ARTIFICIAL INTELLIGENCE
LABORATORY · MONACO

FRONTIER
RESEARCH

PHYSICAL INTELLIGENCE · DECEMBER 2025

From Words to World Models

Babies beat ChatGPT at physics. What that tells us about the next architecture, the next benchmark, and the next capital cycle.



Leonardo Fabbri
Founder & Chief Executive

WE READ THE RESEARCH,
YOU GET THE INSIGHTS

EXECUTIVE SUMMARY

Fluency about physics is not competence at physics. The distinction is now measurable.

An eleven-month-old who has never heard the word *gravity* will stare at a ball that appears to roll through a solid wall. That stare is a measurement. It says the child was running a prediction, and the world broke it. Frontier language models, asked the same question in words, answer correctly and confidently. Asked to predict the same event in a scene, most of them cannot tell the possible from the impossible.

The gap is now documented rather than asserted. **IntPhys 2**, the 2025 successor to the standard intuitive-physics benchmark, tests four properties infants are known to expect — permanence, immutability, spatio-temporal continuity and solidity — using the same violation-of-expectation logic developmental psychologists use on infants. Most state-of-the-art models score at chance, around 50%. Human observers score near-perfect.[1]

On **CausalVQA**, which asks video models what would have happened had something changed, the strongest system evaluated reaches 61.7% against a human baseline of 84.8%. On anticipation questions — what happens next — the gap widens to 38 percentage points.[2] On **NEWTON**, language models answer roughly 50% of object-attribute physical reasoning items correctly where humans answer 84%.[3]

This is not a knowledge deficit that more text will close. It is an architectural one. Next-token prediction optimises for the plausibility of a continuation, not for the accuracy of a state transition. It is not conditioned on actions, it is not required to carry occluded state, and it does not learn to roll a situation forward and choose between imagined futures. A system can be eloquent about solidity and still be unreliable the moment a scene is hidden, moved or interfered with.

The counter-evidence for the architectural reading is precise. When a model is trained to predict masked video *in learned representation space* rather than in pixels, the infant signature appears. V-JEPA reaches 98% zero-shot pairwise accuracy on IntPhys, and separates possible from impossible events on object permanence, continuity and shape constancy at 84–86% against roughly 51% for an untrained network. Pixel-space video prediction and multimodal language models sit only marginally above that untrained baseline.[4]

The efficiency result matters as much as the accuracy result. That capability survives at **128 hours** of unique video and appears in a **115-million-parameter** model. Intuitive physics, on this evidence, is cheap — but only if the prediction target is abstract. It is the representation space, not the parameter count, that is doing the work.

Capital has already repriced against this thesis. Yann LeCun left Meta and raised **$1.03bn** for AMI Labs at a $3.5bn pre-money valuation in March 2026; World Labs raised **$1bn** in February 2026; Prometheus raised **$12bn** in June 2026.[5] The bet is coherent. The timing is not yet underwritten: industrial deployment gates on reliability above 99%, and the best generalist manipulation policies demonstrated in 2026 sit between 50% and 83% on precision tasks, with transfer to new objects and environments largely undemonstrated.[6]

Our conclusion is narrow and, we think, defensible. The next capability cycle will be won by systems that pair a **world model** for grounded, action-conditioned prediction with **language as the interface layer** for abstraction, instruction and explanation. The measurement discipline has to move first: from asking a model whether an event *sounds* right to measuring how *surprised* it is when the event is impossible.

Leonardo Fabbri
Founder & Chief Executive · Artificial Intelligence Laboratory Monaco

The full research edition of **"Babies Beat ChatGPT at Physics"**, published by AI Lab Monaco on 29 December 2025.

CONTENTS

What is in this note

01 The Diagnostic: what infants know before they speak — 05
Core knowledge, the violation-of-expectation paradigm, and what a stare actually measures

02 The Mismatch: why next-token prediction is not physical competence — 08
Action-conditioning, partial observability, and planning by imagination

03 The Evidence: benchmarks and the surprisal gap — 11
IntPhys 2, GRASP, InfLevel, CausalVQA, NEWTON, PHYRE — and what each result licenses

04 The Architecture: prediction in representation space — 15
From world models and MuZero to the JEPA family; why pixels are the wrong target

05 Measurement Discipline: from looking time to surprisal — 18
A protocol for turning a developmental paradigm into a machine-readable metric

06 The Capital Map: who is building what, and on whose balance sheet — 20
Rounds, valuations, platforms and the financing architecture behind physical AI

07 The Reliability Threshold: from demonstration to deployment — 22
What robots actually do in 2026, and the distance between 83% and 99%

08 Risks to the Thesis — 24
Five ways this argument could be wrong, and what would show it

09 Conclusion and Watch List — 25
Dated signals we are tracking through 2027

— References — 26

— Methodology, Scope and Important Disclosures — 27

EXHIBITS

1 · The claim in one chart
2 · The violation-of-expectation paradigm
3 · Surprise as a learning signal
4 · Three ways of predicting, compared
5 · Where the surprise signal comes from
6 · Zero-shot violation-of-expectation accuracy
7 · IntPhys properties: learned vs. untrained
8 · The comprehension–prediction gap
9 · The benchmark landscape
10 · Intuitive physics is cheap, if the target is abstract
11 · Anatomy of a world model
12 · Lineage: five systems that built the idea
13 · A surprisal-gap evaluation protocol
14 · Capital committed to world models and physical AI
15 · The reliability ladder
16 · Watch list to end-2027

KEY JUDGMENTS

Five conclusions, and the evidence each rests on

01 **The deficit is predictive, not verbal — and it is measurable today**

Frontier systems describe physical constraints correctly and fail to anticipate them. Across IntPhys 2, CausalVQA and NEWTON the pattern repeats: model accuracy collapses toward chance precisely where the task requires forward prediction under occlusion or intervention, while humans hold at 84–95%. Verbal benchmarks flatter these systems; predictive benchmarks do not.

02 **The prediction target, not the parameter count, decides whether physics is learned**

Models that predict masked video in learned representation space acquire object permanence, continuity and shape constancy. Models that predict pixels, and multimodal models that predict text, do not — they sit near an untrained network. The result holds at 115M parameters and 128 hours of unique video, which is a statement about architecture rather than scale.

03 **Evaluation should measure surprise, not agreement**

The developmental paradigm translates cleanly: replace looking time with prediction error in latent space, and score the *surprisal gap* between matched possible and impossible trajectories. Matched pairs, latent-space scoring and generalisation splits are what separate a model that has internalised a constraint from one exploiting a surface cue. We regard benchmark hygiene as the binding constraint on honest capital allocation in this category.

04 **The endpoint is a hybrid, not a replacement**

Language is the wrong substrate for grounded prediction and the right substrate for abstraction, instruction and explanation. We expect the leading systems of the next cycle to pair a world model for action-conditioned rollout with a language interface — not to discard either. Positioning that treats this as a winner-take-all contest between LLMs and world models misreads the engineering.

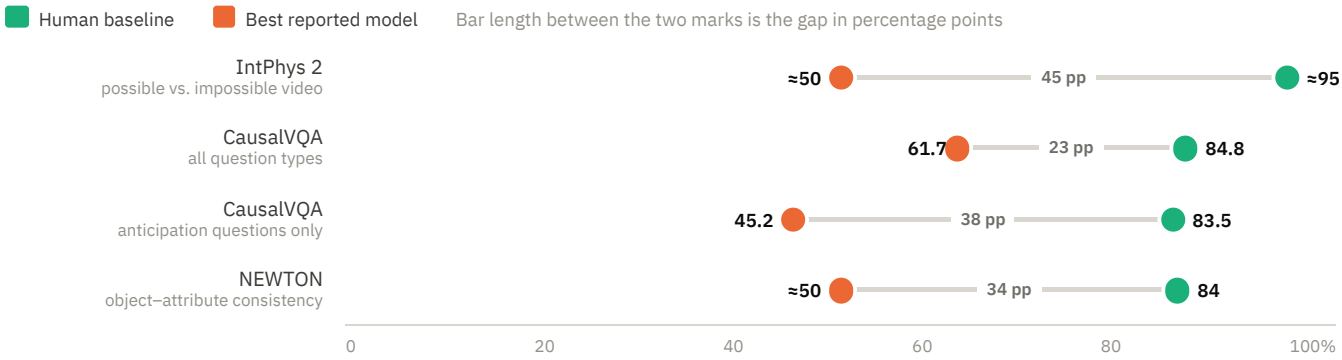
05 **Capital is ahead of deployment economics, and the gap is a reliability number**

More than $14bn of disclosed primary capital went into world-model and physical-AI platforms between February and June 2026 alone. Warehouse-grade deployment requires better than 99% reliability; the best demonstrated generalist manipulation in 2026 runs 50–83% on precision tasks, 3–5× slower than human operators, with transfer largely untested. The thesis can be right and the 2026 vintage still be early.

EXHIBIT 1

The claim in one chart: where machines fall away from people

Accuracy, %. Best reported model versus human baseline on four physical-prediction tasks. Chance is 50% on IntPhys 2; the CausalVQA and NEWTON formats are not two-alternative, so chance is lower.



Source: Bordes et al., *IntPhys 2* (2025); Meta AI, *CausalVQA* (2025); *NEWTON*, Findings of EMNLP (2023). Best model on CausalVQA is Gemini 2.5 Flash. IntPhys 2 and NEWTON model figures are the papers' characterisations ("at chance", "roughly 50%") rather than a single published point estimate. AI Lab Monaco analysis.

01

THE DIAGNOSTIC

What infants know before they speak

Developmental psychology has owned a working instrument for measuring physical expectation for four decades. Artificial intelligence has been slow to borrow it — and the reason it is worth borrowing now is that it measures prediction rather than articulation.

The claim from the developmental literature is narrow and well replicated. Long before they can speak, infants behave as though they hold expectations about how objects behave: that an object continues to exist when it is hidden, that it travels on a connected path rather than teleporting, that two solid objects cannot occupy the same place, and that an unsupported object falls. Elizabeth Spelke's core-knowledge programme and Renée Baillargeon's occlusion studies established the finding; the debate since has been about its origins and its limits, not about whether the expectations are there.

The measuring instrument is the **violation-of-expectation** paradigm. An infant is shown an event that unfolds behind an occluder and then resolves in one of two ways — one consistent with physical constraint, one not. Everything else about the two displays is matched. The dependent measure is looking time. Infants look reliably longer at the impossible outcome.

The inference the paradigm licenses is a prediction claim, not a knowledge claim. The infant is not reporting a belief. The infant generated an expectation, the world contradicted it, and the contradiction had a behavioural cost. That is a system running a forward model and registering error against it — which is precisely the property we want to test for in a machine.

Three features of the paradigm make it useful as a specification for AI evaluation, and it is worth being explicit about them because each maps onto a design choice later in this note.

It is pre-linguistic. Whatever the infant has, it was acquired without instruction, without labels and without a corpus of physical explanations. Whatever a language model has, it was acquired almost entirely from descriptions produced by people who already had the underlying competence.

Surprise is a scalar the system computes about its own prediction. Looking time is a proxy for an internal quantity: the size of the mismatch between what was predicted and what occurred. That quantity has a direct machine analogue — prediction error, or negative log-likelihood — and it can be read out without asking the model anything.

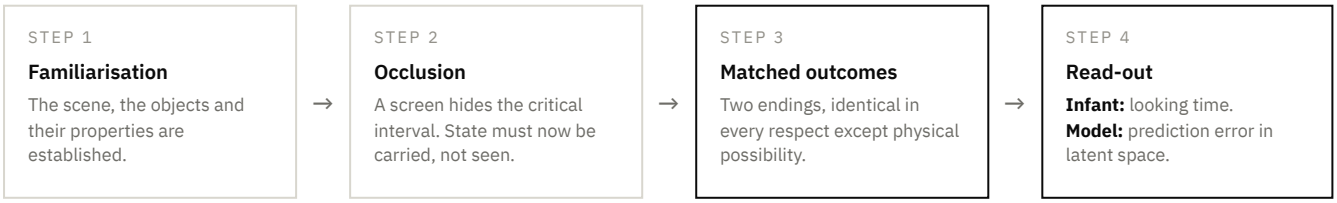
The response to surprise is targeted. This is the part most often left out of the AI retelling, and it is the most diagnostic. Infants do not merely stare; they interrogate the specific constraint that was broken.

Taken together these give us something better than an analogy. They give us a *specification*: a physically competent system should generate expectations without being asked, register a graded error when they fail, and localise the failure well enough to act on it.

EXHIBIT 2

The violation-of-expectation paradigm, and its machine translation

The same experimental spine serves both subjects; only the dependent measure changes.



Source: AI Lab Monaco, after Baillargeon (1987), Spelke et al. (1992) and the machine-readable formulation used in Riochet et al., *IntPhys* (2018) and Garrido et al. (2025).

Surprise is not a stare. It is a query.

In 2015 Aimee Stahl and Lisa Feigenson published a result in *Science* that changed what the paradigm is evidence *for*. Working with eleven-month-olds across four experiments, they showed that violating an expectation does not merely attract attention — it changes what the infant subsequently learns and what the infant subsequently does.

Infants who watched a ball appear to pass through a solid wall learned new facts about that ball. Infants who watched the same ball behave predictably showed no such learning. Given a free choice between the surprising object and a brand-new toy they had never seen, infants chose the surprising object — novelty alone was not the driver.

The decisive detail is what they did with it. When the violation was **support** — the ball appeared to hover unsupported — infants dropped it, testing gravity. When the violation was **solidity** — the ball appeared to pass through a wall — infants banged it, testing rigidity. The exploratory action was matched to the constraint that had been broken.[7]

That specificity is the strongest available evidence that what is being violated is structured, not generic. A novelty detector produces a uniform "something is odd here" response. A system that knows *which* expectation failed produces a targeted probe. As Feigenson put it: infants behave "as though they are taking the occasion to try to figure something out about their world."

For our purposes the implication is a design requirement. A useful benchmark should not only ask whether a model registers that something is wrong. It should ask whether the model localises *what* is wrong — and whether that localisation survives changes to texture, viewpoint, background and object identity that leave the physics untouched.

EXHIBIT 3

Surprise as a learning signal, not an attention signal

Stahl & Feigenson (2015), eleven-month-old infants, four experiments.

MANIPULATION	WHAT THE INFANT SAW	WHAT FOLLOWED
Solidity violated	A ball rolls down a ramp and appears to pass through a solid wall	Enhanced learning about that object; infants **banged** it — a rigidity test
Support violated	A ball appears to remain suspended with no support beneath it	Enhanced learning about that object; infants **dropped** it — a gravity test
Expectation met	The same ball, same scene, physically ordinary outcome	No evidence of learning about the object
Novelty control	Choice between the surprising object and an unseen new toy	Infants preferred the **surprising** object over the merely novel one

Source: Stahl, A. E. & Feigenson, L., "Observing the unexpected enhances infants' learning and exploration", *Science* 348(6230), 3 April 2015, pp. 91–94.

What the paradigm does not license

The weakness is worth stating plainly, because it reappears in the machine setting and is the reason §5 of this note is about protocol rather than architecture.

Longer looking is not self-interpreting. It can reflect perceptual novelty or low-level mismatch rather than a conceptual violation — impossible events are hard to match perfectly against possible ones, and a substantial methodological literature exists for that reason.

The machine analogue is exact: a model can score well on a possible-versus-impossible discrimination by picking up a rendering artefact or a texture cue that correlates with the manipulation and carries no physical content. It looks like physics. It transfers to nothing.

The response in both fields is the same: match the pairs to the frame, vary everything that should be irrelevant, build generalisation splits that make shortcut learning expensive, and report per-property results rather than an aggregate.

Why a developmental result belongs in a technology note

There is a reasonable objection to the whole framing: infants are not a benchmark, and comparing a child to a language model is a category error dressed up as a headline. We accept the objection and think the framing survives it, for a specific reason.

The value of the infant result is not that babies "beat" anything. It is that infants establish an **existence proof with a known budget**. Whatever produces reliable expectations about solidity, continuity, support and containment can be produced without language, without supervision, without a labelled corpus, and inside a small, slow, extremely data-limited learner operating for under a year.

That existence proof has a direct bearing on an engineering argument that is often made in the other direction. If physical expectation required enormous scale, the failure of frontier systems on physical prediction would be a matter of waiting. The infant result says it does not require enormous scale. It says the ingredient is structural.

And, as §3 shows, the machine evidence now agrees. The system that acquires infant-like expectations does so at 115 million parameters and 128 hours of unique video — while systems thousands of times larger, trained on incomparably more data, sit near chance. Scale is not the discriminating variable in the published results. The prediction target is.

There is also a second-order reason to take the developmental framing seriously, which is about evaluation culture rather than architecture. Developmental psychology had to invent methods for testing competence in a subject that cannot be asked questions. Artificial intelligence has the opposite problem: its subjects answer every question fluently, which makes verbal evaluation systematically unreliable. Borrowing methods built for a mute subject is a way of getting an answer that fluency cannot contaminate.

> An infant is an existence proof with a known budget: no language, no labels, no supervision, under a year of data — and reliable expectations about solidity, continuity and support. The constraint that binds current systems is therefore unlikely to be scale.

SECTION 01 · WHAT WE TAKE FORWARD

- Physical competence is testable without language, by measuring the cost of a broken prediction rather than the quality of an explanation.
- Surprise in infants is a *structured* signal — it identifies which constraint failed, and drives a matched investigation. A benchmark should test for that specificity.
- The paradigm is only as good as its controls. Matched pairs and generalisation splits are not refinements; without them the measurement is uninterpretable in either subject.
- The infant result reframes the AI question from “how much more scale?” to “what is being predicted, and in what space?”

02

THE MISMATCH

Why next-token prediction is not physical competence

Ask a frontier model why a ball stops at a wall and the answer is correct, complete and well organised. This tells you the model has an excellent account of how people talk about collisions. It does not tell you the model can say where the ball will be in 400 milliseconds.

We call this the **eloquence trap**, and it is not a rhetorical point — it is a claim about what the training objective optimises. Autoregressive language modelling maximises the likelihood of the next token given the preceding context. It is an extraordinarily effective objective, and the competence it produces is real. But the object it learns is the distribution over *descriptions* of the world, not the transition function *of* the world.

For most tasks that distinction is invisible, because descriptions of a well-understood domain contain almost everything you need. For physical prediction it is decisive, and it shows up in three specific structural mismatches.

1 · ACTION-CONDITIONING

Physics in an embodied setting is not a question about what tends to happen. It is a question about what happens *if I do this*. That requires a model of the form `s' = f(s, a)` — next state as a function of current state and chosen action.

Next-token prediction does not optimise for that object. It learns the statistics of sequences in which causes and effects are both described, which is a different quantity from the effect of an intervention. The model can be extremely well calibrated about how humans narrate consequences and still have no reliable handle on what its own action would do.

2 · PARTIAL OBSERVABILITY

In any real scene most of the relevant state is hidden most of the time. Objects go behind other objects; the agent's own body occludes the workspace; the camera moves. Competence requires *carrying* a belief about the occluded state and updating it as evidence arrives.

This is exactly what the occluder in the violation-of-expectation paradigm tests, and it is exactly what a system that describes visible frames is never required to do. A captioner that is superb on what it can see has no pressure to represent what it cannot.

3 · PLANNING BY IMAGINATION

Choosing an action well means rolling several candidate futures forward and comparing their outcomes against a goal. That is a search over imagined trajectories, and it requires a transition model cheap enough to run many times.

An autoregressive narrator continues the single most plausible story. It is not architecturally organised to branch, evaluate and select — and chain-of-thought, which is sometimes offered as the answer here, is still narration: it produces a plausible sequence of reasoning tokens rather than an evaluated rollout in a state space.

EXHIBIT 4

Three ways of predicting, compared

The differences that matter are in the first two rows. Everything below them follows from what is predicted and in what space.

	NEXT-TOKEN PREDICTION Large language & multimodal models	PIXEL-SPACE PREDICTION Generative video models	LATENT-SPACE PREDICTION World models, JEPA family
What is predicted	The next symbol in a description	The next frame, pixel by pixel	The next **representation** of the scene
What the objective rewards	Plausibility of a continuation	Photometric fidelity, including detail that carries no physics	Accuracy of a state transition under an invariant encoding
Conditioned on actions	No — only on text about actions	Optionally, and usually weakly	Yes — this is the design purpose
Carries occluded state	Only where the text says so	Poorly; the model must re-render what it cannot see	Yes — the latent state is the belief
Supports planning	By narration, not by evaluated rollout	Expensive: each imagined step costs a full frame synthesis	Yes — rollouts are cheap in representation space
Characteristic failure	Fluent, confident, wrong under occlusion and intervention	Capacity spent on texture and lighting; brittle dynamics	Representation collapse; abstractions that discard task-relevant detail
Reported violation-of-expectation performance	**Marginally above a randomly initialised network**	**Marginally above a randomly initialised network**	**98% on IntPhys; 66% GRASP; 62% InfLevel-lab**

Source: AI Lab Monaco analysis. Performance row from Garrido, Ballas, Assran, Bardes, Najman, Rabbat, Dupoux & LeCun, "Intuitive physics understanding emerges from self-supervised pretraining on natural videos", February 2025 — V-JEPA versus VideoMAEv2 (pixel-space) and Qwen2-VL-7B / Gemini 1.5 Pro (multimodal), zero-shot pairwise classification.

The failure has a signature, and the signature is diagnostic

If the eloquence trap is real, it should leave a specific fingerprint in benchmark results: descriptive performance close to human, predictive performance collapsing toward chance, within the same model, on the same videos, in the same evaluation.

That is what CausalVQA reports. On the full set of question types the best model evaluated reaches 61.7% against a human baseline of 84.8%. Restricted to the reasoning subset — excluding purely descriptive items — the same model falls to 53.5%. On anticipation questions, which ask what will happen next, it reaches 45.2% against 83.5% for humans.[2]

The gradient is the evidence. A model that had simply not seen enough video would be uniformly weak. A model that describes well and anticipates badly has learned the surface of the domain and not its dynamics.

The same gradient appears in IntPhys 2, where the manipulation is purely physical and the descriptive shortcut is removed by construction: models that perform competently on ordinary visual understanding drop to chance when asked which of two matched videos is impossible.[1]

> A model that had simply not seen enough of the world would be uniformly weak. A model that describes the world well and anticipates it badly has learned the surface of the domain and not its dynamics. That asymmetry is the whole argument.

Why predicting pixels does not fix it either

The obvious repair is to move from text to video: train the model to predict what the next frame looks like, and physics should fall out of the objective. It is a reasonable idea, it has produced remarkable generative systems, and on this particular test it does not work.

The problem is that pixel-level prediction forces the model to spend capacity on everything in the image, in proportion to how many pixels it occupies rather than how much it matters. A wood grain, a shadow gradient and the trajectory of the object about to fall all compete for the same loss, and the physically irrelevant detail dominates.

The empirical result is unambiguous: a pixel-prediction model evaluated under the same violation-of-expectation protocol performs only marginally above a randomly initialised network.[4] Fidelity is not understanding, and a video model can be beautiful and physically illiterate at the same time.

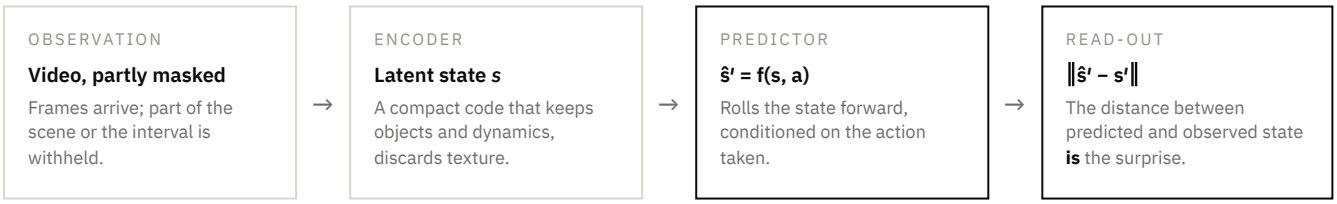
This is worth stating carefully, because the generative video frontier moved a long way in 2025 and 2026 and it is easy to read that progress as evidence on this question. DeepMind's Genie 3 sustains an interactive, navigable environment at 720p and 24 frames per second with roughly a minute of visual memory, and its physics is learned rather than scripted.[8] NVIDIA's Cosmos world-foundation models were trained on the order of 20 million hours of video and had passed two million downloads by January 2026.[9] These are substantial systems.

But rendering a plausible world and predicting a specific one are different tasks. A model can hold a convincing environment for a minute and still fail a matched-pair discrimination an eleven-month-old passes, because nothing in a generative objective penalises the impossible outcome — only the one that looks wrong. The two capabilities are complements, not substitutes.

EXHIBIT 5

Where the surprise signal comes from

Prediction error is measured against a learned representation of the scene, not against its pixels. This is the single design choice that separates the two right-hand columns of Exhibit 4.



If the read-out is in pixel space

Error is dominated by unpredictable surface detail. A physically impossible event and a slightly different shadow produce comparable loss, so the signal that matters is buried.

If the read-out is in representation space

The encoder has already discarded what does not transfer. Remaining error concentrates on object identity, position and dynamics — the quantities physical constraints govern.

Source: AI Lab Monaco, after LeCun (2022) and the I-JEPA / V-JEPA formulation. *a* denotes the action; in purely observational pretraining the predictor is conditioned on masking and temporal offset instead.

SECTION 02 · WHAT WE TAKE FORWARD

Three properties are required for physical competence and none of them is produced by next-token prediction: conditioning on **actions**, carrying **occluded state**, and **rolling futures forward** to choose between them.

Moving the prediction target from text to pixels does not supply them. Moving it to a learned representation does. The rest of this note is about how far that claim has been demonstrated, how it should be measured, and what it is worth.

03

THE EVIDENCE

Benchmarks, and what each result actually licenses

The evidence base is now good enough to separate three claims that are routinely conflated: that models cannot do physics, that scale will fix it, and that one architecture family is measurably above the floor while the others are not. Only the third is supported.

The strongest single published result comes from a February 2025 paper by Garrido and colleagues at Meta's FAIR laboratory, with Yann LeCun as senior author. The setup is deliberately austere: take a video model trained only to predict masked regions of natural video in representation space, give it no physics training, no fine-tuning and no task supervision, and run it through the violation-of-expectation protocol.

V-JEPA reaches **98%** zero-shot pairwise accuracy on IntPhys, **66%** on GRASP and **62%** on InfLevel-lab. A pixel-space video model (VideoMAEv2) and two multimodal language models (Qwen2-VL-7B and Gemini 1.5 Pro) perform only marginally above a randomly initialised network on the same protocol.[4]

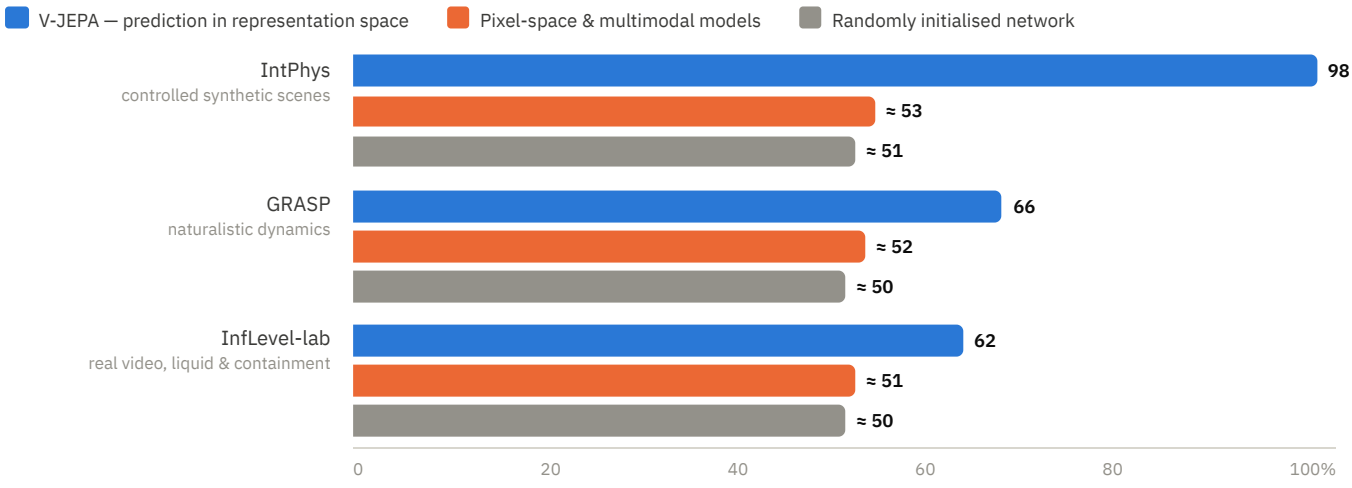
Two things about that result deserve emphasis, and they point in opposite directions.

The first is that 66% and 62% are not good scores. GRASP and InfLevel are harder and more naturalistic than IntPhys, and on those the margin over chance is real but modest. The honest reading is not that physical understanding has been solved. It is that, under a controlled comparison, exactly one architecture family clears the floor at all — and the variable that separates it from the others is the space in which prediction happens.

EXHIBIT 6

Zero-shot violation-of-expectation accuracy, by architecture

Pairwise classification of matched possible/impossible video, %. No physics-specific training or fine-tuning in any condition.



Source: Garrido et al. (February 2025). V-JEPA figures are the reported point estimates with 95% confidence intervals of [95, 99], [64, 68] and [60, 63] respectively. **Note:** the paper reports pixel-space and multimodal baselines as "only marginally above" randomly initialised networks rather than as tabulated point estimates for each benchmark; the orange bars are indicative of that characterisation, not published values. Chance is 50%.

The second point runs the other way. The separation was obtained with *no* physics-specific supervision, no simulator, no annotated trajectories and no fine-tuning on any of the three benchmarks. Whatever produced it was produced by the pretraining objective alone.

That is what makes the comparison informative rather than anecdotal. All four systems saw video. Only one was asked to predict it in a learned representation — and only one acquired the expectations.

The properties that separate, and the ones that do not

Aggregate accuracy hides the finding. Broken out by property, the pattern is the one the developmental literature would predict: the model acquires exactly the expectations infants show, and it acquires them to a similar degree.

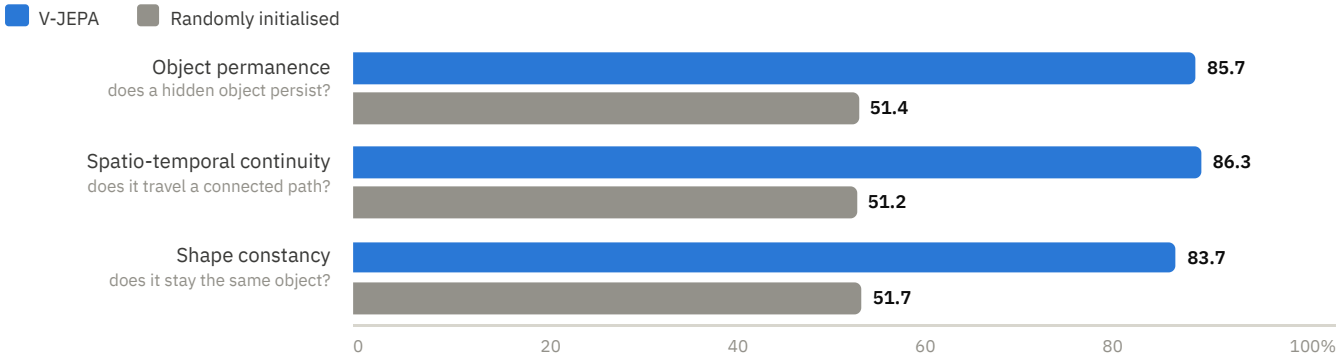
Object permanence, spatio-temporal continuity and shape constancy all separate cleanly from an untrained baseline that sits, as it should, at chance. There is no property on which the untrained network is competitive and none on which the trained model is at the floor.

That uniformity matters for the shortcut objection raised in §1. A model exploiting a rendering artefact would be expected to do well where the artefact is strong and collapse where it is not. Consistent separation across three structurally different violations is weak evidence against a single-cue explanation — weak, not conclusive, which is why §5 argues for protocol rather than trust.

EXHIBIT 7

IntPhys properties: a learned representation versus an untrained one

Pairwise classification accuracy, %. V-JEPA versus a randomly initialised network of identical architecture.



Source: Garrido et al. (2025), per-property results. **Note:** these property-level figures and the paper's headline 98% aggregate IntPhys figure are reported under related but not identical scoring conditions; we present both as published rather than reconciling them.

IntPhys 2 resets the difficulty — and the models fall back to chance

In June 2025 the same laboratory published IntPhys 2, which rebuilt the paradigm inside complex, varied synthetic environments and tested four principles: permanence, immutability, spatio-temporal continuity and solidity.[1]

The result is a corrective. Most state-of-the-art models return to chance — around 50% — while human observers score near-perfect. Basic visual understanding survives the move to complex scenes. Physical discrimination does not.

We read the two papers together rather than against each other. The 2025 result establishes that the *mechanism* works: predicting in representation space produces genuine physical expectations in controlled scenes. IntPhys 2 establishes that the *capability* is not yet robust to scene complexity. Both are true, and any thesis that cites one without the other is selling something.

The comprehension–prediction gap, measured inside one model

The cleanest demonstration of the eloquence trap is a benchmark that asks the same model both kinds of question about the same footage. CausalVQA does exactly that: 1,586 items across 793 paired questions drawn from 779 real egocentric video segments, spanning five question types.[2]

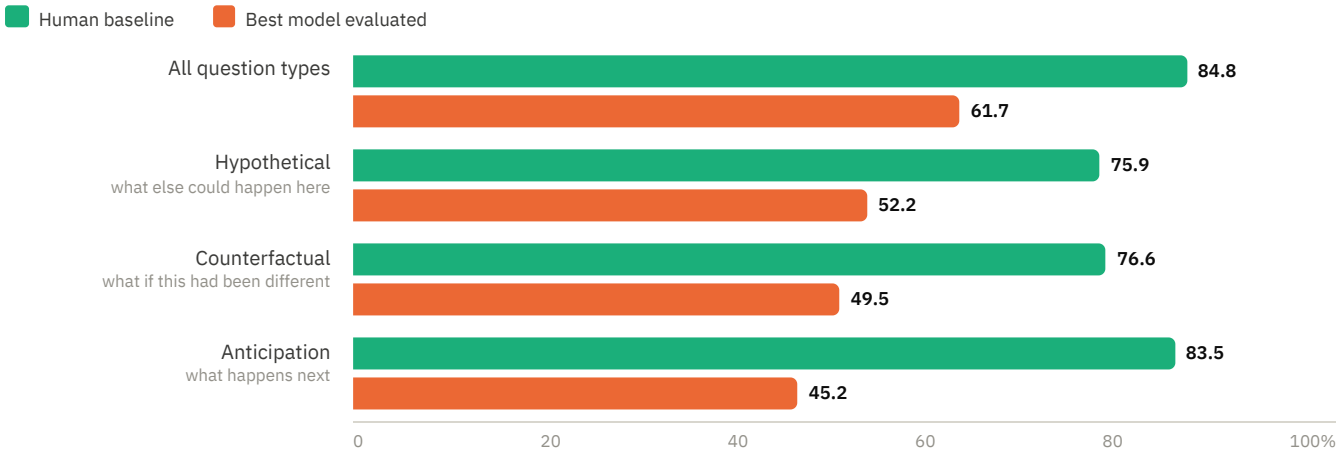
Descriptive items ask what happened. Counterfactual items ask what would have happened had something been different. Hypothetical items extend the scenario. Anticipation items ask what comes next. Planning items ask what steps achieve a goal. The physics is identical across types; only the direction of inference changes.

The result is a clean gradient. Overall, the best model tested reaches 61.7% against a human baseline of 84.8%. On the reasoning subset it falls to 53.5%. On counterfactual items the gap is 27 percentage points; on anticipation items it is 38. Whatever these systems have, it degrades in proportion to how far forward they are asked to run it.

EXHIBIT 8

The gap widens with the distance you have to predict

CausalVQA accuracy, %. Human baseline versus the strongest model evaluated (Gemini 2.5 Flash), by question type.



Source: Meta AI, *CausalVQA: A Physically Grounded Causal Reasoning Benchmark for Video Models* (2025), evaluated on EgoExo4D segments. Gaps: 23.1 pp overall, 23.7 pp hypothetical, 27.1 pp counterfactual, 38.4 pp anticipation.

> Descriptive questions flatter these systems. Predictive questions do not. The steepest degradation is on the one question an embodied agent has to answer continuously: what happens next?

EXHIBIT 9

The intuitive-physics benchmark landscape

What each benchmark can and cannot establish. Read the right-hand column as the constraint on what any result from it licenses.

BENCHMARK	YEAR	WHAT IT MEASURES	HUMAN VS. BEST MODEL	WHAT IT LICENSES — AND WHAT IT DOES NOT
IntPhys	2018	Violation of expectation in controlled synthetic scenes	V-JEPA 98% zero-shot	Establishes the mechanism cleanly. Scenes are simple, so a strong score does not imply competence in complex environments.
PHYRE	2019	Physical puzzles solved by placing one object; measured by AUCCESS	Sample-efficiency, not accuracy	Tests planning under physics rather than perception. Two-dimensional and stylised; transfer to real scenes is not evidenced.
NEWTON	2023	Physical attributes of objects, in language: 2,800 object–attribute pairs, ~160k questions	84% vs. ≈50%	Isolates the verbal channel. Shows language models are inconsistent even about static attributes; says nothing about dynamics.
GRASP **InfLevel**	2023 2022	Naturalistic dynamics; liquids, containment and support in real video	V-JEPA 66% / 62%	The hardest available VoE tests. Margins over chance are real but small — the right place to track progress.
IntPhys 2	2025	Permanence, immutability, continuity and solidity in complex synthetic environments	≈95% vs. ≈50%	The current frontier test. Resets difficulty and shows the 2018 result does not survive scene complexity.
CausalVQA	2025	Counterfactual, hypothetical, anticipation and planning questions on real egocentric video	84.8% vs. 61.7%	Separates description from prediction inside one model. Multiple-choice format, so absolute numbers overstate underlying competence.
MVPBench	2025	Physical understanding via minimal-change question pairs	Human 85–95%	Minimal-change pairing is the strongest available control against shortcut learning. Newest and least characterised of the set.

Source: AI Lab Monaco compilation from Riochet et al. (2018); Bakhtin et al., NeurIPS (2019); *NEWTON*, Findings of EMNLP (2023); Jassim et al. (2023); Weihs et al. (2022); Bordes et al. (2025); Meta AI (2025).

What the evidence base supports, stated conservatively

The published record supports a narrower claim than either the enthusiasts or the sceptics tend to make. It does not show that language models "cannot do physics", and it does not show that world models have solved physical prediction — IntPhys 2 puts them back at chance in complex scenes.

What it shows is a reliable *ordering*. Under matched, controlled comparison, prediction in learned representation space produces physical expectations that prediction in pixel space and in token space do not — reproduced across several benchmarks, model families and two model scales.

SECTION 03 · WHAT WE TAKE FORWARD

- One architecture family clears the floor on violation-of-expectation testing; the others sit within a few points of an untrained network.
- The separation appears on exactly the properties infants show — permanence, continuity, shape constancy — and appears uniformly across them.
- Within a single model, accuracy degrades in step with how far forward the question requires it to predict. That gradient is the cleanest signature in the evidence base.

04

THE ARCHITECTURE

Prediction in representation space

A world model is not a genre of product. It is a specific commitment: learn a compact internal state, learn how that state evolves under your own actions, and use the second to plan. Everything else — the modality, the scale, the interface — is downstream of that commitment.

The idea has a clean lineage and it is older than the current cycle. Ha and Schmidhuber's 2018 paper gave it its name and its most vivid demonstration: an agent that compresses its environment into a latent state, learns the dynamics of that state, and is then trained entirely inside its own imagined rollouts before being returned to the real environment.

MuZero, published in *Nature* in 2020, made the harder point. It mastered Go, chess, shogi and Atari while being told none of the rules, by learning only the parts of the dynamics that mattered for choosing well. The model it learned was not a simulator of the game; it was a representation sufficient for planning. That distinction — *sufficient for planning* rather than *faithful to the world* — is the intellectual core of the field.

DreamerV3, published in *Nature* in April 2025, established robustness. A single fixed configuration learned across more than 150 tasks spanning very different domains, and collected diamonds in Minecraft from scratch without human demonstrations — a long-horizon, sparse-reward problem that had resisted general methods.[10]

What that lineage lacked was a way to build the state representation from ordinary, unlabelled sensory data at scale. Reinforcement-learning world models learn their latents inside a task. The web does not come with tasks.

The Joint-Embedding Predictive Architecture is the proposed answer. Instead of reconstructing the input, encode it, and predict the *encoding* of the part you masked out. The loss lives in representation space, so the model is never penalised for failing to reproduce detail that carries no information about what happens next.

EXHIBIT 10

Intuitive physics is cheap — if the prediction target is abstract

Reported ablations on the V-JEPA violation-of-expectation result. Both figures are far below the scale at which multimodal models still fail the same test.

128 hrs

of unique video from HowTo100M still yields **above 70%** pairwise accuracy on physical properties

115 M

parameters — a compact V-JEPA variant — still reaches **over 85%**, so the effect is not a scale artefact

0

physics supervision, simulators, labelled trajectories or benchmark fine-tuning in any reported condition

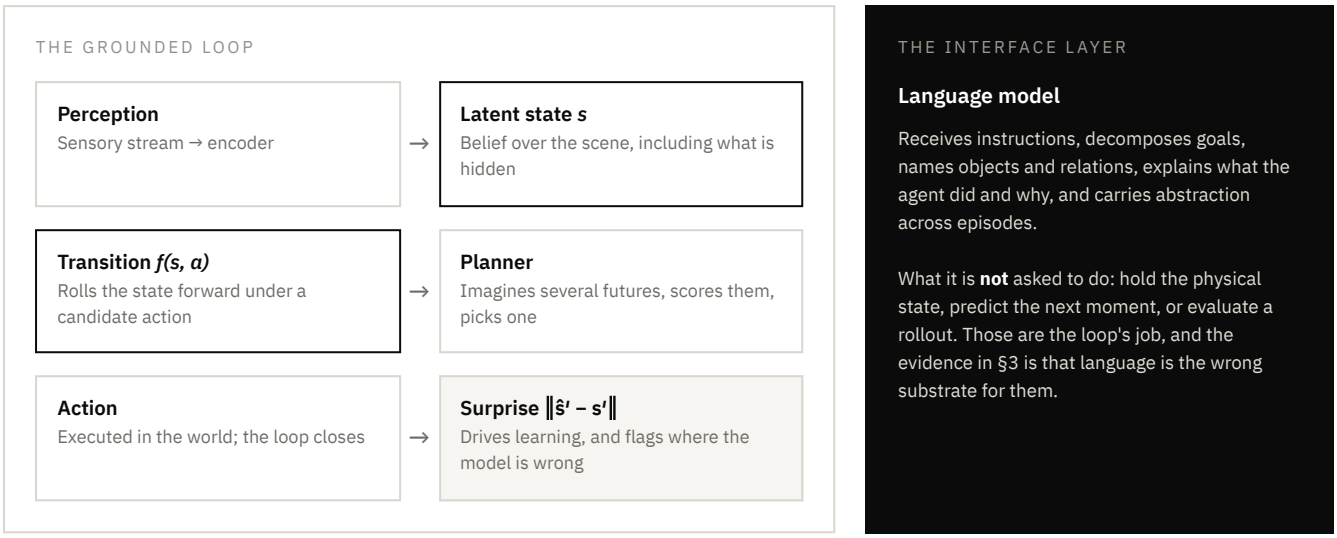
Source: Garrido et al. (2025). **Note:** the comparison models that fail the same protocol include systems several orders of magnitude larger, trained on far more data.

This is the result that should reorder priors. If physical expectation emerged only at frontier scale, the sensible inference would be to wait. It emerges at 115 million parameters and 128 hours of video, in a model with no physics training — and does not emerge in models thousands of times larger with a different prediction target. The discriminating variable is architectural.

EXHIBIT 11

Anatomy of a world model, and where language attaches

The loop on the left runs continuously and fast. Language sits above it as an interface, not underneath it as a substrate.



Source: AI Lab Monaco, synthesising Ha & Schmidhuber (2018), Schrittwieser et al., *Nature* (2020), Hafner et al., *Nature* (2025) and the V-JEPA 2 action-conditioned formulation (2025).

The honest weaknesses of the approach

Two failure modes are intrinsic to predicting in representation space, and both are live research problems rather than settled engineering.

The first is **representation collapse**. If the encoder chooses the space in which it is scored, the trivial solution is to map everything to a constant — a perfect prediction over a worthless representation. The standard answers (asymmetric architectures, stop-gradients, momentum encoders, variance regularisation) work in practice but are not derived from first principles.

The second is **the abstraction may throw away what you need**. Relevance is task-dependent: the surface texture a physics test can ignore is exactly what a grasping policy needs to judge friction. A representation tuned for one downstream use can be silently wrong for another.

Neither weakness undermines the benchmark evidence in §3, which is a comparison between architectures rather than a claim of completeness. Both are reasons to expect the path from a demonstrated mechanism to a deployed capability to be longer than the current funding environment implies.

What the hybrid already looks like in the field

The hybrid in Exhibit 11 is not a forecast. It is a reasonable description of what the leading embodied systems of 2026 are already assembled from, even where the vocabulary differs.

Vision-language-action models are the mainstream instantiation: a language-and-vision backbone supplies goal interpretation and object grounding, while a separately trained policy supplies closed-loop control. Physical Intelligence's π0.5, trained across roughly a hundred homes, is the clearest public example of the pattern generalising to unseen environments.[13] A Goldman Sachs survey of fourteen Chinese robotics developers describes the same convergence from the other direction: single-purpose systems moving toward integrated stacks that pair vision-language-action models *with* world models.[12]

Read that way, the architectural debate is less contested than the public framing suggests. Almost nobody building physical systems proposes removing the language layer. The disagreement is about what sits beneath it.

EXHIBIT 12

Lineage: five systems that built the idea

Each entry solved one problem the previous one left open.

SYSTEM	YEAR	CONTRIBUTION	WHAT IT LEFT OPEN
World Models Ha & Schmidhuber	2018	An agent trained entirely inside its own learned latent dynamics, then returned to the real environment	The latent space was learned per environment; nothing about learning it from open sensory data
MuZero DeepMind, *Nature*	2020	Planning with a learned model without being given the rules — a model sufficient for planning, not faithful to the world	Discrete, fully observed, reward-dense domains; nothing about messy continuous perception
DreamerV3 Hafner et al., *Nature*	2025	One fixed configuration across 150+ tasks; diamonds in Minecraft from scratch with no human data	Still task-situated learning; not a representation acquired from general video
I-JEPA / V-JEPA Meta FAIR	2023 2024–25	Self-supervised prediction of masked **representations** from images and natural video; intuitive physics emerges	Observational only — no actions, so no planning and no closed loop
V-JEPA 2 Meta FAIR	2025	1.2B parameters, >1M hours of video pretraining, then **62 hours** of unlabelled robot data to make the predictor action-conditioned	Manipulation success 65–80% on novel objects; short-horizon; far from deployment reliability

Source: AI Lab Monaco compilation. V-JEPA 2 also reported state-of-the-art results on Something-Something v2 motion understanding, Epic-Kitchens-100 action anticipation, and the Perception Test and TempCompass video question-answering benchmarks.

The 62-hour result, and why it is the interesting one

The V-JEPA 2 disclosure that deserves the most attention is not the parameter count. It is that the transition from a purely observational model to an action-conditioned one required roughly **62 hours** of unlabelled robot interaction data — after which the system performed zero-shot pick-and-place on objects it had never seen, in environments it had never been trained in, at 65–80% success.[11]

Robot data is the binding constraint of the entire embodied-AI category. It cannot be scraped; it has to be generated, on hardware, in real time, and it is therefore the most expensive training data in the industry. Chinese developers are building "data factories" specifically to manufacture it at volume.[12]

If a general representation can be learned from ordinary video — which is abundant and cheap — and only the action-conditioning needs robot data, the cost structure of the field changes materially. That is the commercial content of the architecture argument, and it is why the next two sections are about money and reliability rather than about models.

SECTION 04 · WHAT WE TAKE FORWARD

The design commitment is narrow: a **compact latent state**, a **transition function conditioned on actions**, and **planning by rollout**. Predicting representations rather than pixels is what makes all three affordable.

The open risks are representation collapse and task-blind abstraction. The commercial signal is that the expensive ingredient — robot interaction data — may be needed only for the action-conditioning layer, not for learning the world.

05

MEASUREMENT DISCIPLINE

From looking time to surprisal

The developmental paradigm survives translation almost intact. What changes is the dependent measure: instead of inferring surprise from how long a subject looks, you read it directly off the model's own prediction error. That single substitution is what makes physical competence auditable.

Start from the quantity. For a matched pair of trajectories — one physically possible, one not, identical in every other respect — compute the model's prediction error on each in representation space. Call the difference the **surprisal gap**:

```
Δ = S(impossible) − S(possible)
```

A model that has internalised the constraint assigns systematically higher surprise to the impossible event, so Δ is reliably positive. Pairwise accuracy — the metric reported throughout §3 — is simply the fraction of pairs for which $\Delta > 0$.

Reporting only that fraction throws away most of the information. The *distribution* of Δ tells you whether a model is confidently right, marginally right, or right by coin flip; two models at 66% can be very different systems. We would report the median gap, its spread, and the fraction of pairs where the gap is indistinguishable from zero.

The read-out window matters too, and the literature is explicit about it. Taking the maximum surprise across a video generalises better for classifying a single clip than averaging over it — which makes sense, since a violation is a moment rather than a property of the whole sequence. Any published number should say which was used.

None of this requires the model to answer a question, produce text, or be prompted in any way. That is the point. A verbal benchmark measures a system's account of itself; a surprisal benchmark measures the system.

> A verbal benchmark measures a system's account of itself. A surprisal benchmark measures the system. For a technology whose defining strength is producing convincing accounts, the difference is not academic.

Three things the metric buys that a question cannot

It is architecture-agnostic at the interface. Any system producing a prediction over a learned state can be scored, including systems with no language interface at all — which matters as the field fills with robot policies and video models never built to answer questions.

It is continuous. Accuracy is nearly useless below its decision boundary: two models at chance are indistinguishable. The magnitude of the surprisal gap keeps resolving differences long before the pairwise score moves off 50%.

It can be computed during training. Because the quantity is the model's own prediction error, it can be tracked as a curve rather than measured once at the end. Physical competence becomes something you can watch emerge — which is the condition under which anyone can say what caused it.

THE LIMIT OF THE INSTRUMENT

Surprisal is measured in each model's own representation space, so the raw magnitudes are not comparable between models — only the *sign* of the gap and a standardised effect size within a model are. Any table that ranks systems by absolute surprise is measuring the geometry of their latent spaces, not their grasp of physics. We would report the gap as a within-model standardised difference and let the pairwise accuracy carry the cross-model comparison.

EXHIBIT 13

A surprisal-gap evaluation protocol

Six controls. Each one exists to close a specific route by which a model can score well without having learned anything transferable.

	CONTROL	WHAT IT REQUIRES	WHAT IT DEFENDS AGAINST
1	**Matched pairs to the frame**	The possible and impossible clips differ only in the physical outcome — same assets, lighting, camera, timing	The model separating the classes on a rendering difference that has no physical content
2	**Nuisance randomisation**	Texture, viewpoint, background and object identity vary independently of the manipulation	A cue that happens to correlate with impossibility in this particular dataset
3	**Latent-space scoring**	Surprise measured against the learned representation, with the read-out window stated	Pixel-level error swamping the physical signal; ambiguity about what a number means
4	**Generalisation splits**	Held-out object classes, scenes and violation instances, not just held-out clips	Memorisation of the specific violations that appear in training
5	**Per-property reporting**	Separate results for permanence, continuity, solidity, support and containment	An aggregate that hides competence on one property and chance on four
6	**Two baselines, always**	A randomly initialised network of identical architecture, and a human baseline collected under the same conditions	Reading a modest margin over chance as evidence of understanding

Source: AI Lab Monaco, drawing on the control designs used in IntPhys 2, MVPBench's minimal-change pairing, and the ablation structure of Garrido et al. (2025).

Why this is a capital-allocation question, not a methodological one

It would be easy to file the preceding exhibit under academic hygiene. We think that is the wrong filing, for a straightforward reason: benchmark quality is currently the only independent check on capability claims in a category absorbing billions of dollars a quarter.

Physical AI has no revenue line to validate against. There is no equivalent of daily active users, gross margin or net revenue retention that would tell an allocator whether a robot foundation model is working. The evidence available at the point of underwriting is a demonstration video and a benchmark score.

Demonstration videos are close to uninformative — they are selected, and the selection ratio is never disclosed. That leaves benchmark scores carrying the entire diligence burden, which makes their construction a financial question.

A benchmark that can be passed by exploiting a texture cue will be passed by exploiting a texture cue, and the resulting number will be quoted in a deck. The six controls above are the difference between a number that constrains belief and a number that decorates it.

SECTION 05 · WHAT WE TAKE FORWARD

Physical competence can be measured without asking a model anything, by scoring the gap in its own prediction error between matched possible and impossible events. Report the distribution of that gap, not just its sign; state the read-out window; and always publish both an untrained baseline and a human baseline. In a category with no revenue signal, benchmark construction *is* diligence.

06

THE CAPITAL MAP

Who is building what, and on whose balance sheet

The thesis in this note is no longer contrarian. It is funded. What has not repriced alongside it is the financing architecture: the category is being built almost entirely on venture and growth equity, in volumes that normally attract long-duration capital.

$14.4bn

disclosed primary capital into world-model and physical-AI platforms, **January–June 2026** (four rounds)

$3bn

aggregate venture funding into world-model startups specifically, as reported by **Forbes** in June 2026

1.4m

humanoid units by 2035, from ~20,000 in 2025 — Goldman Sachs Global Investment Research

EXHIBIT 14

Capital committed to world models and physical AI

Disclosed primary rounds. Valuations are pre-money where stated by the company, otherwise as reported.

COMPANY	DATE	ROUND	VALUATION	WHAT IT IS BUILDING, AND WHO BACKED IT
Prometheus	Jun 2026	$12.0bn	~$41bn	An “artificial general engineer” for designing physical systems. Co-founded by Jeff Bezos and Vik Bajaj; among the largest early-stage rounds on record
AMI Labs	Mar 2026	$1.03bn	$3.5bn	World models built on the JEPA thesis. Yann LeCun chairman, Alexandre LeBrun CEO; Cathay, Greycroft, Bezos Expeditions, with Nvidia, Samsung, Temasek and Toyota Ventures
World Labs	Feb 2026	$1.00bn	~$5bn	Spatial intelligence; the Marble world model generates navigable 3D scenes from text or images. AMD, Nvidia, Autodesk, Fidelity, Emerson Collective
Galbot	early 2026	$0.35bn	n/d	Robot “data factories” addressing the interaction-data bottleneck (CNY 2.5bn)
World Labs prior round	Sep 2024	$0.23bn	n/d	Company formation round, for scale reference against the February 2026 raise
2026 disclosed		**$14.4bn**		**Excluding the 2024 comparison round**

Source: TechCrunch (9 March 2026); Reuters / Yahoo Finance (18 February 2026); Goldman Sachs, *Harnessing AI for the Real Economy* (2026); Forbes (30 June 2026). n/d = not disclosed. AI Lab Monaco compilation. Figures are as reported and have not been independently verified.

Read the concentration first. A single round — Prometheus — is roughly 83% of the disclosed 2026 total. Strip it out and the world-model cohort proper raised about $2.4bn in five months, which is a serious but not extraordinary figure for a frontier category. Aggregate headlines in this space are usually one deal wearing a sector's name.

Then read the cap tables. Nvidia, Samsung, Toyota Ventures, AMD, Autodesk and Sea appear across these rounds. Those are not financial allocators reaching for returns; they are strategic investors buying position in a supply chain, and their presence tells you more about where the compute, silicon and manufacturing demand is expected to land than about the probability that any individual company succeeds.

Then read the geography. AMI Labs is in Paris with hubs in New York, Montreal and Singapore; World Labs is in California; Galbot in Beijing. World models are one of the few frontier AI categories where the leading positions are not concentrated in a single jurisdiction — relevant for European and Gulf allocators who have found the language-model layer effectively closed to entry.

The platform layer is already commercial

Alongside the private rounds, the large laboratories have been shipping world-model infrastructure as product rather than as research artefact — which is the more reliable indicator that a category has arrived.

NVIDIA's Cosmos world-foundation models, trained on the order of 20 million hours of real-world video, passed two million downloads by January 2026 and are used by robotics developers including 1X, Agility and Figure, and by autonomous-vehicle programmes at Waabi, XPENG and Uber.[9] DeepMind's Genie 3 sustains interactive environments at 720p and 24 frames per second with roughly a minute of visual memory.[8] World Labs shipped Marble commercially in November 2025 with export into Unreal and Unity.[14]

The strategic read is that the world model is being positioned as the simulation and pretraining substrate for embodied systems — the layer on which robot policies are trained before they touch hardware. That is a different, and considerably nearer, business than autonomous general-purpose robotics.

The financing architecture has not caught up

Goldman Sachs frames the resulting problem precisely, and we think it is the single most important structural observation about the category: digital infrastructure is investable today through familiar instruments — investment-grade bonds, private credit, infrastructure debt, real estate trusts. Industrial AI software is accessible through public exposures and selective M&A. **Physical AI relies almost entirely on venture and growth equity.**[12]

That mismatch has consequences beyond elegance. Late-stage private equity is absorbing venture-stage technology risk. There is no scaled equivalent of credit wrapping for hardware-intensive deployment, and no structure that makes physical-AI capacity financeable the way a data centre is financeable.

The same report notes where capital *is* forming comfortably — logistics automation, predictive maintenance, warehouse robotics — and the common feature is that unit economics sit close to existing underwriting frameworks. Where they do not, funding is forming unevenly or not at all.

For an allocator the practical implication is that exposure to this thesis is currently available in three quite different risk shapes: the platform layer (largely inside listed technology companies), the application layer with settled economics (warehouse and logistics automation), and the frontier layer (private, concentrated, pre-revenue). Only the third is a direct expression of the argument in this note, and it is the one with no financing architecture behind it.

SECTION 06 · WHAT WE TAKE FORWARD

More than **$14bn** of disclosed primary capital went into world-model and physical-AI platforms in the first half of 2026, and the platform layer is already shipping commercially. The thesis is funded.

What is missing is the capital structure. Physical AI is being built on venture and growth equity at volumes that normally require long-duration capital, with no credit wrapper and no proven deployment economics. That is a timing risk, not a thesis risk — but for anyone deploying capital the two are hard to tell apart in the moment.

07

THE RELIABILITY THRESHOLD

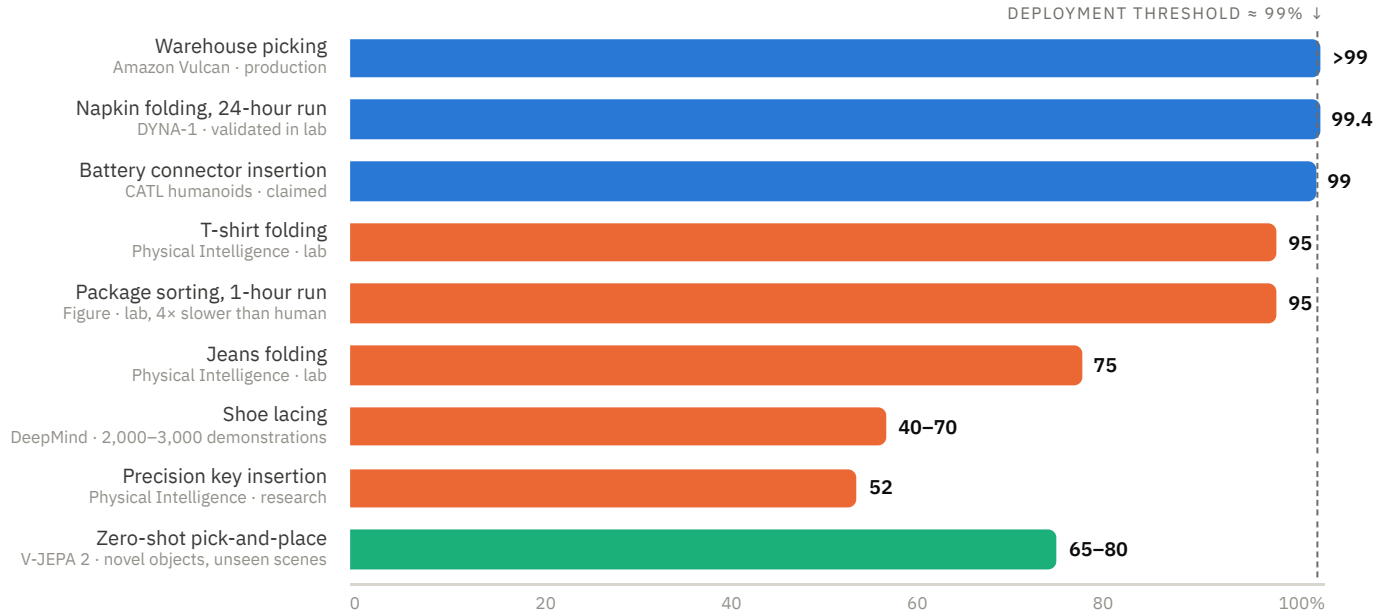
From demonstration to deployment

Everything in this note is an argument about prediction. Here is where prediction is priced. A warehouse will not adopt a system below roughly 99% reliability, and the distance between an impressive demonstration and that number is where the category's value actually sits.

EXHIBIT 15

The reliability ladder

Reported task success rates, %, as assessed in 2026. The dashed line is the reliability generally required for warehouse-grade deployment.



Source: Epoch AI, *Where Autonomy Works: Evaluating Robot Capabilities in 2026*; V-JEPA 2 figure from Meta AI (2025). Colour marks reliability stage rather than vendor: blue at or above threshold, orange below it, green the generalist world-model result shown for comparison — it is the only row measured on objects and environments the system had never seen. Ranges are plotted at their midpoint and labelled in full.

The constraint is transfer, not speed

The intuition most people bring to robot demonstrations is that the machines are too slow. The 2026 evidence says speed is largely a solved commercial problem: systems running three to ten times slower than a human still create value, particularly in domestic settings where the machine can work while the owner sleeps.

What does not clear is generalisation. In the assessment underlying Exhibit 15 the recurring finding is that transfer is rarely demonstrated at all — most results come from systems fine-tuned to a specific task, in a specific environment, on specific objects, with no evidence that performance survives a change to any of the three.[6]

The specifics are instructive. A humanoid that folds towels reliably cannot handle small cutlery. A system that inserts an external battery connector at 99% has no demonstrated path to routing cables inside a chassis. A model given 2,000–3,000 demonstrations of shoe lacing reaches 40–70%.

The reliability gap is therefore not a gap in dexterity. It is a gap in handling situations the demonstrations did not contain — which is exactly the partial-observability and distribution-shift problem §2 identifies as the structural weakness of description-based models.

Which is precisely the claim a world model makes

The argument for the architecture is not that it grips better. It is that a system which predicts state transitions should degrade more gracefully when the scene is not the one it was trained on, because it has learned how things move rather than what this particular thing looked like.

The 65–80% zero-shot result on novel objects in unseen environments is the first published evidence in that direction. It sits well below deployment — and it was measured under conditions none of the higher-scoring rows in Exhibit 15 attempted. Comparing the numbers directly would be a category error; comparing what they were measured on is the whole point.

The data bottleneck, priced

Robot interaction data cannot be scraped. It has to be generated on hardware in real time, which makes it the most expensive training data in the industry and the reason Chinese developers are financing dedicated “data factories” to manufacture it at volume.[12]

Against that, the V-JEPA 2 disclosure is a cost-structure claim as much as a capability one: the general representation came from ordinary video, which is abundant and cheap, and only the action-conditioning required robot data — 62 hours of it.[11]

If that separation holds at scale, the marginal cost of a new embodied capability falls sharply, because the expensive layer becomes thin. If it does not — if action-conditioning turns out to need thousands of hours per domain — then the category's economics look much closer to today's demonstration-by-demonstration reality, and the 2027–2029 window slips.

What would have to be true for the 2027–2029 window

Goldman Sachs places broad commercial deployment of humanoid platforms in 2027–2029, with cost falling on scale and full-stack control.[12] That window is not unreasonable, and it is worth being explicit about what it requires.

It requires generalist policies to move from roughly 80% to above 99% on unconstrained tasks — an order-of-magnitude reduction in failure rate, not an incremental gain. It requires that improvement to come without a proportional increase in robot interaction data, since that data is the sector's binding cost. And it requires the improvement to hold on objects and environments the system has not seen.

Those three requirements are the same requirement stated three ways, and it is the requirement this note has been circling throughout: a model that predicts what happens next, under occlusion and intervention, rather than one that recognises the situation it was trained on. Whether world models deliver it is the open question. That they are aimed at it is not in doubt.

SECTION 07 · WHAT WE TAKE FORWARD

- Deployment gates at roughly 99%. Three tasks clear it; all three are narrow, fine-tuned and environment-specific.
- Speed is not the binding constraint. Transfer is — and transfer is rarely even measured.
- The only result in the ladder obtained on unseen objects in unseen environments is the generalist world-model result, at 65–80%. That is the number to track.

08

RISKS TO THE THESIS

Five ways this argument could be wrong

01

Scale closes the gap after all

The history of the field is a history of structural arguments being overtaken by more compute and more data. Nothing in the evidence base rules out a frontier multimodal model clearing IntPhys 2 without architectural change — it has simply not happened yet. **What would show it:** a large multimodal model reaching human-competitive accuracy on IntPhys 2 or MVPBench with no change to its prediction objective. We would treat that as decisive against this note.

02

The benchmarks are measuring something else

The interpretive weakness of violation-of-expectation applies to machines as much as to infants. A 98% result could rest on a dataset artefact that happens to correlate with impossibility, in which case the ordering in §3 weakens considerably. **What would show it:** the V-JEPA-style advantage failing to replicate under minimal-change pairing, or collapsing when nuisance variables are randomised. MVPBench is currently the best available test of exactly this.

03

Representation-space prediction does not scale

Collapse avoidance in joint-embedding architectures is empirical rather than principled. The methods that stabilise training at 1.2 billion parameters are not guaranteed to hold an order of magnitude up, and a plateau there would leave the architecture argument correct in principle and irrelevant in practice. **What would show it:** published scaling curves that flatten, or frontier-scale JEPA models that underperform their smaller siblings on the same protocol.

04

Physical discrimination does not transfer to control

Telling a possible event from an impossible one is not the same capability as manipulating an object. A model could pass every violation-of-expectation test and remain useless as the substrate for a policy, because control requires precisely the fine-grained detail a good abstraction discards. The 65–80% zero-shot manipulation figure is thin evidence for the bridge. **What would show it:** world-model-based policies failing to beat conventional vision-language-action baselines on held-out objects and environments.

05

The capital cycle closes before the technology arrives

This is the risk we would weight most heavily on a three-year view, and it is independent of whether the science is right. The category has no revenue signal, no credit wrapper, concentrated private positions and late-stage equity absorbing venture-stage technology risk. A funding-window closure delays deployment by years regardless of architectural merit. **What would show it:** down rounds or stalled raises across the world-model cohort, or a retreat by the strategic investors — Nvidia, Samsung, Toyota — currently anchoring it.

OUR CONVICTION, STATED PLAINLY

We hold the *architectural* claim with reasonable confidence: the published comparisons are controlled, and the ordering has reproduced. We hold the *timing* claim with low confidence, and would not underwrite a specific deployment year on the current evidence. Readers should treat §§1–5 as the durable part of this note and §§6–7 as a snapshot with a short half-life.

09

CONCLUSION

What we are watching

The headline that opened this note is a provocation, and it earns its place only because of what sits underneath it. Infants do not beat frontier models at physics in any interesting sense. What they do is demonstrate that the capability in question — reliable expectation about objects, under occlusion, without instruction — is achievable on a budget so small that scale cannot be the missing ingredient.

The machine evidence has now caught up with that observation. Under controlled comparison, the systems that acquire the capability are the ones that predict the world in a learned representation. The systems that describe it, or render it, do not. That is a narrow claim, it is well supported, and it points somewhere specific: toward architectures that carry state, condition on action, and can be surprised.

EXHIBIT 16

Watch list to end-2027

Dated or datable signals, ranked by how much each would move our view.

SIGNAL	HORIZON	WHY IT MATTERS
A model clears 70% on IntPhys 2	2026–27	The single strongest test of whether the mechanism survives scene complexity. Watch equally for *which* architecture does it
MVPBench results under minimal-change pairing	2026–27	The best available check on the artefact objection in §8. A large drop versus IntPhys would weaken the whole ordering
Robot-data hours per new domain, disclosed	ongoing	The V-JEPA 2 result implies thin action-conditioning. If competitors report thousands of hours per domain, the cost argument fails
Transfer reported as standard in robot demonstrations	2026–27	Success on unseen objects in unseen environments, reported alongside headline rates, is the honest-disclosure signal for the sector
AMI Labs' first public model and code release	2026–27	The company has said it will publish research and open-source code. The first release is the cleanest read on whether the thesis is executing
A credit-like structure for physical-AI deployment	2027	The financing gap in §6 closes when hardware-intensive deployment becomes financeable outside equity. Its absence is the timing risk
Humanoid unit shipments against the ~20,000 base	annual	The simplest reality check on the 2027–2029 deployment window and on the 1.4m-by-2035 trajectory

Source: AI Lab Monaco. Horizons are our estimates, not company guidance.

> The question is no longer whether a model can explain why the ball stops at the wall. It is whether the model is surprised when the ball goes through it.

REFERENCES

Sources

Every quantitative claim in this note is traceable to one of the sources below. Where we have relied on a secondary report for a private-market figure, the reporting outlet and date are given rather than the company, because the company has not published the number itself. Preprints are cited by arXiv identifier so that the version we read can be recovered; several have since appeared in peer-reviewed venues.

1. Bordes, F., Garrido, Q., Kao, J. T., Williams, A., Rabbat, M. & Dupoux, E., *IntPhys 2: Benchmarking Intuitive Physics Understanding In Complex Synthetic Environments*, FAIR at Meta, arXiv:2506.09849, 11 June 2025.

2. *CausalVQA: A Physically Grounded Causal Reasoning Benchmark for Video Models*, Meta AI, arXiv:2506.09943, 2025.

3. *NEWTON: Are Large Language Models Capable of Physical Reasoning?*, Findings of the Association for Computational Linguistics: EMNLP 2023, arXiv:2310.07018.

4. Garrido, Q., Ballas, N., Assran, M., Bardes, A., Najman, L., Rabbat, M., Dupoux, E. & LeCun, Y., *Intuitive physics understanding emerges from self-supervised pretraining on natural videos*, arXiv:2502.11831, 17 February 2025.

5. "Yann LeCun's AMI Labs raises $1.03 billion to build world models", *TechCrunch*, 9 March 2026; "AI pioneer Fei-Fei Li's World Labs raises $1 billion in funding", 18 February 2026; Prometheus and Galbot rounds per reference 12.

6. Epoch AI, *Where Autonomy Works: Evaluating Robot Capabilities in 2026*, 2026.

7. Stahl, A. E. & Feigenson, L., "Observing the unexpected enhances infants' learning and exploration", *Science* 348(6230), 3 April 2015, pp. 91–94.

8. Google DeepMind, *Genie 3*, research preview, 2025.

9. NVIDIA, *Cosmos World Foundation Models*, announced CES 2025; adoption and download figures as reported January 2026.

10. Hafner, D., Pasukonis, J., Ba, J. & Norouzi, M., "Mastering diverse control tasks through world models", *Nature*, April 2025, doi:10.1038/s41586-025-08744-2.

11. Meta AI, *Introducing the V-JEPA 2 world model and new benchmarks for physical reasoning*, June 2025.

12. Goldman Sachs, *Harnessing AI for the Real Economy*, 2026.

13. Physical Intelligence, π0.5, 2025, as assessed in reference 6.

14. World Labs, *Marble*, commercial release November 2025.

15. Ha, D. & Schmidhuber, J., *World Models*, arXiv:1803.10122, 2018.

16. Schrittwieser, J. et al., "Mastering Atari, Go, chess and shogi by planning with a learned model", *Nature* 588, 2020, pp. 604–609.

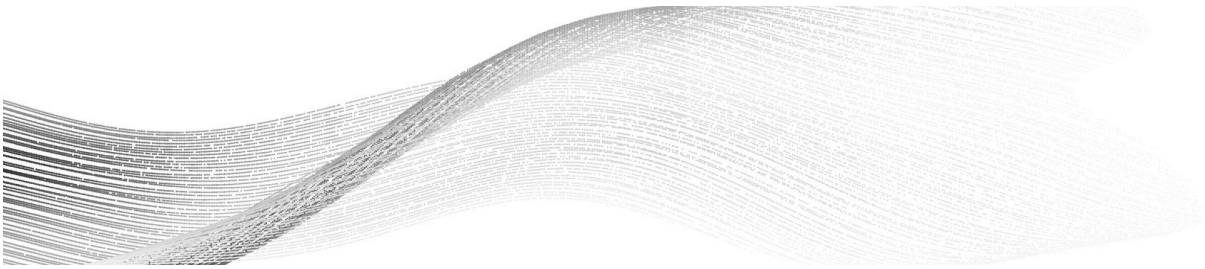
17. Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L. & Girshick, R., *PHYRE: A New Benchmark for Physical Reasoning*, NeurIPS 2019, arXiv:1908.05656.

18. Riochet, R. et al., *IntPhys: A Framework and Benchmark for Visual Intuitive Physics Reasoning*, 2018.

19. "VCs Pour $3 Billion Into AI's Next Big Bet: World Models", *Forbes*, 30 June 2026.

20. Fabbri, L., "Babies Beat ChatGPT at Physics: From Words to World Models", Artificial Intelligence Laboratory Monaco, 29 December 2025.

21. Spelke, E. S., Breinlinger, K., Macomber, J. & Jacobson, K., "Origins of knowledge", *Psychological Review* 99(4), 1992; Baillargeon, R., "Object permanence in 3½- and 4½-month-old infants", *Developmental Psychology* 23(5), 1987.



METHODOLOGY, SCOPE AND IMPORTANT DISCLOSURES

How this note was made

PUBLICATION HISTORY

First published **29 December 2025** as “Babies Beat ChatGPT at Physics”. This research edition was **updated in August 2026**; sources dated after December 2025 were added in that update and are cited with their own dates. The cover carries the original publication date.

SCOPE AND METHOD

All quantitative claims are drawn from primary sources — peer-reviewed papers, preprints, laboratory technical reports, company disclosures and named press reporting — and each is cited. Where a figure is a characterisation rather than a published point estimate (for example, “models perform at chance”), the exhibit footnote says so. Where two figures from the same paper are reported under related but non-identical protocols, we present both as published rather than reconciling them. We have not independently verified company-reported funding figures, valuations or success rates.

WHAT THIS NOTE IS NOT

This is a research note on the state of a technical field and its commercial context. It is not investment research, and it is not a recommendation to buy, sell or hold any security or private interest. AI Lab Monaco is not a registered investment adviser and nothing here should be read as investment, legal, accounting or tax advice. Named companies are discussed as examples of technical or commercial developments, not as recommendations. Any reader considering a financial decision in this category should take independent professional advice.

FORWARD-LOOKING STATEMENTS

Sections 06, 07 and 09 contain estimates, expectations and dated projections. These reflect our judgment at the date of publication, rest on assumptions we have tried to make explicit, and are subject to change without notice. Section 08 sets out, in the clearest terms we can, the conditions under which we would consider the central argument falsified. Views expressed are those of the author and are as of the August 2026 revision.

ABOUT AI LAB MONACO

The Artificial Intelligence Laboratory Monaco is an independent research and advisory organisation based in Monaco. It publishes research on frontier artificial intelligence, advises decision-makers, and convenes the Artificial Intelligence Monaco conference. Its editorial principle is stated simply: we read the research, you get the insights.

HOW TO CITE THIS NOTE

Fabbri, L. (2025). *From Words to World Models: why babies beat ChatGPT at physics, and what it means for the next capability cycle.* Frontier Research. Artificial Intelligence Laboratory Monaco, December 2025; research edition revised August 2026.

CORRECTIONS

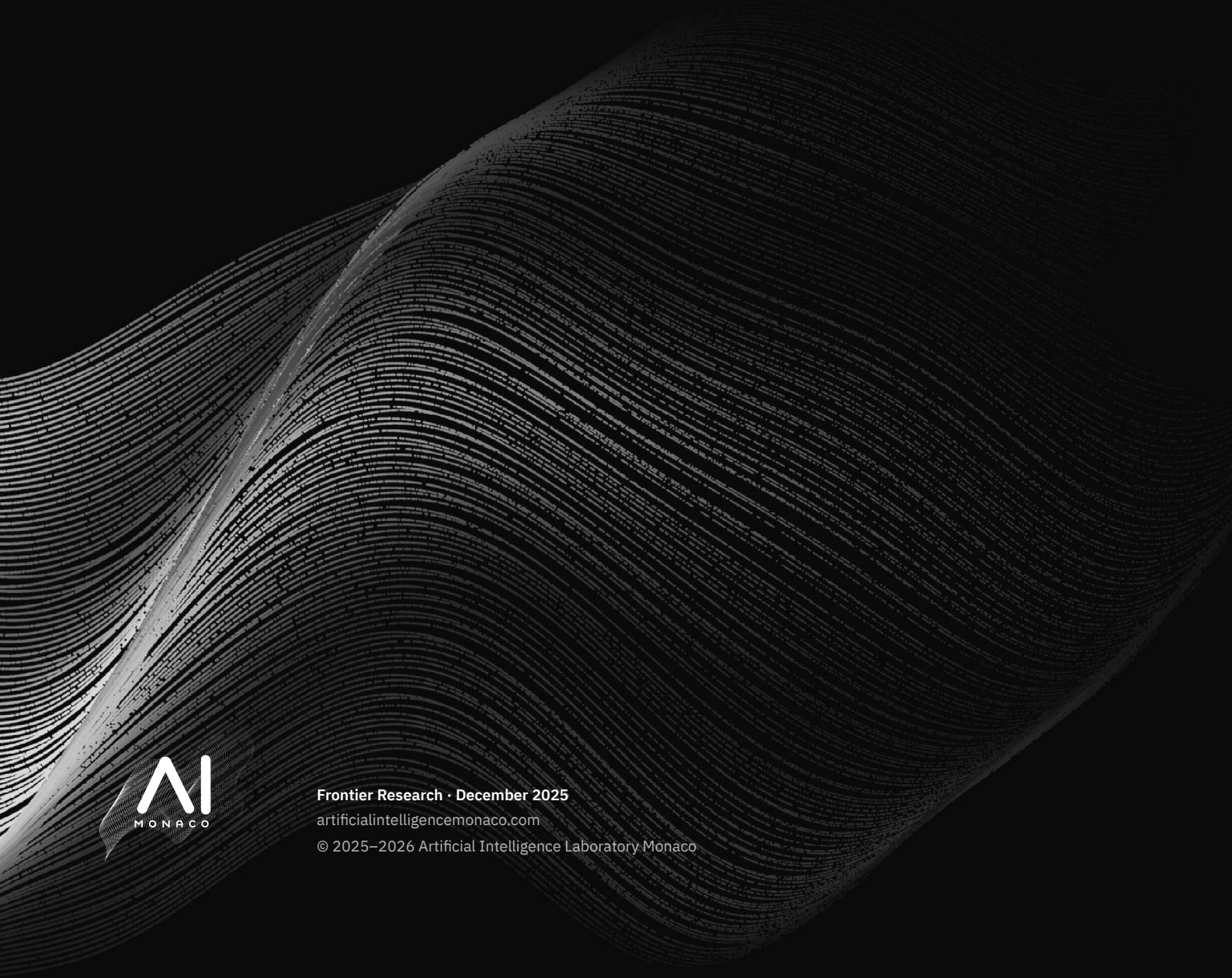
We correct errors of fact openly and date every revision. If you believe a figure, citation or characterisation in this note is wrong, please write to us — particularly if you are an author of work cited here.

Artificial Intelligence Laboratory Monaco
Frontier Research · December 2025
artificialintelligencemonaco.com

© 2025–2026 Artificial Intelligence Laboratory Monaco. All rights reserved. No part of this publication may be reproduced or redistributed without prior written consent. The AI Monaco name and mark are the property of the Artificial Intelligence Laboratory Monaco.

ARTIFICIAL INTELLIGENCE LABORATORY MONACO

We read the research, **you get the insights.**



Frontier Research · December 2025

artificialintelligencemonaco.com

© 2025–2026 Artificial Intelligence Laboratory Monaco