as an action-closed loop

A Physical AI system is not simply a model attached to a robot or vehicle. It is a control architecture embedded in the real world. It must observe through sensors, recover the physical state of the environment from imperfect evidence, reason over what might happen next, choose a course of action, and execute that action through a bounded control path.

That broader view matters because the unit of competition is changing. In digital AI, the model is often the product. In Physical AI, the model is only one part of a larger system that includes sensing, simulators, policy training, orchestration, safety, edge deployment, and feedback from live operations.

A useful intuition comes from human systems. Prior experience and learned models shape what we believe is plausible. Our senses—not vision alone—provide incomplete and noisy evidence. We then form a best estimate of what is actually happening, reason about what to do—especially in situations we have not seen before—and take action. Physical AI follows the same basic logic, but inside engineered control loops with explicit safety, latency, and power constraints.

A natural counterargument is that sufficiently large end-to-end models may eventually learn much of this internally, without an explicitly separated recovery capability. This paper does not argue that such internal learning is impossible. It argues that a modular physics-based recovery layer remains valuable because it can exploit richer sensing, improve observability before higher-level reasoning begins, and transfer across downstream world models, reasoning systems, and embodiments instead of requiring each of them to relearn the same recovery problem from scratch. Human systems illustrate the point: our

decisions are not driven by vision alone, but by integrated sensory evidence and prior models of the world.

This is why the strongest pattern in Physical AI is not a single magical brain. It is a loop of co-dependent capabilities. In some architectures the policy layer may itself be implemented as a VLA or embodied agent—for example, Alpamayo in autonomous driving—but the execution path still ends in bounded planning, control, and safety logic. Better reasoning does not remove the need for better sensing. Better world generation does not remove the need for bounded control. And more compute does not remove the need for a reliable estimate of the actual world during real-world operation.

02. Preliminaries: world models, recovery, reasoning, action

Physical AI refers to intelligent systems that must perceive, reason, decide, and act in the real world under the constraints of physics and real-world uncertainty. Unlike purely digital AI systems, Physical AI systems operate inside a loop with the external world: they observe through sensors, form an internal understanding of the environment, choose actions under physical and operational limits, and drive those actions through real actuators or control systems.

For this paper, the architecture is easiest to understand through four essential capabilities:

- World models - learned priors, simulation engines, and plausible future world states.
- Physical state recovery - the recovery of the actual current world from incomplete, noisy, or ambiguous observations.
- Reasoning systems - comparison of alternatives, risks, and intentions at decision time given the recovered state estimate.
- Action - bounded actuation in the physical world, which is what separates Physical AI from purely digital AI.

Observation is the substrate for the loop. Sensing supplies the evidence; physical state recovery interprets what is actually there; reasoning selects among available choices; action changes the world and produces new observations.

In that framing, Atomathic is differentiated primarily around a modular physics-based recovery layer and reasoning over difficult observations: recovering meaningful physical structure before higher-level policy and control begin.

03. Physical AI 1.0: data, simulation, world models

Physical AI 1.0 is defined by three ideas. First, scale matters outside language. Second, real-world experience can be supplemented by simulation and synthetic data. Third, world models can help machines learn environment dynamics before they ever act in those environments.

LeCun’s world-model program helps clarify why this matters. In A Path Towards Autonomous Machine Intelligence, the world model is tasked both with completing missing information about the current world state and predicting plausible future states. Subsequent JEPA work pushes this idea further by learning

predictive abstractions in representation space rather than reconstructing sensory input at full detail. This paper is aligned with that direction, but argues that in practical Physical AI deployments the missing-information problem deserves explicit architectural emphasis as physical state recovery, especially when observations are sparse, noisy, nonvisual, or physically ambiguous. [1][2][3]

NVIDIA's Cosmos platform is one of the clearest public formalizations of this idea. It is presented as a developer-first platform for Physical AI with world foundation models, reasoning models, controllable world generation, guardrails, and accelerated data processing. [4][5][6][7][8] The result is a practical flywheel: collect logs, curate them, train world models, generate new variations, run simulation at scale, and use the expanded distribution to improve downstream policy models.

That is a major step forward. Physical systems cannot rely on unsafe trial-and-error in the open world. Rare events are too expensive to wait for, too dangerous to collect at scale, or too sparse to cover with purely real-world data. World models and simulation make those tails more learnable.

04. Why inference-time reasoning matters

At the same time, the AI story itself has changed. The frontier is no longer only training-time scale. It is also inference-time reasoning: models that deliberate, compare alternatives, expose intermediate rationale, and improve robustness by spending more compute at decision time. [12][18][23]

Inference-time reasoning does not usually mean a deployed model is learning new weights. The parameters remain fixed. What changes is how much computation the system spends at decision time: longer deliberation, deeper search, richer prompting, or more structured intermediate reasoning before a decision is emitted. [23] In physical systems, the quality of the observations strongly shapes which hypotheses, causal chains, and action branches are even visible to the model.

That shift is especially important in Physical AI. A robot may need to decide whether a garment is fragile, whether a grasp has failed, or whether an operator handoff is safer than a retry. An autonomous vehicle may need to decide whether a pedestrian will continue crossing, whether a hand signal overrides the light phase, or whether a strange scene requires a more conservative maneuver. But reasoning is only as good as the state estimate it reasons over. If the observation is incomplete or distorted, even an excellent reasoning model can become confidently wrong.

05. A clearer control architecture for Physical AI

The most useful way to think about modern Physical AI is not as one monolithic model, but as a stack of cooperating capabilities. Each one has a distinct responsibility, and the boundaries between them matter. The practical design rule is simple: reasoning should influence control, but it should not directly actuate. In robust systems, reasoning emits intent, constraints, explanations, or action candidates. Planning, control, and safety logic still own the final conversion to physical motion.

Capability	Primary job	Representative examples	Why it matters / where Atomathic fits
World Models	Encode priors, generate futures, and create synthetic scenario coverage.	Cosmos-Predict2.5, Cosmos-Transfer2.5, Omniverse- and AlpaSim-based workflows.	Expand coverage, anticipation, and training efficiency. They are strongest when conditioned on a better estimate of the current world.
Physical State Recovery	Recover the actual state of the world from incomplete or ambiguous observations.	Physics-grounded recovery, reconstruction, and state estimation beneath the public world-model narrative.	Without a trustworthy state estimate, all downstream reasoning inherits upstream ambiguity. This is the core gap the paper highlights, and the clearest place where Atomathic differentiates. World-model priors can support recovery, but they do not eliminate the need for faithful recovery from the observations themselves.
Reasoning Systems	Interpret scenes, compare alternatives, explain intent, and propose candidate actions.	Cosmos-Reason2, Isaac GR00T, Alpamayo, domain-specific policy stacks.	Make behavior more selective and more interpretable under uncertainty. They can only reason over the branches that the recovered state makes visible.
Action	Translate intent into executable motion while enforcing constraints and guardrails.	Reference controllers, robot control stacks, DRIVE AV safety architecture, bounded execution on Thor-class hardware.	Action closes the loop in the physical world. It is what makes Physical AI physical rather than merely descriptive or predictive.

Table 1. A capability-first view of Physical AI. The central claim is that physical state recovery sits between world observation, learned priors and reasoning at decision time, while action closes the loop in the real world.

06. The missing question: what if the world is not properly observed?

This is the hinge of the paper. The strongest version of the current stack still leaves a crucial question unresolved: what happens when the world is not cleanly observed in the first place?

Recent work in end-to-end driving is explicitly moving toward long-tail and out-of-distribution evaluation. The Waymo Open Dataset for End-to-End Driving (WOD-E2E) was created to benchmark challenging long-tail scenarios. Newer work on appearance robustness, structural flaws in closed-loop driving, and causal de-confounding makes clear that robustness remains an active research problem rather than a solved one. [19][20][21][22]

Those efforts are important. But benchmarking what is hard is not the same as recovering what is unobserved. A benchmark can reveal that a system struggles under difficult conditions. It does not by itself solve cases where the observation basis is fundamentally limited. A microscope with insufficient resolution cannot infer invisible structure simply by labeling the case as difficult; it still needs a stronger observation or reconstruction layer.

07. Why video-first world models are powerful but incomplete

The critique of camera-only or video-first systems is not that they are unimportant. They are indispensable. The critique is that they assume the relevant physical state is sufficiently encoded in visual observations or downstream perception tokens.

That assumption holds best when the world is visible, well calibrated, and richly observed. It weakens when reality is only partially observed, when weak targets disappear behind stronger ones, when geometry creates blind regions, when clutter masks structure, or when the sensing process itself creates ambiguity that cannot be resolved by shallow smoothing.

A system can become excellent at predicting what the world should look like next and still struggle when the input itself is structurally degraded. In those cases, encoding physics from visual data is not the same thing as recovering the actual state under poor observability.

08. Atomathic and the physical state recovery thesis

This is where Atomathic fits architecturally. It is best understood not merely as an autonomy company, a defense company, or a radar company, but as a Physical AI sensing-and-reasoning company whose core capability is physical state recovery: the recovery of meaningful physical structure from sparse, noisy, overlapping, or degraded measurements before higher-level policy and control begin.

That matters because not all ill-posed problems are equal. Some are merely under constrained. Others are measured through a flawed lens. In the second class, simply taking more downstream samples with the same effective lens does not fully solve the problem. The right answer is stronger recovery under physics-based constraints.

Atomathic therefore complements rather than duplicates the broader Physical-AI stack. World foundation models such as Cosmos help build models of possible worlds. Embodied policy models such as Alpamayo and GR00T help machines choose among possible actions. Atomathic addresses the lower and more foundational problem of estimating the actual world when measurements are incomplete or structurally ambiguous.

09. Physical AI 2.0

Physical AI 1.0 made world models, synthetic data, simulation, and reasoning central. That shift is real, and NVIDIA has helped define it publicly. [4][5][12]

Physical AI 2.0 adds a new center of gravity: physical state recovery between world models and reasoning. The winning systems in this next phase will not only simulate better worlds or produce better reasoning traces. They will combine those capabilities with a stronger estimate of the actual world during real-world operation under incomplete observation, and then close the loop through real-world action.

This is the practical summary of the architecture: world models encode priors and plausible future world states. Physical state recovery improves what the system can know about the actual world. Reasoning systems determine what the system should do. Action turns that decision into physical consequence and produces new observations for the next loop.

Conclusion

Vision is extraordinarily powerful and, in many deployments, the dominant sensor modality. But it is not a complete observation basis for every physical environment. World models provide synthetic experience and dynamic priors. Physical state recovery improves what a system can know about the actual world. Reasoning systems improve action selection under uncertainty. Deployment platforms bring these capabilities into robots and vehicles.

But all of those layers depend on the quality of the state estimate beneath them. That is why the next missing capability is not just more reasoning. It is better observation translated into a more trustworthy estimate of physical reality, followed by bounded action in the real world. Stated plainly, vision and language-first world models are not sufficient on their own to satisfy the state goals of Physical AI.

This is the core argument of Physical AI 2.0: a complete embodied stack must connect possible worlds, the actual world under incomplete observation, reasoning about what to do, and action in the physical world. Read this way, the paper is not a rejection of world models; it is an attempt to sharpen one of their own foundational responsibilities while making the observation-action loop explicit.

Atomathic

Appendix A. NVIDIA's stack: world models, reasoning, orchestration, and deployment

NVIDIA's public stack is useful not just because of the individual products, but because it shows how the capabilities connect. Cosmos serves as the broad Physical-AI platform. Cosmos-Reason2 is an open reasoning VLM for Physical AI and robotics. Cosmos-Predict2.5 and Transfer2.5 handle future-state generation and controllable world generation. Isaac GR00T is a humanoid embodied policy / VLA model and data-pipeline effort for generalized humanoid skills. Alpamayo is the AV-specific reasoning-and-action stack. OSMO orchestrates Physical-AI workflows. NIM packages models as production-oriented APIs. Jetson Thor and DRIVE AGX Thor provide the edge compute substrate for robots and vehicles during deployment. DRIVE Hyperion provides a validated AV reference platform for sensors, compute, and software. [4][6][7][8][9][11][12][13][14]

Component	Contribution	Best interpretation in the capability stack
Cosmos-Reason2	Reasoning VLM for Physical AI and robotics.	Interpretation and planning layer that can propose intent or next steps; not a low-level actuator interface.
Cosmos-Predict2.5	Unified text / image / video-to-world model for future-state generation.	Possible-world generator for anticipation, data generation, and synthetic coverage.
Cosmos-Transfer2.5	Controllable world model that conditions generation on structured inputs.	Realism and diversity engine for controlled scenario variants.
Isaac GR00T	Humanoid embodied policy / VLA model and data-pipeline effort for generalized humanoid skills.	Robot-side embodied policy and action layer.
Alpamayo	Open VLA family for reasoning-based autonomous driving.	AV-side embodied policy and action layer.
OSMO / NIM / deployment stack	Workflow orchestration, model serving, and edge compute for deployment.	Operational layer that turns models into repeatable, deployable systems.

Appendix Table A1. A compact interpretation of the current NVIDIA Physical-AI stack as of April 2026.

Appendix B. Worked examples: robot folding and autonomous driving

The same control architecture becomes easier to see when mapped onto two concrete examples: a humanoid robot folding laundry and an autonomous vehicle navigating long-tail road scenes. The point is not that the embodiments are identical. It is that the same architectural logic reappears across them.

Capability	Robot folding / laundry process	Autonomous driving
World Models	Cosmos-Predict2.5 and Transfer2.5 expand coverage by generating new lighting, viewpoints, clutter, and garment presentations around the same workflow.	Cosmos-Predict2.5 and Transfer2.5 expand AV coverage by creating new weather, lighting, traffic, and sensor-appearance variations around the same base scenarios.
Physical State Recovery	The robot must recover garment state from partial views, slip, occlusion, and ambiguous contact before higher-level policy can act reliably.	The vehicle must recover the current scene from clutter, ambiguity, occlusion, and sensing limitations before reasoning can become dependable.
Reasoning Systems	Cosmos-Reason2 and Isaac GR00T interpret garment state, fragility, task intent, and failure modes, then propose actions such as fold, re-grasp, set aside, or ask for help.	Cosmos-Reason2 and Alpamayo-style reasoning interpret road-user intent, gestures, occlusions, and unusual hazards, then propose actions such as yield, stop, nudge, or replan.
Action	Jetson Thor and robot control stacks execute the bounded motion plan in the physical world.	DRIVE AGX Thor, DRIVE Hyperion, and AV control stacks execute bounded vehicle motion in the physical world.

Appendix Table B1. The same explanatory sequence - world models, physical state recovery, reasoning, and action - is portable across embodiments even though the sensors, tasks, and safety envelopes differ.

Atomathic

References

[1] LeCun, Y., "A Path Towards Autonomous Machine Intelligence," OpenReview, 2022. https://openreview.net/forum?id=BZ5a1r-kVsf

[2] Assran et al., "Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture," CVPR 2023 / arXiv:2301.08243. https://arxiv.org/abs/2301.08243

[3] Assran et al., "V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning," arXiv:2506.09985, 2025. https://arxiv.org/abs/2506.09985

[4] NVIDIA Cosmos documentation, Introduction. https://docs.nvidia.com/cosmos/latest/introduction.html

[5] Agarwal et al., "Cosmos World Foundation Model Platform for Physical AI," arXiv:2501.03575, 2025. https://arxiv.org/abs/2501.03575

[6] NVIDIA Cosmos-Reason2 documentation. https://docs.nvidia.com/cosmos/latest/reason2/index.html

[7] NVIDIA Cosmos-Predict2.5 documentation. https://docs.nvidia.com/cosmos/latest/predict2.5/index.html

[8] NVIDIA Cosmos-Transfer2.5 documentation. https://docs.nvidia.com/cosmos/latest/transfer2.5/index.html

[9] NVIDIA Isaac GR00T developer page. https://developer.nvidia.com/isaac/gr00t

[10] NVIDIA Isaac Lab developer page. https://developer.nvidia.com/isaac/lab

[11] NVIDIA Alpamayo developer page. https://developer.nvidia.com/drive/alpamayo

[12] Wang et al., "Alpamayo-R1: Bridging Reasoning and Action Prediction for Generalizable Autonomous Driving in the Long Tail," arXiv:2511.00088, 2025. https://arxiv.org/abs/2511.00088

[13] NVIDIA OSMO platform page. https://developer.nvidia.com/osmo

[14] NVIDIA NIM documentation. https://docs.nvidia.com/nim/index.html

[15] NVIDIA Jetson Thor page. https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-thor/

[16] NVIDIA DRIVE Hyperion / in-vehicle computing page. https://www.nvidia.com/en-us/solutions/autonomous-vehicles/in-vehicle-computing/

[17] Hwang et al., "EMMA: End-to-End Multimodal Model for Autonomous Driving," arXiv:2410.23262, 2024. https://arxiv.org/abs/2410.23262

[18] Waymo, "Scaling Laws of Motion Forecasting and Planning," 2025. https://waymo.com/research/scaling-laws-of-motion-forecasting-and-planning/

[19] Xu et al., "WOD-E2E: Waymo Open Dataset for End-to-End Driving in Challenging Long-tail Scenarios," arXiv:2510.26125, 2025. https://arxiv.org/abs/2510.26125

[20] Wang et al., "Benchmarking and Bridging Appearance Robustness in End-to-End Autonomous Driving," arXiv:2602.12563, 2026. https://arxiv.org/abs/2602.12563

[21] Gerstenecker et al., "PlanT 2.0: Exposing Biases and Structural Flaws in Closed-Loop Driving," arXiv:2511.07292, 2025. https://arxiv.org/abs/2511.07292

[22] Tang et al., "CausalVAD: De-confounding End-to-End Autonomous Driving via Causal Intervention," arXiv:2603.18561, 2026. https://arxiv.org/abs/2603.18561

[23] Guo et al., "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," arXiv:2501.12948, 2025. https://arxiv.org/abs/2501.12948

Public sources current as of April 2026.