

A Ceiling Score Is Not a Physical Inference: Why Pretrained Video Encoders Look Physically Competent on One Benchmark and Near Chance on Its Successor, What a Clean Scene Lets a Readout Read Off, and the Rankability Checks a Physics Evaluation Owes

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Pretrained video and image encoders are now chosen as the perception front end of world models for manipulation partly on the strength of evaluations that are said to measure physical understanding: violation-of-expectation benchmarks, outcome-prediction tasks, and linear probes for properties such as mass and friction. The published verdicts disagree sharply. A 2025 study reports that a model trained to predict masked video in a learned representation space separates possible from impossible events on the original IntPhys benchmark with 98 percent accuracy, and concludes that intuitive physics can emerge without built-in core knowledge; the same model scores 53.75 percent on IntPhys 2, and no model tested there clears 60 percent overall against a human majority vote of 96.44 percent. A September 2026 study finds that in a clean, fixed-camera scene eight representations, including a randomly initialised network and raw pixels, all land within $0.02 R^2$ of the ceiling on a task that requires inferring mass and friction, and spread across $0.50 R^2$ once camera, lighting and clutter vary. In this analysis the author reviews and synthesises published results to show that these verdicts do not answer one question. A physics score can certify four different things: that the outcome is legible in the input, that a latent property was identified from interaction history, that the identified property is used, and that the instrument can rank models at all. The positive results concentrate on the first. On the latent-property benchmarks (Physion+ +, ComPhy and ExPhy) the negative results concentrate where the second and third are required, while CALIPER's swap control shows a fitted readout using the object's own interaction history with every representation tested, an untrained one included; IntPhys 2 is the exception, a near-chance result on largely object-intrinsic properties in more complex scenes. Almost no benchmark has demonstrated the fourth. The paper consolidates 30 published measurements in a table, separates the shortcut a clean scene offers a linear readout from the different failure on harder violation-of-expectation scenes, states a rankability audit as an algorithm, and proposes five tests that would settle what remains open.

Keywords: intuitive physics; video representation learning; world models; benchmark validity; linear probing; violation of expectation; latent physical properties; robot manipulation

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

A light box and a heavy box of the same shape look the same until something pushes them. Any system that is to predict what a push will do must therefore either be told the object's mass and friction or infer them from what it has already seen the object do. That requirement is old in robotics, where controllers adapt to unobserved payloads and terrain from a short history of their own interactions (Kumar et al., 2021), where pushing policies adapt a learned dynamics model online because latent friction and mass distribution cannot be read off appearance (Lee et al., 2026), and where a robot that poked objects more than 100,000 times learned forward and inverse models from images of the results (Agrawal et al., 2016). It is now being asked of pretrained visual encoders. Frozen self-supervised features serve as the state space for world models that plan pushes and pick-and-place (Zhou et al., 2024; Assran et al., 2025), and representations trained on egocentric human video are offered as general perception modules for manipulation (Nair et al., 2022; Majumdar et al., 2023).

Whether these representations carry physical understanding is contested, and the contest is unusually sharp because both sides cite measurements on overlapping models. On one side, Garrido et al. (2025) report that V-JEPA, a video model trained to predict masked regions in a learned representation space (Bardes et al., 2024), is reliably more "surprised" by physically impossible videos than by matched possible ones, reaching 98 percent zero-shot pairwise accuracy on IntPhys, while pixel-prediction models and multimodal language models sit near chance; they read this as evidence that intuitive physics can be acquired from video without innate core knowledge. Interpretability work on video encoders finds physical variables becoming linearly accessible in a band of intermediate layers (Joseph et al., 2026), a probe vector from that band can push plausibility judgements in either direction (Alam, 2026), and on the Physion outcome-prediction benchmark a probed V-JEPA 2 comes within about one

percentage point of human accuracy (Li et al., 2026). On the other side, the successor benchmark built by an overlapping group of authors finds every model it tests near chance, V-JEPA included (Bordes et al., 2025); a benchmark that requires inferring mass, friction, elasticity or deformability finds that models trained in standard regimes do not learn to make those inferences (Tung et al., 2023); video generators trained on synthetic mechanics generalise by retrieving the most similar training case rather than by abstracting a law (Kang et al., 2024); and a probing study of V-JEPA 2 and VideoMAE-v2 finds them near chance on IntPhys 2 intuitive-physics probes while they encode camera motion well (Musa et al., 2026).

The newest entry, the anchor of this analysis, takes the measuring instruments rather than the models as its object. Mehta and Baviskar (2026) build a calibrate-then-predict task: a box whose mass and friction are hidden is hit by a puck at two known speeds, the model then sees a third hit cut off at the moment of impact, and a linear readout fitted on frozen features has to say how far the box will travel. They find that in a clean, fixed-camera scene every representation they test, a randomly initialised ViT and raw pixels among them, scores within $0.02 R^2$ of every other and close to the ceiling, and that the same task separates the same representations by $0.50 R^2$ once camera pose, lighting, floor tint and distractors are resampled for every clip. Linear probes, by their account, track none of this ordering.

The cost of misreading this literature is concrete. Encoder choices for robot world models are being made now, partly on benchmark evidence, and the same benchmarks are used to argue a scientific claim about whether physical knowledge needs to be built in. If the scores certify physical inference, both uses are sound. If a score can be reached by reading the answer off the input, neither is.

The paper's claim is that the disagreement is mostly not about models. A physics score can certify four different things, and the studies on each side have measured different ones. First, outcome legibility: the outcome the benchmark asks about is present in, or a simple function of, the frames the model receives. Second, identification: a latent property that no single frame reveals has been inferred from the object's interaction history. Third, use: the inferred property is what drives the model's prediction or the controller's choice. Fourth, rankability: the instrument separates models that differ in the first three, rather than giving every model, including an untrained one, the same score. The positive results concentrate on properties whose violation is visible in the frames, such as object permanence and support; the negative results concentrate where identification or use is required, as with collisions, solidity, mass and friction; and rankability has been demonstrated by very few instruments. The anchor study is valuable chiefly as a demonstration of the fourth point, and the paper argues

that its own scope is narrower than its title.

The strongest objection belongs here. The IntPhys 2 collapse cannot be blamed on a clean-scene shortcut, because the clean, fixed-camera subset of IntPhys 2 is also near chance (Bordes et al., 2025). That is correct, and it is why the paper separates two failure modes the literature runs together: a benchmark whose answer is legible to any injective feature map, so that every model scores at the ceiling, and a benchmark on which every model scores at chance for reasons not yet identified. On IntPhys 2 those failures include immutability, a property of the same object-intrinsic kind that V-JEPA passes on IntPhys, while permanence is only partly spared. Both make an instrument unable to rank; they call for different remedies.

The rest of the paper is organised as follows. Section 2 reads the anchor study's design and its own limits. Section 3 states the case that physical competence has emerged in video encoders, and Section 4 the case against. Section 5 defines the four things a physics score can certify. Sections 6 to 10 take outcome legibility, latent-property identification, probing, nuisance variation and strategy in turn. Section 11 consolidates 30 measured values in Table 1 and Figure 1. Section 12 states a rankability audit as Algorithm 1. Section 13 gives the strongest case against this reading, Section 14 proposes tests, and Sections 15 to 17 state what is not known, the limitations and the conclusion. The evidence base is 43 published sources.

2. What the anchor study measures, and what its design rules out

The CALIPER task is simulated in MuJoCo with one rigid box per episode (Mehta and Baviskar, 2026). Mass is drawn log-uniformly and sliding friction independently, and the box's size, colour and aspect ratio are drawn independently of both, so appearance carries no information about the physics. A puck of fixed mass strikes the box head-on twice at known speeds; these two calibration clips are shown in full. A third strike at a new speed is shown only up to the moment of contact, and the regression target is how far the box then slides. The authors chose collisions rather than applied forces because a fixed force either leaves heavy objects stationary or sends light ones out of view.

The readout is deliberately thin. Each clip is mapped to one vector by a frozen encoder, reduced by principal components under a component budget shared across conditions so that adding a clip never adds parameters, concatenated with the commanded speed, and fitted by ridge regression with an inner cross-validated penalty. Four nested conditions follow: the commanded speed alone, the truncated query clip, the query plus the object's own calibration clips, and the query plus another episode's calibration clips assigned by a derangement. The last condition is the study's key control. Any

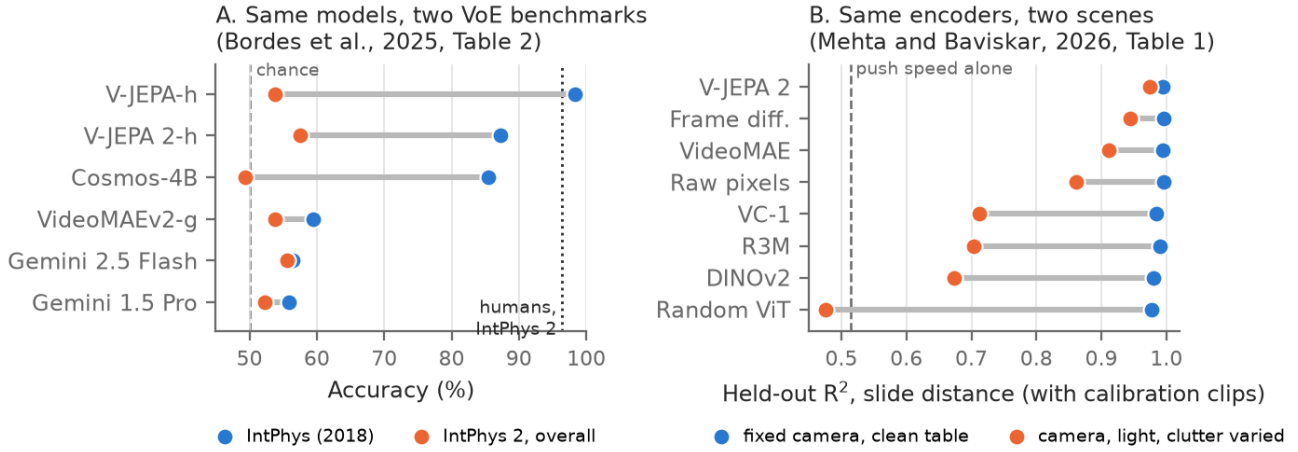


Figure 1. A. The same six models on the original IntPhys benchmark and on IntPhys 2, as reported in Table 2 of Bordes et al. (2025); IntPhys 2 values are best-of-runs over hyperparameters, the dashed line is chance and the dotted line is the human majority-vote accuracy on IntPhys 2. B. Held-out R^2 for predicting slide distance with the object’s own calibration clips, for eight representations in a fixed-camera clean scene and in a scene with camera, lighting, floor and distractors resampled per clip, from Table 1 of Mehta and Baviskar (2026); the dashed line is what the commanded push speed alone achieves in the varied scene.

accuracy that survives replacing the object’s own history with another object’s cannot have come from inferring this object’s physics.

The task itself behaves as designed. The commanded speed alone reaches an R^2 of 0.524 in the clean scene; the truncated query clip adds nothing over it, which confirms that the frames before contact carry no usable physics; the calibration clips raise every representation to between 0.977 and 0.996; and the swap returns every representation to between about 0.48 and 0.50. The readout is therefore using the object’s own interaction history.

What the clean scene cannot do is rank. The with-calibration column spans 0.020 R^2 across V-JEPA 2, VideoMAE (Tong et al., 2022), VC-1, DINOv2, R3M, a randomly initialised ViT, raw pixels and frame differences, and raw pixels (0.996) tie V-JEPA 2 (0.995). The authors’ explanation is mechanical: with a fixed camera, where the box ends up after each calibration slide is a fixed function of pixel coordinates, so a linear readout can recover the slide distance from the pixels and from any injective transformation of them, a random projection included. Nothing has to be inferred about mass or friction, because the readout can learn a mapping from the observed calibration slides and the commanded speed to the query slide, which is a sufficient statistic for the task. The study’s erasure results point the same way: removing a decoded mass direction never returns prediction to the query-only floor, which the authors read as the readout depending on how far the object slid rather than on an identified mass.

In the nuisance scene, where camera pose, lighting, floor tint and two distractor objects are resampled for every clip, the spread across representations grows to 0.498. V-JEPA 2 keeps 0.974, frame differences 0.944, VideoMAE 0.912 and raw pixels 0.861; VC-1, R3M and DINOv2 fall

to 0.712, 0.703 and 0.673; and the random ViT drops to 0.476, below the 0.514 that the commanded speed alone achieves in that scene. The swap still removes every gain, so the ordering is not a new shortcut. The ordering also holds at every training-set size the authors report; at 75 training episodes it runs from 0.716 for V-JEPA 2 to 0.017 for the random ViT. Translated into a control-shaped question, choosing the push speed predicted to reach a goal distance, V-JEPA 2 misses by a median of 4.0 mm and reaches 82 percent of goals within 10 mm, while the random ViT misses by 19.6 mm, matching the 19.0 mm of a policy that ignores the object altogether.

The study’s own limits matter for how far its conclusion carries, and the authors state most of them. The environment has rigid boxes and one contact event per clip. A frozen encoder with a ridge readout measures what is linearly available, which is a lower bound on what fine-tuning could extract, and each encoder is scored under a shared readout rather than its own prediction machinery. The nuisance scene varies viewpoint, lighting and clutter but not object category or contact type, and its wider camera also makes targets smaller, so the drop cannot be assigned to one factor. Erasure removes only linearly decodable structure. The perturbation benchmark the study audits, in which 13 of 20 task-property cells moved the simulated outcome by less than the authors’ threshold, is the authors’ own earlier design, and they state that they do not know how common such defects are in published benchmarks. The study is a two-author preprint, and its generalisation claim, “Whether a benchmark can rank models is an empirical property” (Mehta and Baviskar, 2026), is a methodological claim supported by one task family. The paper treats its numbers as established for that setting and its conclusion about benchmarks in general as a hypothesis the rest of the evidence can test.

One omission in the anchor's related work is worth recording, because it bears on novelty. The calibrate-then-predict design, in which a model sees a few videos of the same objects interacting and must use their hidden properties to answer a question about another video, is the design of ComPhy (Chen et al., 2022), which reports that video reasoning models of the time failed to capture hidden mass and charge. What CALIPER adds is the swap control, the frozen-encoder comparison and the two-scene contrast, not the identification task itself.

3. The case that physical competence has emerged

The strongest positive evidence is Garrido et al. (2025). Using the violation-of-expectation (VoE) paradigm borrowed from infant research, they compute a surprise signal as the error between a model's predicted representation of upcoming frames and the representation of what actually happens, and ask whether the impossible video in a matched pair produces more surprise. V-JEPA trained on natural video reaches 98 percent on IntPhys and 62 percent on InfLevel; untrained V-JEPA initialisations sit near chance; and pixel-space video predictors and multimodal language models are much closer to chance. The V-JEPA variants they train on HowTo100M or the full VideoMix2M mix, including a 115-million-parameter model and one trained on a single week of unique video, are significantly above chance; the exception is their data-source ablation, where training only on the motion-focused SSv2 videos "leads to almost chance-level performance" (Garrido et al., 2025). The design includes two of the controls the anchor study asks for: an untrained network of the same architecture as a random-features baseline, and matched possible and impossible clips that share everything except the violation.

The positive results are not confined to one group or one instrument. On Physion, which asks whether two cued objects will touch as a scene unfolds (Bear et al., 2021), a frozen V-JEPA 2 with a trained readout reaches 73.2 percent against 74.2 percent for humans (Li et al., 2026). V-JEPA 2 itself, post-trained as an action-conditioned world model on under 62 hours of unlabelled robot video, plans zero-shot reaching, grasping and pick-and-place on Franka arms in two labs; in the second lab it reaches 100 percent on reaching and 80 and 50 percent on pick-and-place with a cup and a box, against 80, 0 and 0 percent for a Cosmos-based planner (Assran et al., 2025; NVIDIA et al., 2025). DINO-WM plans pushes and other tasks zero-shot by predicting future DINOv2 patch features (Zhou et al., 2024; Oquab et al., 2024).

Interpretability results add internal evidence. Joseph et al. (2026) report a sharp transition at intermediate depth in several video encoders, which they call the Physics Emergence Zone, at which physical variables become accessible; scalar quantities such as speed and acceleration are readable from early layers, while motion

direction becomes readable only in that zone and is carried by a high-dimensional circular population code rather than a single factor. A follow-up uses a probe's weight vector from that zone in VideoMAE as a steering direction and reports that it moves IntPhys plausibility judgements in either direction, and only when injected inside the zone (Alam, 2026). Esmati et al. (2026) recover the internal trajectories of video diffusion models for real videos and report that physical plausibility is linearly decodable from the denoising transformer's states with up to about 81.27 percent average accuracy across IntPhys and InfLevel (WAN-1.3B), above V-JEPA and VideoMAE baselines, while the signal is absent from the autoencoder latent fed into the transformer.

Read together, this is a coherent position: predictive training on video yields representations from which physical regularities can be read, and the regularities are good enough to plan with.

4. The case against

The most direct counter-evidence comes from IntPhys 2, built to address the saturation of IntPhys that Garrido et al. had reported (Bordes et al., 2025). It keeps the VoE quadruplet design, adds richer scenes, camera motion and four principles (permanence, immutability, spatio-temporal continuity and solidity), and releases a held-out split whose labels are withheld. Every model it tests is near chance, among them the pixel-reconstruction model VideoMAEv2 (Wang et al., 2023) and the world model Cosmos (NVIDIA et al., 2025). The best predictive model, V-JEPA 2-h, scores 57.51 percent overall and 56.40 percent on the held-out split; the original V-JEPA-h, which scores 98.30 percent on IntPhys, scores 53.75 percent; Cosmos-4B drops from 85.42 to 49.41 percent. Humans, scored by majority vote, reach 96.44 percent. These model numbers are the best of about a dozen hyperparameter runs per column, so they are optimistic, and when the task is changed from comparing a matched pair to classifying a single video, performance stays closer to chance. The fixed-camera subset, the condition closest to earlier benchmarks, is also near chance, and the V-JEPA models do better with a moving camera, which the authors note "may be explained by slightly different testbeds for both conditions" (Bordes et al., 2025).

Benchmarks that require inferring hidden properties report the same direction. Physion++ builds scenarios in which predicting the outcome depends on mass, friction, elasticity or deformability that can be inferred only by watching objects move, and finds that models trained in standard regimes do not spontaneously make those inferences, that models with object and physical-state structure do better, and that all models correlate poorly with human predictions (Tung et al., 2023). ComPhy reports the same for hidden mass and charge (Chen et al., 2022). ExPhy, a 2026 benchmark with

24,000 simulated scenes labelled for mass, friction and restitution, reports that accurate trajectory forecasting does not necessarily imply accurate recovery of the properties that govern it (Wang et al., 2026). Earlier diagnostic benchmarks pointed the same way: CLEVRER found models strong on descriptive questions about collisions and weak on explanatory, predictive and counterfactual ones (Yi et al., 2020), and PHYRE found learning agents far from efficient on simple mechanics puzzles (Bakhtin et al., 2019).

Generative video models, often proposed as world simulators, have been tested by asking whether their continuations obey physics. Diffusion video models trained on a simulator of object motion and collision generalise perfectly in distribution, but their out-of-distribution velocity errors are an order of magnitude larger, and when they generalise they appear to reference the closest training case, prioritising colour over size, size over velocity and velocity over shape (Kang et al., 2024). Physics-IQ finds the physical understanding of several current video generators severely limited and unrelated to their visual realism, although some cases are solved (Motamed et al., 2025). Morpheus, built from 130 real videos governed by conservation laws, finds the same struggle (Tragoudaras et al., 2025), and an evaluation anchored on 124 clips of real free-fall events finds two commercial generators producing the release and impact events in more than 93 percent of cases but often substantially late, while Cosmos-Predict-2.5 and MAGI-1 frequently leave the scene in its pre-event state (Arriaga Santana et al., 2026). Multimodal language models perform at or below chance on the intuitive-physics level of GRASP while humans average 80 percent (Jassim et al., 2023), and PhysBench, with 10,002 entries across 75 models, finds physical understanding limited (Chow et al., 2025).

The internal evidence shows the same drop. Musa et al. (2026), probing V-JEPA 2 and VideoMAE-v2 layer by layer, find camera motion decodable above 0.90 ROC AUC at 60 to 70 percent of network depth and anomaly detection above 0.60, but intuitive-physics probes near chance. Those probes are trained and tested on IntPhys 2, and the authors set their near-chance result against earlier probing that succeeded on the original IntPhys (Joseph et al., 2026), the benchmark also behind the steering result (Alam, 2026). Esmati et al. (2026) find the same for their own probes: on IntPhys 2 every model they probe, diffusion model or encoder, falls well below its IntPhys and InfLevel accuracy, to "between roughly 55% and 68%". The probe results therefore reproduce the behavioural drop from IntPhys to IntPhys 2 rather than contradicting each other.

Finally, the closest match to human accuracy hides a mismatch in strategy. Li et al. (2026) treat eight probe initialisations per encoder and the human raters as populations. V-JEPA 2's 73.2 percent on Physion sits beside a 26.4 percent model-human disagreement rate

against 4.8 percent between humans, and an agreement kappa of about 0.48 against 0.91. V-JEPA 2 beats humans on the linking scenario by 11.8 points and trails them on rolling and sliding by 11.8 points and on dominoes by 10.5, the scenarios the authors associate with forward simulation; because some rolling and sliding trials are rendering-sensitive, the authors treat that gap as descriptive of the benchmark split rather than as evidence for a mechanism. All three architectures share error patterns with each other that humans do not.

5. Four things a physics score can certify

The two cases in Sections 3 and 4 cite different numbers for overlapping models because their instruments certify different things. This section stipulates four, stated flatly because they are definitions.

Outcome legibility (O). The quantity the benchmark scores is present in the input at test time, or is a simple function of it, so a readout or a surprise signal can reach the right answer by detecting what is visible. A ball that vanishes behind an occluder and does not reappear is a violation legible in the frames; so is a box's final position when the camera does not move. Legibility is a property of the stimulus and the readout jointly, not of the model.

Identification (I). A property that no single frame reveals, such as mass, friction or elasticity, has been inferred from how the object behaved earlier. Identification requires an interaction history and a target that depends on the property in a way the history constrains.

Use (U). The identified property is what drives the prediction or the action, so that removing it, or replacing the history it was inferred from, degrades the output. Identification without use is possible; so, through shortcuts, is a correct output without identification.

Rankability (R). The instrument separates systems that differ in O, I or U. A benchmark that gives the same score to every system, whether all at the ceiling or all at chance, cannot rank them, whatever its face validity. Rankability is checked by baselines that should fail, by the spread between them and the strongest system, and by effect-size audits of the manipulations the benchmark applies.

Every result in Sections 3 and 4 can be placed against these four. The VoE results of Garrido et al. (2025) are strongest for properties whose violation is legible in the frames. Their per-property analysis compares five runs of a ViT-L V-JEPA trained on HowTo100M with twenty untrained networks. On IntPhys the trained model reaches 85.7 percent on permanence, 86.3 on continuity and 83.7 on shape constancy, against about 51 to 52 for the untrained networks. Its other significant gains are on permanence, continuity, support, gravity and inertia in GRASP and on permanence in InfLevel, and its non-significant results are on colour constancy, solidity and

collision in GRASP and on gravity and solidity in InfLevel. The authors themselves note that violations involving interactions between objects are near chance. That pattern is what O without I predicts. The calibrate-then-predict results of Mehta and Baviskar (2026) establish I and U for the readout on every representation, through the swap, and show that R fails in the clean scene. IntPhys 2 is a benchmark at which R fails at the floor rather than the ceiling. And the Physion result of Li et al. (2026) shows that an accuracy match at the population level does not establish a strategy match, which is a statement about U.

Rankability is the benchmark-level form of a construct-validity worry that has been raised about AI benchmarks generally: a score that stands in for a broad capability is only as good as the argument that the task requires that capability (Raji et al., 2021). The four-way split makes the argument checkable for physics benchmarks in particular, because each of O, I and U has a control that would fail if the certificate did not hold.

The decomposition matters because the two uses of these benchmarks need different certificates. The scientific claim that physical understanding can be learned from video without innate structure needs I and U, since core-knowledge accounts of infant cognition are about expectations that go beyond what is visible. The engineering claim that one encoder is a better perception front end for a manipulation world model needs R, under the nuisance conditions the robot will meet. Neither is delivered by O alone.

The scientific claim deserves a closer look, because it is the one most often drawn from the positive results. Garrido et al. (2025) frame their finding against the view that a set of innate systems has to be built in for intuitive physics to develop, and conclude that predicting missing parts of video in a learned representation space is sufficient to acquire it. As stated, the conclusion is about the properties their benchmarks test, and their own per-property analysis bounds it: permanence, continuity, shape constancy, support, gravity and inertia show significant gains in at least one benchmark, while solidity and collision do not, and the authors attribute the latter partly to frame-rate limits and to interactions being rare in training video. A sufficiency claim is refuted by one clear failure on a property within its scope and is not established by successes on others. What the evidence currently supports is narrower and still interesting: that predictive training on unlabelled video produces a surprise signal that detects object-intrinsic violations far above an untrained network of the same architecture. Whether the same training produces expectations about interactions, about properties revealed only by interactions, or about scenes as complex as IntPhys 2's, is open. The paper takes no position on core knowledge in infants; it notes only that the machine evidence cited against it has so far been strongest where the visible input carries most of the

answer.

A second consequence concerns how the four certificates combine. They are not ordered, and a benchmark can deliver one without another. A benchmark can be rankable without testing identification: CALIPER's nuisance scene ranks representations by how invariantly they expose displacement, which a fitted readout then uses to identify the object's properties, so the ranking certifies O-under-nuisance for the representation and I for the readout. A benchmark can test identification without being rankable: Physion++ requires latent-property inference, and if every model fails it equally, it cannot order them, although it remains informative about the population. Reading a score therefore requires knowing which certificate the design supports and for which component of the system, encoder or readout, it supports it.

6. When the answer is in the frame

The clean-scene result of Mehta and Baviskar (2026) is a clean instance of outcome legibility defeating rankability. The readout never has to infer mass or friction, because a fixed camera makes the calibration slides' lengths a fixed linear pattern in pixel coordinates and the query slide is a smooth function of those lengths and the commanded speed. Any representation that preserves pixel geometry, which an untrained ViT does, gives a linear readout the same access. This is the general phenomenon Geirhos et al. (2020) call shortcut learning: a decision rule that succeeds on the benchmark's distribution and fails when the testing conditions change.

It would be a mistake to read the IntPhys-to-IntPhys 2 collapse as the same shortcut, and the evidence says so directly. VoE benchmarks were built against exactly this failure. IntPhys scores a plausibility judgement over well-matched possible and impossible videos generated in a game engine (Riochet et al., 2018), IntPhys 2 keeps its quadruplet structure and describes it as mitigating low-level perceptual biases (Bordes et al., 2025), and Garrido et al. (2025) report an untrained-network baseline near chance. A random network does not pass the original IntPhys; a predictive model trained on video does. So whatever V-JEPA exploited on IntPhys, it was not the pixel-level legibility that lets a random ViT pass CALIPER's clean scene.

What, then, explains 98.30 percent on one benchmark and 53.75 percent on its successor? The published evidence does not settle it, and the paper states the candidates rather than choosing among them. The first is outcome legibility at a higher level: the three IntPhys properties on which V-JEPA does best, permanence, continuity and shape constancy, are violated by an object disappearing, jumping or changing shape, events that any good predictor of the next representation should register as large error, so a model needs to predict

persistence and smooth motion, not to simulate interactions. The per-property pattern in Garrido et al. (2025), with permanence and continuity strong and collision and solidity at chance, fits this candidate. The second is scene complexity: IntPhys 2's richer scenes may make the prediction error of a normal clip large enough to swamp the extra error from a violation, which would explain why scoring single clips is harder than scoring matched pairs. The third is that IntPhys 2 tests solidity, which the earlier results already placed at chance. The three candidates make different predictions about which IntPhys 2 conditions should be easiest, and IntPhys 2's own per-condition table reports the permanence condition as the easiest for both models and humans and a few permanence and continuity subsets above 65 percent, which is the direction the first candidate predicts. The subsets are small and the reported values are best-of-runs, so this is weak evidence.

The general lesson is narrower than the anchor's title. Clean scenes do not prevent ranking in general; IntPhys's clean scenes separated a trained V-JEPA from an untrained one. What a clean scene does is make outcome legibility likely for any readout that has access to geometry, which matters most when the readout is trained on the evaluation task, as linear probes and outcome regressors are, and least when the score is zero-shot and computed from the model's own predictions, as VoE surprise is. A fitted readout can learn the legible shortcut from the evaluation's own training split; a surprise signal can only exploit what the model's pretraining taught it to predict.

This distinction bears on the Physion results as well. Physion's object-contact task is answered by attaching a readout to a frozen encoder and training it on labelled scenes (Bear et al., 2021), and the population analysis of Li et al. (2026) treats different random initialisations of that readout as the model population. A fitted readout on Physion is exposed to the same risk CALIPER identifies: whatever regularity in the training scenes predicts contact, physical or not, the readout can find. Li et al.'s attribution result, that features of the unobserved outcome rather than visible scene properties predict where models and humans diverge, is consistent with the readout having found such regularities in scenarios where they exist and failing where only forward simulation would do. It does not show which regularities were used. A swap-style control on Physion, for instance training the readout on one scenario family and testing it on another with the same contact statistics but different dynamics, would separate the two, and the published evaluations do not report one. The general point is that zero-shot surprise and fitted readouts fail in different ways, so a benchmark that uses both, as the encoder literature does, needs both kinds of control.

7. Latent properties: the part almost no benchmark scored

Identification of hidden properties is what separates a physical model from a motion extrapolator, and it has been tested far less often than permanence or continuity. Physion++ was built because, in the words of its authors, earlier benchmarks "typically do not require an understanding that objects have individual physical properties" (Tung et al., 2023). ComPhy and ExPhy were built for the same reason (Chen et al., 2022; Wang et al., 2026). In each case the verdict was negative for standard models.

CALIPER is the first of these designs, as far as this review found, to compare frozen pretrained encoders under a shared readout with a control that tests whether the interaction history is actually used. Its result on identification is more positive than the earlier literature: every representation gains about 0.50 R^2 from the object's own calibration clips and loses it under the swap (Mehta and Baviskar, 2026). But the identification is done by the readout, fitted on the evaluation's own episodes, and the representation contributes only the ability to report how far the object slid in each clip. In the nuisance scene the representations differ in whether they can report that distance invariantly to camera and lighting, not obviously in whether they encode mass. The authors draw the same conclusion from their erasure result: erasing a decoded mass direction never removes the calibration benefit, which suggests the readout relies on a sufficient statistic of the interaction rather than on an identified mass. CALIPER is therefore evidence that pretrained video encoders give a fitted readout invariant access to displacement, which is useful for a world model, and not direct evidence that the encoders themselves infer mass and friction.

Robotics practice reached a similar division of labour independently. Rapid Motor Adaptation trains an adaptation module to infer an environment code from a short history of states and actions (Kumar et al., 2021), and MetaPusher meta-learns a dynamics model that adapts online because, as its authors put it, the relevant properties "cannot be inferred from perception alone" (Lee et al., 2026). In both, identification is the job of a component that sees interaction history, not of the visual encoder. If that is the right architecture, the benchmark question for encoders is not whether they infer mass but whether they report the observable consequences of interactions faithfully under the conditions a robot meets. That is a narrower and more testable claim than "physical understanding".

8. Probes: decodable, readable, used

Linear probes are the second standard instrument, and the methodological literature on them in language models applies with little change. A probe's accuracy reflects both the representation and the probe's own

capacity, which control tasks were designed to separate (Hewitt and Liang, 2019). Decodability is not use: removing a property from a representation and measuring the behavioural cost answers a different question, and conventional probing accuracy did not track task importance in BERT (Elazar et al., 2021; Ravfogel et al., 2020). Belinkov (2022) reviews these and other limitations. A 2026 study adds that the common practice of ablating a single probe direction does not remove the target, because a probe refitted on the ablated activations recovers its accuracy from the orthogonal complement; replacing the ablation with iterative nullspace projection scored against random subspaces reversed some conclusions about causal inertness (Jiang, 2026).

CALIPER contributes two visual-physics instances (Mehta and Baviskar, 2026). One is about the probe's input rather than the representation: concatenating frames instead of averaging them raises the random ViT's friction probe from 0.361 to 0.672, above VC-1 (0.390) and DINOv2 (0.291) under averaging, while leaving its slide prediction unchanged. A ranking of encoders by probe accuracy can therefore be a ranking of aggregation conventions. The other is about erasure: raw pixels and frame differences both decode mass at about 0.98 in the clean scene, yet erasing the decoded direction costs frame differences 0.306 R^2 and raw pixels 0.001, and under nuisance variation the same erasure costs raw pixels 0.350. The authors attribute the difference to redundancy: raw pixels encode the target's position at every frame, so one direction can be removed and the property recovered from the rest. Whether a decoded property is load-bearing depends on what else the scene leaves available.

These findings bear directly on the internal evidence in Sections 3 and 4. The Physics Emergence Zone result (Joseph et al., 2026), the steering result (Alam, 2026), plausibility decoding from diffusion states at up to about 81.27 percent (WAN-1.3B; Esmati et al., 2026) and the near-chance intuitive-physics probes of Musa et al. (2026) are all probe results, on different encoders, layers, aggregation rules and stimulus sets. Their apparent disagreement is mostly a difference of benchmark: the high values come from IntPhys and InfLevel and the near-chance values from IntPhys 2, where Esmati et al.'s own probes also drop to roughly 55 to 68 percent. Because the other choices are not held fixed across the studies, the probing literature gives no reason to compare their values directly even on one benchmark. Two of the four report interventions: Joseph et al. (2026) use targeted attention ablations and report that motion direction requires coordinated multi-feature intervention to control, and Alam (2026) steers plausibility judgements. The steering result's readout is IntPhys, whose best-detected violations Section 6 argues are legible as abrupt changes. A probe that reads "plausibility" may read "something abrupt happened",

which is a property of the stimulus. None of these probe results establishes identification or use of a latent property, and none reports Hewitt-style selectivity against random-label control tasks or erasure scored against matched random subspaces, although Joseph et al. (2026) compare their layer-wise pattern against control tasks and their subspace overlaps against random baselines, and Alam (2026) compares the steering direction with a random unit vector.

9. Nuisance variation is what ranks

The single factor that turned CALIPER's flat clean-scene scores into a wide ordering was resampling camera, lighting, floor and clutter for every clip (Mehta and Baviskar, 2026). The robotics literature had found the same factor mattering for policies. Burns et al. (2023) compared 15 pretrained models and found that representations designed for manipulation did not necessarily hold up under small changes in lighting, scene texture or distractor objects, and that the emergent segmentation ability of a ViT's attention heads predicted out-of-distribution performance better than ImageNet accuracy, in-domain accuracy or shape bias. Hansen et al. (2022) found a learning-from-scratch baseline with data augmentation competitive with frozen pretrained representations such as R3M across several control benchmarks, attributing the gap to a domain mismatch that fine-tuning alleviates. CortexBench found no pretrained representation universally dominant across 17 embodied tasks (Majumdar et al., 2023). And probing of visual foundation models for 3D awareness, which its authors define to include representing a surface consistently across views, found large variation between models (El Banani et al., 2024).

Two points follow. First, the clean-scene benchmarks that dominate encoder selection measure the condition in which representations are least distinguishable. Second, nuisance variation interacts with physics in a way that is not yet separable. CALIPER's ordering under nuisance groups by pretraining objective, with the two video-prediction models first, but frame differences (0.944), a feature with no learned invariance at all, beat VideoMAE (0.912) and every per-frame encoder, and the authors suspect that their per-clip camera change is small enough for a difference image to localise the moving object. They did not test larger viewpoint changes. So the nuisance ordering certifies invariance to the nuisances sampled, at the magnitudes sampled, not physical inference as such.

The V-JEPA 2 paper reports a related deployment fact: its action-conditioned model must implicitly infer the action coordinate frame from a monocular camera and is sensitive to camera positioning (Assran et al., 2025). And on IntPhys 2 the V-JEPA models did better with a moving camera than with a fixed one (Bordes et al., 2025), which, although the authors suspect the two conditions' testbeds differ slightly, suggests that camera

motion is not uniformly a handicap. The direction of the nuisance effect depends on the task and the readout.

10. Accuracy is not strategy

Physion was designed to compare models with people, and its authors reported that vision models with object-centric representations did better than others while all fell short of humans, and that graph networks given true physical state did substantially better and made more human-like predictions (Bear et al., 2021). Their reading was that extracting physical state from pixels is the bottleneck. Five years later, a probed V-JEPA 2 matches human accuracy on the same benchmark within about a point while disagreeing with humans on a quarter of trials, outperforming them on a geometric scenario and underperforming them on rolling and dominoes (Li et al., 2026). The authors' attribution analysis finds unobservable outcome features, rather than visible scene properties, predicting where models and humans diverge, and they read this as models relying on scene statistics more than on forward simulation.

That result is the population-level counterpart of CALIPER's swap. Both ask whether a correct answer was reached by the route the benchmark is meant to measure, and both find that accuracy alone cannot say. The generative-model results make the same point from the other side: a model that interpolates within its training distribution perfectly and fails out of distribution by an order of magnitude is not using the law that governs both (Kang et al., 2024). The common lesson is that a benchmark claiming to measure physical inference owes a control that would fail if the inference were not made: a swapped history, an out-of-distribution parameter range, a strategy comparison, or an untrained baseline.

The four controls are not interchangeable, and each catches a different route to a correct answer. An untrained baseline catches answers available to any feature map that preserves the input's geometry, which is the clean-scene route in CALIPER. A swapped history catches answers available without the specific object's evidence, such as a prior over typical slide distances. An out-of-distribution parameter range catches answers available by retrieving a near training case, which is the route Kang et al. (2024) document. A strategy comparison catches answers available from scene statistics that happen to correlate with the outcome in the benchmark's distribution, which is the route Li et al. (2026) suggest. A benchmark that reports only one of these has excluded only one alternative. None of the benchmarks reviewed here reports all four, and most of the encoder comparisons that inform world-model design report none.

There is also a selection effect in how strategy evidence is reported. When accuracy is the only published number, a model that reaches human accuracy by a non-

human route is indistinguishable from one that reaches it by forward simulation, and the field's attention follows the accuracy. The Physion result illustrates the cost: a gap of about one point at the aggregate level conceals a difference of more than twenty points in item-level agreement (Li et al., 2026). Item-level human data exist for Physion and Physion++ because their designers collected them (Bear et al., 2021; Tung et al., 2023). Several newer benchmarks reviewed here report aggregate human accuracy only, as IntPhys 2 and GRASP do (Bordes et al., 2025; Jassim et al., 2023).

11. Consolidated evidence

Table 1 collects the measured values on which the argument rests, one row per cited result. All values are as reported by the named source. Rows from Bordes et al. (2025) are best-of-runs over hyperparameters except the held-out value, which was run once. Rows from Mehta and Baviskar (2026) are held-out R^2 , probe R^2 , erasure cost or median goal error on 2,000 simulated episodes, except the audit count, which comes from the authors' earlier benchmark.

Figure 1 draws the two contrasts that carry the argument, each within a single study so that no number is compared across instruments.

Panel A shows models that differ by up to about 42 points on IntPhys collapsing into a band between about 49 and 58 on IntPhys 2. Panel B shows the opposite collapse: eight representations indistinguishable at the ceiling in the clean scene, separating by about half the R^2 range in the varied one. Both are failures of rankability, at different ends of the scale, and they call for different remedies: a ceiling collapse calls for removing the legible shortcut, a floor collapse for asking whether the benchmark tests something the models represent at all.

12. A rankability audit

The anchor study proposes three checks a benchmark should report: an effect-size audit of its manipulations run with no model, the spread between a random-features baseline, raw pixels and the best model, and a swap or erasure behind any claim that a property is used (Mehta and Baviskar, 2026). The evidence reviewed here supports those three and adds two, one for floor collapse and one for strategy. Algorithm 1 states the resulting audit as a decision procedure that a benchmark designer can run before any model comparison is published.

```
Algorithm 1. Rankability audit for a physics benchmark.
Input: benchmark B (stimuli, labels, readout protocol),
       nuisance generator N, candidate models M.
1. Effect audit: for each manipulation in B, simulate
   its effect on the scored outcome with no model.
   Drop cells whose effect is below measurement noise.
2. Floor and ceiling baselines: score an untrained
   network of each architecture, raw pixels, and a
   readout given only the action or commanded input.
```

Table 1. Measured results on physical competence of pretrained visual and video models, by source.

Source	System	Setting	Measure	Value
Garrido et al., 2025	V-JEPA	IntPhys, zero-shot VoE	pairwise acc.	98%
Garrido et al., 2025	V-JEPA	InfLevel, zero-shot VoE	pairwise acc.	62%
Garrido et al., 2025	V-JEPA vs untrained	GRASP support	acc.	98.1 vs 58.4
Garrido et al., 2025	V-JEPA	GRASP solidity, collision	vs untrained	not significant
Bordes et al., 2025	V-JEPA-h	IntPhys / IntPhys 2	acc.	98.30 / 53.75
Bordes et al., 2025	V-JEPA 2-h	IntPhys / IntPhys 2	acc.	87.22 / 57.51
Bordes et al., 2025	V-JEPA 2-h	IntPhys 2 held-out	acc.	56.40
Bordes et al., 2025	Cosmos-4B	IntPhys / IntPhys 2	acc.	85.42 / 49.41
Bordes et al., 2025	Gemini 2.5 Flash	IntPhys 2 easy / overall	acc.	64.42 / 55.63
Bordes et al., 2025	Humans	IntPhys 2 overall	majority vote	96.44
Li et al., 2026	V-JEPA 2 vs humans	Physion OCP	acc.	73.2 vs 74.2
Li et al., 2026	V-JEPA 2 vs humans	Physion OCP	disagreement	26.4% vs 4.8%
Li et al., 2026	V-JEPA 2 vs humans	dominoes / linking	acc. gap	-10.5 / +11.8 pp
Mehta and Baviskar, 2026	8 representations	clean, with calibration	R ² spread	0.020
Mehta and Baviskar, 2026	8 representations	nuisance, with calibration	R ² spread	0.498
Mehta and Baviskar, 2026	raw pixels / V-JEPA 2	clean, with calibration	R ²	0.996 / 0.995
Mehta and Baviskar, 2026	V-JEPA 2 / random ViT	nuisance, with calibration	R ²	0.974 / 0.476
Mehta and Baviskar, 2026	push speed alone	nuisance	R ²	0.514
Mehta and Baviskar, 2026	V-JEPA 2 / random ViT	nuisance, goal reaching	median miss	4.0 / 19.6 mm
Mehta and Baviskar, 2026	random ViT	friction probe, mean vs concat	R ²	0.361 vs 0.672
Mehta and Baviskar, 2026	raw pixels	mass erasure, clean vs nuisance	R ² cost	0.001 vs 0.350
Mehta and Baviskar, 2026	own earlier benchmark	perturbation audit	null cells	13 of 20
Esmati et al., 2026	WAN-1.3B diffusion states	IntPhys + InfLevel probe	avg. acc.	81.27%
Musa et al., 2026	V-JEPA 2, VideoMAE-v2	CameraBench probe	ROC AUC	above 0.90
Musa et al., 2026	V-JEPA 2, VideoMAE-v2	IntPhys 2 probe	ROC AUC	near chance
Assran et al., 2025	V-JEPA 2-AC vs Cosmos	Lab 2 pick-and-place, cup	success	80% vs 0%
Jassim et al., 2023	MLLMs vs humans	GRASP intuitive physics	acc.	at or below 50 vs 80%
Arriaga Santana et al., 2026	Runway, Veo	real free-fall clips	events produced	above 93%
Kang et al., 2024	video diffusion	mechanics, OOD vs ID	velocity error	an order of magnitude or more higher
Wang et al., 2026	trajectory forecasters	ExPhy, 24,000 scenes	property recovery	not implied by accuracy

If best model - best baseline < margin: B is not rankable for M; report this and stop.

- Legibility test: re-score all models with nuisances from N resampled per clip. If the spread across M grows by more than the margin, the clean score was set by the scene; report both.
- Use test: for any claim that M uses property P, swap the interaction history that reveals P across episodes, or erase P against matched random subspaces and refit the probe. Claim use only if the score falls toward the baseline.
- Floor diagnosis: if all M are near chance, test the same M on a subset whose violation is legible in the frames. If they pass it, B tests a property M does not represent; if they fail it, suspect the protocol.
- Strategy test: where human data exist, report item-level agreement with humans and among models, not only accuracy.

Output: rankability verdict, the spread at each step, and which of O, I, U each score can certify.

Steps 1, 2 and 4 are the anchor's checks. Step 3 operationalises Section 9's finding that nuisance variation is what ranks. Step 5 addresses the IntPhys 2 pattern, in which a near-chance result says nothing about whether the cause is the models or the protocol until the protocol is shown to pass models on easier items; IntPhys 2's own easy split and its permanence subsets are a partial version of this step. Step 6 follows Li

et al. (2026). Several published benchmarks already perform parts of the audit. VoE benchmarks satisfy part of step 2 through matched quadruplets and, in Garrido et al. (2025), an untrained baseline. Physion++ and ComPhy are designed to require identification, which is the precondition for step 4. What none of the benchmarks reviewed here reports in full is the audit as a gate before ranking.

13. The strongest case against this reading

The best objection is that the decomposition is too fine to be useful: that physical understanding just is the ability to predict physical outcomes, so a model that predicts them under nuisance variation and plans successful robot actions has it, whatever its route. On this view the V-JEPA 2 planning results (Assran et al., 2025) and the nuisance-scene ordering (Mehta and Baviskar, 2026) are the evidence that matters, and the strategy mismatches and probe disputes are academic.

The objection is right about engineering and wrong about the scientific claim. For choosing an encoder, a score under realistic nuisance variation is exactly the right evidence, and this paper's recommendation agrees with the objection: rank under nuisance, at the

deployment's magnitudes, with a control-shaped readout. But the claim that intuitive physics emerges without core knowledge (Garrido et al., 2025) is a claim about identification and expectation, and it is that claim the IntPhys 2 floor, the interaction-property failures and the strategy mismatch bear on. The objection also understates how little of the positive evidence involves latent properties at all. The robot tasks in which V-JEPA 2 succeeds are reaching, grasping and pick-and-place with image goals, where the relevant variables are visible.

A second objection is that the anchor's clean-scene result is trivial: of course a fixed camera makes displacement legible, and nobody ranks encoders on a single-box scene. The first half is true, and the anchor study says so. The second half is an empirical claim about practice that the anchor asserts and this review could not check: Mehta and Baviskar (2026) state that encoder physics evaluations are "almost always" run in a clean, fixed-camera scene, but the one perturbation benchmark they audit is their own. How common the defect is in published benchmarks is unknown.

A further objection is that the anchor's evidence is simulated and its nuisance scene synthetic, so its ordering may say nothing about real robots. The objection has force; nothing in CALIPER is measured on real video. The closest real-world evidence points the same way without settling the question. Burns et al. (2023) found that a property measured on pretrained models, the emergent segmentation ability of their attention heads, predicted performance under lighting, texture and distractor shifts in simulation and also predicted real-world performance on a bimanual setup after offline training on 50 demonstrations. That is evidence that robustness to visual nuisances measured cheaply can forecast robustness on hardware. It is not evidence about physical inference, and whether CALIPER's nuisance ordering forecasts anything on hardware is untested.

A fourth objection is that IntPhys 2's near-chance results reflect hyperparameter or protocol mismatch rather than model limitations, given that the same models score high on IntPhys. Step 5 of Algorithm 1 is the test for this, and the evidence is incomplete: the permanence and continuity subsets show some models above 65 percent, which suggests the protocol can register competence, but those subsets are small.

14. Tests that would settle the open part

Test 1: CALIPER's nuisance scene with larger viewpoint changes and contact variety. The anchor varies viewpoint modestly and contact type not at all. Sweeping viewpoint change from zero to large, and adding rolling, toppling and multi-contact events, would show whether the nuisance ordering is about invariant displacement reporting, which frame differences nearly

match, or about representations of interaction. Prediction under this paper's reading: frame differences fall sharply with viewpoint change while the video-prediction models degrade gracefully.

Test 2: a swap control on VoE. For VoE benchmarks, the analogue of CALIPER's swap is to give the model a context clip of a different scene before the test clip and check that surprise depends on the test clip's own history. More directly useful is a VoE set in which the violation is defined relative to a latent property established earlier in the clip, such as a heavy object that should not move when struck. Garrido et al.'s pattern predicts V-JEPA near chance on such items.

Test 3: matched identification and use on Physion + +. Run frozen encoders with CALIPER's readout and swap on Physion + + scenarios. This would place the 2023 negative result and the 2026 positive result on one instrument.

Test 4: a joint probe audit. Take the four probe studies in Section 8, fix layer, aggregation and stimulus set, and report selectivity against control tasks and erasure costs against matched random subspaces with refitting. Running every encoder on both IntPhys and IntPhys 2 under this protocol would show whether the drop between them belongs to the encoders or to the probes.

Test 5: encoder ranking by downstream control under nuisance. For a fixed world-model architecture and planner, compare encoders on push tasks whose outcome depends on mass and friction, under the nuisance variation of Test 1, and check whether the CALIPER nuisance ordering predicts control success. If it does, the cheap readout is a valid proxy; if it does not, the field needs the expensive evaluation.

15. What is not known

It is not known why V-JEPA scores 98.30 percent on IntPhys and 53.75 percent on IntPhys 2. It is not known whether any pretrained video encoder identifies mass or friction, as distinct from reporting the displacements from which a fitted readout identifies them. It is not known how many published encoder evaluations share the clean-scene defect the anchor demonstrates. It is not known whether the probe results on physical plausibility, high on IntPhys and near chance on IntPhys 2, would survive common controls. It is not known whether the CALIPER nuisance ordering predicts control success. No study reviewed here measures whether a representation's physical competence under nuisance variation transfers from simulation to real scenes.

16. Limitations

The analysis rests on published results, and several of its key sources are 2026 preprints without peer review, including the anchor, the strategy comparison and two of the four probe studies. The anchor's numbers come from one simulated task family with rigid boxes. The paper

compares results across benchmarks only qualitatively and draws each quantitative contrast in Figure 1 within a single study; cross-study comparison of accuracies on different stimuli is not licensed by any of the sources. The decomposition into legibility, identification, use and rankability is this paper’s stipulation; other cuts are possible, and Section 6’s three candidate explanations of the IntPhys 2 collapse are hypotheses, not findings. The review did not cover physics-informed architectures that build in object or state structure, except as baselines, and did not cover the infant-cognition literature beyond what the cited studies report.

17. Conclusion

The question whether pretrained video encoders understand physics has been answered both ways because the instruments answering it certify different things. Scores near the ceiling on benchmarks whose violations are legible in the frames certify that a model predicts persistence and smooth motion. Scores near chance on benchmarks that require hidden properties certify that those are not yet captured, and scores near chance on IntPhys 2, where most of the tested principles concern single objects, show that success in simpler scenes does not carry over, for reasons not yet identified. A clean, fixed-camera scene can make a hidden-property task legible to any geometry-preserving feature map, so that a random network ties the best encoder, while a varied scene separates the same representations by half the range. A benchmark’s ability to rank is an empirical property, and it can be checked before any ranking is published: by effect audits, by baselines that should fail, by nuisance resampling, by swaps and matched erasures, and by comparing strategies as well as accuracies. Until those checks are reported, a ceiling score is evidence about the scene, not about a physical inference.

References

- Agrawal, P., Nair, A., Abbeel, P., Malik, J., & Levine, S. (2016). Learning to Poke by Poking: Experiential Learning of Intuitive Physics. *NeurIPS 2016*. arXiv:1606.07419. doi:10.48550/arXiv.1606.07419.
- Alam, N. (2026). Causal Physics Steering in Video World Models via Concept Activation Vectors. arXiv:2605.24322. doi:10.48550/arXiv.2605.24322.
- Arriaga Santana, E. M., Rosas Scull, J., Nunez Rico, E. S., & Escalante, H. J. (2026). Foresight at the Event Boundary: Evaluating Physical Prediction in Video World Models. arXiv:2609.36531. doi:10.48550/arXiv.2609.36531.
- Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., et al. (2025). V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. arXiv:2506.09985. doi:10.48550/arXiv.2506.09985.
- Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L., & Girshick, R. (2019). PHYRE: A New Benchmark for Physical Reasoning. *NeurIPS 2019*. arXiv:1908.05656. doi:10.48550/arXiv.1908.05656.
- Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., & Ballas, N. (2024). Revisiting Feature Prediction for Learning Visual Representations from Video. arXiv:2404.08471. doi:10.48550/arXiv.2404.08471.
- Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H.-Y. F., Pramod, R. T., et al. (2021). Physion: Evaluating Physical Prediction from Vision in Humans and Machines. *NeurIPS 2021 Datasets and Benchmarks*. arXiv:2106.08261. doi:10.48550/arXiv.2106.08261.
- Belinkov, Y. (2022). Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*. arXiv:2102.12452. doi:10.48550/arXiv.2102.12452.
- Bordes, F., Garrido, Q., Kao, J. T., Williams, A., Rabbat, M., & Dupoux, E. (2025). IntPhys 2: Benchmarking Intuitive Physics Understanding In Complex Synthetic Environments. arXiv:2506.09849. doi:10.48550/arXiv.2506.09849.
- Burns, K., Witzel, Z., Hamid, J. I., Yu, T., Finn, C., & Hausman, K. (2023). What Makes Pre-Trained Visual Representations Successful for Robust Manipulation? arXiv:2312.12444. doi:10.48550/arXiv.2312.12444.
- Chen, Z., Yi, K., Li, Y., Ding, M., Torralba, A., Tenenbaum, J. B., & Gan, C. (2022). ComPhy: Compositional Physical Reasoning of Objects and Events from Videos. *ICLR 2022*. arXiv:2205.01089. doi:10.48550/arXiv.2205.01089.
- Chow, W., Mao, J., Li, B., Seita, D., Guizilini, V., & Wang, Y. (2025). PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding. arXiv:2501.16411. doi:10.48550/arXiv.2501.16411.
- El Banani, M., Raj, A., Maninis, K.-K., Kar, A., Li, Y., Rubinstein, M., Sun, D., Guibas, L., Johnson, J., & Jampani, V. (2024). Probing the 3D Awareness of Visual Foundation Models. *CVPR 2024*. arXiv:2404.08636. doi:10.48550/arXiv.2404.08636.
- Elazar, Y., Ravfogel, S., Jacovi, A., & Goldberg, Y. (2021). Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *TACL*. arXiv:2006.00995. doi:10.48550/arXiv.2006.00995.
- Esmati, P., Nath, S., Hofmann, K., Nowrouzezahrai, D., Kahou, S. E., & Mirmehdi, M. (2026). The Invisible Hand of Physics: When Video Diffusion Models Know More Than They Show. arXiv:2606.05328. doi:10.48550/arXiv.2606.05328.
- Garrido, Q., Ballas, N., Assran, M., Bardes, A., Najman, L., Rabbat, M., Dupoux, E., & LeCun, Y. (2025). Intuitive physics understanding emerges from self-supervised pretraining on natural videos. arXiv:2502.11831. doi:10.48550/arXiv.2502.11831.
- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*. arXiv:2004.07780. doi:10.48550/arXiv.2004.07780.
- Hansen, N., Yuan, Z., Ze, Y., Mu, T., Rajeswaran, A., Su, H., Xu, H., & Wang, X. (2022). On Pre-Training for Visuo-Motor Control: Revisiting a Learning-from-Scratch Baseline. arXiv:2212.05749. doi:10.48550/arXiv.2212.05749.
- Hewitt, J., & Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. *EMNLP 2019*. arXiv:1909.03368. doi:10.48550/arXiv.1909.03368.
- Jassim, S., Holubar, M., Richter, A., Wolff, C., Ohmer, X., & Bruni, E. (2023). GRASP: A novel benchmark for evaluating language GRounding And Situated Physics understanding in multimodal language models. arXiv:2311.09048. doi:10.48550/arXiv.2311.09048.
- Jiang, Y. (2026). Refit the Probe: Single-Direction Ablation Is Not a Necessity Test. arXiv:2606.00926. doi:10.48550/arXiv.2606.00926.
- Joseph, S., Garrido, Q., Balestriero, R., Kowal, M., Fel, T., Bakhtiari, S., Richards, B., & Rabbat, M. (2026). Interpreting Physics in Video World Models. arXiv:2602.07050. doi:10.48550/arXiv.2602.07050.
- Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., & Feng, J. (2024). How Far is Video Generation from World Model: A Physical Law Perspective. arXiv:2411.02385. doi:10.48550/arXiv.2411.02385.
- Kumar, A., Fu, Z., Pathak, D., & Malik, J. (2021). RMA: Rapid Motor Adaptation for Legged Robots. *RSS 2021*. arXiv:2107.04034.

- doi:10.48550/arXiv.2107.04034.
- Lee, D., Golestan, S., Singh, J., Zhong, Z., Kapoutsis, A., & Chamzas, C. (2026). MetaPusher: Meta Learning and Planning for Nonprehensile Manipulation of Unseen Objects with Rapid Online Adaption. arXiv:2609.21122. doi:10.48550/arXiv.2609.21122.
- Li, F., Zheng, S., Yin, Z., Cui, J., Ji, L., & Liu, J. (2026). Human-Level Accuracy, Non-Human Strategies: Revealing Model-Human Divergence in Video Physical Reasoning. arXiv:2609.22788. doi:10.48550/arXiv.2609.22788.
- Majumdar, A., Yadav, K., Arnaud, S., Ma, Y. J., Chen, C., Silwal, S., et al. (2023). Where are we in the search for an Artificial Visual Cortex for Embodied Intelligence? arXiv:2303.18240. doi:10.48550/arXiv.2303.18240.
- Mehta, A., & Baviskar, R. (2026). CALIPER: Clean Scenes Cannot Rank Physical Inference in Pretrained Visual Representations. arXiv:2609.08250. doi:10.48550/arXiv.2609.08250.
- Motamed, S., Culp, L., Swersky, K., Jaini, P., & Geirhos, R. (2025). Do generative video models understand physical principles? arXiv:2501.09038. doi:10.48550/arXiv.2501.09038.
- Musa, S. S., Forghani, F., Thasarathan, H., Joseph, S., Kowal, M., & Derpanis, K. G. (2026). What, Where, and How: Probing Spatiotemporal Representations in Video Foundation Models. arXiv:2609.01551. doi:10.48550/arXiv.2609.01551.
- Nair, S., Rajeswaran, A., Kumar, V., Finn, C., & Gupta, A. (2022). R3M: A Universal Visual Representation for Robot Manipulation. arXiv:2203.12601. doi:10.48550/arXiv.2203.12601.
- NVIDIA, Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., et al. (2025). Cosmos World Foundation Model Platform for Physical AI. arXiv:2501.03575. doi:10.48550/arXiv.2501.03575.
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., et al. (2024). DINOv2: Learning Robust Visual Features without Supervision. TMLR. arXiv:2304.07193. doi:10.48550/arXiv.2304.07193.
- Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. ACL 2020. arXiv:2004.07667. doi:10.48550/arXiv.2004.07667.
- Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the Everything in the Whole Wide World Benchmark. arXiv:2111.15366. doi:10.48550/arXiv.2111.15366.
- Riochet, R., Ynocente Castro, M., Bernard, M., Lerer, A., Fergus, R., Izard, V., & Dupoux, E. (2018). IntPhys: A Framework and Benchmark for Visual Intuitive Physics Reasoning. arXiv:1803.07616. doi:10.48550/arXiv.1803.07616.
- Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. arXiv:2203.12602. doi:10.48550/arXiv.2203.12602.
- Tragoudaras, A., Zhang, C., Cherniavskii, D., Vozikis, A., Nijdam, T., Prinzhorn, D. W. E., Bodracska, M., Sebe, N., Zadaianchuk, A., & Gavves, E. (2025). Evaluating Newtonian Mechanics in Video Generative Models with Real Physical Systems. arXiv:2504.02918. doi:10.48550/arXiv.2504.02918.
- Tung, H.-Y., Ding, M., Chen, Z., Bear, D., Gan, C., Tenenbaum, J. B., Yamins, D. L. K., Fan, J. E., & Smith, K. A. (2023). Physion + +: Evaluating Physical Scene Understanding that Requires Online Inference of Different Physical Properties. arXiv:2306.15668. doi:10.48550/arXiv.2306.15668.
- Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., & Qiao, Y. (2023). VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. arXiv:2303.16727. doi:10.48550/arXiv.2303.16727.
- Wang, R., Wu, Y., Zhang, X., & Qi, M. (2026). ExPhy: A Benchmark for Explicit Physical Property Learning in Multi-Object Trajectory Forecasting. arXiv:2608.20009. doi:10.48550/arXiv.2608.20009.
- Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., & Tenenbaum, J. B. (2020). CLEVRER: CoLLision Events for Video REpresentation and Reasoning. ICLR 2020. arXiv:1910.01442. doi:10.48550/arXiv.1910.01442.
- Zhou, G., Pan, H., LeCun, Y., & Pinto, L. (2024). DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning. arXiv:2411.04983. doi:10.48550/arXiv.2411.04983.