KNOWING WHEN NOT TO ANSWER: SELF-ASSESSED MEASURABILITY AND THE LIMITS OF PHYSICS TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

A world model that has learned physics should carry something to physics it has not seen, and the field tests this by training on one system and evaluating on another. We show that this test usually does not measure what it is taken to. The reason is structural: a latent world model compresses a field, evolves it and decompresses, so its error contains both what the dynamics got wrong and what the round trip lost, and which term dominates depends on something rarely reported, namely how far the physics moves between the two frames being compared. When the field barely moves and the codec is lossy, the error measures the codec, and a model that has learned nothing transferable still scores well. The fix is a check the model runs on itself before acting, with no labels and no access to the future. Let ρ be its reconstruction error divided by the observed frame-to-frame change: below 1 the comparison is attributable to dynamics, at or above 1 it is not. The threshold is derived rather than fitted, and over 52 conditions the rule makes no errors; two stress tests and two weaker checks leave it intact. Applied to a standard PDE benchmark it disqualifies the protocol rather than the models: of 42 cross-family pairs across eight datasets, none is measurable at the single-step horizon those benchmarks use. This is not undertraining, nor simply a weak codec. Training the source to convergence improves *in-domain* attributability by over an order of magnitude and cross-domain attributability by under $2\times$, while refitting the codec on the target buys attributability only by destroying the skill being measured. What survives is narrower and real. Inside one physics family transfer occurs and is measurable, in three of six directions among three acoustic geometries, one beating codec-matched persistence by 0.568 $[0.530, 0.606]$ at every horizon and replicating across seeds; across families, nowhere. The consequence is a world model that knows when not to answer, whose failure mode is decidable rather than probabilistic: we placed conditions on the boundary itself and the classes still did not overlap. It also predicts the size of its own forecast error well enough to be calibrated, at 90.6% coverage for a nominal 90% once the interval accounts for shift between families.

1 INTRODUCTION

A world model that has learned physics should carry something to physics it has not seen. The field tests this the obvious way: train on one system, evaluate on another, and report the error. Lower error is taken to mean that something transferred.

This test is confounded in a way that is easy to miss and easy to check for, and correcting it changes what a standard benchmark reports: not that transfer fails, but that it works in one place and not another.

The confound is structural rather than incidental. A latent world model compresses a field, evolves the compressed state, and decompresses. Its output error therefore contains two terms: what the encode-decode round trip loses, and what the dynamics prediction gets wrong. These are not separable by inspection, and their relative size depends on something the experimenter rarely reports, namely how much the physics moves between the two frames being compared. When the field

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

changes little over the evaluation interval and the round trip is lossy, the reported error is dominated by reconstruction. A model that has learned nothing transferable then posts a competitive score, and the score is a measurement of its codec.

How we found it. This project began as a study of cross-domain transfer on a standard PDE benchmark, organised around the hypothesis that a seven-primitive "physical vocabulary", read from the published governing equations and fixed before any transfer experiment, would predict which transfers succeed. It appeared to work: zero-shot error fell monotonically with vocabulary overlap across five held-out families, with residual structure in the direction the schema predicted.

Then we re-evaluated the same checkpoints on the same targets against the trivial baseline of copying the previous frame (Murphy, 1992). The model was worse than copying on every target tested, *including the domain it was trained on*. The monotone relationship held among numbers that measured reconstruction, and the schema had been fitted, in effect, to a codec.

We report that sequence rather than only its surviving half, because the defect belongs to the standard protocol rather than to our particular error. Section 9 reports the original hypothesis as untested rather than refuted.

Contributions.

1. **A criterion** that decides, for a given model and evaluation interval, whether a field-space error can be attributed to dynamics at all (Section 4), with a deployable estimate that uses no labels and no access to the future. What the evidence supports is *normalised reconstruction error*: a single-frame competitor ties it on every condition, and the ratio is preferred because its threshold is derivable rather than fitted (Section 4.5).
2. **A measurement of a standard benchmark under that criterion** (Section 5): across 42 cross-family source-target pairs spanning eight PDE datasets, not one is attributable at the single-step horizon those benchmarks use.
3. **A positive transfer result and its boundary** (Section 6). Within a physics family transfer occurs, is measurable and replicates across seeds, in three of the six directions we could run. Across families, in every pair we evaluated, it does not.
4. **Two capabilities that field-space error cannot see** (Sections 7–8): prediction of the *difference* an intervention causes, and a per-window estimate of the model's own forecast error.

2 Related Work

Physics foundation models and benchmarks The Well (Ohana et al., 2024) is the corpus we measure. Walrus (Polymathic AI Collaboration, 2025a;b), GPhyT (Herde et al., 2025), PhysiX (Nguyen et al., 2025) and multiple-physics pretraining (McCabe et al., 2023; Subramanian et al., 2023; Herde et al., 2024) establish that joint exposure to diverse PDE families yields transferable representations, reported as field-space error. Our contribution is prior to theirs: we ask whether a field-space error on a given interval is attributable to dynamics at all, a question about the protocol rather than about any model. Operator learning (Lu et al., 2021; Li et al., 2021) and graph-network simulators (Sanchez-Gonzalez et al., 2020; Pfaff et al., 2021; Battaglia et al., 2018) supply the architectural context.

Latent world models Latent world models trace to early proposals (LeCun, 2022; Ha & Schmidhuber, 2018) with modern instantiations (Hafner et al., 2023), and our latent predictive objective follows JEPA (Assran et al., 2023; Bardes et al., 2024; Garrido et al., 2024). The confound we identify is a property of the latent factorisation itself, so it applies to the whole class.

Baselines, calibration and coverage Persistence as a reference forecast is standard in numerical weather prediction (Murphy, 1992) and underused in learned PDE surrogates. We report proportions with exact Clopper-Pearson intervals (Clopper & Pearson, 1934), compare classifiers with McNemar's exact test (McNemar, 1947), and bootstrap by trajectory block (Künsch, 1989; Efron & Tibshirani, 1994). For the error-magnitude predictor we use split conformal intervals (Vovk et al.,

2005; Angelopoulos & Bates, 2023); holding out an entire physics family deliberately breaks exchangeability, and the resulting shortfall is the expected behaviour rather than a defect (Barber et al., 2021). Calibration is reported with the Brier score (Brier, 1950) and expected calibration error (Guo et al., 2017). Declining to answer is long established as selective prediction, in which a model rejects low-confidence inputs to trade coverage against risk (Geifman & El-Yaniv, 2017; 2019). The gate we study is not of that kind. It never consults the model's confidence and is not tuned to a coverage target; it asks whether the evaluation interval resolves the model's own representation error, so a confident model is refused whenever the measurement is degenerate.

3 Model and Protocol

Every ρ here contains a particular model's reconstruction error, so we state the model before using it. LATH-v1 is a latent world model of 816k parameters. The encoder embeds the field in 4×4 patches; the latent core is a message-passing network over the real grid geometry rather than a generic sequence; the transition is residual and zero-initialised, so an untrained model begins as an exact copy of the previous latent and any movement away from persistence is learned. The decoder inverts the patch embedding. Training runs 59,700 steps at batch 8 under AdamW with learning rate 10^{-4}. Evaluation windows cover each trajectory uniformly and are identical across all conditions compared. Section 5.1 refits this codec to measure how much of the central result it accounts for, and Section 10 reports how much the same quantities move across training seeds.

4 The Criterion

4.1 Attributability

Let x_t be the field at step t, E and D the encoder and decoder, and f the model's one-step forecast. The reported error of a latent world model, $\|f(x_t) - x_{t+1}\|$, is bounded below by what the codec cannot represent: even a perfect dynamics model returns $D(E(x_{t+1}))$, not x_{t+1}. Define the **representation-to-change ratio**

$$\rho = \frac{\|D(E(x_{t+1})) - x_{t+1}\|}{\|x_{t+1} - x_t\|}. \tag{1}$$

The numerator is the error an oracle dynamics model would still make. The denominator is the entire signal the evaluation is asking about. When $\rho \geq 1$, the irreducible reconstruction term is as large as the quantity being predicted, and no comparison of forecast errors on this interval can be attributed to dynamics. This is a statement about the *measurement*, not about the model: a model may be excellent and be measured on an interval where $\rho \geq 1$, and the resulting number is uninformative either way.

4.2 A deployable estimate

ρ cannot be computed at decision time, since it requires x_{t+1} and its reconstruction. We substitute the two frames already in hand:

$$\hat{\rho} = \frac{\|D(E(x_t)) - x_t\|}{\|x_t - x_{t-1}\|}. \tag{2}$$

Both quantities are available to a deployed model before it acts. The substitution is justified when the reconstruction error and the rate of change are both slowly varying, and Section 4.4 measures how far that holds. The decision rule is $\hat{\rho} \geq 1 \Rightarrow$ **abstain**. The threshold is not fitted. It is the point at which redrawing the field costs what the physics moves, and it was frozen before any outcome was observed.

4.3 It decides correctly, and the classes do not overlap

Across 52 conditions, six checkpoints over three source datasets evaluated on nine targets spanning eight physics families, the rule makes **zero errors**. The two outcome classes are separated: attributable conditions reach at most $\hat{\rho} = 0.659$, non-attributable ones begin at 1.091, a gap of $1.66\times$ straddling the threshold.

model	Brier	ECE
fitted curve (t, s free)	0.0140	0.0235
fitted curve (t fixed at 1)	0.0011	0.0057
frozen rule, hard 0/1	0.0000	0.0000
constant at base rate	0.2500	0.0000

Table 1: Frame separation was used to move 58 conditions onto the decision boundary ($\rho \in [0.5, 2.0]$), giving a base rate of 0.500 – a balanced band, where a wrong call is possible and a probability would mean something. The frozen rule is still exact there. A fitted curve does not beat it, and independently locates the threshold at $t = 1.0077$, confirming the derived value of 1 from data that never saw it.

Conditions are not independent samples, and we report the bound at the unit the generalisation claim is made over. With zero errors on **eight held-out physics families**, the exact 95% Clopper-Pearson upper bound on the error rate is **37%**; at the finer unit of physics-family pairs it is 15%, and over all 52 conditions 6.8%. We quote the first.

4.4 Two stress tests, and two weaker checks

Separate the frames. If $\hat{\rho} \approx \rho$ only because consecutive frames barely differ, the result is an artifact of near-static data. Increasing the frame separation sixteenfold drives the median disagreement between estimate and target from 0.65% to 6.87%, with a maximum of 13.8%, and leaves the decision correct on **all 96 conditions** across the three separations.

Populate the boundary. Since ρ falls monotonically with separation, the separation is a knob that moves conditions through the threshold. Sweeping it placed **58 conditions in the band** $\rho \in [0.5, 2]$ **with a base rate of exactly** 0.500, 29 either side, so a wrong call is possible and a probability would be meaningful. The rule is exact there (Table 1). A fitted curve does not improve on a hard decision; even one with its threshold pinned at 1 reaches only Brier 0.0011.

The second of these is why we describe the failure mode as **decidable rather than probabilistic**. We did not fail to find a calibration curve. We constructed the region where one would be needed and found it was not.

Two weaker checks, reported as weaker. A logistic fitted to the band places the threshold at $t = 1.0077$, within 0.8% of the derived value. This is not independent of the paragraph above: it is the same fit on the same band, so we count it once. Separately, a ridge regression on pre-outcome features, held out by family over 224 conditions, is $1.26\times$ worse at predicting ρ (RMSE 0.069 against 0.055 in log space) and $1.9\times$ worse on median absolute error, while making the *identical* decision on every condition. That is a null on the decision rather than a win, and we report it as one. What it supports is only that the derived threshold needed no data to match a fitted one.

4.5 What is doing the work

$\hat{\rho}$ is a ratio, and either half might carry it. Held out by family, the numerator alone reaches 0.887 and the denominator alone 0.750, against a majority-class baseline of 0.875; McNemar's exact test against the ratio gives $p = 0.031$, 0.008 and 0.016 respectively, with **zero losses** on any condition.

One competitor is not beaten. Reconstruction error normalised by *field amplitude*, a single-frame quantity, makes the identical call on all 52 conditions ($b = c = 0$, $p = 1.000$). **What the evidence supports is normalised reconstruction error, not the ratio specifically.** We report the reconstruction-to-change form as primary because its threshold is derivable rather than fitted, and say so rather than claiming an accuracy advantage it does not have.

5 What the Criterion Says About a Standard Benchmark

Applying the test to eight datasets from The Well: across **42 cross-family source-target pairs, none is attributable at the single-step horizon** (Table 2). This is a statement about the benchmark

study	pairs	measurable	structural cells
Study A (raw inputs)	21	0	0
Study B (input-standardised)	21	0	0
total	**42**	**0**	**0**

Table 2: Track A. Every cross-domain pair ever measured. $\rho^{(u)}_{RTC}$ min 1.092, median 14.667, max 57.296; **none below 1**.

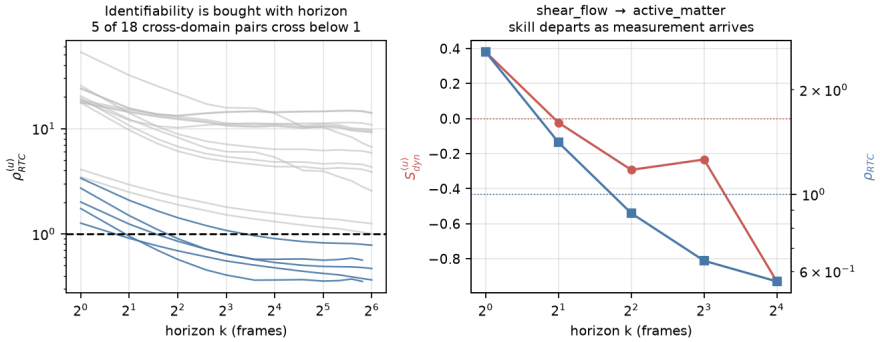

Figure 1: Identifiability is bought with horizon, unevenly. Lengthening the evaluation interval increases the denominator of eq. 1; 5 of 18 cross-domain pairs fall below $\rho = 1$ at some $k \leq 64$ and 13 never do.

protocol, prior to any question of whether transfer occurred. Even the single best cross-domain pair costs 9% more to redraw than its physics moves in one step. The numerator is a property of this codec rather than of the benchmark, and Section 5.1 refits it to establish what that dependence is worth.

Lengthening the evaluation interval increases the denominator and buys attributability back, but unevenly: 5 of 18 pairs cross below 1 at some $k \leq 64$ (95% CI $[9.7\%, 53.5\%]$), and 13 never do (Figure 1).

A closed-form argument predicts which targets can never become attributable at any horizon, by bounding ρ below once the field decorrelates from its past. It ranks targets by difficulty but its binary call is not separable from a constant "never", so we report it in Appendix B as a mechanism rather than as a validated test.

5.1 Measurability can be purchased, and the price is the result

Since ρ's numerator is this model's reconstruction error, a codec that had seen the target would shrink it. Refitting the codec on target *reconstruction*, using no target future frame and no labels, does exactly that: ρ falls below 1 on four of five pairs attempted, and fails only on `viscoelastic_instability` ($48.8 \rightarrow 4.27$). So $\rho \geq 1$ is not a property of the benchmark alone.

What the repair costs is the reason it is not a remedy. The transition is frozen throughout, verified by a core drift of exactly zero, so the model does not change; the codec changes, and with it the codec-matched persistence baseline S_{dyn} is measured against.

target	ρ strict	ρ refit	S_{dyn} strict	S_{dyn} refit
`MHD_64`	2.026	0.505	-0.348	-1.815
`acoustic_maze`	18.649	0.633	$+0.268^{*}$	-1.233
`active_matter`	2.617	0.270	-0.372	-4.380

* unidentifiable at $\rho = 18.6$, so not a score.

direction	seed	S_{dyn} at k=1 (95% CI)	ρ at k=1 (95% CI)	qualifies at k
incl→disc	43†	$+0.916$ $[0.907, 0.924]$	0.073 $[0.064, 0.084]$	1, 2, 4, 8
incl→maze	42	$+0.568$ $[0.530, 0.606]$	0.491 $[0.462, 0.524]$	1, 2, 4, 8
incl→maze	43	$+0.557$ $[0.517, 0.597]$	0.487 $[0.459, 0.521]$	1, 2, 4, 8
maze→disc	42	$+0.591$ $[0.542, 0.640]$	2.395 $[2.035, 2.826]$	**none**
maze→disc	43	$+0.609$ $[0.572, 0.643]$	2.096 $[1.760, 2.505]$	4
maze→incl	42	$+0.420$ $[0.329, 0.515]$	2.673 $[2.166, 3.230]$	**none**
disc→incl	42†	-0.094 $[-0.164, -0.003]$	1.705 $[1.340, 2.135]$	**none**
disc→maze	42†	-0.337 $[-0.372, -0.296]$	2.767 $[2.523, 3.051]$	**none**
across families (best of 7, k=8)	42	$+0.007$ $[-0.040, 0.066]$	0.502 $[0.414, 0.569]$	**none**

Table 3: All six directions among three acoustic geometries, plus the best cross-family cell for contrast. A cell qualifies at a horizon only if its whole S_{dyn} interval exceeds 0 and its whole ρ interval falls below 1; intervals are trajectory-block bootstrap, $b = 2000$. Three of the six qualify, from two of the three source models. maze→disc qualifies at k=4 on one seed and misses on the other, and is the only direction whose verdict is seed-dependent; maze→incl fails on seed 43 as well ($+0.504$ at $\rho = 2.306$). The acoustic sources were trained on the valid split, as The Well ships no train split for incl or maze; targets are different datasets and untouched.

This is the better-codec counterfactual. The objection to Section 5 is that $\rho \geq 1$ reports a weak codec, not an unmeasurable interval. The refit above *is* that counterfactual: ρ falls below 1 on four of five pairs, and skill does not follow it down.

Giving the target a codec that can render it does not reveal transfer that measurement error was hiding. It establishes that once the target is renderable, stepping the frozen transition is *worse than not stepping at all*: every refit S_{dyn} is far below the persistence baseline in its own codec. The adaptation converts "we cannot tell" into "we can tell, and there is nothing there".

The dichotomy is total across the ladder. Of the **fifteen** conditions that preserve zero-shot evaluation, with no target future frame and no labels, **none** reaches both $\rho < 1$ and $S_{\text{dyn}} > 0$. Of the **eleven** that read target future frames, **eight** do. What purchases attributability *and* skill is supervision, and supervision is not transfer. We report those eleven as a supervised ceiling, not as transfer, and note that the frozen core reaches at most 14% of it.

6 Where Transfer Survives

6.1 Across physics families

For a model trained *at* the evaluation horizon, so that compounding autoregressive error cannot be the explanation, none of seven cross-family targets achieves both an attributable measurement and positive skill over codec-matched persistence. The best reaches $S_{\text{dyn}} = +0.007$ with a 95% interval of $[-0.040, +0.066]$, selected as the best of 28 (target, horizon) looks (Table 3). Requiring an interval rather than a point estimate matters here: two cells land on the correct side of zero and neither survives its own interval.

6.2 Within a physics family

Three acoustic scattering geometries, inclusions, maze and discontinuous, give six directions, and we ran all six. Three qualify. incl→disc reaches $S_{\text{dyn}} = +0.916$ $[0.907, 0.924]$ at $\rho = 0.073$ and qualifies at every horizon. incl→maze reaches $+0.568$ $[0.530, 0.606]$ at $\rho = 0.491$ on one seed and $+0.557$ $[0.517, 0.597]$ at $\rho = 0.487$ on the other, qualifying at every horizon on both. maze→disc qualifies at $k = 4$ on one seed ($+0.155$ $[0.075, 0.233]$) and misses narrowly on the other (-0.015 $[-0.152, 0.117]$). The remaining three fail, and maze→incl fails on both seeds (Table 3).

The largest number is not the strongest transfer. The $+0.916$ exceeds this model's own in-domain score, $+0.878$ on the same seed, which should not happen. Motion is almost identical to

family	action-cond.	shuffled	degrades?	propagation
force	0.4718	1.3203	yes	0.3308
viscosity	0.1848	0.2392	yes	0.3163
ic_perturb	–	–	–	0.0621

Table 4: L3 mechanism ratio (< 1 = real, ≥ 1 = autocomplete). Competence gate $S_{dyn}^{(u)} = 0.6085$, PASSED. The shuffled control must degrade the ratio; if it does not, the model never used the action and the conditioning claim is void.

the training domain, 32.4% of field RMS against 32.8%, so the difference is in the numerator: the codec reconstructs `disc` $3.4\times$ better than the data it was trained on. It is an easier field, not stronger transfer, and it is a single seed.

These models occupy the region in-domain, which is what makes the cross-family negative interpretable: attributability and skill are jointly achievable, so their joint absence across families is a fact about transfer, not about the metric. Every *cross-family* cell with $\rho < 1$ had no skill, consistent with the criterion selecting conditions that are measurable **and worthless**; here $\rho = 0.49$ and the skill is large. **Attributability is necessary and not sufficient**, and both halves are now demonstrated.

A pattern we cannot explain, and a prediction that failed. Every direction *into* `incl` fails, and every direction *out of* `disc` fails. We registered one mechanism in advance, that transfer runs harder to easier, which predicts `maze`→`incl` succeeds. It fails on both seeds. We report the prediction and its refutation rather than substituting a second account fitted after the fact.

The in-distribution reading, and why it fails. Two obstacle geometries from one solver may reasonably be called in-distribution generalisation rather than transfer. That reading makes a testable prediction, and it fails: a descriptor distance over amplitude, motion and spectral slope puts the three within-family pairs at a median of 2.12 and the thirty-three cross-family pairs at 1.89, so the within-family pairs are, if anything, *further* apart than the cross-family ones (Appendix F). Proximity in this space does not predict transfer. What it leaves open is the name: whether the boundary is best called "physics family", "solver" or "generating process" is a question this experiment raises and does not answer.

Limits. Six directions from three geometries, not a rate over families. The `incl` and `maze` sources used the `valid` split, as The Well ships no `train` split for either, while `disc` used its `train` split ($80{,}000$ windows), so its rows are not protocol-matched with the other two. Each target is a different dataset and untouched; the in-domain figures are train-on-test and used only for the comparison above, and in-domain ρ spans more than $12\times$ across the three geometries, so in-domain S_{dyn} is not a quality ranking.

7 Intervention: A Capability Field-Space Error Cannot See

Given paired counterfactual simulations branching from an identical state, the question "what difference does this action make" is separable from "what happens next". We generate such pairs on a validated pseudospectral Navier-Stokes solver (11 of 11 validation checks, Appendix C) and train an action-conditioned model whose null branch is the same weights with the action channels zeroed. At the branch point the two branches share a state exactly, so only a model that *receives* the action can predict the difference. The mechanism ratio $\|[f(x,a) - f(x,0)] - \delta\|/\|\delta\|$ is **0.472** $[0.423, 0.550]$ for a localised force and **0.185** $[0.181, 0.189]$ for a viscosity change, where 1.0 would indicate no mechanism.

The model composes: on an intervention type combining both, never seen together in training, it scores **0.401** $[0.398, 0.405]$ with 400 of 400 pairs below 1, and shuffling those actions degrades it to **1.131** with only 120 of 400 below 1. It fails on an unseen action *geometry* (**5.72**), and there shuffling changes the number not at all (5.726 against 5.722, overlapping intervals). The contrast between the two shuffles is the claim: composition is a mechanism the model uses, and the geometry failure is an action **ignored** rather than mispredicted.

held-out family	RMSE($\log F$)	coverage
held-out linear on $(R_t, d_{back}, \mathrm{rms})$	0.591	–
constant (global mean)	1.523	–
acoustic	0.467	95.2%
shear	0.547	87.7%
mhd	0.514	90.3%
buoyant convection	0.591	83.6%
active matter	0.766	85.2%
radiative	0.535	84.7%
viscoelastic	0.830	80.2%
planetswe (never used in development)	0.059	100.0%

Table 5: Predicting the model's own forecast error before making it, on 95,390 per-window rows across 8 physics families, every number held out by whole family. Spearman $+0.822$ against the realised error; ρ_{hat} alone reaches only $+0.268$, so measurability and error magnitude are distinct predictions. Split conformal at nominal 90% realises 88.2% overall; holding out a whole family breaks the exchangeability the guarantee assumes, and the group-robust widening estimated from the training families alone ($\times 1.11$ at the median) brings it to 90.6%.

A control ladder localises the capability further. Both arms track action *scale* monotonically, and both fail on *direction* asymmetrically: a localised force is recognised in presence and position but not direction (sign cosine -0.088, against -0.990 for viscosity). The decisive rung is the constant mean action. Replacing the true force with its mean **breaks** the prediction (1.624, only 12 of 100 pairs below 1), while replacing the true viscosity with its mean **still reads as mechanism** (0.252, 100 of 100 below 1). The viscosity arm is therefore largely insensitive to *which* action it receives, only that one of roughly the right scale is present. The two arms are not equally supported and we do not quote them as a pair.

8 Predicting Its Own Error

Attributability is binary and answers "will this measurement mean anything". A different question, "how large will my error be", has a continuous answer and admits calibration in the ordinary sense.

From the same forward pass on the same windows, we record the model's realised forecast error F alongside the pre-outcome quantities available at decision time. Held out by whole physics family over **95,390 per-window rows**, a linear estimate of $\log F$ achieves RMSE **0.591** against a constant's 1.523, which is $2.58\times$ sharper, with Spearman $+0.822$ against the realised error (Table 5). The two predictors are distinct, and measurably so: $\hat{\rho}$ alone predicts error magnitude at Spearman $+0.268$.

Split conformal at nominal 90% realises **88.2%** coverage, ranging 80% to 100% across held-out families. Holding out an entire physics family breaks the exchangeability the conformal guarantee assumes (Barber et al., 2021), so the nominal rate does not apply. The size of that violation is measurable without touching held-out data: refitting while leaving out each *training* family in turn gives the inflation needed. The median factor, $1.11\times$, yields **90.6% at a nominal 90%**; the worst-case factor, $1.62\times$, over-covers at 96.0%. We report both.

`planetswe`, shallow-water flow on a rotating sphere and the one family withheld from all development, is predicted **best of the eight** (RMSE 0.059, coverage 100%).

9 The Vocabulary Hypothesis, Reported as Inconclusive

The hypothesis that motivated this work, that shared physical primitives predict transfer, cannot be tested here for a structural rather than a statistical reason: no cross-family cell is attributable at one step and none is both attributable and skillful at any horizon tested, so **with the outcome constant across every eligible cell, no predictor of it can be evaluated**. We report the leg as untested rather than refuted, and give the frozen registration we can no longer adjudicate in Appendix E.

10 LIMITATIONS

Seed dependence, measured. A second training seed produces **zero decision flips across eight targets** and no rule errors, and the separation margin widens. But in-domain $\hat{\rho}$ moves 0.15% while **cross-domain $\hat{\rho}$ moves by up to** $3.17\times$. Every ρ magnitude here is a per-run quantity; every decision built from one is not. We therefore quote no minimum cross-domain ρ: on one seed the closest cell is `rayleigh_benard` at 1.303, on the other `active_matter` at 1.979.

Independence units. The 52 conditions are three source datasets crossed with eight target families. Bounds are quoted at the family level, five times weaker than the condition count would suggest and the honest unit for the generalisation claim.

Replication. Both single-checkpoint sections were re-run on a second seed. Within-family transfer replicates within 1.9% and qualifies at every horizon on both. Every qualitative claim in Section 7 replicates: the force mechanism ratio to 0.7%, the compose-versus-ignore dissociation exactly, and the force/viscosity asymmetry. Only the viscosity action-conditioning ratio moves materially (22%, $0.185 \rightarrow 0.145$), the same arm Section 7 identifies as having clustered controls, so we report it as the less supported of the two. Seeds are not cosmetic here: `maze→disc` reads -0.015 $[-0.152, 0.117]$ at $k = 4$ on one seed and $+0.155$ $[0.075, 0.233]$ on the other, which flips its verdict. That is why no single-seed negative is quoted against a two-seed positive anywhere in this paper.

$\hat{\rho}$ runs the target through the codec. The claim is "no target future frame and no labels", not "no model evaluation".

Unfamiliarity is not the mechanism. Every condition the rule refuses is one the model was not trained on, so a detector firing on distribution shift alone would score as well on those. The discriminating case is the converse: a *familiar* condition the rule refuses. A deliberately undertrained model (150 steps, 48 windows) evaluated on **its own training domain** is refused at $\hat{\rho} = 9.190$, correctly, its true ρ being 9.003. A second undertrained model, equally familiar with its own domain, is accepted at 0.159. Both carry zero distribution shift, so a familiarity detector must treat them alike; the rule separates them by $58\times$ on codec quality alone, and decides all eight conditions in that run correctly. This establishes that familiarity is not *sufficient* for acceptance, not that the two are uncorrelated.

One architecture, and what that does not excuse. The confound belongs to the latent factorisation and so applies to the class, but the measurement of it is one codec. Undertraining is the cheapest version of that objection and we can exclude it. Against a 150-step checkpoint on the same source and targets, full training improves *in-domain* ρ by $15.7\times$ ($9.00 \rightarrow 0.57$), crossing the threshold, while its best cross-domain gain is $1.35\times$; on the second seed those figures are $13.8\times$ and $1.74\times$, and `turbulent_radiative` moves the wrong way on both. Training longer buys attributability where the model already lives and little elsewhere. The two budgets were scored under different window protocols, so we read the contrast in effect sizes and not the individual ratios. Section 5.1 rules out the other cheap reading. A larger or differently designed encoder is untested; a third-party checkpoint would settle it.

Horizon semantics. Two tools mean different things by k, frame gap in one and autoregressive passes in the other, and they never share an axis. Appendix A gives the mapping and the artifact provenance.

11 CONCLUSION

A metric with no baseline is not a measurement. We give the check that makes a field-space error interpretable, show that a standard benchmark fails it at the horizon it uses, and report what survives: transfer inside a physics family and not across one, in three of six directions we could test and from two of three source models, an action-conditioned model that composes interventions it never saw together, and a per-window error estimate sharpest on the one family withheld from development. The criterion needs only a latent model, a reconstruction and two consecutive states, so any latent world model in any domain can compute it about itself. For an agent that plans with a learned model, that is the difference between acting on a rollout and knowing not to.

AI USE STATEMENT

Large language model tools were used for drafting assistance and for LaTeX formatting. Experiments in this project were executed by an autonomous agent under author direction; every number reported here was verified by the authors against the logged artifacts in `results/canonical/`, and every claim was checked against an audit that asks of each one what its independence unit is, whether it carries an interval, and whether the quantity was measured on the thing it is quoted about. Two findings did not survive that audit and are reported as retracted rather than omitted.

ETHICS STATEMENT

This work uses The Well, a public benchmark of simulated PDE data, and simulations from a solver we wrote. It involves no human subjects, personal data or human-annotated labels. The intended use is to let a physics surrogate decline targets it cannot model, which we regard as risk reducing. The main foreseeable misuse is the opposite reading, treating a passing measurability check as a licence to trust a rollout. We state explicitly that attributability is necessary and not sufficient: Section 6 exists precisely because a valid measurement and a useful model are different things.

REPRODUCIBILITY STATEMENT

Section 4 gives the criterion in closed form with its threshold and the reason the threshold is 1. Every table carries the artifact it was generated from, every interval states its resampling unit and bootstrap count, and every proportion carries an exact Clopper-Pearson interval at the independence unit named in Section 10. Artifacts backing the claims here carry a content digest over all source files, so two runs with the same digest provably ran the same code. Code and configuration are in the supplementary material.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference*, 10(2):455–482, 2021.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video (V-JEPA). *arXiv preprint arXiv:2404.08471*, 2024.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.

Quentin Garrido, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv:2403.00504*, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4878–4887, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

David Ha and Jürgen Schmidhuber. World models. In *NeurIPS*, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Kappeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for PDEs. In *NeurIPS*, 2024.

Maximilian Herde et al. Towards a physics foundation model. *arXiv preprint arXiv:2509.13805*, 2025.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

Yann LeCun. A path towards autonomous machine intelligence, 2022. OpenReview / Meta AI position paper, Version 0.9.2.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *ICLR*, 2021.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Holden Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, et al. Multiple physics pretraining for physical surrogate models. *arXiv preprint arXiv:2310.02994*, 2023.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Allan H. Murphy. Climatology, persistence, and their linear combination as standards of reference in skill scores. *Weather and Forecasting*, 7(4):692–698, 1992.

Tung Nguyen, Arsh Koneru, Shufan Li, and Aditya Grover. PhysiX: A foundation model for physics simulations. *arXiv preprint arXiv:2506.17774*, 2025.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, et al. The well: a large-scale collection of diverse physics simulations for machine learning. In *NeurIPS Datasets and Benchmarks*, 2024.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning mesh-based simulation with graph networks. In *ICLR*, 2021.

Polymathic AI Collaboration. Walrus: A cross-domain foundation model for continuum dynamics. *arXiv preprint arXiv:2511.15684*, 2025a.

Polymathic AI Collaboration. Physics steering in Walrus, 2025b. Technical blog post, `https://polymathic-ai.org/blog/walrus/`.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *ICML*, 2020.

Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael W. Mahoney, and Amir Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. In *NeurIPS*, 2023.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

quantity	value
deployable rule correct	48 / 48
headline Spearman	+0.9999
numerator R_t/R_{oracle} median	0.9999
denominator $\Delta_{back}/\Delta_{fwd}$ median	1.0011

Table 6: L2, over the 48 development conditions. The deployable rule $\hat{\rho}_{RTC} \geq 1$ uses two frames already in hand and no ground truth. **Both decomposition halves sit at 1 at this horizon**, so the headline correlation is near-tautological here and is never reported alone; Section 4.4 therefore tests the estimate at sixteen times the frame separation, where the two quantities disagree by up to 13.8% and the decision is still correct on all 96 conditions.

A Horizon Semantics and Artifact Provenance

Horizon semantics. Two tools mean different things by k, frame gap in the sweep and autoregressive passes in the rollout, and they never share an axis. Window populations differ between them by construction, a ρ from one is not comparable to a ρ from the other, and we verified that the two agree exactly on their shared windows. The axis is frame count throughout; the same k spans a $160\times$ range in physical time across datasets, which we do not control for.

Provenance. Artifacts generated before this project recorded a content digest name a commit taken from a working tree with uncommitted changes, so they cannot be tied to an exact revision. Every artifact backing a claim in the main text was regenerated afterwards and carries a digest over the content of every source file.

B A Closed-Form Floor on ρ

Since $\|x_{t+k} - x_t\|$ saturates at $\sqrt{2}\,\sigma_{\text{field}}$ once the field decorrelates from its past, and the oracle reconstruction error is roughly horizon-independent, ρ has a floor at $\rho_{\text{floor}} \approx R_{\text{oracle}}/(\sqrt{2}\,\sigma_{\text{field}})$. This predicts which targets can never become attributable at any horizon.

Over 45 cross-domain conditions the bound ranks targets by difficulty (Spearman $+0.554$, $p = 8 \times 10^{-5}$) and declares 53% permanently unreachable. Its binary call, however, is not separable from a constant "never", which scores 13 of 18 against the rule's 15 (McNemar $p = 0.727$). We therefore report the floor as a mechanism with a consistent failure direction, and not as a validated test. The ranking result is the part we would defend.

C Solver Validation

The intervention experiments of Section 7 run on a pseudospectral Navier-Stokes solver we wrote, which passes 11 of 11 validation checks including Taylor-Green decay to 1×10^{-8}, RK4 order measured at 4.04, and inviscid energy drift at 2.2×10^{-16}.

One of the original eleven checks was replaced. It compared a single Fourier mode at two resolutions, returned exactly 0.0, and could never have returned anything else. A check that cannot fail is not a check, and we report the replacement rather than the count it inflated.

D The Deployable Estimate and Its Decomposition

Table 6 is scored over the 48 development conditions. The 52 conditions quoted throughout the body are those 48 together with the four `planetswe` confirmation cells, which were generated after a window-selection guard landed and are therefore reported separately here.

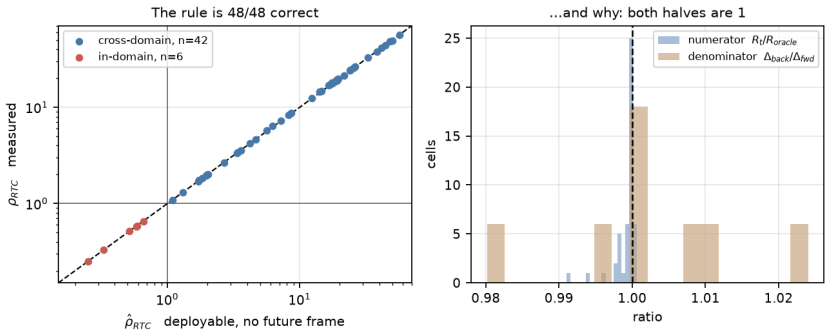

Figure 2: The deployable rule across the development conditions. Both halves of the decomposition sit near 1 at the single-step horizon, which is why Section 4.4 tests the estimate at sixteen times the frame separation rather than resting on the single-step agreement.

E The Vocabulary Hypothesis in Full

The hypothesis that motivated this work, that shared physical primitives predict transfer, cannot be tested here, for a structural rather than a statistical reason. No cross-family cell is attributable at one step (Section 5), and none is both attributable and skillful at any horizon tested (Section 6). **With the outcome constant across every eligible cell, no predictor of it can be evaluated**, and the overlap score is such a predictor. We report the leg as untested rather than refuted.

A registration we can no longer adjudicate. We registered four predictions in a frozen specification before the corresponding runs existed, each placing a target in a high- or low-error band. They are not wrong but *unadjudicable*: all four targets sit at $\rho \geq 1$ at the horizon the registration specified, so the quantity predicted was never measured. We record this rather than dropping it.

F Dataset Descriptor Distance

Each dataset is summarised by three scale-free descriptors computed on its `valid` split: field amplitude (RMS), motion (mean frame-to-frame change divided by field RMS) and the slope of the spatial power spectrum. Distance is the Euclidean norm between standardised descriptor vectors, so it is a statement about the data, not about any model. All $\binom{9}{2} = 36$ pairs are listed, sorted by distance; **w** marks a within-family pair.

The three within-family pairs have median distance $\mathbf{2.12}$ and the thirty-three cross-family pairs $\mathbf{1.89}$. The closest cross-family pair is nearer than two of the three within-family pairs. This is the basis for the claim in Section 6 that proximity in descriptor space does not predict which pairs transfer: the two directions that qualify at every horizon run between datasets 0.36 and 2.12 apart, while `shear_flow` fails to `active_matter` at 1.09, to `rayleigh_benard` at 1.36, to `viscoelastic_instability` at 1.89 and to `MHD_64` at 2.03.

pair	d		pair	d	
incl–disc	0.36	w	MHD_64–viscoel	1.92	
disc–rayleigh	0.53		incl–active	1.99	
incl–rayleigh	0.55		shear–MHD_64	2.03	
incl–radiative	0.90		incl–maze	2.12	w
rayleigh–radiative	1.05		maze–MHD_64	2.22	
shear–active	1.09		shear–radiative	2.40	
disc–radiative	1.09		maze–radiative	2.44	
MHD_64–rayleigh	1.16		rayleigh–viscoel	2.45	
disc–shear	1.31		maze–disc	2.48	w
shear–rayleigh	1.36		MHD_64–active	2.52	
incl–shear	1.53		maze–rayleigh	2.62	
active–viscoel	1.62		maze–active	2.81	
disc–MHD_64	1.69		active–radiative	2.89	
incl–MHD_64	1.71		disc–viscoel	2.98	
MHD_64–radiative	1.79		incl–viscoel	2.99	
disc–active	1.80		viscoel–radiative	3.08	
active–rayleigh	1.85		maze–viscoel	3.62	
shear–viscoel	1.89		maze–shear	3.65	

G BENCHMARK IDENTIFIABILITY

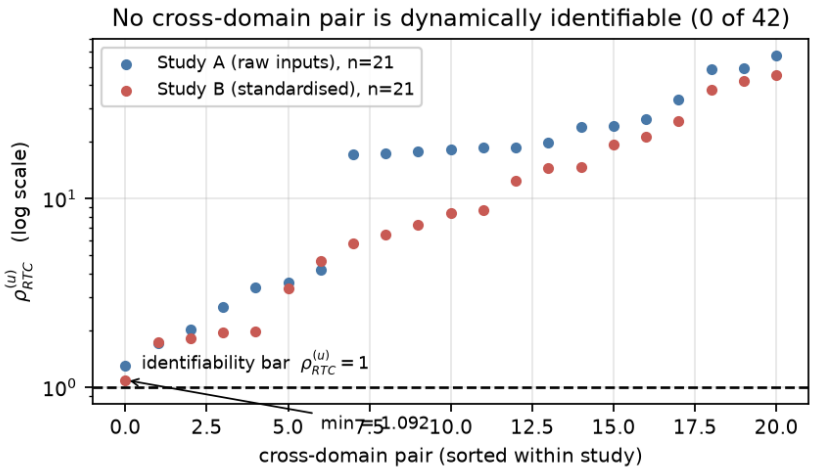

Figure 3: Every cross-domain pair measured, ordered by ρ. None falls below 1 at the single-step horizon; the minimum is 1.092.