DOES SURPRISE MEASURE INTUITIVE PHYSICS? THE SCORE DEPENDS ON HOW THE MODEL IS READ

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent-predictive video models are said to acquire intuitive physics when they are more surprised by an impossible event than by a possible one (Garrido et al., 2025). That surprise is a single scalar, the difference in prediction error between two otherwise identical clips. The benchmark that defines that score chose the measure deliberately, naming one alternative protocol (Weihs et al., 2022) and reporting that the alternative had not yielded evidence of understanding (Bordes et al., 2025). Nothing has established what the difference contains. We decompose the squared form exactly into magnitude, direction and a selection term, and show it is not a fixed property of the model. We measure it on V-JEPA 2 (Assran et al., 2025) and V-JEPA 2.1 (Mur-Labadia et al., 2026). Both are publicly released latent-predictive encoders, and only V-JEPA 2.1 exposes intermediate depths to read at. Reading V-JEPA 2.1 at the four encoder depths its predictor was trained to read, with nothing attached on top, the reported accuracy spans 10.18 percentage points with the model, the clips and the readout all unchanged. Across the two encoders every term keeps its sign and none keeps its size, so the two standard normalisations disagree about which model is more direction-driven. A motion-only observer touching neither model shows that most or more of what the score detects is available from movement alone. We then move the score using a regulariser carrying no physics. An untrained random projection preserves accuracy at 88.39% against the frozen encoder's 89.55%. Training a small network in its place, under the coefficients published with VICReg (Bardes et al., 2022), drives it to 46.57%, a point estimate below chance whose 95% confidence interval contains chance. Raising one of those three coefficients recovers 94.62%. A violation-of-expectation score is not readable as physics evidence until six quantities are reported beside it. We specify those six and show what each one catches.

1 INTRODUCTION

A video model is said to have acquired intuitive physics when it is more surprised by an impossible event than by a possible one. Two beliefs sit behind that inference. First, that the difference in prediction error between two otherwise identical clips measures the model's surprise. Second, that this single scalar is a property of the model, so that comparing it across models compares how much intuitive physics each one has acquired. We show that the second belief does not hold, and that the first cannot be checked without it.[1]

The benchmark that defines it named one alternative protocol, reported that the alternative had not yielded evidence of understanding, and chose the scalar deliberately (Bordes et al., 2025). In the nine papers surveyed in Section 2, the two whose conclusion is physical understanding carry no trained readout between them. If the scalar does not mean what it is taken to mean, nothing else in those two papers supports the conclusion.

The number moves, and nothing about the model moves with it. Hold the encoder and the predictor frozen. Place a small network on top and train it with an objective containing no physics, no labels and no reference to the task. At the coefficients published with that objective, paired

[1] An anonymised code archive accompanies this submission at `https://anonymous.4open.science/r/voe-score-decomposition-401C/`.

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.

physics classification falls to 46.57%, a point estimate below chance whose 95% confidence interval of $[39.45, 53.70]$ contains chance. Raise one of the three coefficients and the same architecture, at the same width and parameter count, reaches 94.62%. The frozen model underneath does 89.55% throughout and never changes. That is a swing of 48 percentage points in the number the field publishes, driven by a choice nobody reports. Reading depth does the same more quietly. Read at the four trained reading depths with nothing attached, the reported accuracy spans 10.1838 percentage points with the model, the clips and the readout all unchanged. Only the reading place moves. Four depths do not make a curve, so we claim no ordering among them.

A score that moves under reading choices is still useful, but only if those choices are reported. None of the papers surveyed states the depth it reads at, and none reports a readout objective. The quantity is treated as though it had one value per model. It has one value per model *and reading*.

Our contributions.

1. **An exact decomposition of the squared surprise score** into magnitude, direction and selection, computed at the predictor with no trained readout in the path, on two public encoders. No paper in the surveyed pool decomposes a violation-of-expectation score. The effect lives in direction, since the direction term is positive on 9 of 9 violation types while the magnitude term is positive on 3 of 9 and negative when pooled, the opposite of the intuitive reading.
2. **The measurement that the score is not a property of the model**, across the four trained reading depths of one frozen model and across 23 readout settings at four invariance weights, ten of them measured on the accuracy scale and spanning 46.57 to 97.68 percent with no physics content in the change.
3. **A reporting requirement of six quantities, with a demonstration of what each catches**, applied to our own runs. The score needs a rule for which positions it reads, and four are available. Under one of them an untrained baseline outperforms the trained readout, so that rule supports no claim about the model.

Section 7 states them in full. They are a checklist rather than a decomposition of one another, because three are bounded on $[0, 1]$ and three are not, so they do not compose into a single residual.

Relation to known results on readout capacity. Kolesnikov et al. (2019) establish that a linear readout is adequate for evaluation, which is easily read as evidence against Section 5.5. Its one readout-size experiment returned a null, and its main variation was the encoder. We hold the readout's size fixed to the byte and vary the loss it is trained to minimise. Section 6 gives that comparison and two others. That an evaluation choice can move a headline is not a new worry, having been reported for pretext-task proxies in vision (Kolesnikov et al., 2019), downstream architecture in speech (Zaiem et al., 2023), and representation-quality protocols in images (Marks et al., 2024). What we add is an exact account of what the moving quantity is made of, on the metric that carries the physics claim.

2 THE PRACTICE WE EXAMINE

A violation-of-expectation evaluation shows a video model two clips that are pixel-identical until a divergence frame, after which one obeys a physical principle and the other does not. The model's prediction error on each clip is read as its surprise, and the reported quantity is the difference between the two. A larger error on the impossible clip is taken as evidence that the model held an expectation which that clip violated.

The benchmark's own statement, and the alternative it set aside. Bordes et al. (2025) state the inference directly, that "we split a video into overlapping windows … For each window, the model predicts the target part based on the context, and the prediction error measures the model's surprise. Comparing surprise across videos probes the model's understanding." They named one alternative, on a prior benchmark, and set it aside, because there "models trained for action recognition are probed by measuring how out of distribution impossible videos are. This has however not yielded

evidence of understanding in models." The scalar is a considered choice about what counts as evidence, and we treat it as one.

The word "probe" carries two senses and the field uses both. In representation learning a probe is a readout *trained on labels*. In ordinary English it means to examine. Both papers use it in the ordinary sense while describing a prediction-error method, and Garrido et al. (2025) write "without any additional adaptation, to probe the model's understanding," using the word and disclaiming the trained head in one breath. A reader who imports the technical sense imports a guarantee that was never offered.

Two facts recorded from each of nine papers. We read the full text of each of nine papers and recorded two facts about each, whether it reports a prediction-error surprise score and whether it reports the accuracy of a readout trained on labels. Table 2 in Appendix A gives them per paper. **The two rarely appear together.** Garrido et al. (2025) and Bordes et al. (2025) report a surprise score and train no readout at all. Assran et al. (2025) and Joseph et al. (2026) train readouts and report no surprise score, using them to measure capability rather than physics.

Two rows need care. Esmati et al. (2026) appears to report both and does not, because its "surprise" is a probe delta taken when a block is perturbed, the architecture having no predictive future representation to take a prediction error from. A diffusion model cannot produce this scalar at all, which is a fact about the metric. Internò et al. (2026) reports both kinds, scoring VideoMAE by masked reconstruction and separately training a linear probe on DINOv3 features. Those are two different backbones, so the two readings never meet on one model.

Two papers rest on the scalar alone. No paper in this nine-paper pool reports a prediction-error surprise score and a trained-readout accuracy *on the same model*. The two papers whose conclusion is physical understanding carry zero trained readouts between them, and Garrido et al. (2025) concludes that "prediction in a learned representation space is sufficient to develop an understanding of intuitive physics" on the scalar alone. Neither paper reports a trained readout, so the scalar carries the conclusion by itself.

Scope of the pool. Nine papers reflects the size of this literature. The pool is not exhaustive, so it shows the practice is common without showing it is universal. It does not establish that papers which probe are wrong, and Esmati et al. (2026) disclaims that inference itself, writing that linear probes "demonstrate decodability but do not imply that the model explicitly represents physical laws." The claim is about what is reported *together*, never about who is correct.

3 What the score is, formally

A violation-of-expectation score is a distance between a prediction and an observation, and a distance is not defined until somebody chooses a space to measure it in. We analyse the squared form, which is what our measurements use. Garrido et al. (2025) defines its surprise with an L_1 norm, so the split below covers the squared form only. The field reports such a score as a property of the model when it is a joint property of the model and the reading choice. Nothing below is a new result.

The squared form we analyse. For a prediction $\boldsymbol{p} \in \mathbb{R}^d$ and an observed target $\boldsymbol{a} \in \mathbb{R}^d$ at one position,

$$s(\boldsymbol{p}, \boldsymbol{a}) = \frac{1}{d} \sum_{j=1}^{d} (p_j - a_j)^2 = \frac{1}{d} \left\| \boldsymbol{p} - \boldsymbol{a} \right\|^2 . \tag{1}$$

The reduction is a mean over the feature axis rather than a sum, and $1/d$ is constant, so every property of a squared Euclidean distance transfers to s. A distance needs a space and the model supplies none. The predictor we study (Mur-Labadia et al., 2026) concatenates four encoder depths, and in the checkpoint we load each is 1408 wide, so reading a score means slicing one depth out of 5632. That is a required argument with four legal values, no default, and no paper in our pool reports which it used. The step is definitional. Whether it matters is empirical.

The decomposition is the square expanded. Write mean_d for the mean over the feature axis. The two clips of a twin pair share a context, so the frozen predictor emits the same tensor for both arms, which we verify rather than assume, and the shared term cancels position by position,

$$s_{\mathrm{imp}} - s_{\mathrm{pos}} = \underbrace{\mathrm{mean}_d\left(\boldsymbol{a}_{\mathrm{imp}}^2 - \boldsymbol{a}_{\mathrm{pos}}^2\right)}_{\mathrm{NORM}} \underbrace{-2\,\mathrm{mean}_d\left(\boldsymbol{p}\left(\boldsymbol{a}_{\mathrm{imp}} - \boldsymbol{a}_{\mathrm{pos}}\right)\right)}_{\mathrm{ALIGN}}. \tag{2}$$

NORM compares the magnitudes of the two targets and never touches the prediction, while ALIGN is the only term the prediction enters. That asymmetry falls out of the algebra. A third term is forced because *our* readout averages over the top q fraction by score, taken separately in each arm. No published protocol does this, Garrido et al. (2025) offering a mean or a maximum over the window sequence, so the term is forced by our aggregation rather than by the field's. Write T_{imp} and T_{pos} for each arm's *support*, the set of positions that arm's own top-q selection keeps. Write mean_T for a mean over the positions in T rather than over the feature axis. The reported number averages each arm over its own support,

$$\Delta = \mathrm{mean}_{T_{\mathrm{imp}}}\, s_{\mathrm{imp}} - \mathrm{mean}_{T_{\mathrm{pos}}}\, s_{\mathrm{pos}}. \tag{3}$$

Equation 2 relates the two arms at one position. The two means in Equation 3 run over different positions, so the identity does not substitute into Δ as it stands. Write $M = \mathrm{mean}_{T_{\mathrm{imp}}}\, s_{\mathrm{pos}}$ for the possible arm's score read on the impossible arm's positions. Adding M and subtracting it again leaves Δ unchanged, and splits it into two differences that are each well posed,

$$\Delta = \left(\mathrm{mean}_{T_{\mathrm{imp}}}\, s_{\mathrm{imp}} - M\right) + \left(M - \mathrm{mean}_{T_{\mathrm{pos}}}\, s_{\mathrm{pos}}\right). \tag{4}$$

The first bracket averages both arms over one support, which is the condition Equation 2 needs, so it is NORM and ALIGN averaged over T_{imp}. The second bracket holds one quantity, s_{pos}, read over two supports, so the impossible arm never enters it. Substituting the first bracket and naming the second P,

$$\Delta = \mathrm{mean}_{T_{\mathrm{imp}}}\left(\mathrm{NORM} + \mathrm{ALIGN}\right) + \underbrace{\mathrm{mean}_{T_{\mathrm{imp}}}\, s_{\mathrm{pos}} - \mathrm{mean}_{T_{\mathrm{pos}}}\, s_{\mathrm{pos}}}_{P}. \tag{5}$$

NORM and ALIGN keep the meaning they had at a single position. P measures how far the two selections differ and is zero when they agree.

Opposite shifts carry no information when Δ is near zero. The three terms partition Δ exactly, so their shifts must sum to the shift in Δ. Where Δ barely moves the shifts nearly cancel, and two terms moving in opposite directions is forced by the arithmetic. A claim that some intervention shifted weight from one term to another is therefore empty in that regime, whatever the measured shift. The constraint is derived, and Section 5.5 declines exactly such a claim on our own data.

A readout is a choice of metric. Insert a map g before the score, applied alike to the prediction $\boldsymbol{p}$ and its observed target $\boldsymbol{a}$ from Equation 1, and it becomes $s_g = (1/d')\|g(\boldsymbol{p}) - g(\boldsymbol{a})\|^2$, a different distance on the same pair. For linear $g = \boldsymbol{A}$ and $\boldsymbol{e} = \boldsymbol{p} - \boldsymbol{a}$, $\|\boldsymbol{A}\boldsymbol{e}\|^2 = \boldsymbol{e}^\top \boldsymbol{M}\boldsymbol{e}$ with $\boldsymbol{M} = \boldsymbol{A}^\top\boldsymbol{A}$, so the readout chooses the metric. Preserving the score up to one overall scale needs $\boldsymbol{M} = k\boldsymbol{I}$, which forces full column rank, and a map from $\mathbb{R}^{1408}$ into $\mathbb{R}^{256}$ has rank at most 256. So no linear readout from $\mathbb{R}^{1408}$ into $\mathbb{R}^{256}$ preserves it. Such a map sends a subspace of at least $1408 - 256 = 1152$ dimensions to zero, and it discards whatever part of the error lies there. The readout reads an error lying entirely inside that subspace as zero surprise, while the error itself is non-zero. This is a counting fact rather than an empirical finding, and it does not describe our trained head, which is nonlinear and attention based, though to first order that head acts through a local Jacobian of the same shape.

The objective pulls on the score. The readout is trained on two views of one clip, the frozen predictor's output and the frozen encoder's target. Its invariance term is a mean squared error between the two readout outputs. The objective treats every position of every clip in the batch as one sample, so with N samples and output width d',

$$\mathcal{L}_{\mathrm{inv}} = \frac{1}{Nd'}\sum_{i=1}^{N}\sum_{j=1}^{d'}\left(g(\boldsymbol{p})_{ij} - g(\boldsymbol{a})_{ij}\right)^2 = \frac{1}{N}\sum_{i=1}^{N} s_g(\boldsymbol{p}_i, \boldsymbol{a}_i). \tag{6}$$

The invariance term equals the mean of the score over the batch. The two are the same arithmetic on the same two tensors, the readout's output on the prediction and its output on the target. Across the two registered settings the invariance coefficient is held at VICReg's published value of 25 while only the variance and covariance coefficients move, so the pressure described below is present in both. Since s_g, the score read through the readout, vanishes only when $g(\boldsymbol{p}) = g(\boldsymbol{a})$, gradient descent on its batch mean drives that score toward zero in both arms of every pair, and Δ is built from those scores. The readout is trained to erase the quantity being decomposed. That is a direction of pressure rather than an endpoint. Training stops short of the zero-score solution that term points at, and the variance and covariance terms oppose collapse, so the magnitude of the erasure is measured and never derived.

The reading of the two terms is a hypothesis. Reading the variance term as magnitude and the covariance term as direction is a hypothesis, recorded as one in our implementation. Nothing above derives it, and we do not argue for it. Nothing here shows that the model lacks physical understanding. It shows that this readout does not isolate it.

4 METHOD

4.1 THE CORPUS

The stimuli are generated twins of the kind Section 2 describes. Generating them rather than collecting them is what makes the controls below possible, so the corpus is control apparatus and nothing here is a claim about scale.

Nine violation types are used, inheriting the taxonomy of Bordes et al. (2025) with Permanence dropped in full, because a prediction-error readout is structurally blind to a principle that changes whether an object exists at all. That blindness is untested, so nothing here rests on it. Permanence was measured on an earlier cohort, where it was the largest contributor, and dropped before this corpus was built. A sign count is the number of the nine types where the difference between the impossible and possible clips comes out positive, written k of 9 and referred to $\mathrm{Binomial}(9, 0.5)$. That reference assumes nine independent trials at a null probability of one half. The nine types are disjoint pair sets, which supports it, but they share one frozen model, one readout and one generator, so a bias at any of those levels moves all nine signs together. Their pair counts also run from 51 to 952, so the nine are not equally precise. We use the reference as a scale rather than an exact null, and no conclusion here rests on a p value taken from it alone.

The generator produced 3476 twin identities and 3427 clear the perceptual contrast bar. Per-type counts run from 51 pairs to 952. Table 1 and Figure 1 in Appendix A give the counts and one twin per type. Three properties hold, and the third was measured on a separate cohort. The shared prefix is byte-identical within every pair and stored losslessly. Every violation type passes a human visibility gate. Every type also draws on the same mix of object shapes by construction, so no type can be told apart by the objects in it rather than by the physics. That marginal was measured on a 1152-pair cohort with equal per-type counts, not on the corpus reported here, whose counts are unequal by design.

4.2 HOW THE TERMS ARE NORMALISED

The magnitude, direction and selection terms of Section 3 are each divided by the mean of the impossible and possible arms' output mean-squares, which makes them dimensionless. A raw difference scales with the square of the readout's output scale, and the readouts we measure span a factor of 253 in that scale. Without dividing, their raw differences would span roughly the square of that for reasons that have nothing to do with physics.

4.3 THE READOUTS AND THE DEPTH AXIS

Three readout conditions appear here, and the encoder and predictor are frozen in all three. The first attaches no readout, so the model's output is read directly and gives the frozen model's own number. The second is a random projection from 1408 dimensions down to 256, every weight left at its initial value, so it compresses without having learned anything. The third is a network

with that same 256-dimensional output, trained with the VICReg objective (Bardes et al., 2022) at variance and covariance coefficients including the published ones. Every trained readout has the same parameter count, so none has more capacity than another, and all start from byte-identical weights at one seed. Only the objective's coefficients differ, and we train no readout anywhere on labels. A probe is a readout trained on labels, so accuracy through these readouts is a property of the representation handed over rather than of a classifier fitted to the answer.

V-JEPA 2.1's predictor consumes the concatenation of encoder layers 9, 19, 29 and 39, and emits those four levels at 5632 features. We call these the four trained reading depths, since the predictor was trained to read exactly these four and no labels enter at any of them. Each depth carries its own trained normalisation, measured in the released weights rather than inferred, which is what makes the four readings comparable. The same comparison is not available on V-JEPA 2 for a structural reason. That encoder builds a single normalisation fitted to its final layer, so pulling an intermediate layer out sends the result through a normalisation that does not belong to it. A reading that moved across depths on V-JEPA 2 could not be told apart from that mismatch.

4.4 The pinned divergence frame, and why a sliding window does not apply

The divergence frame is pinned at frame 9 or later of 16, so the violation always lands in the half the predictor is asked to anticipate. That control separates asking a model to predict a violation from handing it one.

The obvious temporal control, sliding a window across the clip, does not port to this metric, and the design above is why. A window falling entirely inside the shared prefix sees byte-identical twins, so its difference is structurally zero. A window straddling the onset hands the violation to the conditioning context, where the readout is given the difference rather than asked to anticipate it. A sliding window therefore measures onset alignment, a different quantity, and averaging across window positions mixes the two. The eligibility rule that enforces the pinning drops any pair whose divergence reaches the frames the model reads as context. No violation type is held out of the 3427-pair cohort reported here.

5 Results

5.1 The reported number moves with reading depth alone

Reading the same frozen model at each of its four trained reading depths, with an identity readout and no trained head anywhere, changes nothing about the model and nothing about the clips. Table 7 in Appendix E gives every value per depth.

Three quantities from one model disagree about how much changed. Accuracy, the number the field reports, spans 10.1838 percentage points and peaks at depth 29. The raw gap spans 68.33 percent and is largest at depth 39, while the gap divided by the representation's own scale spans only 3.88 percent and the selection share falls monotonically by a factor of 2.96. So depth 39 carries the largest raw gap while depth 29 carries the best accuracy, and size and directional reliability come apart inside one frozen model with no readout, training or regulariser in the path. The three share columns sum to exactly 1.0000 at every depth, so reporting shares alone removes the base. A share cannot be read without its denominator.

Limits of the depth table. No cited paper states which depth it reads, so no published result is impugned by this table. No mechanism for the peak is offered. Four depths of one model is not a curve, and we do not draw one.

5.2 The effect is learned, and it lives in direction

At layer 29 under per-arm aggregation on the 3427-pair cohort, with no readout in the path, the trained model reads a pooled gap of 0.20630975 with all nine violation types signed. Four untrained initialisations read 0.00000575, 0.00000930, 0.00002743 and -0.00003028, with three or four types signed, which is chance on nine sign tests. Figure 2 in Appendix B plots them per type.

The direction term is positive on nine of nine types. The magnitude term is positive on three of nine and its pooled value is negative. An impossible clip producing a *larger* predicted representation would show up as a positive magnitude term, and pooled that term is negative. Solidity is the clearest single case, its direction term reading 0.51588338 over 471 pairs while its magnitude term reads -0.10492901 on the same pairs. Whatever the score measures, it is not the size of the predicted representation. Figure 3 in Appendix B plots both terms per type.

The decomposition does not remove motion. Motion enters the score through the direction term at a correlation of $+0.4416$.

5.3 How much of the score plain motion explains

A motion-only observer reading pixels through a frozen DINOv2 image encoder, touching neither model under test, reaches 0.7660 paired accuracy over all 3427 pairs. Appendix H gives its encoder, tracker, feature set and readout in full. Adding the trained direction channel, the direction term of Section 3 used as a further input, raises that to 0.8862 on V-JEPA 2 and 0.9378 on V-JEPA 2.1, increments of $+0.1202$ and $+0.1719$, the first with a 95% confidence interval of $[+0.1054, +0.1350]$, from a one-sample t on the per-pair differences over these same 3427 pairs rather than over seeds. Subtracting each encoder's worst untrained floor leaves separation residuals of $+0.1118$ and $+0.1631$. On the encoders' own statistic the motion-only observer separates the twins at AUROC 0.8587, against V-JEPA 2.1's 0.8955 and V-JEPA 2's 0.5708. So on V-JEPA 2.1 plain movement explains most of what the score detects, and on V-JEPA 2 more than the score does, which closes the alternative explanation on one encoder and reverses it on the other. The two statistics coincide for a sign-based score and not for a fitted baseline. Every paired accuracy here is an integer count over 3427 pairs, so one pair is worth 0.000292 and we round to 10^{-4}.

The direction channel alone separates the two encoders. Without motion features it reaches 0.9087 on V-JEPA 2.1 and only 0.7683 on V-JEPA 2, barely above the baseline it is meant to improve on. Only the direction channel differs between the two encoders, and the motion-only baseline does not use it. That baseline is therefore identical under both by construction, and its equality says nothing about what adding the direction channel achieved. These are paired-accuracy deltas and we report one scale throughout, since the same quantity also has an AUROC form and a value on a 3376-pair cohort that drops one violation type, neither comparable to these. These increments sit about one order of magnitude above their untrained floors, where the decomposition terms sit three to five orders above theirs. Figure 4 in Appendix C plots them.

5.4 The two encoders disagree, and the decomposition reveals it

On the same 3427-pair cohort through one code path, V-JEPA 2 separates the twins at AUROC 0.5708 with 5 of 9 types signed and V-JEPA 2.1 at layer 29 at 0.8955 with 9 of 9, over untrained floors by $+0.0926$ and $+0.4159$. Removing the whole-field offset then changes the conclusion on one model and sharpens it on the other. On V-JEPA 2.1 the localised residue survives at 0.8637, 95% confidence interval $[0.8523, 0.8751]$, clearing its floor of 0.4762 by $+0.3875$ with 9 of 9 signed. On V-JEPA 2 it falls to 0.4946, 95% confidence interval $[0.4780, 0.5112]$, which contains chance, clearing its floor of 0.4927 by $+0.0019$ with 3 of 9. The pattern does not hold across both encoders, and a single reported score would not have shown that. Both are percentile bootstrap intervals over 4000 resamples, a method our protocol does not pre-register and which was therefore chosen after the data were collected.

The two standard normalisations disagree, and both are legitimate. By share of its own gap, V-JEPA 2 reads $+163.889$ percent on direction against V-JEPA 2.1's $+114.525$, making V-JEPA 2 the more direction-dominated of the two. Dividing each term by its own model's output mean-square reverses that, with direction reading 0.011514 against 0.093897, about eight times smaller, and the selection separation residual reading 0.000273 against 0.005551, about twenty times smaller. V-JEPA 2's shares read large only because they divide a gap that is itself small. Neither form may be quoted without the other.

5.5 A REGULARISER CARRYING NO PHYSICS MOVES THE SCORE ACROSS CHANCE

We now intervene, holding the encoder and predictor frozen so that only what a readout reports can move. A random, untrained projection from 1408 dimensions to 256 classifies at 88.39 percent against the frozen model's 89.55, with effect sizes of $+1.2304$ and $+1.2293$ agreeing to two decimals, so the width is not a bottleneck. Training a network of that same width under the coefficients published with VICReg (Bardes et al., 2022) drives it to 46.57 percent, a point estimate below chance whose 95% confidence interval $[39.45, 53.70]$ contains chance. Table 5 in Appendix D gives every setting with its seeds, intervals and effect sizes.

The finding lands on our own registered configuration first. VICReg's published coefficients are the field's default and are also what this paper pre-registered. Over 12 seeds the readout retains -0.49 percent of the frozen model's dimensionless gap, 95% confidence interval $[-1.25, +0.27]$ percent of that same gap, with 3 of the 12 seeds retaining a positive share. This is a bound from the interval and never a failed rejection, because a p value above a threshold establishes nothing.

The swing spans 51 percentage points across knobs nobody reports. Raising the variance coefficient recovers the signal completely, and the setting at variance 2500 and covariance 1 reaches 94.62 percent with an interval lying entirely above the frozen model's. Across the ten settings measured on the accuracy scale the field's own accuracy runs from 46.57 percent to 97.68 percent, from below chance to better than the frozen encoder, with no physics content anywhere in the change. Those two ends sit at different invariance weights, and holding that weight at its published value moves the accuracy from 46.57 to 94.62 percent on the variance coefficient alone. Across 17 readouts at the published invariance weight the retained gap spans 0.03 to 59.17 percent of the frozen model, 4 of the 17 invert the gap, and none exceeds the model. The risk named in advance was that accuracy would hold while the latent gap collapsed. Accuracy collapses with the gap, so the gap tracks what the field publishes.

Accuracy and share of the gap can point in opposite directions. The setting at variance 2500 and covariance 1 exceeds the frozen model on accuracy and on effect size while retaining only 53.37 percent of the scale-free gap. Both are true, because accuracy depends on the gap's mean relative to its spread rather than on the gap's size, so a readout can shrink the gap while shrinking its variance more.

Limits of the regulariser result. One claim this does not support is that the regulariser moves the split between magnitude and direction. The fair test shifted that split by 0.0114654 against a noise band of 0.0151749, inside noise, and we do not claim it. Section 3 gives the prior reason, that at this retention the shift could not have been informative whatever its size. The invariance weight is not monotone, so no monotonicity claim rests on it, and Appendix D gives the three readings that show it. No mechanism is established, since all 17 readouts trained at invariance weight 25, and nothing here reaches beyond one readout architecture.

6 ENGAGING THE NEAREST COMPETING CLAIMS

Three lines of work sit close to ours. Each asks a different question.

They varied readout size, we vary the loss. Kolesnikov et al. (2019) tested one readout size and found it made little difference. Zaiem et al. (2023) varies the size of the downstream readout and finds it reorders the ranking of evaluated speech models. Both train on labels. We hold the readout's size fixed to the byte and vary the loss it is trained to minimise, with no labels. We cite them as the published basis for the objection that readout size can create an artifact, which our matched-capacity design answers.

Randomising the whole network and randomising only the readout ask different questions. Garrido et al. (2025) randomise the entire network and obtain near-chance performance, establishing that the signal is learned. We randomise only the readout, on a trained and frozen model, and the untrained projection in Section 5.5 nearly preserves accuracy. Their question is whether training

produced the signal. Ours is whether a readout creates or destroys it. We do not re-run their control and we take from their work that the signal is learned, so we carry no independent confirmation of that.

Two different decompositions. Joseph et al. (2026) reparameterise motion into polar coordinates and probe the encoder for the parts. Their decomposition is of motion, recovered from the encoder with a trained probe. Ours is of the surprise difference itself, computed at the predictor with no probe at all, and they never compute a surprise difference, so they cannot have found its magnitude component negative. A trained probe can recover a quantity the readout never touches, so their result and ours can both hold while a violation-of-expectation number remains uninterpretable.

Only one released model can support a depth comparison. The depth result of Section 5.1 comes from one model, and a reader is entitled to ask whether that is a property of video models or a quirk of one checkpoint. Not every result here comes from one model, since the decomposition, the untrained floor and the global-versus-local split are measured on both encoders and Section 5.4 reports that they disagree. The depth comparison itself requires a predictor reading several encoder depths each with its own trained normalisation, and Section 4.3 gives the reason. V-JEPA 2.1 is the only released model we found that qualifies.

7 The requirement, and conclusion

No violation-of-expectation score is readable as physics evidence until six quantities sit beside it, each naming the aggregation it was computed under, and they fall into three kinds. One is a base, discriminability, and it takes two forms. With a readout in the path it is the class separation that readout delivers *per unit of its own output*, set against the frozen model's same ratio, which is what Section 5.5 reports. With no readout in the path there is no such output to divide by, so it is the model's own separability and nothing is divided, which is what Table 3 reports. The two are not interchangeable and a paper reporting one should say which. Three are the decomposition, asking whether the effect is one thing, everywhere, on the same positions. Two are alternative explanations, asking whether an untrained network or plain movement could produce the result without physics.

1	discriminability	base	whether the twins separate at all
2	untrained floor	alternative	an untrained network reaching the same number
3	magnitude vs direction	decomposition	a larger prediction, or one elsewhere
4	global vs local	decomposition	a whole-field offset, or a localised residue
5	selection term	decomposition	which positions the score sums over
6	motion-only baseline	alternative	plain movement explaining the result

They work by triangulation, which is why all six are needed. No single knob moves all six, and our own intervention demonstrates that rather than asserting it. The regulariser coefficient reaches four of them, readout discriminability, the magnitude and direction split, the selection term and the global and local split, because all four are computed on the readout's outputs. It structurally cannot reach the other two. No readout choice makes an untrained network learn physics, and none stops a motion-only observer from explaining the effect. The two quantities that do not pass through the readout stay load-bearing precisely while the four that do are being steered.

Reporting all four selection rules disqualified one of them. The score sums over the top q fraction of positions, that fraction has to be selected somewhere, and four rules are available. Table 4 reports every one. Under one of them the worst of four untrained initialisations buys more accuracy than the trained term does, so the separation residual is negative, and its direction channel alone sits at 0.4552, below chance. The other three keep a clear margin.

The sixth quantity is the one that most often goes unasked. A requirement that did not ask for it would let a motion statistic pass as physics evidence. Section 5.3 reports it.

Nothing here shows that a published number is wrong. What it shows is that the number does not carry enough information to be checked, and we report all six quantities for our own runs.

ETHICS STATEMENT

The applicable content here is narrow. This work involves no human subjects. The human visibility gate applied to each violation type is a perceptual check on rendered stimuli rather than a study. All stimuli are synthetic and generated by the authors, so no privacy, licensing or dataset-release question arises.

The one topic that genuinely applies is research integrity. We criticise a measurement practice, never any author or group, and we land our first finding on our own pre-registered configuration, which classifies at chance. We have reported that result in the same terms we would report it of anyone else.

REPRODUCIBILITY STATEMENT

Section 4 specifies the corpus, the three readout conditions and the depth axis. Appendix F reports the positive control on the measurement pipeline itself, which is the run that establishes the instrument detects a planted effect before any null we report is read. Appendix C reports all six required quantities for our own runs. Appendix D reports every readout setting measured rather than the subset shown in Section 5. Appendix H gives seeds, hyperparameters, hardware and run times, and Appendix G gives the generator detail for the clip pairs including the visibility gate.

An anonymised code archive accompanies this submission at `https://anonymous.4open.science/r/voe-score-decomposition-401C/`.

THE USE OF LARGE LANGUAGE MODELS

In this work, we used generative AI tools to generate synthetic data sets, to help develop theoretical models and conceptual frameworks, to formulate mathematical claims and provide critical ingredients for proving them, to assist in the writing of the formal section's derivations, to refine hypotheses, to design and give feedback on research methodology and experiments, to implement methods, to clean and reformat datasets, to support qualitative and thematic data analysis, and to interpret results. We did not use them to propose hypotheses or to assist with translation. We have reviewed all AI-assisted work.

Idea selection and every go or no-go decision was made by the authors, and the argument's overall structure came from an author. Substantially all analysis and experiment code was written with AI assistance, and every interpretation was reviewed by the authors. We take responsibility for the final content of this work, including the text, the claims and the artifacts produced with the aid of generative AI.

This project carries verification apparatus that most papers do not, and it is the direct answer to how the AI-assisted work was checked. Every reported figure is tied to its producing artifact by a number ledger. Load-bearing sentences are recorded as executable claims that are re-run rather than re-read. Minimum effect sizes were registered before the runs that test them, and the measurement pipeline itself carries positive controls, one of which is reported in Appendix F.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments, 2025.

Parsa Esmati, Somjit Nath, Katja Hofmann, Derek Nowrouzezahrai, Samira Ebrahimi Kahou, and Majid Mirmehdi. The invisible hand of physics: When video diffusion models know more than they show, 2026.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos, 2025.

Christian Internò, Alexander Pondaven, Habon Issa, Fabio Pizzati, Francesco Pinto, Markus Olhofer, Ivan Laptev, Philip Torr, Eero P. Simoncelli, Barbara Hammer, and David Klindt. GEOPHYS: The geometry of physical plausibility, 2026.

Sonia Joseph, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake Richards, and Mike Rabbat. Interpreting physics in video world models, 2026.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1920–1929, 2019. doi: 10.1109/cvpr.2019.00202.

Markus Marks, Manuel Knott, Neehar Kondapaneni, Elijah Cole, Thijs Defraeye, Fernando Perez-Cruz, and Pietro Perona. A closer look at benchmarking self-supervised pre-training with image classification. *International Journal of Computer Vision*, 2024.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles?, 2025.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning, 2026.

Shuyao Shang, Yuqi Wang, Ruopeng Gao, Xu Chen, Tieniu Tan, Lue Fan, and Zhaoxiang Zhang. PhiZero: A world model built around physical language, 2026.

Luca Weihs, Amanda Rose Yuile, R. Baillargeon, Cynthia Fisher, G. Marcus, Roozbeh Mottaghi, and Aniruddha Kembhavi. Benchmarking progress to infant-level physical reasoning in AI. *Transactions on Machine Learning Research*, 2022, 2022.

Jianhao Yuan, Fabio Pizzati, Francesco Pinto, Lars Kunze, Ivan Laptev, Paul Newman, Philip Torr, and Daniele De Martini. LikePhys: Evaluating intuitive physics understanding in video diffusion models via likelihood preference, 2025.

Salah Zaiem, Youcef Kemiche, Titouan Parcollet, Slim Essid, and Mirco Ravanelli. Speech self-supervised representations benchmarking: a case for larger probing heads. *Computer Speech and Language*, 2023.

A THE CORPUS AND THE SURVEYED POOL

Table 1: The corpus, by violation type, sorted by generated count. Valid pairs clear the perceptual contrast bar.

violation type	generated	valid	not valid
shape swap	965	952	13
collision	804	804	0
size change	512	512	0
solidity	471	471	0
interpenetration	303	274	29
two place	155	154	1
continuity	153	153	0
occluder gap skip	57	51	6
immutability	56	56	0
total	**3476**	**3427**	**49**

Table 2: The nine-paper pool. *Surprise* is whether the paper reports a prediction-error surprise score. *Readout* is whether it reports the accuracy of a readout trained on labels. *Same model* applies only where both are present, and asks whether the two readings are taken on one model. The last column names what that paper's conclusion rests on. Each row is taken from the full text. Esmati et al. (2026) reports a quantity it calls surprise, and it is a probe delta taken when a block is perturbed rather than a prediction error, because a diffusion model carries no predictive future representation to take one from.

paper	surprise	readout	same model	what carries the claim
Garrido et al. (2025)	yes	no	—	the scalar, alone
Bordes et al. (2025)	yes	no	—	the scalar, by choice
Motamed et al. (2025)	no	no	—	continuation fidelity
Shang et al. (2026)	no	no	—	generation benchmarks
Yuan et al. (2025)	no	no	—	a likelihood preference
Assran et al. (2025)	no	yes	—	readouts, for capability
Joseph et al. (2026)	no	yes	—	trained readouts
Esmati et al. (2026)	no	yes	—	a probe delta, not a prediction error
Internò et al. (2026)	yes	yes	**no**	trajectory geometry, two backbones

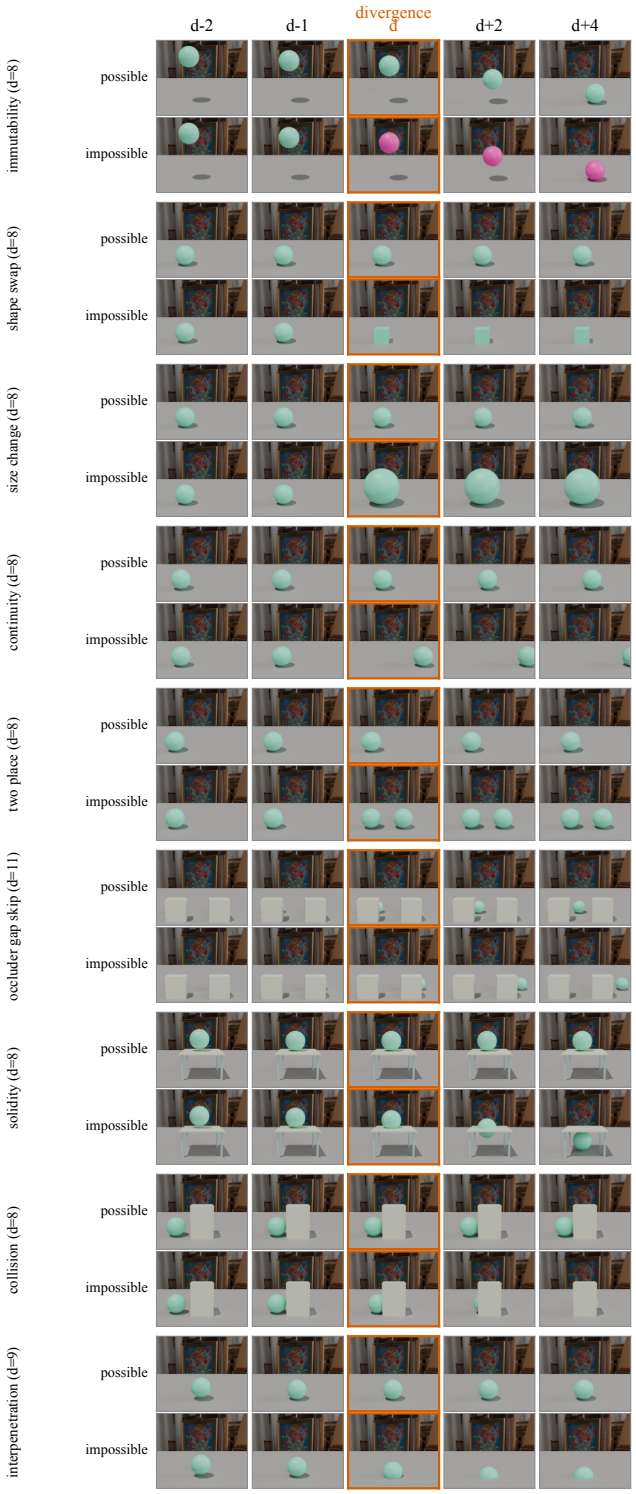


Figure 1: One paired twin per violation type. The upper strip of each row is the possible clip and the lower strip is its impossible twin. The marked column is the first frame at which the two clips differ *visibly*. The generator measures that as the first frame whose mean absolute difference clears its render-noise floor, and the pinning control in Section 4.4 is stated on the same frame. For eight of the nine types it is also the first frame that differs at all, and the pair is byte-identical before it. `occluder gap skip` is the one exception, because its clips begin to differ two frames earlier behind an occluder. Its two columns before the mark therefore carry a difference too small to see.

B PER-TYPE FIGURES

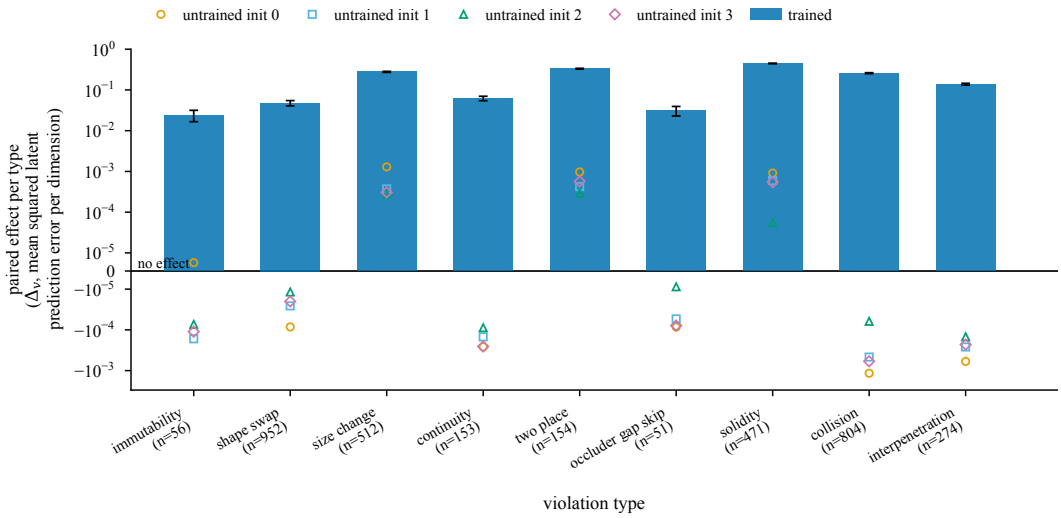


Figure 2: Per-type effect for the trained model against four untrained initialisations, at layer 29 under per-arm aggregation, in units of mean squared latent prediction error per feature dimension. The marked line is the no-effect line, not a chance line, because that quantity is a prediction error rather than an accuracy. The trained readout signs all nine types. The untrained readouts sign three or four, which is what chance gives on nine sign tests. Per type the closest floor sits between two and three orders below its trained value, from 2.16 to 2.75 decades with a median of 2.38 over the nine types. The wider separation quoted elsewhere belongs to the pooled figures and not to any single type.

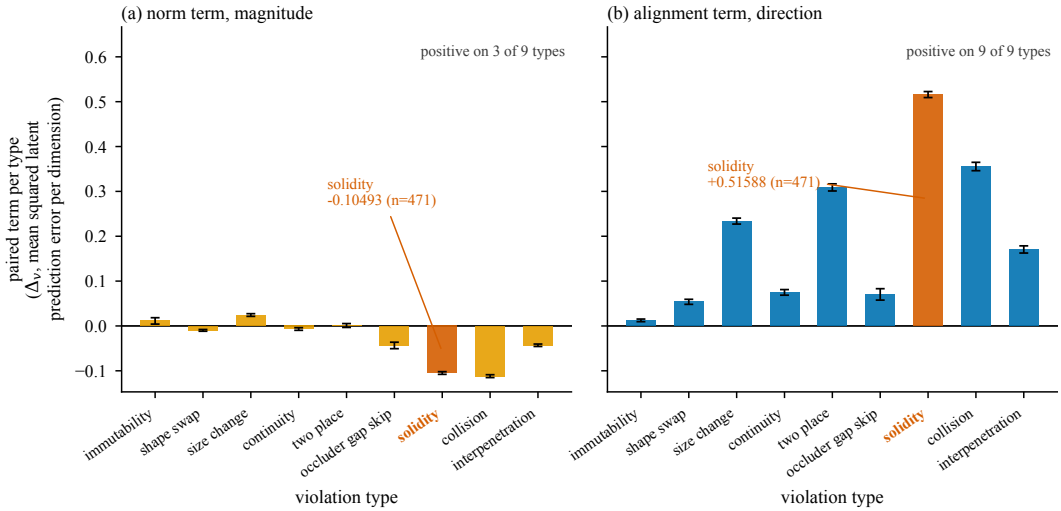


Figure 3: The magnitude term (left) and the direction term (right) per violation type, sharing a y axis, with a zero line. The left panel straddles zero and the right panel does not. Error bars are 95% normal intervals on each type's paired mean, $1.96\,\mathrm{sd}/\sqrt{n}$, and are not bootstrap intervals.

C ALL SIX REQUIRED QUANTITIES, FOR OUR OWN RUNS

A paper demanding six quantities beside every score has to report them for its own measurements, or the requirement reads as a demand rather than a standard. Table 3 does that. Every row is computed under one aggregation, per-arm, which the caption names once rather than in a column that would carry the same word six times.

Table 3: The six required quantities with their values, units and aggregations, on both encoders, over 3427 pairs. V-JEPA 2 is read at its final layer and V-JEPA 2.1 at layer 29, both per-arm, which are the only cells at which each is interpretable. Values are rounded to the precision the cohort supports, 10^{-4} for a paired statistic over 3427 pairs; full precision sits in the released artifacts. Rows 3 and 5 are trained minus the worst of four untrained initialisations, a separation residual rather than a trained value. The worst floor is the one most favourable to the null, which is the largest floor for a positive term and the smallest for a negative one. The magnitude and selection floors sit at 10^{-6} or below and move nothing at this precision. Both direction floors do move the fourth decimal, at 2.8×10^{-5} on V-JEPA 2.1 and 1.3×10^{-4} on V-JEPA 2, so the subtraction matters on that term for both encoders. The AUROC rows report the paired sign fraction, which is the same statistic as the paired accuracy we report elsewhere. These runs carry no readout, so row 1 is the model's own separability rather than the per-unit-of-output form Section 7 defines for a readout. Row 6 gives the motion-only baseline first and then the increment the trained direction channel adds to it, and Section 5.3 carries both increments with their intervals.

quantity	units	V-JEPA 2, final	V-JEPA 2.1, layer 29
discriminability	AUROC	0.5708, k 5/9	0.8955, k 9/9
untrained floor	AUROC resid.	$+0.0926$	$+0.4159$
magnitude vs direction	pred. error resid.	-0.0556 / $+0.1342$	-0.0439 / $+0.2362$
global vs local	AUROC	0.4946 $[0.4780, 0.5112]$	0.8637 $[0.8523, 0.8751]$
selection term	pred. error resid.	$+0.0032$, k 7/9	$+0.0140$, k 8/9
motion-only baseline	paired acc.	0.7660, then $+0.1202$	0.7660, then $+0.1719$
depth comparison	—	available on V-JEPA 2.1 only, see Section 4.3	

Three rows are AUROC or paired accuracy, bounded on $[0, 1]$, and three are prediction error, which is unbounded. Those two families are not comparable to each other, and the global-versus-local row reports an interval that contains chance on V-JEPA 2 and does not on V-JEPA 2.1, which is the disagreement Section 5.4 turns on.

Table 4: The four selection rules, on V-JEPA 2.1 at layer 29 over 3427 pairs. The score sums over the top q fraction of positions, and these are the four ways that fraction can be chosen. The per-arm rule chooses separately in each arm, and the other three share one support between the arms. Values are paired accuracy. The worst floor is the largest of four untrained initialisations, and the residual is the trained value minus it, taken at full precision rather than from the rounded columns beside it. Under `cpos` the residual is negative and the direction channel alone falls below chance, which is the variant the requirement disqualifies. The decomposition's selection term exists only under the per-arm rule, because a shared support makes the prediction cancel exactly.

selection rule	trained	worst floor	residual	direction alone
`perarm`	$+0.1719$	$+0.0088$	$+0.1631$	0.9087
`common`	$+0.1590$	$+0.0082$	$+0.1509$	0.8982
`pooled`	$+0.1587$	$+0.0044$	$+0.1544$	0.8658
`cpos`	$+0.0400$	$+0.1917$	-0.1517	0.4552

The depth comparison is structurally unavailable on V-JEPA 2 rather than omitted. One reading place is one point, and Section 4.3 gives the normalisation reason.

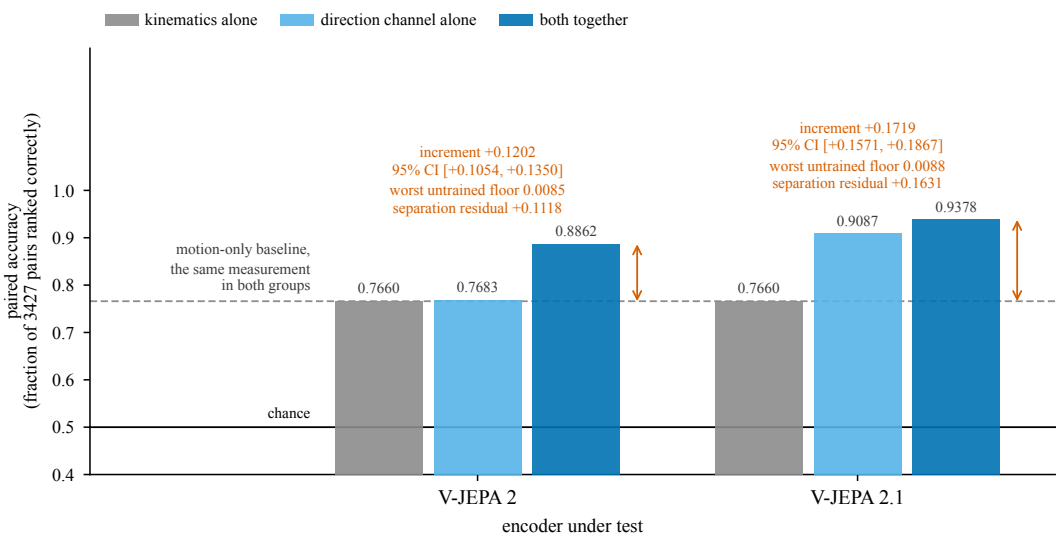


Figure 4: Motion alone, the trained direction channel alone, and the two together, for each encoder. Values are paired accuracy over 3427 pairs, with each increment annotated beside its interval, its worst untrained floor and its separation residual. The motion-only bar is the same measurement in both groups. That equality is guaranteed by construction rather than observed, since the baseline is computed from the pairs alone and the direction channel enters only when the two are fitted together. On V-JEPA 2 the first two bars are nearly equal, which is the measurement and not a rendering fault.

D EVERY READOUT SETTING MEASURED

Section 5.5 shows a handful of readout settings chosen to make the argument legible. The natural objection is that the settings were chosen for the answer they give. Ten trained conditions carry per-seed accuracy, and Table 5 reports the ones the main text names. The remaining settings were measured only as the dimensionless gap, never on the accuracy scale, so Section 5.5 reports them by their span and by the four that invert the gap rather than as accuracies they do not have. Table 6 lists every one of the twenty-three, so the selection the main text made is visible rather than described. The invariance weight is not monotone. At weight 25 no readout of the 17 passes the model on the gap, at weight 5 one reaches 112.83 percent, and weight 1 sits below weight 25 on both settings. The two discarded as collapsed are named next.

Two conditions may not be read as measurements and are marked as such. They are variance 1 with covariance 25, at a variance-to-covariance ratio of 0.04, and variance 0.01 with covariance 1000, at 1×10^{-5}. In both the head collapsed, so its outputs carry no signal to retain and the retention figure is not a measurement of anything. Quoting them would inflate the reported spread by about two orders of magnitude. What is withdrawn is the retention figure and not the accuracy, which is why variance 1 with covariance 25 still carries an accuracy, a seed count and an interval in Table 5.

One quantity we reported in earlier drafts is withdrawn here and the reason is recorded. Retention expressed as a percentage of the frozen model's raw gap is not well defined, because it divides a difference measured in the head's 256-dimensional output space by one measured in the frozen model's 1408-dimensional space, and the valid heads span a factor of 253 in output scale. The tell was retention above 100 percent, which cannot mean a head kept more than all of the signal. Every retention figure we report divides the paired difference by the mean of the two arms' output mean-squares, which is dimensionless and free of both output scale and feature count.

E THE FOUR-DEPTH MEASUREMENT IN FULL

Table 7 above gives accuracy, effect size, the three term shares, the raw gap and the scale-free gap at each of the four trained reading depths. Depth 29 carries both the best accuracy and the *smallest* raw gap of the four, 0.20630975 against 0.21528177, 0.26043160 and 0.34728547.

Table 5: Paired physics classification accuracy through readouts of identical width, over the same 3427-pair cohort, with the readout reading V-JEPA 2.1 at layer 29 per-arm throughout. The first row has no readout in the path and the second is untrained, so neither carries seeds or an interval. The *inv.* column is the VICReg invariance weight; 25 is the value published with that objective and is the one the main text means by the published weight. Intervals are 95% over the stated seed count, from a one-sample t on the per-seed accuracies. That interval is normal-theory and is not bounded to the accuracy scale, which is why the weight-5 row's upper bound reads 100.12; the bound is not attainable and we print it rather than silently clipping it. The 97.68 that the main text quotes as the top of the measured range is that row. One condition was measured three times at the published weight. The variance 2500, covariance 1 row is the twelve-seed run, and two earlier four-seed runs of the same condition read 93.75 and 93.54. We print the twelve-seed run because it is the best powered of the three, and disclose the others rather than pooling them, because pooling runs of one condition changes a reported number on no new evidence. The registered variance 25, covariance 1 setting was re-trained and re-measured at twelve fresh seeds in that same run, reading 48.67 on a 95% confidence interval of $[41.16, 56.18]$, so the collapse replicates while its point estimate moves. That setting is unstable across seeds, spanning 33.50 to 70.06 at a standard deviation of 11.82, against the recovering setting's 2.04. Both settings of this paper's own registered configuration classify at chance and one sits below it. The accuracies in this table are per-run and are not pooled across runs of one condition. The dimensionless-gap shares quoted in Section 5.5 are means over every measurement of a condition, so the two scales follow different pooling policies.

readout	inv.	seeds	accuracy	95% CI	effect size d
none, the frozen model	—	one pass	**89.55**	—	$+1.2293$
untrained, random initialisation	—	—	**88.39**	—	$+1.2304$
variance 2500, covariance 1	25	12	**94.62**	[93.32, 95.92]	$+1.3123$
variance 50, covariance 50	25	4	81.24	[78.12, 84.37]	$+0.8497$
variance 1, covariance 25, registered	25	12	**51.92**	[46.87, 56.98]	$+0.0547$
variance 25, covariance 1, registered	25	12	**46.57**	[39.45, 53.70]	-0.1050
variance 2500, covariance 1	5	4	**97.68**	[95.25, 100.12]	$+1.2834$
variance 50, covariance 50	5	4	95.12	[93.86, 96.37]	$+1.2644$
variance 2500, covariance 1	1	4	94.35	[91.27, 97.42]	$+1.2649$
variance 50, covariance 50	1	4	83.07	[77.63, 88.50]	$+0.9236$
variance 2500, covariance 1	0	4	90.69	[84.31, 97.07]	$+1.1152$
variance 50, covariance 50	0	4	75.24	[70.25, 80.23]	$+0.6737$

One claim stood here and is withdrawn. We promised a per-pair spread column and per-type sign counts behind each row, and read the peak as driven by consistency rather than by magnitude. We print neither column anywhere and no measurement we report supports that reading, so the claim rested on data a reader could not check. Section 5.1 states that no mechanism for the peak is offered, and we do not supply one.

Four depths of one model is not a curve. The table is reported as four points and no functional form is fitted to it.

F THE POSITIVE CONTROL ON THE MEASUREMENT PIPELINE

With no readout in the path, over all 3427 valid pairs, the decomposition reproduces the independently certified frozen gap of 0.20630975155008058 as 0.20630974545008957, through a different code path and a different job. The absolute difference is 6.100×10^{-9}, so the two agree through their seventh significant figure and diverge at the eighth.

A planted effect at the pre-registered minimum size of 0.009358 is recovered with the correct sign at empirical power 1.0000 against a target of 0.8, on the pooled paired statistic. That control is pooled, so it carries no per-type sign count. The twins share a context, so the predictor emits the same tensor for both arms, and the maximum absolute *difference* between the two arms' predictions measures exactly zero both before and after planting. That establishes that the plant never reached the conditioning context, so the selection term remains a pure selection term under the plant.

Table 6: Every readout setting measured, as a share of the frozen model's dimensionless gap. *inv*, *var* and *cov* are VICReg's invariance, variance and covariance weights, and 25, 25, 1 is the published triple. Seventeen sit at the published invariance weight and six at the other three. Four of the seventeen carry a negative share, meaning the readout inverts the physics rather than merely losing it. One setting in the table exceeds the frozen model, at invariance weight 5. The two collapsed conditions are excluded and named below, and Table 5 gives the ten that also carry per-seed accuracy.

inv	var	cov	share	inv	var	cov	share
25	1	1	−0.60	25	500	1	13.92
25	5	1	0.03	25	1000	1	26.92
25	1000	0.01	0.04	25	50	50	31.25
25	25	1	−0.49	25	2500	1	53.37
25	50	1	−2.22	25	100	25	38.06
25	100	1	−0.83	25	25	100	3.37
25	10	10	0.29	25	10000	1	34.25
25	5	25	0.04	25	2500	10	59.17
25	250	1	6.25	0	2500	1	18.49
0	50	50	13.57	1	2500	1	40.87
1	50	50	20.77	5	2500	1	70.81
5	50	50	112.83				

Table 7: One frozen model, identical clips, four trained reading depths, no trained head in the path. Accuracy is paired, chance is 50. The raw gap is in units of mean squared latent prediction error. The three term shares sum to exactly 1.0000 at every depth. The scale-free gap column divides the raw gap by the mean of the two arms' output mean-squares. That is the denominator every retention figure we report uses, so one convention governs the gap throughout, and at depth 29 this column equals the frozen reference those retention figures divide by.

depth	**accuracy**	effect size	ALIGN	NORM	P	raw gap	scale-free gap
9	**79.3697**	0.93	1.0064	−0.1631	+0.1567	0.26043160	0.0800878541
19	**87.2775**	1.17	1.0463	−0.1398	+0.0935	0.21528177	0.0813604292
29	**89.5535**	1.23	1.1453	−0.2130	+0.0677	0.20630975	0.0822424598
39	**87.4526**	1.20	1.0788	−0.1317	+0.0529	**0.34728547**	0.0831928455

This is the run that licenses reading a null on the pooled statistic as a fact about a readout rather than a fact about the instrument. No arm certified the sign count at the pre-registered minimum size, so a null on a sign count is reported with that limit attached rather than under this licence.

G THE GENERATOR, AND THE GATES EACH CLIP PASSES

Clips are generated in pairs that are byte-identical until a divergence frame pinned at frame 9 or later of 16. Frames are stored losslessly, because lossy coding leaks a diverging suffix backward into the encoded prefix and would destroy the property the whole design rests on. The shared prefix was verified byte-identical rather than assumed.

Each violation type passes a human visibility gate before it enters the corpus, admitting a type only where people identify the violation at near ceiling. The shape marginal is equalised across violation types by construction, and it was measured on the 1152-pair generator cohort, where every type carries 128 pairs and each type's marginal reads cube 20, cylinder 31, mesh 34 and sphere 43. That is the cohort on which the appearance control was verified. The shipped 3427-pair corpus has per-type counts that are unequal by design, so its marginal cannot take those values, and we have not measured it. A perceptual contrast floor is applied per pair, and Table 1 reports how many pairs of each type cleared it.

H RUN DETAIL

We report what the run records carry. Where a quantity is not in a run record we say so, rather than quoting a default from the source tree, because a default is what the code would do and not what the run did.

Readout training. Batch size 16, two epochs, 24000 clips, an input width of 1408 into an output width of 256, on one NVIDIA B200 under CUDA. The twelve-seed rows use seeds 0 to 11 and the four-seed rows use seeds 0 to 3. Wall clock was 3461.2 and 3354.5 seconds for the two 24-head runs, and 1739.6 and 1298.9 seconds for the two 8-head runs. The optimiser, its learning rate, its weight decay and the gradient clip are not recorded in any run manifest. We do not quote the source tree's defaults for them, because the epoch count recorded by the runs already differs from the default sitting in that file. The objective reduces each batch by treating every position of every clip as one sample. That gives the covariance term enough samples to estimate a 256×256 matrix, and it treats positions within a clip as independent. Both settings use the same reduction, so it cannot move the comparison between them.

Measurement. Cohort 3427 requested and 3427 measured, read at layer 29 under per-arm aggregation, with a global selection fraction of 0.0678 held frozen across both arms. The maximum absolute difference between the two arms' predictions measured exactly zero. Wall clock ran from 2929.2 to 2992.8 seconds for the 8-head measurements and from 3822.3 to 3876.1 seconds for the 24-head ones, on one NVIDIA B200.

Matched capacity. All readouts share one seed and one parameter count across settings, so capacity is held fixed while the objective's coefficients vary. The run manifest records this rather than asserting it, since at every seed the two settings carry a byte-identical initialisation fingerprint and the 24-head run manifest's matched-capacity defect list is empty. The untrained control runs at zero epochs, which leaves every weight at initialisation. Two coefficient settings returned byte-identical numbers under that control, which is the expected result rather than a defect, because with no training the coefficients cannot act and a shared seed makes the two networks the same network. The control fails if they differ.

The motion-only baseline. The sixth required quantity is measured by an observer that touches neither model under test. Frames are read by a frozen `facebook/dinov2-base` image encoder, image-only and with no video pretraining, which is what makes it comparable to training-free tracking baselines rather than a second learned model. Its patch features form one grid per frame, at a patch stride of 32 pixels and 12 frames per second, and a single object is tracked across that grid by cosine similarity between patch features, under a smoothness penalty of 0.05 and with no confidence threshold for declaring a track lost. The tracker does not re-anchor when an object changes appearance, and that is deliberate, because several violation types change appearance at the divergence frame and a re-anchoring tracker would absorb the violation instead of reporting it. From each trajectory we compute 29 kinematic features over two temporal scopes, the whole clip and the segment after the divergence frame, giving 58 in total. They cover speed and its variation, acceleration, turning angle, autoregressive residual, smoothness, contact, path geometry, pairwise interaction and track quality. The readout is a ridge logistic regression whose penalty is chosen by cross-validation over five grouped folds stratified by violation type, at seed 0, on the same stimulus set as every other measurement here, with its frame hashes verified. The penalty grid itself is not in the run record.