Human-Level Accuracy, Non-Human Strategies: Revealing Model–Human Divergence in Video Physical Reasoning

Fanhong Li[1], Shurui Zheng[1], Zi Yin[1], Junbo Cui[2], Lei Ji[3], and Jia Liu[1⋆]

[1] Tsinghua University, Beijing, China
{lifh21,zhengsr23,z-yin21}@mails.tsinghua.edu.cn,
liujiaTHU@tsinghua.edu.cn
[2] ModelBest Inc., Beijing, China
cuijb2000@gmail.com
[3] Microsoft, Beijing, China
leiji@microsoft.com

Abstract. Video foundation models now reach human-level accuracy on physical-reasoning benchmarks, yet such tasks require predicting unobserved physical outcomes. Do these models perform human-like forward simulation, or do they exploit statistical regularities in visible scenes? Accuracy alone cannot distinguish these strategies. We introduce a distributional evaluation framework that treats model seeds and human raters as populations, enabling comparison of consensus, uncertainty, and strategy. On the Physion benchmark, we evaluate three ViT-L architectures (V-JEPA2, VideoMAEv2, DINOv2). V-JEPA2 narrows the accuracy gap to ∼1 percentage point (73.2% vs. 74.2%), yet model–human disagreement reaches 26.4%, far exceeding human–human disagreement (4.8%), with substantially lower agreement ($\kappa \approx 0.48$ vs. 0.91). The divergence is scenario-structured: models outperform humans on geometric reasoning (linking, +11.8 pp), while human advantages appear on causal-chain and other future-dynamics scenarios in the released split (e.g., dominoes, −10.5 pp). Strategy fingerprinting confirms all three architectures share non-human strategies while none aligns with humans. Attribution analysis suggests that unobservable outcome features, rather than visible scene properties, predict this divergence, consistent with models relying more on scene-level statistical regularities than on explicit forward simulation, a systematic divergence that accuracy alone cannot reveal. Code is available at https://github.com/fanhong-li/model-human-divergence.

Keywords: Physical Reasoning · Human–AI Comparison · Video Understanding · Distributional Evaluation

1 Introduction

Physical reasoning, *i.e.* predicting how objects move, collide, and interact, is a core cognitive ability that develops early in infancy [6, 18, 23]. A distinguishing

⋆ Corresponding author.

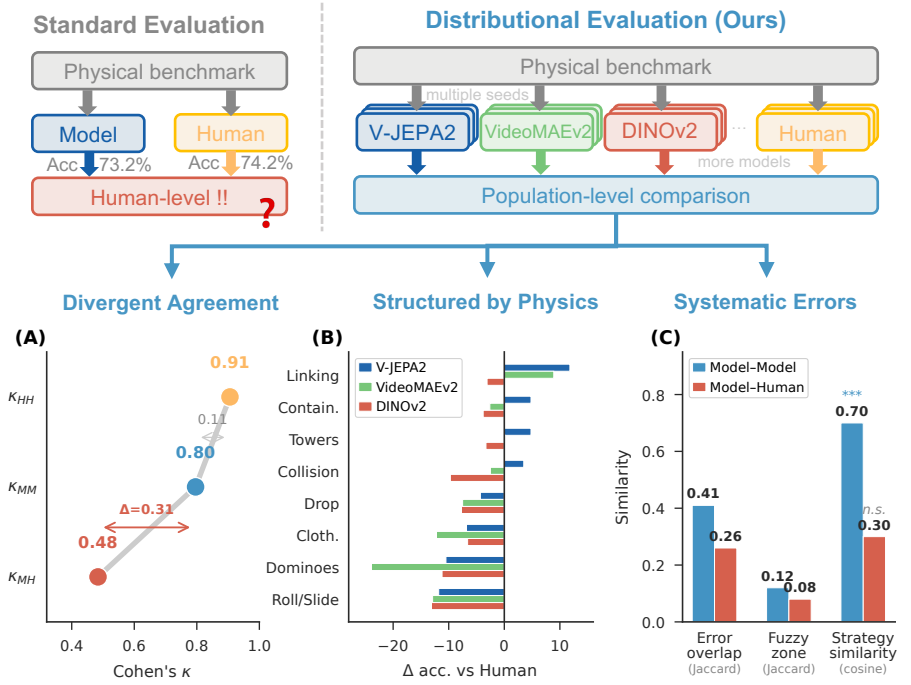

Fig. 1: Overview. *Top:* Standard evaluation compares a single model's aggregate accuracy against human mean accuracy, masking systematic differences. Our distributional evaluation compares model populations (3 architectures × 8 seeds) against the full human rater population (∼100 raters × 1200 trials). *Bottom:* Three key findings. (A) Agreement hierarchy: model–model agreement ($\kappa_{MM} = 0.80$) far exceeds model–human agreement ($\kappa_{MH} = 0.48$), which in turn falls well below human–human agreement ($\kappa_{HH} = 0.91$). (B) Released-split scenario-specific accuracy gaps include a model advantage on linking (+11.8 pp) and human advantages on rolling/sliding (−11.8 pp) and dominoes (−10.5 pp). (C) Models share systematic errors with each other (error overlap Jaccard = 0.41, strategy cosine = 0.70) but not with humans (Jaccard = 0.26, cosine = 0.30). All Jaccard values are averages across model pairs.

feature of physical reasoning is that it requires predicting unobserved outcomes: given partial observation of a scene, humans mentally simulate forward in time to anticipate what will happen next [6]. Recent self-supervised video models have made striking progress on physical reasoning benchmarks. V-JEPA [5] reaches 98% zero-shot accuracy on violation-of-expectation tasks [11], and V-JEPA2 [2] matches human-level performance on contact prediction [7]. But do these results reflect human-like forward simulation, or a qualitatively different strategy that extracts statistical regularities from visible scenes without simulating unobserved dynamics?

A single accuracy number cannot distinguish between these possibilities. Two systems can achieve identical scores while making errors on entirely different

samples [12,13], implying fundamentally different underlying strategies. When a physical-reasoning benchmark contains a mix of scenarios, some solvable from geometric cues alone, others requiring forward simulation of gravitational dynamics or causal chains, aggregate accuracy conflates performance on both types, masking systematic differences in *which* scenarios each system solves. Distinguishing these possibilities requires comparing not just how often models and humans are correct, but *on which samples* they agree, how confidently they respond, and whether their errors cluster in the same places.

Our Approach. We address this with a distributional evaluation framework that treats model seeds and human raters as *populations* rather than point estimates. For models, multiple random probe initializations on the same frozen encoder produce a population of predictions per sample; for humans, $\sim$100 independent raters per stimulus provide the reference population. This enables three levels of analysis: consensus comparison via inter-rater agreement metrics, uncertainty decomposition via entropy analysis, and strategy fingerprinting via physical-feature attribution. Critically, by evaluating three ViT-L architectures (V-JEPA2 [2], VideoMAEv2 [25], DINOv2 [20]) under a unified protocol, identical backbone size, frame sampling, and probe architecture, we can test whether divergence from human behavior is architecture-specific or shared across the evaluated frozen representations.

Key Findings. Applying this framework to the Physion benchmark [7] reveals converging evidence at three levels. *Existence:* V-JEPA2 achieves 73.2%$\pm$0.95%, within 1 pp of human accuracy (74.2%), yet model–human disagreement (26.4%) is 5.5$\times$ higher than human–human disagreement (4.8%; $\kappa = 0.48$ vs. 0.91). *Structure:* divergence is scenario-structured: models outperform humans on geometric scenarios (linking +11.8 pp), while human advantages appear on causal-chain and other future-dynamics scenarios in the released split (dominoes -10.5 pp; rolling/sliding -11.8 pp), consistently across all three architectures. *Source:* in our attribution analysis, observable scene properties carry little accessible signal for where models and humans diverge, whereas unobservable outcome features do; this pattern is consistent with divergence tied to forward simulation rather than scene encoding alone.

Contributions.

- A *distributional evaluation framework* that treats model seeds and human raters as populations, enabling consensus, uncertainty, and strategy comparison beyond aggregate accuracy (§3).
- A three-level characterization of model–human divergence on physical reasoning, from existence (systematic disagreement despite matched accuracy) through structure (released-split scenario differences) to source (unobservable outcome features predict divergence better than observable features in the attribution analysis), established across three ViT-L architectures under fair comparison (§5, §6).

– Evidence that, for the evaluated frozen ViT-L representations and supervised probes, divergence is more strongly associated with unobserved physical outcomes than with visible-scene features, a limitation that accuracy-based evaluation cannot detect.

2 Related Work

Physical Reasoning Benchmarks. Physion [7] introduced eight physical scenarios with per-stimulus human judgments (∼100 raters), establishing frozen-encoder probing for binary contact prediction. Physion++ [24] extends this to latent physical properties. Other benchmarks include IntPhys [22], CLEVRER [26], and CoPhy [4]. Garrido *et al.* [11] showed that V-JEPA acquires intuitive physics through self-supervised pretraining but revealed weaknesses on inter-object interactions. These benchmarks provide rich human response data, yet prior work uses only aggregate accuracy for model–human comparison, without examining whether model performance reflects the same reasoning mechanisms as humans.

Simulation vs. Pattern Matching in Physical Reasoning. Cognitive science distinguishes two accounts of physical reasoning. The *simulation* account [6] proposes that humans predict physical outcomes by running approximate, probabilistic simulations of scene dynamics, an "intuitive physics engine" analogous to game-engine physics. This view is supported by systematic biases in human judgments that match simulation noise [18], and has been formalized as a core component of human-like reasoning [19]. The alternative, *pattern matching*, holds that many physical judgments can be made from static or early-dynamic cues without forward simulation. In deep learning, evidence is mixed: Piloto *et al.* [21] showed that a graph network trained on physical dynamics develops violation-of-expectation responses, while Garrido *et al.* [11] found that self-supervised video models acquire some intuitive physics but fail on inter-object interaction, a capacity closely tied to forward simulation. These are scattered observations; no prior work has systematically tested, in a controlled setting, whether the locus of model–human divergence lies in processing visible information or in reasoning about unobserved futures.

Beyond-Accuracy and Distributional Evaluation. Geirhos *et al.* [13, 14] demonstrated that vision models with similar accuracy exhibit fundamentally different error patterns, advocating for error consistency analysis. This "beyond accuracy" perspective has been applied to image classification [14] and robustness evaluation [15], but not to physical reasoning. Complementarily, psychometric traditions, Item Response Theory [10] and inter-rater reliability (Fleiss' κ, Krippendorff's α), provide tools for comparing agent populations. Recent work applies these ideas to LLM evaluation [9] and human-AI complementarity [3]. We unify both lines: error consistency analysis to quantify *what* diverges and population-level agreement metrics to quantify *how much*, applied for the first time to physical reasoning where the simulation-vs-cue-based-reasoning question gives these tools a specific theoretical target.

Probing Frozen Representations. Linear probing [1, 8, 16] is standard for representation evaluation. Concurrent with our work, Joseph *et al.* [17] identify a "Physics Emergence Zone" at intermediate layers of V-JEPA2, a pattern we independently confirm. While their work focuses on *what* physical information is represented, we address *whether* model and human populations use it in the same way.

3 Distributional Evaluation Framework

Standard evaluation compares a single model prediction against ground truth, yielding accuracy. Our approach treats both model instantiations and human raters as *populations*, enabling richer comparisons at the consensus, uncertainty, and strategy levels.

3.1 Population Representation

Let $N = 1200$ denote test samples, $K = 8$ model seeds (random probe initializations on a frozen encoder), and $H_i \approx 100$ human raters for sample i. We define:

- $\mathbf{M} \in \{0,1\}^{N \times K}$: model predictions, where $M_{ik} = 1$ if seed k predicts contact for sample i.
- $\mathbf{R} \in \{0,1\}^{N \times H_{\max}}$: human responses ($H_i \approx 100$ varies by sample; unused entries are masked).
- $y \in \{0,1\}^N$: ground-truth labels.

From these, we derive population-level consensus:

$$p_m^{(i)} = \frac{1}{K} \sum_{k=1}^{K} M_{ik}, \qquad p_h^{(i)} = \frac{1}{H_i} \sum_{j=1}^{H_i} R_{ij}, \tag{1}$$

representing the fraction of each population predicting contact. The majority-vote predictions are $\hat{y}_m^{(i)} = \mathbb{1}\left[p_m^{(i)} > 0.5\right]$ and $\hat{y}_h^{(i)} = \mathbb{1}\left[p_h^{(i)} > 0.5\right]$.

3.2 Three-Level Agreement Hierarchy

We decompose agreement into three levels, each computed via split-half bootstrap (1,000 iterations) to produce confidence intervals:

Level 1: Human–Human. Randomly split H raters into two halves; each half produces a majority vote. Cohen's κ between the two halves, averaged over bootstrap iterations, yields κ_{HH}.

Level 2: Model–Model. Randomly split K seeds into two halves; each half produces a majority vote. Cohen's κ between the two halves yields κ_{MM}.

Level 3: Model–Human. Cohen's κ between the model majority vote $\hat{y}_m$ and the human majority vote $\hat{y}_h$ yields κ_{MH}. To obtain a confidence interval comparable to the other two levels, we bootstrap κ_{MH} by jointly resampling seeds and raters (1,000 iterations).
The expected ordering $\kappa_{\mathrm{HH}} \gg \kappa_{\mathrm{MM}} \gg \kappa_{\mathrm{MH}}$ holds if models form a coherent population that differs systematically from the human population. Deviations from this ordering are diagnostic: $\kappa_{\mathrm{MM}} \approx \kappa_{\mathrm{HH}}$ would suggest model seeds are as reliable as human raters; $\kappa_{\mathrm{MM}} \approx \kappa_{\mathrm{MH}}$ would suggest model internal disagreement fully accounts for model–human divergence.

3.3 Uncertainty Decomposition

For each sample, we compute the binary entropy of each population's consensus:

$$\mathcal{H}_m^{(i)} = H(p_m^{(i)}), \qquad \mathcal{H}_h^{(i)} = H(p_h^{(i)}), \tag{2}$$

where $H(p) = -p\log_2 p - (1-p)\log_2(1-p)$. This induces a four-quadrant classification at threshold τ (we use $\tau = H(0.25) \approx 0.81$ bits, corresponding to 25%/75% consensus):

- **Q1** ($\mathcal{H}_m < \tau$, $\mathcal{H}_h < \tau$): Both populations certain, "easy" samples.
- **Q2** ($\mathcal{H}_m \geq \tau$, $\mathcal{H}_h < \tau$): Model uncertain, humans certain; model blind spots.
- **Q3** ($\mathcal{H}_m < \tau$, $\mathcal{H}_h \geq \tau$): Model certain, humans uncertain; confident divergence.
- **Q4** ($\mathcal{H}_m \geq \tau$, $\mathcal{H}_h \geq \tau$): Both uncertain, genuinely ambiguous.

The asymmetry ratio $A = |Q2|/|Q3|$ quantifies the directionality of information access: $A \gg 1$ indicates systematic model blind spots, while $A \ll 1$ indicates models are confidently committed where humans are uncertain.

3.4 Strategy Fingerprinting

To characterize *how* each population makes decisions, we extract physical features from the Physion HDF5 simulation data organized into three temporal layers: **Layer A** (static scene properties: mass, friction, object counts, geometry), **Layer B** (observable dynamics up to the cutoff frame: speeds, kinetic energies, collision counts, trajectory statistics), and **Layer C** (unobservable outcome features: post-cutoff contact timing, causal chain length, prediction horizon). We use this layered framework at two levels of granularity depending on the analysis goal.

For *strategy fingerprinting* (§5.3), we use a 37-feature core subset spanning all three layers ($6\mathrm{A} + 24\mathrm{B} + 7\mathrm{C}$), selected to capture representative physical quantities at coarse granularity. For each population, the regression target is the length-N correctness vector over the same 1,200 stimuli (1 if majority vote matches ground truth, 0 otherwise), so model and human fingerprints are estimated from directly aligned item-level behavior. We train Ridge regression

(L2-penalized least squares) to predict this per-sample correctness vector from the 37 physical features. We choose Ridge over logistic regression because the MSE objective yields coefficient vectors reflecting each feature's marginal linear contribution to accuracy variance, providing a more graded signal for pairwise cosine comparison than log-likelihood-optimized coefficients. The resulting $\boldsymbol{\beta} \in \mathbb{R}^{37}$ serves as a "strategy fingerprint," and pairwise cosine similarity between fingerprints quantifies strategy alignment. For *feature attribution* (§6), we use the full 89-feature set ($19\mathrm{A} + 47\mathrm{B} + 23\mathrm{C}$), a systematic expansion that adds zone properties, probe/force information, collision detail breakdown, and composite outcome indices. The expanded set enables layer-wise attribution of model–human divergence. The 37-feature subset has direct semantic equivalents for 32/37 features in the 89-feature set; the five without equivalents are camera-relative or derivative quantities excluded from the systematic extraction (full correspondence table in supplementary material).

We note that strategy fingerprints capture *correlational* structure between physical features and accuracy patterns, not causal mechanisms; we use "strategy" as shorthand for this statistical signature.

4 Experimental Setup

4.1 Dataset: Physion

Physion [7] consists of 8 physical scenarios (*collision*, *containment*, *dominoes*, *drop*, *linking*, *rolling/sliding*, *towers*, *clothiness*), each containing 150 test samples (1,200 total) and ∼1,000 training samples (8,005 total). Each 5-second video is rendered from a physics simulator (ThreeDWorld); the task is binary contact prediction: will two designated objects come into contact? Human judgments from ∼100 Amazon Mechanical Turk raters per stimulus provide the reference population, totaling 120,450 raw responses across all 1,200 test samples.

The HDF5 simulation files provide complete physical state trajectories (per-frame object positions, velocities, collisions, physical properties), enabling extraction of the layered physical features used for strategy fingerprinting and attribution analysis (§3.4).

4.2 Models and Fair Comparison Protocol

A central design goal is *fair cross-model comparison*. Prior work compared models with different ViT sizes (Base/Large/Giant), frame rates, and probe architectures, confounding these factors with model quality. We standardize:

Unified backbone. All models use **ViT-Large** (∼307M parameters):

- **V-JEPA2** [2]: video joint-embedding predictive architecture.
- **VideoMAEv2** [25]: video masked autoencoder.
- **DINOv2** [20]: image self-distillation, applied per-frame.

Table 1: Cross-model comparison on Physion (all ViT-L, unified protocol, E2E attentive probes). V-JEPA2 significantly outperforms alternatives ($p < 0.001$, McNemar). All models show $> 5\times$ the disagreement rate of human–human baseline (4.8%).

Model	Seeds	Test Acc.	Train Acc.	Disagr.	κ_{MH}
V-JEPA2	8	**73.16±0.95**	90.60±0.56	26.4	0.484
VideoMAEv2	8	67.65±0.72	85.41±0.47	30.0	0.397
DINOv2	8	66.90±0.73	91.96±0.38	30.2	0.399
Human	∼100	74.2	—	4.8	0.906 (HH)

Unified evaluation. 8 input frames, fractional frame step 5.75 (∼1.5s temporal coverage, matching the human observation window), same E2E attentive probe architecture (depth 4, 16 heads), identical hyperparameter search grid (5 learning rates × 2 weight decay values = 10 probe configurations per seed), 20 training epochs. Best head selected by training accuracy (standard Physion protocol). A supplementary head-selection audit shows that choosing the test-oracle head would change accuracy by only 0.36pp, so probe selection cannot explain the much larger model–human disagreement.

Multi-seed evaluation. Each model is trained with **8 random seeds** (probe initialization only; encoders remain frozen). This enables population-level analysis: each seed produces an independent prediction vector over all 1,200 test samples.

4.3 Human Population Data

For each test sample, ∼100 independent raters provide binary contact predictions. We use the full response matrix $\mathbf{R} \in \{0,1\}^{1200\times H}$ (where H varies by scenario, averaging ∼100) for all distributional analyses. The human majority vote serves as the reference prediction; the response fraction $p_h^{(i)}$ quantifies human consensus per sample. All human data are drawn from the published Physion benchmark [7]; the original data collection protocol, including IRB approval and informed consent procedures, is described therein.

5 Results: Convergent Scores, Divergent Solutions

5.1 Surface Equivalence

Finding 1: *V-JEPA2 achieves* 73.2%±0.95%*, within 1.0pp of human accuracy (*74.2%*), but this apparent convergence is misleading.*

Table 1 summarizes overall accuracy. V-JEPA2 leads by ∼5–6pp over alternatives (all differences significant). DINOv2 shows the largest train–test gap (25.1pp), suggesting partial memorization, though its per-scenario error patterns remain qualitatively consistent with the other models (Table 2). The V-JEPA2–human gap (1.0pp) might suggest human-level reasoning; the remainder of this section shows why this conclusion is premature.

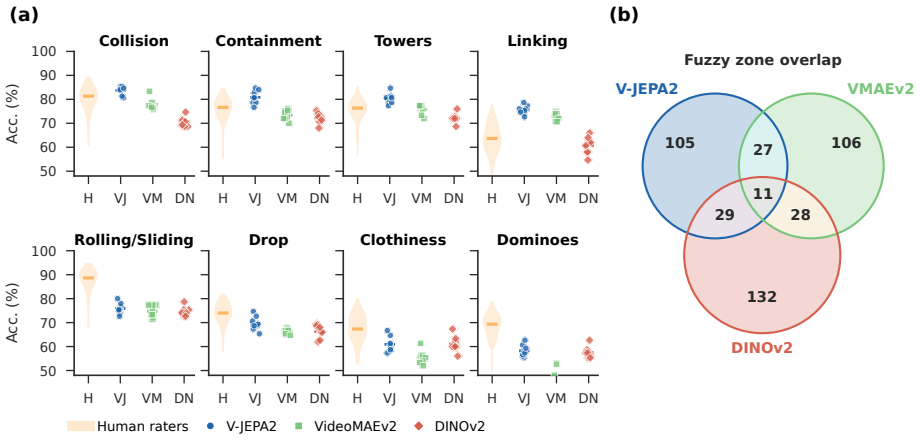

Fig. 2: Models are consistent but differently consistent than humans. (a) Per-scenario variability of human raters (gray violins, ∼100 raters) versus model seeds (colored dots, 8 seeds per model). Human rater distributions are 2–3× wider ($\sigma = 5$–8%) than model seed distributions ($\sigma = 1$–3%), yet model clusters are *offset* from the human median on most scenarios. (b) Fuzzy zone overlap: each model's "fuzzy zone" (near-chance internal agreement) contains 14–17% of samples, but pairwise Jaccard overlap is only ∼0.12, meaning models are uncertain about different samples.

The agreement hierarchy. The three-level agreement hierarchy (§3.2), estimated via split-half bootstrap, yields:

$$\underbrace{\kappa_{\text{HH}} = 0.906\ [.89,.92]}_{\text{human–human}} \gg \underbrace{\kappa_{\text{MM}} = 0.796\ [.77,.82]}_{\text{model–model}} \gg \underbrace{\kappa_{\text{MH}} = 0.484\ [.44,.49]}_{\text{model–human}} \tag{3}$$

Gap 1 ($\kappa_{\text{HH}} - \kappa_{\text{MM}} = 0.110$) is moderate, reflecting probe initialization noise. Gap 2 ($\kappa_{\text{MM}} - \kappa_{\text{MH}} = 0.312$) is 2.8× larger: models are self-consistent but converge to a different solution than humans. Since κ_{MM} is estimated from $K = 8$ seeds (split 4+4) with ties broken toward negative, Gap 2 is likely a *lower bound* on true model–human divergence.

5.2 Distributional Divergence

Finding 2: *Model populations are 2–3× more internally consistent than humans ($\sigma = 1$–3% vs. 5–8%), yet converge to systematically different solutions.*

Variability comparison. Figure 2 visualizes the core finding. Model seeds cluster tightly ($\sigma = 1$–3%) compared to human raters ($\sigma = 5$–8%), but the clusters are systematically *offset* from the human median on most scenarios. Inter-seed agreement is 84.2%±1.0%, meaning ∼16% of samples yield different predictions across probe initializations; yet the model–human agreement rate (73.6%) is

Table 2: Per-scenario accuracy (mean±std across seeds) for three models and humans. Δ = V-JEPA2 − Human. Best model per scenario in **bold**.

Scenario	V-JEPA2	VideoMAEv2	DINOv2	Human	Δ
Collision	**83.8±1.7**	77.8±2.2	70.6±1.8	80.3	+3.5
Containment	**80.8±2.6**	73.4±1.9	72.2±2.1	76.0	+4.8
Towers	**80.3±2.1**	75.8±1.9	72.2±1.9	75.5	+4.8
Linking	**75.6±1.8**	72.7±1.4	60.7±3.4	63.8	+11.8
Roll/Slide	75.9±2.1	**74.8±2.4**	74.6±1.8	87.7	−11.8
Drop	**69.7±2.8**	66.4±1.0	66.2±2.5	74.0	−4.3
Clothiness	60.9±3.0	55.5±2.6	**61.1±3.1**	67.7	−6.8
Dominoes	**58.2±2.3**	44.8±3.6	57.5±2.2	68.7	−10.5
Overall	**73.2±0.95**	67.7±0.72	66.9±0.73	74.2	−1.0

substantially lower, confirming that the model–human gap exceeds internal variability.

Fuzzy zone analysis. Each model places 14–17% of samples in its "fuzzy zone" (3–5 of 8 seeds correct), but the pairwise Jaccard overlap between fuzzy zones is only ∼0.12: different models are uncertain about largely *different* samples (Figure 2b).

5.3 Structured Divergence

Finding 3: *Divergence is scenario-structured: models excel on geometry/linking, while human advantages appear in several released-split future-dynamics scenarios, and this structure is captured by strategy fingerprints.*

Scenario-level patterns. Table 2 reveals consistent patterns across architectures. V-JEPA2 outperforms humans on collision, containment, towers, and linking (spatial/geometric scenarios); all three models underperform on *rolling/sliding*, *dominoes*, and *clothiness* in the released split. The largest model advantage is V-JEPA2's +11.8 pp on linking; the largest released-split human advantage is −11.8 pp on rolling/sliding. Because a subset of rolling/sliding ledge trials is rendering-sensitive, we treat this gap as descriptive of the reviewed benchmark split rather than as standalone evidence for a specific gravitational-dynamics mechanism.

Notably, scenario rankings *partially* differ across models: DINOv2 (image-only) matches V-JEPA2 on clothiness (61.1% vs. 60.9%) and dominoes (57.5% vs. 58.2%), suggesting these scenarios rely more on spatial than temporal cues. But DINOv2 collapses on linking (60.7%, −14.9pp vs. V-JEPA2), where temporal chain dynamics are critical. At the 45-subtask level, rank reversals occur across models (*e.g.* on Dominoes 4-Mid, DINOv2 reaches 95% while VideoMAEv2 reaches 43%), with gaps up to 52 pp on the same subtask, confirming qualitatively different strategies. To make these aggregate reversals inspectable

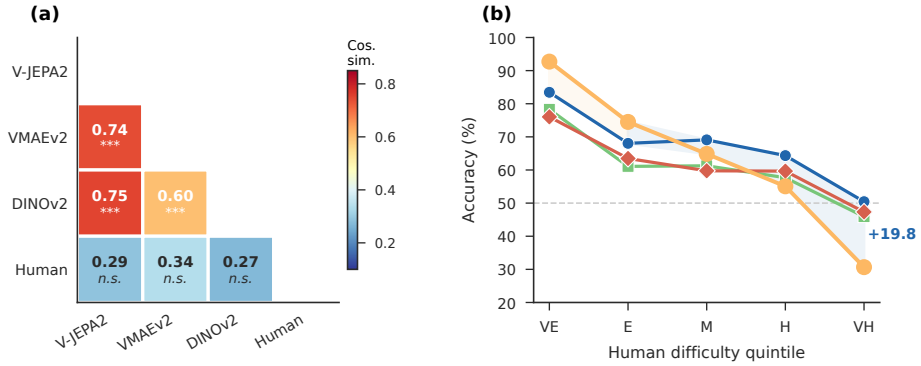

Fig. 3: Strategy and difficulty divergence. (a) Strategy fingerprint similarity matrix: cosine similarity between Ridge-regression strategy vectors (37 physical features → per-sample accuracy). All three models form a tight cluster (cosine = 0.60–0.75, $p < 0.001$), but *no model's strategy is significantly similar to humans* ($p > 0.1$ for all). (b) Difficulty alignment reversal: model accuracy declines less steeply than human accuracy as stimuli get harder. On the hardest human-defined quintile, V-JEPA2 scores +19.8 pp higher while remaining near chance ($r = 0.38$).

at the stimulus level, the supplementary material adds representative high-confidence disagreement cases selected from real Physion trials. These cases cover both directions of model–human reversal: human-correct/model-wrong samples whose correct label depends on unobserved future dynamics, and model-correct/human-wrong samples where humans are misled by plausible but incorrect trajectories. The qualitative panels show that divergence is not merely random label noise; they give concrete examples of disagreements around future contact, causal chains, and ambiguous trajectories.

Strategy fingerprinting. Using the 37-feature core subset defined in §3.4, Figure 3a shows the pairwise cosine similarity between strategy fingerprints. Two patterns emerge: (i) all model pairs are significantly similar (cosine 0.60–0.75, all $p < 0.001$), forming a tight cluster of shared non-human strategies; (ii) *no model aligns with humans* (cosine 0.27–0.34, all $p > 0.1$). The resulting behavioral fingerprints form a shared non-human cluster. The top model coefficients (`cutoff_target_moving`, `speed_trajectory_var`) relate to dynamics at the observation cutoff, while human correctness is more strongly associated with scene-level properties (`n_distractors`, `total_kinetic_energy_max`) that receive little weight in the model fingerprints.

Future outcomes are associated with divergence. For the 3.2% of stimuli where the target already contacts the zone at the observation cutoff, model–human disagreement is just 10.5% (baseline: no forward reasoning required); for the remaining stimuli requiring future prediction, disagreement rises to 26.8% (26.8/10.5 =

2.6×, $p < 0.05$). Crucially, prediction horizon distance does not predict disagreement ($r = -0.065$, $p = 0.12$): clothiness has the shortest horizon yet the highest disagreement, suggesting that divergence depends more on *which type* of reasoning is required than on how far forward the prediction extends.

5.4 Difficulty Alignment and Error Structure

Finding 4: *Models and humans find different stimuli difficult: model–human difficulty correlation is moderate ($r = 0.38$), and on the hardest stimuli for humans, models score +19.8 pp higher.*

Difficulty alignment. On "very hard" samples (human accuracy <50%), humans score 30.7% while V-JEPA2 scores 50.4% (+19.8 pp). This apparent model advantage is a useful diagnostic precisely because it is counterintuitive: the bin is selected by human difficulty, not by physical simplicity or model performance. Therefore, the gap should be read as difficulty misalignment rather than stronger physical reasoning: V-JEPA2 is near chance on these samples, while the entropy analysis shows that model confidence is often high where humans are uncertain. The model–human difficulty correlation is moderate ($r = 0.38$), leaving ∼86% of model difficulty variance unexplained by human difficulty.

Error overlap structure. Pairwise agreement between model majority votes is only ∼72–77%, substantially less than human–human agreement (95.2%). Error consistency analysis (Geirhos *et al.* [13]) reveals that model–model error overlap (avg. Jaccard = 0.41) far exceeds model–human overlap (avg. Jaccard = 0.26): models share error patterns with each other but not with humans.

Robustness on Physion++. We replicate the analysis on Physion++ [24], which tests latent physical properties (mass, friction, elasticity, deformability) across 10 scenarios. This robustness check asks whether the disagreement pattern is tied only to the original Physion item distribution or persists when the benchmark stresses different physical properties. Under the matched supplementary protocol, V-JEPA2 reaches 70.44% accuracy on Physion++, while model–human disagreement rises from 27.0% on Physion (cf. 26.4% under the main-text protocol) to 50.1% on Physion++; the deformability condition also reverses the apparent ranking between model and human performance. Because the protocol and stimuli are not identical to the main benchmark, we report Physion++ as supplementary evidence rather than a replacement for the main analysis, but it supports the conclusion that the divergence is not a narrow artifact of the original Physion split (details in supplementary material).

6 What Drives Divergence?

The distributional analysis in §5 establishes *that* models and humans diverge. We now ask which classes of physical features are associated with this divergence, testing whether the accessible signal lies primarily in visible information or in unobserved future outcomes.

Feature attribution. Using the full 89-feature set defined in §3.4, which expands the 37-feature strategy fingerprinting subset to enable layer-wise attribution, we separate features into **A** (19 static/scene), **B** (47 observable dynamics), and **C** (23 unobservable outcome). This analysis directly tests whether model–human divergence can be explained by visible scene information available before the prediction cutoff. Across Ridge regression, random forest, and gradient boosting models evaluated by 5-fold cross-validation, the best observable-feature model (A+B, 66 dimensions) still yields *negative* $R^2 = -0.32$ for predicting model–human divergence $|p_m - p_h|$ (AUROC = 0.63). We cannot rule out other nonlinear relationships, but this suggests observable features carry little accessible information about divergence under this protocol. Unobservable outcome features (C, 23 dimensions) achieve $R^2 = 0.29$ and AUROC = 0.70; for "both-certain disagree" samples, AUROC reaches 0.85. This asymmetry is consistent with divergence associated with forward outcomes rather than scene encoding alone. We note that Layer C features are correlated with task outcome by construction, so this attribution is suggestive rather than conclusive; disentangling the two would require interventional experiments beyond the scope of this work. The supplementary material reports all regressors and targets to make the negative observable-feature result and the positive unobservable-feature result transparent.

Calibration and confidence. Entropy decomposition reveals a notable asymmetry ($A = |Q2|/|Q3| = 0.31$–0.39 across models; 0.32 for V-JEPA2): models are more often confident where humans are uncertain (Q3) than uncertain where humans are certain (Q2). This does not imply superior reasoning: models confidently commit to answers on samples where humans remain split, consistent with a qualitatively different strategy rather than additional knowledge. However, within disagreement samples, overconfidence is prevalent: V-JEPA2 is the best-calibrated (ECE = 0.166, only 3 high-confidence errors vs. 16–24 for alternatives), but all models assign confident predictions to samples they get wrong. Full calibration and surprise analysis details are in supplementary material.

7 Discussion

Accuracy as an Incomplete Metric. Our central finding is that near-identical accuracy (73.2% vs. 74.2%) coexists with systematic behavioral divergence structured by scenario and future-prediction demands. The three-level agreement hierarchy quantifies this: Gap 1 ($\kappa_{\text{HH}} - \kappa_{\text{MM}} = 0.110$) is moderate, but Gap 2 ($\kappa_{\text{MM}} - \kappa_{\text{MH}} = 0.312$) is 2.8× larger, representing irreducible divergence between two populations that converge to *different* solutions.

Why Do Models Converge to Different Solutions? The strategy fingerprint analysis is consistent with *information access asymmetry* at the behavioral level: human correctness patterns reflect scenario-specific priors over object dynamics and causal chains that generalize from real-world experience, whereas the evaluated

frozen representations show a more shared cue-based fingerprint that succeeds where geometry is sufficient but struggles where future dynamics are critical. We cannot attribute this pattern uniquely to objective, architecture, training data, or the supervised probe protocol; a supplementary recurrent-model check supports the same signature within this protocol. The Jaccard overlap of $\sim$0.12 between model fuzzy zones suggests architecture-specific blind spots within this broader pattern. The difficulty alignment reversal should also be read as difficulty misalignment rather than model superiority: on stimuli hardest for humans ($<$50% accuracy), V-JEPA2 is near chance but still +19.8 pp above the human bin average, implying that the factors making stimuli hard differ between populations.

Implications for Benchmark Design. Our framework reveals two limitations of current practice. First, *reporting accuracy alone is insufficient*: we advocate including κ_{MH} and disagreement rate as standard metrics. Second, *multi-seed evaluation should be standard*: the $\sim$15% inter-seed disagreement rate means any single-seed result is a noisy estimate, and population-level comparison requires multiple runs. The $\sim$5–6pp gap between V-JEPA2 and alternatives only became apparent with unified ViT-L backbones.

Confident Divergence, Not Superiority. The entropy asymmetry ($A = 0.31$–0.39 across models; §6) shows that Q3 (model certain, humans uncertain) outnumbers Q2 by about 3$\times$, but model accuracy on Q3 samples is not above chance for the hardest cases. Models do not "know more"; they commit confidently to a different answer. This is consistent with a cue-based strategy that can produce confident outputs even on genuinely ambiguous stimuli, without matching human uncertainty.

Scene-Level Cues vs. Forward Simulation. Our findings are consistent with a dual-process account [6, 18]: models excel on geometry-driven scenarios (linking +11.8 pp), whereas human advantages appear on causal-chain judgments, including dominoes (-10.5 pp). We report rolling/sliding (-11.8 pp) as a released-split descriptive gap, not a standalone mechanism diagnostic, because ledge trials are rendering-sensitive. The cutoff-contact analysis (§5.3) suggests that the relevant distinction is not temporal extrapolation distance alone, but whether the future dynamics needed for the decision are accessible to the frozen representation and probe. We emphasize that our analysis provides behavioral and attributional constraints on these accounts, not direct evidence of internal mechanisms; forward-simulation-like representations may exist within models but remain untapped by this evaluation protocol.

Broader Applicability. Our distributional framework applies wherever per-sample human judgments are available alongside multiple model runs. The agreement hierarchy, entropy decomposition, and strategy fingerprinting are general tools that could reveal similar convergence illusions in other domains.

Limitations. Our framework requires multiple human raters per stimulus, so it applies most directly to crowd-sourced benchmarks. The main analysis uses one binary contact benchmark; Physion++ provides robustness evidence but not a complete benchmark sweep. Some released-split scenario subtypes are item-quality sensitive, so no single scenario gap is a definitive mechanism diagnostic. We evaluate ViT-L frozen representations with supervised probes, so larger 1B+ models, fine-tuned systems, generative/violation-of-expectation protocols, or interactive settings may yield different divergence patterns. Training data and benchmark familiarity are potential confounds that we do not causally isolate. Finally, strategy fingerprints are correlational summaries based on the 37-feature core subset (89-feature set for attribution); richer feature sets or interventional experiments may reveal additional structure.

8 Conclusion

Our analysis yields a simple diagnostic: when a model matches human accuracy on a physical-reasoning benchmark, the next question should not be "has the gap closed?" but "do model and human errors cluster in the same places?" On Physion, the answer in our setting is no. The κ_{MH} of 0.48, compared to κ_{HH} of 0.91, quantifies a behavioral gap that persists across three ViT-L architectures and is structured by the physical demands of each scenario: models excel where geometric cues suffice, while human advantages appear in several future-dynamics and causal-chain scenarios.

The locus of this gap, as indicated by feature attribution, lies downstream of scene encoding: observable scene properties carry little accessible predictive signal for where models and humans diverge within the scope of this evaluation ($R^2 = -0.32$), whereas unobservable outcome features are more predictive ($R^2 = 0.29$). These results do not show that models lack all relevant physical information; rather, they show that human-level accuracy can coexist with a different treatment of unobserved outcomes, a pattern shared across the evaluated architectures and invisible to accuracy-based evaluation.

These findings motivate two concrete next steps: applying distributional evaluation wherever per-sample human judgments are available, and testing interactive or generative forward-simulation protocols that separate future-dynamics errors from task-understanding or perceptual-ambiguity effects.

Acknowledgements

We thank the creators of the Physion benchmark for making their data and human response data publicly available.

References

1. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644 (2018), `https://arxiv.org/abs/1610.01644`, accessed 1 July 2026

2. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M.J., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Arnaud, S., Gejji, A., Martin, A., Hogan, F.R., Dugas, D., Bojanowski, P., Khalidov, V., Labatut, P., Massa, F., Szafraniec, M., Krishnakumar, K., Li, Y., Ma, X., Chandar, S., Meier, F., LeCun, Y., Rabbat, M., Ballas, N.: V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. CoRR **abs/2506.09985** (2025), https://doi.org/10.48550/arXiv.2506.09985, accessed 1 July 2026
3. Bansal, G., Nushi, B., Kamar, E., Horvitz, E., Weld, D.S.: Is the most accurate ai the best teammate? optimizing ai for teamwork. In: AAAI Conference on Artificial Intelligence (2021)
4. Baradel, F., Neverova, N., Mille, J., Mori, G., Wolf, C.: Cophy: Counterfactual learning of physical dynamics. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020), https://openreview.net/forum?id=SkeyppEFvS, accessed 1 July 2026
5. Bardes, A., Garrido, Q., Ponce, J., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv:2404.08471 (2024)
6. Battaglia, P.W., Hamrick, J.B., Tenenbaum, J.B.: Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences **110**(45), 18327–18332 (2013). https://doi.org/10.1073/pnas.1306572110
7. Bear, D., Wang, E., Mrowca, D., Binder, F.J., Tung, H., Pramod, R.T., Holdaway, C., Tao, S., Smith, K.A., Sun, F., Li, F., Kanwisher, N., Tenenbaum, J., Yamins, D., Fan, J.E.: Physion: Evaluating physical prediction from vision in humans and machines. In: Vanschoren, J., Yeung, S. (eds.) Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual (2021), https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/d09bf41544a3365a46c9077ebb5e35c3-Abstract-round1.html, accessed 1 July 2026
8. Belinkov, Y.: Probing classifiers: Promises, shortcomings, and advances. Comput. Linguistics **48**(1), 207–219 (2022). https://doi.org/10.1162/COLI_A_00422
9. Burnell, R., Schellaert, W., Burden, J., Ullman, T.D., Martínez-Plumed, F., Tenenbaum, J.B., Rutar, D., Cheke, L.G., Sohl-Dickstein, J.N., Mitchell, M., Kiela, D., Shanahan, M., Voorhees, E.M., Cohn, A.G., Leibo, J.Z., Hernández-Orallo, J.: Rethink reporting of evaluation results in ai. Science **380**, 136 – 138 (2023)
10. Embretson, S.E., Reise, S.P.: Item response theory for psychologists. Lawrence Erlbaum Associates, Mahwah, NJ (2000)
11. Garrido, Q., Ballas, N., Assran, M., Bardes, A., Najman, L., Rabbat, M., Dupoux, E., LeCun, Y.: Intuitive physics understanding emerges from self-supervised pretraining on natural videos. CoRR **abs/2502.11831** (2025). https://doi.org/10.48550/ARXIV.2502.11831
12. Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R.S., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nat. Mach. Intell. **2**(11), 665–673 (2020). https://doi.org/10.1038/S42256-020-00257-Z
13. Geirhos, R., Meding, K., Wichmann, F.A.: Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/9f6992966d4c363ea0162a056cb45fe5-Abstract.html, accessed 1 July 2026

14. Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F.A., Brendel, W.: Partial success in closing the gap between human and machine vision. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 23885–23899 (2021), https://proceedings.neurips.cc/paper/2021/hash/c8877cff22082a16395a57e97232bb6f-Abstract.html, accessed 1 July 2026
15. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T.L., Parajuli, S., Guo, M., Song, D.X., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8320–8329 (2021)
16. Hewitt, J., Liang, P.: Designing and interpreting probes with control tasks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 2733–2743. Association for Computational Linguistics (2019). https://doi.org/10.18653/V1/D19-1275
17. Joseph, S., Garrido, Q., Balestriero, R., Kowal, M., Fel, T., Bakhtiari, S., Richards, B., Rabbat, M.: Interpreting physics in video world models. arXiv preprint arXiv:2602.07050 (2026), https://arxiv.org/abs/2602.07050, accessed 1 July 2026
18. Kubricht, J.R., Holyoak, K.J., Lu, H.: Intuitive physics: Current research and controversies. Trends in cognitive sciences **21 10**, 749–759 (2017)
19. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. CoRR **abs/1604.00289** (2016), http://arxiv.org/abs/1604.00289, accessed 1 July 2026
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. **2024** (2024), https://openreview.net/forum?id=a68SUt6zFt, accessed 1 July 2026
21. Piloto, L.S., Weinstein, A., Battaglia, P.W., Botvinick, M.M.: Intuitive physics learning in a deep-learning model inspired by developmental psychology. Nature Human Behaviour **6**, 1257 – 1267 (2022)
22. Riochet, R., Castro, M.Y., Bernard, M., Lerer, A., Fergus, R., Izard, V., Dupoux, E.: Intphys 2019: A benchmark for visual intuitive physics understanding. IEEE Trans. Pattern Anal. Mach. Intell. **44**(9), 5016–5025 (2022). https://doi.org/10.1109/TPAMI.2021.3083839
23. Spelke, E.S., Breinlinger, K., Macomber, J., Jacobson, K.: Origins of knowledge. Psychological Review **99**(4), 605–632 (1992)
24. Tung, H., Ding, M., Chen, Z., Bear, D., Gan, C., Tenenbaum, J., Yamins, D., Fan, J.E., Smith, K.A.: Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA,

USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/d3e8011c912e651ab2a76e7935a1e464-Abstract-Datasets_and_Benchmarks.html, accessed 1 July 2026

25. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae V2: scaling video masked autoencoders with dual masking. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 14549–14560. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.01398
26. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: CLEVRER: collision events for video representation and reasoning. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020), https://openreview.net/forum?id=HkxYzANYDB, accessed 1 July 2026