GEOPHYS: The Geometry of Physical Plausibility

Christian Internò[1*] **Alexander Pondaven**[2] **Habon Issa**[3] **Fabio Pizzati**[4]
Francesco Pinto[5] **Markus Olhofer**[6] **Ivan Laptev**[4] **Philip Torr**[2]
Eero P. Simoncelli[7,8] **Barbara Hammer**[1] **David Klindt**[3]

[1] Bielefeld University [2] University of Oxford [3] Cold Spring Harbor Laboratory
[4] MBZUAI [5] Independent [6] Honda Research Institute EU
[7] New York University [8] Flatiron Institute, Simons Foundation

Abstract

While humans can identify physically implausible events within milliseconds, machine learning approaches addressing the same problem are extremely slow and expensive. They either rely on external multimodal-LLM judges or require ad-hoc modifications to the training procedure. In this work, we argue that indicators of physical plausibility are implicitly captured by five geometric properties of the per-frame embeddings produced by frozen image encoders. In aggregate, we call them GEOPHYS. First, we show that these signals correlate with human EEG responses to two forms of object-permanence violations. Second, GEOPHYS robustly discriminates physically implausible videos from realistic ones, achieving state-of-the-art physics-violation detection: 98.3% on LikePhys [1] and 93.3% on IntPhys2 [2], whereas V-JEPA 2, GPT-4o, Gemini, and twelve modern video diffusion models perform near chance. Third, used as a best-of-N verifier for physical alignment during video generation, GEOPHYS lifts MAGI-1 24B from 50.01% to 64.50% on PhysicsIQ [3] at $1.5\times$ lower wall-clock and $4.65\times$ lower memory than the V-JEPA 2 world-model verifier. Ultimately, GEOPHYS demonstrates that physical plausibility in videos can be assessed by leveraging the emergent geometric properties of temporal features extracted from image encoders.

1 Introduction

Humans readily detect physical violations directly from visual perception: infants register surprise on seeing impossible events [4, 5], and adult brains register such events rapidly, without deliberation [6, 7]. In machine learning, by contrast, detecting physical violations in video has been achieved using physics-targeted model training [8, 9], billion-parameter video-pretrained world models [10, 11], multi-modal language models prompted to act as judges [12], neuro-symbolic pipelines that parse scenes and run physical simulators [13, 14, 15] or hand-crafted forensic checks for physical consistency in generated imagery [16, 17, 18, 19]. All these approaches share the core assumption that signals for physical plausibility must be explicitly induced into, or reasoned about by a model. In this work, we challenge that assumption by asking a simpler question:

Does a vision model trained on still images, with no video training or physics supervision, contain a geometric signature of physical plausibility?

A long line of work in machine learning and computational neuroscience suggests that good visual representations should make natural transformations geometrically simple, mapping nonlinear pixel-space motion onto approximately linear feature-space trajectories. This thread runs from temporal

*Correspondence to: christian.interno@uni-bielefeld.de
Project page: https://christianinterno.github.io/GeoPhys/
Code: https://github.com/ChristianInterno/GeoPhys

slowness and tangent-propagation [20, 21, 22] through Lie-group representations of transformation [23] to learned linearization [24, 25, 26], and connects to the neuroscience finding that primate visual cortex straightens natural-video trajectories relative to pixel space [27, 28, 29, 30]. Together, these results motivate a concrete prediction about physical plausibility:

Hypothesis.
Vision models learn to linearize natural video dynamics, mapping physically plausible videos to smooth, locally predictable feature-space trajectories disrupted by physical violations.

This link between geometry and physical plausibility is purely statistical. We do not argue that frozen backbones represent physics or reason about it [31, 7, 32]; we argue that the geometry of feature-space trajectories is disrupted during physics-violations in existing benchmarks [2, 1, 7] reliably enough to be useful. We find that this prediction holds across four distinct backbones, none trained on video or physics: self-supervised transformers (DINOv2 [33] & v3 [34]), a recurrent ventral-stream CNN (CORnet-S [35]), and a ResNet with a fixed Gabor V1 front-end (VOneNet [36]).

Building on our insights, we introduce GEOPHYS, a training-free score that captures the temporal evolution of feature trajectories. It computes a set of kinematic statistics on feature space across frames: speed and its variation, turning-angle curvature, acceleration, and the residual of a linear auto-regressive predictor. These combine into a single scalar physical plausibility score. We apply the same score across three settings, each testing a property a useful signal should have: **(i)** It must correspond to a real phenomenon rather than an artifact of pretraining statistics. We test this by examining the score's temporal response to physical violations and its alignment with human neural responses. **(ii)** It must discriminate plausible from violated video at scale. We test this via detection on standard physics-violation benchmarks. **(iii)** It must transfer beyond passive measurement to active control. We test this via inference-time best-of-N selection for physically plausible video generation.

Contributions: **(1) Biologically grounded**. Lacking pretraining on physics objectives, geometric signals could flag invalid videos for reasons unrelated to physical violations per se. We thus validate per-frame GEOPHYS signals are directionally meaningful, aligning with known surprise-detection signatures of human EEG responses to matched stimuli [6, 7, 37]. **(2) Detection**. We achieve SoTA physics-violation detection: 98.3% on LikePhys [1] and 93.3% on IntPhys2 [2]. Every individual backbone exceeds all published baselines, including V-JEPA 2 [11], GPT-4o [38], Gemini [39], and twelve video diffusion models. **(3) Generation**. We offer the signal as an inference-time verifier for test-time scaling. It closes more of the gap to the oracle ceiling than any prior method on a matched candidate pool. On PhysicsIQ [3], it lifts MAGI-1 24B [40] from 50.01% to 64.50%. This comes at $1.5\times$ lower wall-clock and $4.65\times$ lower memory than the VJEPA-2 baseline WMReward [10].

2 Related work

Physics evaluation in video models. Physical-plausibility benchmarks include PhysicsIQ [3] (real-world fidelity), LikePhys [1] (12 dynamic scenarios), and IntPhys2 [2] (four core-knowledge properties), after earlier Physion [41], IntPhys [42], and GRASP [43]. Three baseline families dominate: multimodal-LLM judges [12, 38, 39], video diffusion models scored by likelihood [44, 45, 46], and video-pretrained encoders [11, 47]. All require video-scale pretraining, billion-parameter inference, or both. GEOPHYS requires neither: every backbone is a frozen image encoder, sub-billion parameters, no video pretraining.

Verifier models for video generation. Inference-time methods rerank a fixed generator's candidates with a verifier [48]. WMReward [10], the SoTA on PhysicsIQ, uses V-JEPA 2 [11] (1.1B parameters, video-pretrained). Other verifiers use multimodal-LLM judges [49], video-feature backbones [47], learned video-quality models [50, 51], or 3D scene-geometry consistency [52]. All draw the signal from a video-pretrained, language-scale, or 3D-estimation system. We use the opposite: a frozen image encoder with no learned ranker. This places physics-aware generation in the test-time scaling literature [53, 54, 55, 56, 57, 58, 59], where verifier cost trades off against selection quality.

Geometric analysis of feature trajectories. The straightening lineage [20, 21, 23, 22, 24, 28, 29, 30, 26] studies feature-trajectory geometry in biological and artificial vision. Predictive coding offers a complementary view [60, 61]: perception minimises feature-space prediction error, motivating

GEOPHYS's residual signal. For AI-vs-natural-video detection, ReStraV [62] uses DINOv2 curvature and Grab-3D [63] uses 3D scene-geometry consistency. GEOPHYS is the first to apply frozen-image-encoder trajectory geometry to physical plausibility itself, across four backbones (ViTs [33] and V1-like CNNs [36]), five kinematic statistics, and three settings: neural alignment (4.1), physics-violation benchmarks (4.2), and best-of-N selection at generation time (4.3).

3 A GEOmetrical signal of PHYsical plausibility

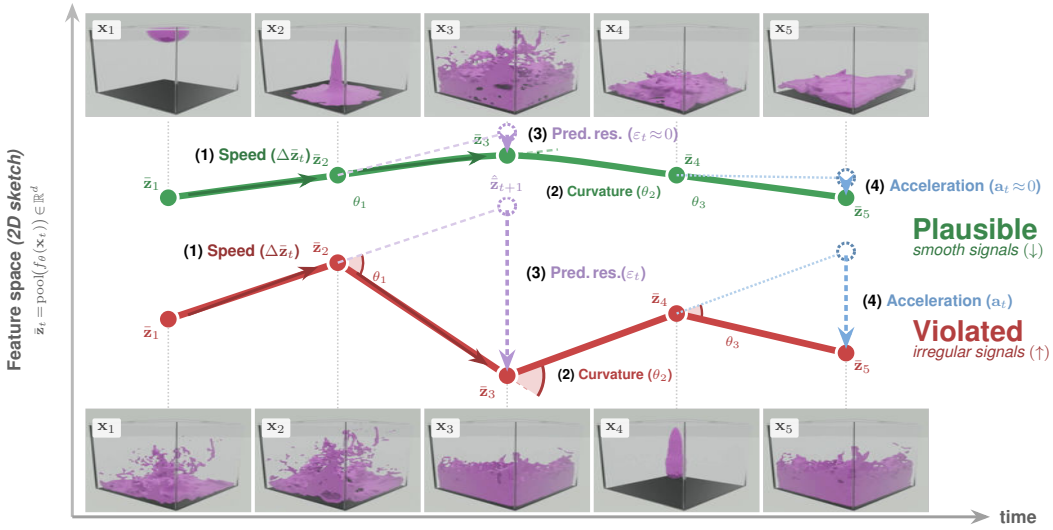

Figure 1: **GEOPHYS signals of plausible vs. violated dynamics in frozen feature space.** Paired counterfactuals from LikePhys [1], rendered in Blender [64]. A frozen backbone maps each frame $\mathbf{x}_t$ to a pooled feature $\bar{\mathbf{z}}_t$, yielding a trajectory in representation space (sketched in 2D). **Plausible** videos produce smooth trajectories; **Violated** (no momentum conservation) ones show irregular. **(1)** speed, **(2)** curvature, **(3)** prediction residual, and **(4)** acceleration.

GEOPHYS is applied to synthetic and AI generated videos [65], which systematically violate the physical laws inherent in real data [3, 2]. Given a T-frame video $\mathbf{V} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ and a frozen visual backbone f_θ, each frame yields spatial embeddings $Z_t = f_\theta(\mathbf{x}_t) = \{\mathbf{z}_{t,n}\}_{n=1}^N \in \mathbb{R}^{N\times d}$, containing N tokens of dimension d. Spatial average-pooling extracts a single feature vector per frame, $\bar{\mathbf{z}}_t = \frac{1}{N}\sum_{n=1}^N \mathbf{z}_{t,n} \in \mathbb{R}^d$, defining the video as a discrete feature-space trajectory: $\Gamma_\theta(\mathbf{V}) = (\bar{\mathbf{z}}_1, \ldots, \bar{\mathbf{z}}_T)$. We evaluate temporal irregularities via finite differences on $\Gamma_\theta(\mathbf{V})$, using kinematic terms ("velocity", "speed", "acceleration") strictly as geometric analogies in representation space, not as physical motion in pixels or world coordinates.

3.1 GeoPhys signals

First-order motion: speed. The trajectory's first-order kinematics are defined by the frame-to-frame feature displacement, $\mathbf{v}_t = \bar{\mathbf{z}}_{t+1} - \bar{\mathbf{z}}_t$ $(t = 1, \ldots, T-1)$. *Its magnitude, $s_t = \|\mathbf{v}_t\|_2$, measures the representational shift between consecutive frames ((**1**), $\Delta\bar{\mathbf{z}}_t$ in Fig. 1).* Because stable video dynamics preclude abrupt step-size fluctuations, we summarize speed instability via its temporal standard deviation: $\phi_{\text{speed}}(\mathbf{V}) = \mathrm{std}\big(\{s_t\}_{t=1}^{T-1}\big)$. A large ϕ_{speed} flags erratic representational changes, such as objects jumping, disappearing, or moving inconsistently.

Curvature and angle consistency. Constant speed does not guarantee a straight trajectory: representations can bend sharply between frames, marking locally nonlinear transformations. Curvature is the central diagnostic in biological vision [28]: humans and macaque V1 populations perceptually straighten natural videos while curving unnatural ones. In artificial networks, straightening requires specific conditions (SSL objectives [30], adversarial robustness [8]) and is absent from standard classifiers [27]. *We measure local curvature by the turning angle between consecutive displacements ((**2**), θ_t in Fig. 1):* $\theta_t = \arccos\!\left(\frac{\langle \mathbf{v}_t,\, \mathbf{v}_{t+1}\rangle}{\|\mathbf{v}_t\|_2\, \|\mathbf{v}_{t+1}\|_2}\right), \quad t = 1, \ldots, T-2$. Locally linear

transformations yield small θ_t. Physical violations break this linearity: a backbone expects natural dynamics (balls bouncing off walls), so an implausible "wall-pass" frame diverges sharply from the smooth pre-impact trajectory, bending θ_t. We capture this with two statistics: the mean turning angle $\phi_{\text{curv}}(\mathbf{V}) = \frac{1}{T-2}\sum_{t=1}^{T-2}\theta_t$ and its standard deviation $\phi_{\text{ang}}(\mathbf{V}) = \text{std}\big(\{\theta_t\}_{t=1}^{T-2}\big)$, separating consistent trajectories from erratic ones. Empirically, all four GEOPHYS's backbones straighten plausible videos relative to violated ones (Appendix A).

Second-order motion: acceleration. *A plausible sequence should also avoid abrupt changes in the feature-space displacement itself.* We capture this with the second-order difference $\mathbf{a}_t = \mathbf{v}_{t+1} - \mathbf{v}_t = \bar{\mathbf{z}}_{t+2} - 2\bar{\mathbf{z}}_{t+1} + \bar{\mathbf{z}}_t \quad (t = 1, \ldots, T-2)$. This quantity is large when the trajectory abruptly changes speed or direction (**(4)** ,$\mathbf{a}_t$ in Fig. 1). We summarize it as $\phi_{\text{accel}}(\mathbf{V}) = \frac{1}{T-2}\sum_{t=1}^{T-2}\|\mathbf{a}_t\|_2^2$. While ϕ_{curv} isolates directional bending, ϕ_{accel} captures the full second-order instability of the feature trajectory, including both changes in direction and changes in step magnitude.

Local linearity residual. A complementary signal, inspired by Goroshin et al. [24]'s linearisation objective, asks whether the next feature lies in the affine subspace spanned by the previous H. When local dynamics are linear, the trajectory is locally low-dimensional: the next feature is a predictable linear combination of its predecessors. Physical violations that break this subspace (abruptly reversing fluids, spontaneously accelerating objects) spike the residual at the violation frame. We fit a linear auto-regressive predictor $\hat{P}_H : \mathbb{R}^{Hd} \to \mathbb{R}^d$ on past windows, $\hat{\bar{\mathbf{z}}}_{t+1} = \hat{P}_H\big([\bar{\mathbf{z}}_{t-H+1}; \ldots; \bar{\mathbf{z}}_t]\big)$ for $t = H, \ldots, T-1$, and take the residual $\varepsilon_t = \bar{\mathbf{z}}_{t+1} - \hat{\bar{\mathbf{z}}}_{t+1}$ ($[\cdot;\cdot]$: concatenation). Geometrically, ε_t is the component of $\bar{\mathbf{z}}_{t+1}$ orthogonal to the H-step linear span (**(3)**, ε_t in Fig. 1). We summarise it as $\phi_{\text{perr}}(\mathbf{V}) = \frac{1}{T-H}\sum_{t=H}^{T-1}\|\varepsilon_t\|_2$; low values indicate a low-dimensional manifold.

Together, these five scalar signals form the temporal geometric descriptor used by GEOPHYS:

$$\Phi_{\text{temp}}(\mathbf{V}) = \big(\phi_{\text{curv}}, \phi_{\text{ang}}, \phi_{\text{speed}}, \phi_{\text{accel}}, \phi_{\text{perr}}\big). \tag{1}$$

For all five signals, larger values indicate less regular feature-space dynamics.

3.2 Vision backbones and GEOPHYS pipeline

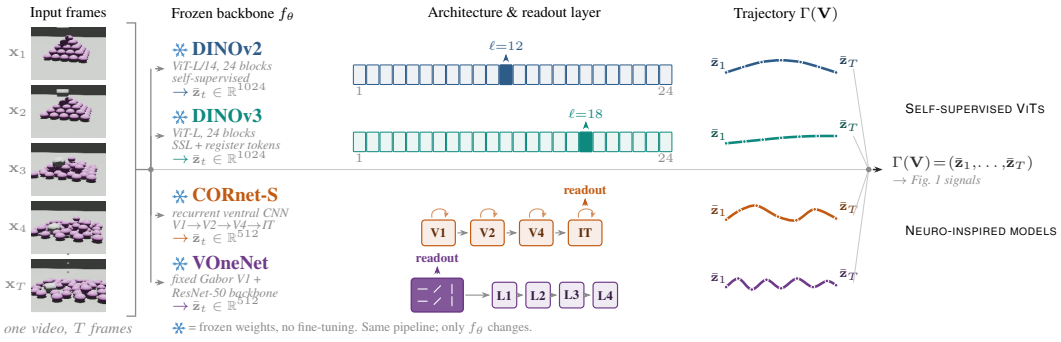

Figure 2: **GEOPHYS pipeline.** A single video $\mathbf{V} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ feeds each backbone f_θ unchanged. Each backbone reads out at the layer $\ell^\star$ selected on a held-out validation split (Appendix B); the per-frame embedding $\bar{\mathbf{z}}_t = \text{pool}\big(f_\theta^{(\ell^\star)}(\mathbf{x}_t)\big)$ is spatial-pooled and stacked across time into the trajectory $\Gamma(\mathbf{V}) = (\bar{\mathbf{z}}_1, \ldots, \bar{\mathbf{z}}_T)$, on which the geometric signals of Fig. 1 operate.

From frames to temporal feature trajectories. GEOPHYS operates on any frozen vision backbone without retraining or modification. Given a backbone f_θ and a held-out validation split, we select a readout layer $\ell^\star$ that maximizes the curvature gap between plausible and violated videos (Fig. 3), yielding the frozen feature map $f(\cdot) = f_\theta^{(\ell^\star)}(\cdot)$. For a video $\mathbf{V} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, each frame produces N spatial tokens of dimension d: $Z_t = f(\mathbf{x}_t) = \{\mathbf{z}_{t,n}\}_{n=1}^N \in \mathbb{R}^{N \times d}$. Spatial pooling collapses each grid into a single vector $\bar{\mathbf{z}}_t = \text{pool}(Z_t) \in \mathbb{R}^d$. Stacking these across time forms the trajectory $\Gamma(\mathbf{V}) = (\bar{\mathbf{z}}_1, \ldots, \bar{\mathbf{z}}_T)$, upon which GEOPHYS's signals (Sec. 3) operate. This pipeline remains identical across all backbones, changing only the choice of f_θ (Fig. 2).

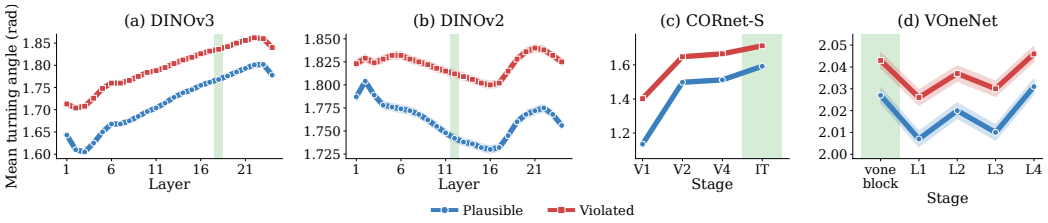

Figure 3: **Mean curvature across layers** Blue: plausible. Red: violated (65 video pairs). Violated lies above plausible at every layer of every backbone ($p < 10^{-8}$, t-test). Green band: readout layer.

Backbone selection. Hypothesis 1 originates in the V1-straightening literature [27, 28], so testing it requires backbones that span V1-like and non-V1-like representations. We use four frozen backbones in two architecture classes. **DINOv2** [33] and **DINOv3** [34] are self-supervised ViT-L/14s already shown to discriminate AI-generated from natural video via the curvature signal [62]. **CORnet-S** [35] (recurrent V1→V2→V4→IT ventral-stream CNN [66]) and **VOneNet** [36] (ResNet-50 with a fixed Gabor V1 front-end [67, 68]) are architecturally explicit models of primate visual cortex.

Readout via curvature. For each backbone, we select the readout layer $\ell^\star$ that maximises the curvature delta $\Delta\phi_{\text{curv}} = \bar{\phi}^{-}_{\text{curv}} - \bar{\phi}^{+}_{\text{curv}}$ between violated and plausible videos on a held-out LikePhys validation split [1]. By Hypothesis 1, $\ell^\star$ best reflects the distinction GEOPHYS exploits. The geometric and task-driven criteria coincide: for every backbone, $\ell^\star$ also maximises detection accuracy (Appendix B). The gap is positive across all 57 layer×backbone combinations ($p < 10^{-8}$, paired t-test; $d \in [0.22, 0.58]$; Fig. 3), with CORnet-S V1 producing the largest effect and DINOv2/3 rising with depth. For backbone b with readout $\ell^\star$, we write $u^{(b)}_{\phi}(\mathbf{V}) = \phi\big(\Gamma_{f_\theta}(\mathbf{V})\big)$ for the value of signal ϕ on $\mathbf{V}$'s trajectory through that backbone.

Decision rule. Each backbone uses the single signal that scores best on the LikePhys [1] held-out split: angle consistency for DINOv2/v3, speed variation for CORnet-S, and acceleration for VOneNet (Table 7). GEOPHYS maps a video to a single scalar, its standardized signal z_b, with larger values indicating less plausible dynamics. We predict the video with the larger z_b as violated, and read $|z_b|$ as confidence. Because the backbones catch different violations (Fig. 5), we combine them by **Majority** ($\geq 3/4$ agree) or **OR** (the most confident, $b^\star = \arg\max_b |z_b|$).

4 Evaluation

We test GEOPHYS in three studies, applying the same score unchanged throughout, to show that feature-space geometry reflects physical plausibility, that the resulting signal is biologically meaningful, and that it is useful for physics-aligned video generation. **Study 1 (Sec. 4.1)** compares GEOPHYS signals to EEG visual working-memory recordings from humans viewing matched violation-of-expectation stimuli [6, 7, 37], testing whether the per-frame score tracks neuro-biological responses. **Study 2 (Sec. 4.2)** discriminates plausible from violated videos on LikePhys [1] and IntPhys2 [2]. **Study 3 (Sec. 4.3)** deploys the score as a training-free best-of-N verifier on PhysicsIQ [3].

4.1 Study 1: Detecting object permanence violations in models and brains

Setup. The violation-of-expectation (VOE) paradigm uses paired valid/invalid stimuli to elicit reliable surprise responses to physically impossible events in infants [4, 5] and, in adult EEG, rapid shifts in the contralateral delay activity (CDA), a marker of visual working memory that scales with the number of tracked objects [6, 69, 70]. Across two experiments, we test whether GEOPHYS per-frame signals $(\theta_t, s_t, \mathbf{a}_t, \varepsilon_t)$ reliably track the same time course on two object-permanence scenarios: *Create* (one object enters an occluder, two emerge) and *Vanish* (two enter, one emerges). Experiment 1: we use the public stimulus subset from [6] ($n = 16$ subjects) to visualize temporal alignment. Experiment 2: we render additional matched pairs with ADEPT [13] and measure the valid − invalid signal difference pre- and post-occlusion.

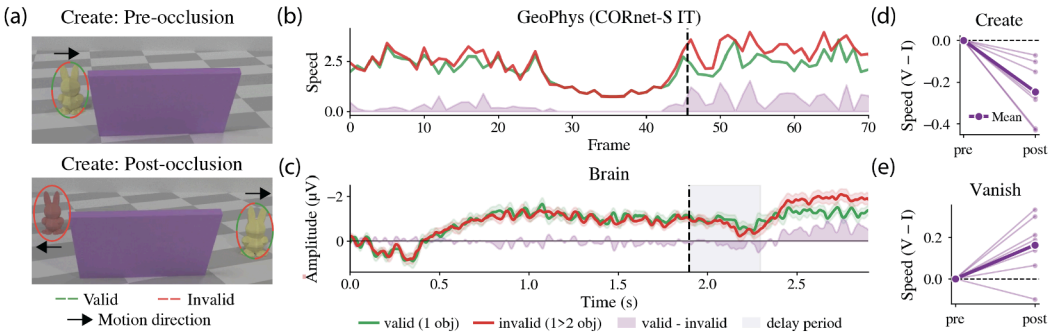

Figure 4: **Model and brain comparison.** (a) Example VOE stimuli for the Create scenario: in valid videos, one object enters and exits occlusion; in invalid videos, one enters but two exit. (b–c) Valid and invalid Create signals from GEOPHYS CORnet-S IT speed (b) and EEG contralateral delay activity (from [6]; c). Both are elevated for the invalid condition after occlusion offset (vertical dashed line). (d–e) Mean valid − invalid GEOPHYS CORnet-S IT speed pre- and post-occlusion for Create (d) and Vanish (e) scenarios (rendered with ADEPT). Known EEG VOE delay period in gray [6].

Findings. In Create (Fig. 4a), GEOPHYS CORnet-S IT speed (Fig. 4b) and the human CDA (Fig. 4c) both rise rapidly and persistently in the invalid condition after the violation; in Vanish (Fig. 14a), both decrease. The two signals diverge during occlusion (∼1.2–1.9 s), where the CDA stays elevated, consistent with abstract object tracking beyond the visual input [6], whereas GEOPHYS approaches zero. Quantitatively, GEOPHYS tracks object number in Create ($t(6) = 4.501$, $p = 0.0041$; Fig. 4d) and Vanish ($t(6) = -2.926$, $p = 0.0264$; Fig. 4e). Across backbones (Table 5), CORnet-S IT speed, acceleration, and prediction error are the only signals showing notable post-occlusion divergence across Create and Vanish. Per-backbone and ensemble detection on all stimuli is reported in Tab. 6.

→ **Takeaway 1.** GEOPHYS signals align with human EEG responses to object-permanence violations [6] and scale with object number, consistent with physical-violation perception.

4.2 Study 2: Physics violation detection

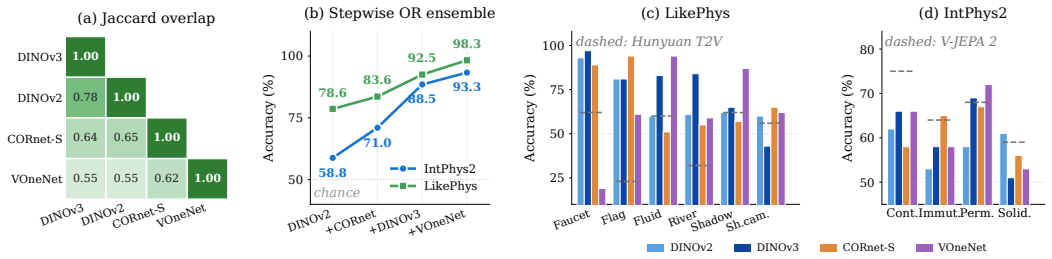

Figure 5: **Backbones are complementary. (a)** Jaccard overlap of correctly-flagged sets on LikePhys. **(b)** Greedy stepwise ensemble reaches 98.3% on LikePhys and 93.3% on IntPhys2. **(c)** Each backbone wins a different physics domain of LikePhys. **(d)** Each backbone wins a different condition in IntPhys2. Dashed: best baseline (Hunyuan T2V [45] in c, V-JEPA 2 [11] in d).

Setup. LikePhys [1] and IntPhys2 [2] present matched pairs $(\mathbf{V}^+, \mathbf{V}^-)$ sharing initial conditions, with one physically violated; the task is to identify it. The per-backbone scalar is the signed delta $\delta_b = u^{(b)}_{\phi_b}(\mathbf{V}^-) - u^{(b)}_{\phi_b}(\mathbf{V}^+)$; we z-normalise it to z_b, whose sign identifies the violated video, and aggregate by Majority or OR exactly as defined in Sec. 3.2. 95% CIs: 1K-resample bootstrap (LikePhys [1] per-scenario, IntPhys2 [2] per-pair).

Table 1: **Physics violation detection.** Pairwise accuracy (%, ↑ 0% 100%). Underline: best baseline per column. **Bold**: best GEOPHYS backbone per column.

	LikePhys [1]													IntPhys2 [2]					
	Ball coll.	Ball drop	Block slide	Cloth drape	Faucet	Flag	Fluid	Pendulum	Pyramid	River	Shadow	Sh. cam.	*M.Avg.*		Contin.	Immut.	Perman.	Solid.	*M.Avg.*
AnimDiff [71]	**37**	40	**35**	47	**38**	**21**	**37**	48	**39**	56	**28**	44	39.2	Cosmos [72]	54	51	52	55	49.4
AnimSDXL [71]	**37**	**33**	62	53	46	**17**	40	**30**	**31**	54	57	68	44.0	Qwen-VL [49]	54	57	54	51	52.3
ZeroScope [73]	47	45	53	61	54	**16**	**38**	42	**27**	60	60	58	46.7	Gemini 1.5 [39]	55	57	56	57	52.3
ModelScope [73]	48	47	53	60	60	**16**	**38**	**33**	**21**	60	67	62	47.1	GPT-4o [38]	57	60	60	57	53.8
Mochi [74]	60	50	68	71	62	**17**	40	**35**	**16**	56	42	60	48.1	VidMAEv2 [47]	65	55	64	60	53.8
CogVidX-5B [46]	57	58	**38**	61	64	**19**	47	**35**	**17**	60	77	70	50.2	V-JEPA [11]	57	56	68	52	53.8
CogVidX-2B [46]	60	62	40	64	64	**19**	50	**37**	**17**	60	82	68	51.8	Gemini 2.5 [39]	55	64	64	56	55.6
Wan2.1-1B [44]	57	47	**33**	**34**	40	67	67	80	60	**22**	77	40	52.0	V-JEPA 2 [11]	75	59	60	59	57.5
LTX Vid. [75]	63	42	65	53	68	51	72	**32**	81	60	**25**	52	55.3						
CogVidX1.5 [46]	**38**	50	48	47	46	77	78	68	74	**32**	77	**38**	56.2						
Wan2.1-14B [44]	58	43	**38**	**37**	44	66	85	87	57	**36**	83	40	56.2						
Hunyuan [45]	67	52	75	64	62	**23**	60	78	50	**32**	62	56	56.4						
†: *trained with leave-one-scene-out cross-validation on same features of* GEOPHYS *(Dinov3)*																			
Lin. probe†	85	**38**	85	75	75	65	45	61	85	45	45	45	62.4 ±7.1	Lin. probe†	51	60	59	53	55.5 ±4.1
GEOPHYS (V-JEPA 2)	92	68	72	75	92	76	93	81	**97**	66	72	57	78.3 ±8.4	GEOPHYS (V-JEPA 2)	67	58	60	53	59.5 ±5.5
GEOPHYS individual																			
DINOv2 [33]	**100**	45	**100**	92	93	81	60	89	**100**	61	62	60	78.6 ±8.2	DINOv2 [33]	62	53	58	**61**	58.8 ±7.1
DINOv3 [34]	**100**	**38**	**100**	92	**97**	81	83	87	**100**	**84**	65	43	80.8 ±9.7	DINOv3 [34]	**66**	58	69	51	60.5 ±4.5
CORnet-S [35]	80	**83**	**100**	78	89	**94**	51	86	**100**	55	57	**65**	78.2 ±9.2	CORnet-S [35]	58	**65**	67	56	61.1 ±4.6
VOneNet [36]	90	82	**100**	**88**	**19**	61	**94**	**90**	**100**	59	**87**	62	77.6 ±9.0	VOneNet [36]	66	58	**72**	53	**61.7** ±4.2
GEOPHYS ensembles																			
Majority	**100**	77	**100**	**95**	**99**	**96**	93	**96**	**100**	86	83	78	90.9 ±6.3	Majority	78	75	83	75	77.5 ±3.7
OR	**100**	88	**100**	**100**	**100**	**100**	**97**	**99**	**100**	**100**	**97**	92	**98.3** ±2.1	OR	94	91	95	93	**93.3** ±2.8
														Human	–	–	–	–	96.4

IntPhys2 baselines: per-subset best runs (Bordes et al. [2], Tab. 3); M.Avg. = their Main-set overall (Tab. 2).

Data: **LikePhys** [1] provides 650 matched pairs rendered in Blender [64] across 12 scenarios in four domains (rigid-body, continuum, fluid, optical). Each pair shares appearance and differs only by a controlled violation (reversed gravity, ground penetration, energy non-conservation, temporal disorder), probing whether models encode physical dynamics. **IntPhys2** [2], tests four core-knowledge properties (*permanence*, *solidity*, *continuity*, *immutability*) via 506 photorealistic 3D pairs in Unreal Engine [76], with fixed and moving cameras and structured as quadruplets where possible/impossible videos share initial conditions and swap roles across occluder configurations.

Findings: **LikePhys** [1] ranks matched pairs via video diffusion models' (DMs) likelihood preference: a DM judges plausibility by assigning higher likelihood to the plausible video. All 12 DMs evaluated score 39–56% (near chance, Table 1, left). A leave-one-scene-out linear probe trained on DINOv3 features reaches 62.4% (Table, †). This probe controls for feature quality. GEOPHYS gain is attributable to trajectory geometry, not the backbone alone. Every GEOPHYS backbone exceeds both: DINOv3 L18 80.8% (angle consistency), DINOv2 L12 78.6%, CORnet-S IT 78.2% (speed variation), VOneNet V1 77.6% (acceleration). Applying the same framework to V-JEPA 2 features yields 78.3%, showing GEOPHYS is not backbone-specific. Each backbone specialises in a different signal (full breakdown in Appendix D); majority vote reaches 90.9%±6.3 and OR 98.3%±2.1, within 1.7 points of the per-scenario ceiling (per-domain breakdown in App. E).

In **IntPhys2** [2], V-JEPA 2 (57.5%), Gemini-2.5 Flash (55.6%), GPT-4o (53.8%), and a linear probe (55.5%, Table, †) all remain within 8 points of chance (Table 1, right). Every GEOPHYS backbone exceeds prior SoTA: VOneNet V1 61.7%, CORnet-S V1 61.1%, DINOv3 L12 60.5%, DINOv2 L12 58.8%. Majority vote reaches 77.5% (+20 over V-JEPA 2) and OR reaches 93.3%, within 3.1 points of human performance (96.4%); per-difficulty and per-camera splits are in App. E.

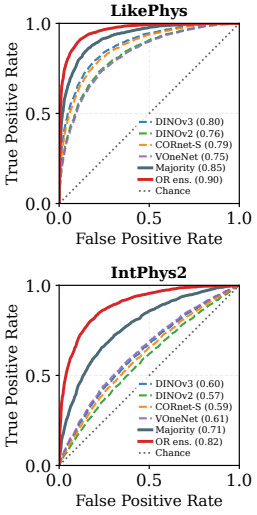

Figure 6: **ROC for physics violation detection.** LikePhys (top) and IntPhys2 (bottom).

ROC analysis. Pairwise accuracy tests within-pair ranking only. ROC analysis (Fig. 6) additionally tests cross-pair ranking under a global threshold on the continuous score: z_b for each individual backbone, $\sum_b z_b$ for Majority, and $\max_b |z_b|$ for OR. Smooth curves at every operating point indicate the score is aligned with implausibility, not locally fitted to the pairwise rule. AUCs reach 75–80% on LikePhys (Majority 85%, OR 90%) and 57–61% on IntPhys2 (Majority

71%, OR 82%); the ensemble lift over the best single backbone is +10 points on LikePhys and +21 points on IntPhys2.

Backbone analysis. Each backbone covers a different set of violations (Fig. 5). Stepwise addition lifts the OR ensemble to 98.3% on LikePhys (from 78.6%) and 93.3% on IntPhys2 (from 58.8%). On LikePhys, CORnet-S wins flag, VOneNet wins fluid and shadow, and DINOv2/v3 capture faucet and river, exceeding 70 points on a single scenario (Faucet: DINOv3 97 vs VOneNet 19). On IntPhys2, V1-like layers preserve signal while DINOs require a mid-layer; mid-level features beat late ones across all backbones (Appendix B). The per-condition winners follow: VOneNet on permanence (+4 over V-JEPA), CORnet-S on immutability (+1 over Gemini 2.5), DINOv2 on solidity (+2 over V-JEPA 2). Only on continuity does V-JEPA 2 give an edge, but majority vote (78%) and OR (94%) recover better performance.

→ **Takeaway 2.** GeoPhys unlocks physics plausibility detection: every backbone surpasses all SoTA baselines on both benchmarks (LikePhys [1] 98% & IntPhys2 [2] 93%).

4.3 Study 3: Improved Physically Plausible Video Generation

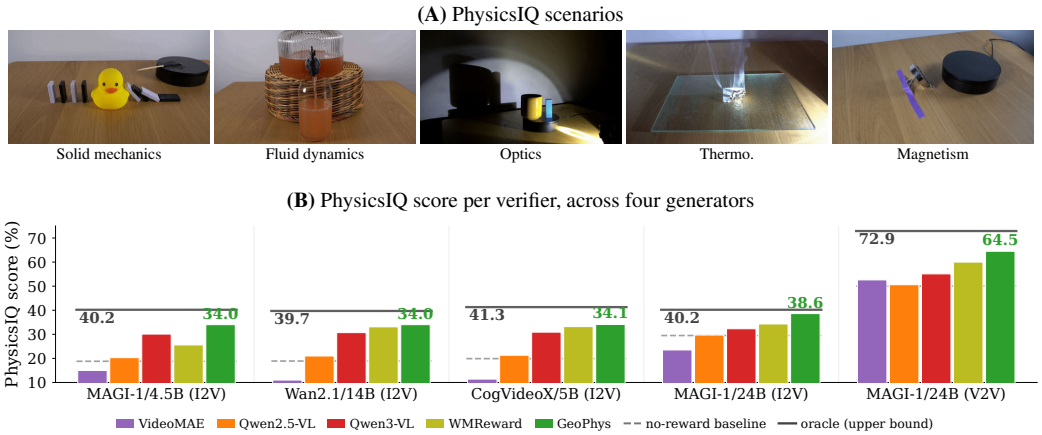

Figure 7: **PhysicsIQ benchmark and GeoPhys performance. (A)** The five PhysicsIQ scenario categories. **(B)** Best-of-N (N=16) PhysicsIQ score for five verifiers on five generators; dashed strokes: no-verifier baseline, solid: oracle upper bound.

Setup. We follow the test-time scaling setup of [48]: a generator samples candidates, a verifier ranks them, and an oracle bounds the achievable ceiling. We test all three on PhysicsIQ [3], which provides 198 scenarios, each with a 3 s conditioning video and a held-out 5 s ground truth. Outputs are scored by four motion-mask metrics aggregated into a single PhysicsIQ score (Appendix F). For each generator, we sample N=16 candidates per scenario. The verifier selects one. GeoPhys reuses the per-backbone signals and OR rule defined in Sec. 3.2 unchanged. We evaluate four I2V generators (MAGI-1 4.5B distill [40], CogVideoX-5B [46], Wan2.1 14B [44], MAGI-1 24B) and one V2V setting (MAGI-1 24B). For scaling curves, we re-rank random subsets of size N from the same pool (Fig. 8). N=1 is no-reranking, reported as mean±std across 198 scenarios.

Baselines. We compare four inference-time BoN baselines. **VideoMAE(BoN)** [47] uses pixel-space

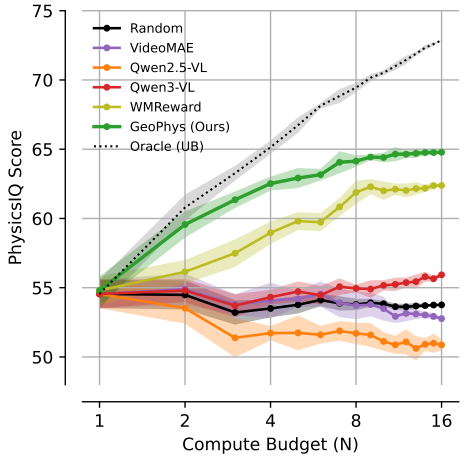

Figure 8: **Test-time scaling on MAGI-1 24B (V2V).** PhysicsIQ score as a function of the candidate budget $N \in \{1, \ldots, 16\}$. GeoPhys scales closest to oracle.

Table 2: **PhysicsIQ results** (N=16). For each generator, four motion-mask metrics (Sp. IoU, Sp-Temp. IoU, W. Sp. IoU, MSE) and PhysicsIQ score (%); Δ over the baseline ( ).

	MAGI-1 4.5B (I2V)					Wan2.1 14B (I2V)					CogVideoX-5B (I2V)				
Verifier	Sp	SpT	W.Sp	MSE	PhyIQ	Sp	SpT	W.Sp	MSE	PhyIQ	Sp	SpT	W.Sp	MSE	PhyIQ
Baseline	.143	.133	.070	.019	18.8	.143	.133	.070	.018	18.9	.143	.133	.070	.008	19.9
+ VideoMAE	.122	.044	.042	.029	15.0	.122	.044	.042	.013	11.0	.122	.044	.042	.008	11.4
+ Qwen2.5-VL	.168	.132	.072	.018	20.3	.168	.132	.072	.011	21.0	.168	.132	.072	.008	21.3
+ Qwen3-VL	.227	.193	.117	.015	30.1	.227	.193	.117	.009	30.7	.227	.193	.117	.007	30.9
+ WMReward [10]	.148	.205	.082	.012	25.6	.235	.210	.130	.008	33.1	.235	.210	.130	.007	33.2
+ GEOPHYS (Ours)	**.225**	**.232**	**.128**	**.007**	**34.0**	**.225**	**.232**	**.128**	**.007**	**34.0**	**.225**	**.232**	**.128**	**.006**	**34.1**
Oracle (Upper bound)	*.304*	*.245*	*.161*	*.017*	*40.2*	*.297*	*.236*	*.158*	*.010*	*39.7*	*.302*	*.246*	*.162*	*.007*	*41.3*

	MAGI-1 24B (I2V)					MAGI-1 24B (V2V)				
Verifier	Sp	SpT	W.Sp	MSE	PhyIQ	Sp	SpT	W.Sp	MSE	PhyIQ
Baseline	.245	.155	.142	.011	29.5	.413	.265	.288	.003	50.01
+ VideoMAE	.205	.122	.119	.016	23.5	.397	.270	.276	.003	52.6
+ Qwen2.5-VL	.262	.134	.156	.013	29.7	.401	.238	.273	.003	50.6
+ Qwen3-VL	.260	.174	.157	.011	32.3	.415	.282	.291	.003	55.1
+ WMReward [10]	.234	.232	.148	.009	34.3	.430	.314	.312	.003	62.29
+ GEOPHYS (Ours)	**.276**	**.235**	**.181**	**.008**	**38.6**	**.472**	**.336**	**.346**	**.003**	**64.50**
Oracle (Upper bound)	*.306*	*.224*	*.195*	*.009*	*40.2*	*.522*	*.363*	*.393*	*.002*	*72.9*

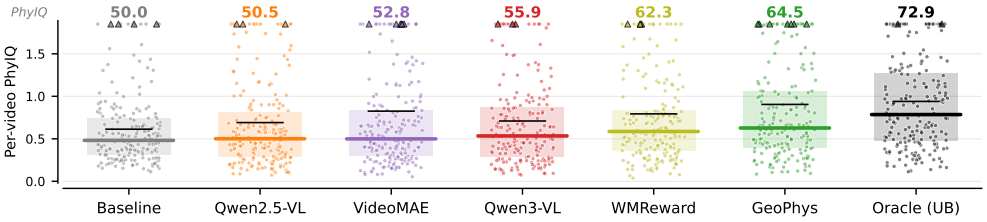

Figure 9: **PhysicsIQ distributions on V2V** (MAGI-1 24B, N=16). Each point is one scenario; the coloured bar is the median, the black the mean, the shaded band the IQR, and triangles are outliers. GEOPHYS shifts the whole distribution towards the Oracle ceiling, not only the mean.

reconstruction error on masked spatiotemporal patches as a surprise score. **Qwen2.5-VL(BoN)** and **Qwen3-VL(BoN)** [49] prompt a multimodal-LLM with a binary physics-plausibility question and use the positive-token logit as the score. **WMReward(BoN)** [10] uses V-JEPA 2 latent prediction error [11], a 1.1B-parameter video-pretrained world model. We also report (supervised) *Oracle* upper bound: the candidate with the highest PhysicsIQ score per video.

Findings. GEOPHYS outperforms every inference-time verifier baseline on all matched-pool generators (Table 2). On MAGI-1 4.5B it scores 34.0 (+15.2 over no verifier). The next-best verifier, Qwen3-VL, scores 30.1. On Wan2.1 14B, CogVideoX-5B, and MAGI-1 24B it scores 34.0, 34.1, and 38.6, within 5.7, 7.2, and 1.6 points of the Oracle upper bound respectively. No backbone dominates: DINOv3 L18 angle consistency (+10.32) and VOneNet V1 acceleration (+10.26) lead on MAGI-1 4.5B, with CORnet-S IT speed variation (+7.64) and DINOv2 L12 angle consistency (+3.06) trailing; the OR ensemble reaches +15.21 (App. H). GEOPHYS also scales sharply on MAGI-1 24B V2V (Fig. 8). As N grows, Qwen2.5-VL and Qwen3-VL plateau near baseline. WMReward [10] reaches ∼62 at N=12. GEOPHYS rises to ∼64. The per-scenario distributions (Fig. 9) show the gain is broad rather than concentrated: GEOPHYS raises both the median and the upper quartile above every other real selector and towards the Oracle, so most scenarios improve, not just a few large wins (GEOPHYS Qualitative examples in App. G).

Video-quality metrics. We score the V2V MAGI-1 24B (N=16) on four video-quality metrics against PhysicsIQ (Fig. 10; full ± SE in Table 11), each with a scenario-level standard error (SE): FVD [77] (Fréchet distance between the real and generated clip distributions in I3D feature space), LPIPS [78] (per-frame perceptual distance to the real continuation), FVMD [79] (Fréchet distance between tracked-keypoint motion features), and VBench [80] (a no-reference quality suite; we report VBench-Q, the mean of its six quality dimensions). FVD and LPIPS measure fidelity to the real continuation, which a plausible video matches more closely; both track PhysicsIQ ($\rho = -0.96, -0.89$), and GEOPHYS lead on both. FVMD does not track PhysicsIQ ($\rho = -0.32$): its SE of 30 to 47 spans the entire 44-point range, leaving GEOPHYS (162 ± 47) and the baseline

(152 ± 38) indistinguishable. VBench-Q is flat ($\rho = +0.07$, $\leq 0.4\%$ spread within SE), so generic quality is unchanged; the full breakdown is in Appendix H.2. Thus GEOPHYS's gains are confined to the physics-fidelity metrics (FVD, LPIPS); generic quality (VBench-Q) is unchanged and motion distribution (FVMD) is unresolvable. The only VBench dimension that moves, dynamic degree, falls as PhysicsIQ rises ($\rho \approx -0.87$), so GEOPHYS suppresses spurious motion and selects for physical plausibility, not only visual quality.

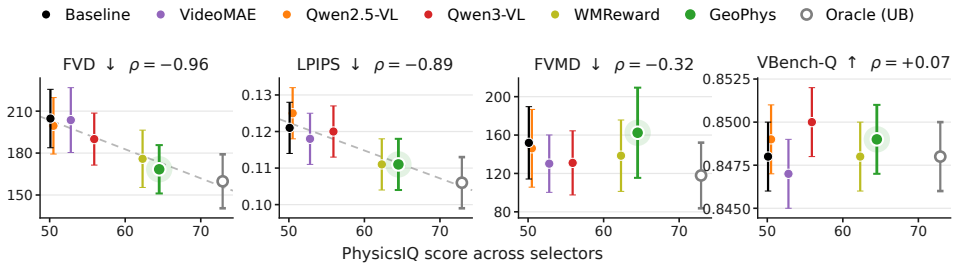

Figure 10: **Video-quality metrics** (MAGI-1 24B, V2V, N=16). Each metric against the PhysicsIQ score; bars are scenario-level standard errors. Full $\pm$ SE values in Table 11.

Compute footprint. GEOPHYS's verifier path scores at 0.25 s/video and 1.2 GB VRAM with a single backbone (DINOv3 ViT-L), and 1.0 s/video and 2.0 GB with the ensemble (Fig. 11; full breakdown in Appendix I). WMReward takes 1.5 s/video and 9.3 GB. GEOPHYS reaches better PhysicsIQ at 1.5× lower wall-clock and 4.65× lower memory. Because test-time scaling compounds the per-candidate cost, this matters: at the same wall-clock budget GEOPHYS supports roughly 5× more candidates than WMReward, compounding the per-candidate quality gain.

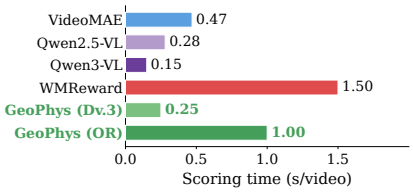

Figure 11: **GEOPHYS inference time per video** (N=16, single H100).

→ **Takeaway 3.** Test-time compute scales physically plausible video generation, but only with the right verifier. GEOPHYS closes the gap to the oracle ceiling faster than world-model and other verifiers, at 4.65× less compute, using only frozen image encoders.

5 Discussion and Conclusion

Summary. The geometry of feature trajectories through frozen image encoders is a strong, training-free signal for physical plausibility. A single score tracks human EEG responses to object-permanence violations (Sec. 4.1); sets SoTA detection on LikePhys (98.3%) and IntPhys2 (93.3%), surpassing twelve diffusion models, MLLM judges, and video-pretrained world models (Sec. 4.2); and lifts MAGI-1 from 50.01% to 64.50% on PhysicsIQ at 1.5× lower wall-clock and 4.65× lower memory than V-JEPA 2 verifiers (Sec. 4.3). For test-time scaling, our results extend the literature [55, 56, 57] to physically plausible video under a much weaker verifier. Best-of-N exploits the asymmetry between generation and verification, reminiscent of the asymmetry that, in its formal limit, separates P from NP: generating a physically plausible continuation is hard, recognising one need not be. The fact that a frozen image encoder, orders of magnitude cheaper than the generator and trained on neither video nor physics, recovers most of the oracle gap suggests that for many applications the bottleneck is not sophisticated verification but having any verifier with the right inductive bias.

Implications for neuroscience. Peri-occlusion, per-frame GEOPHYS signals reproduce the time course and load-scaling of the contralateral delay activity (CDA) [6, 7], an EEG marker of visual working memory. The encoders see only static images, with no working-memory, object-permanence, or physics objective. One reading: the CDA reflects IT cortex object-load signatures, which CORnet-S IT is modelled after. IT recognises objects [81]; the CDA scales with object count, not feature complexity [82]. Geometric features suffice to mimic the CDA's surface properties without claiming a shared mechanism. Replication on larger VOE corpora is the natural next step.

Limitations. We do not claim frozen backbones represent or simulate physics; GEOPHYS is a correlate of plausibility, not an implementation. One shortcoming is that our signals are time-symmetric: a backwards-played video traces the same feature-space trajectory as the forward original, consistent with evidence that video encoder features are broadly time-symmetric [83]. Moreover, the test-time-scaling result is bounded by candidate diversity: PhysicsIQ's oracle ceiling is below 100%, so selection is constrained by what generators produce. We see GEOPHYS as a diagnostic for current generators, not a substitute for genuine world-models [31, 32].

6 Acknowledgements

We would like to thank Andrea Castellani, Sebastian Schmitt, Xavier Bonet-monroig, Linus Ekstrøm, Riccardo Cadei for insightful discussions and feedback. Christian Internò acknowledges funding from the Honda Research Institute Europe. This work was performed with assistance from the US National Institutes of Health Grant S10OD028632-01.

References

[1] 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. *arXiv preprint arXiv:2510.11512*, 2025. (Cited on pages 1, 2, 2, 2, 3, 5, 5, 5, 6, 6, 7, 7, 7, 8, and 21.)

[2] Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025. (Cited on pages 1, 2, 2, 2, 3, 5, 6, 6, 7, 7, 7, 7, 8, 18, 21, and 22.)

[3] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles?, 2025. URL https://arxiv.org/abs/2501.09038. (Cited on pages 1, 2, 2, 3, 5, 8, and 22.)

[4] Francesco Margoni, Luca Surian, and Renée Baillargeon. The violation-of-expectation paradigm: A conceptual overview. *Psychological Review*, 131(3):716, 2024. (Cited on pages 1 and 5.)

[5] Kirsty Dunn and J Gavin Bremner. Investigating looking and social looking measures as an index of infant violation of expectation. *Developmental Science*, 20(6):e12452, 2017. (Cited on pages 1 and 5.)

[6] Halely Balaban, Kevin A Smith, Joshua B Tenenbaum, and Tomer D Ullman. Electrophysiology reveals that intuitive physics guides visual tracking and working memory. *Open Mind*, 8: 1425–1446, 2024. (Cited on pages 1, 2, 5, 5, 5, 6, 6, 6, 6, 10, 20, and 21.)

[7] Shari Liu, Kirsten Lydic, Lingjie Mei, and Rebecca Saxe. Violations of physical and psychological expectations in the human adult brain. *Imaging Neuroscience*, 2:imag–2–00068, 02 2024. ISSN 2837-6056. doi: 10.1162/imag_a_00068. URL https://doi.org/10.1162/imag_a_00068. (Cited on pages 1, 2, 2, 2, 5, 10, and 21.)

[8] 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. URL https://arxiv.org/abs/2502.11831. (Cited on pages 1 and 3.)

[9] Peiyao Wang, Weining Wang, and Qi Li. Physcorr: Dual-reward dpo for physics-constrained text-to-video generation with automated preference selection, 2025. URL https://arxiv.org/abs/2511.03997. (Cited on page 1.)

[10] Jianhao Yuan, Xiaofeng Zhang, Felix Friedrich, Nicolas Beltran-Velez, Melissa Hall, Reyhane Askari-Hemmat, Xiaochuang Han, Nicolas Ballas, Michal Drozdzal, and Adriana Romero-Soriano. Inference-time physics alignment of video generative models with latent world models. *arXiv preprint arXiv:2601.10553*, 2026. (Cited on pages 1, 2, 2, 9, 9, 9, 9, and 26.)

[11] 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. URL https://arxiv.org/abs/2506.09985. (Cited on pages 1, 2, 2, 2, 6, 7, 7, and 9.)

[12] Saman Motamed, Minghao Chen, Luc Van Gool, and Iro Laina. Travl: A recipe for making video-language models better judges of physics implausibility, 2025. URL https://arxiv.org/abs/2510.07550. (Cited on pages 1 and 2.)

[13] Kevin Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Joshua B. Tenenbaum, and Tomer D. Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. (Cited on pages 1 and 5.)

[14] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4509–4517, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819. (Cited on page 1.)

[15] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/d09bf41544a3365a46c9077ebb5e35c3-Paper.pdf. (Cited on page 1.)

[16] Hany Farid. Lighting (in)consistency of paint by text, 2022. URL https://arxiv.org/abs/2207.13744. (Cited on page 1.)

[17] Eric Kee, James F. O'brien, and Hany Farid. Exposing photo manipulation from shading and shadows. *ACM Trans. Graph.*, 33(5), September 2014. ISSN 0730-0301. doi: 10.1145/2629646. URL https://doi.org/10.1145/2629646. (Cited on page 1.)

[18] Sarah Barrington and Hany Farid. Distinguishing authentic from ai-generated explosions using spatiotemporal dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 10659–10667, June 2026. (Cited on page 1.)

[19] Hany Farid. Perspective (in)consistency of paint by text, 2022. URL https://arxiv.org/abs/2206.14617. (Cited on page 1.)

[20] Peter Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2): 194–200, 1991. (Cited on page 2 and 2.)

[21] Laurenz Wiskott and Terrence J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 04 2002. ISSN 0899-7667. doi: 10.1162/089976602317318938. URL https://doi.org/10.1162/089976602317318938. (Cited on page 2 and 2.)

[22] Patrice Simard, Bernard Victorri, Yann LeCun, and John Denker. Tangent prop: A formalism for specifying selected invariances in an adaptive network. In *Advances in Neural Information Processing Systems*, volume 4, pages 895–903, 1991. (Cited on page 2 and 2.)

[23] Rajesh Rao and Daniel Ruderman. Learning lie groups for invariant visual perception. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1998. URL https://proceedings.neurips.cc/paper_files/paper/1998/file/277281aada22045c03945dcb2ca6f2ec-Paper.pdf. (Cited on page 2 and 2.)

[24] Ross Goroshin, Michael Mathieu, and Yann LeCun. Learning to linearize under uncertainty. 2015. URL https://arxiv.org/abs/1506.03011. (Cited on pages 2, 2, and 4.)

[25] Olivier J. Hénaff and Eero P. Simoncelli. Geodesics of learned representations. *CoRR*, abs/1511.06394, 2015. URL https://api.semanticscholar.org/CorpusID:2208884. (Cited on page 2.)

[26] Xueyan Niu, Cristina Savin, and Eero P Simoncelli. Learning predictable and robust neural representations by straightening image sequences. *Advances in Neural Information Processing Systems*, 37:40316–40335, 2024. (Cited on page 2 and 2.)

[27] Olivier J Hénaff, Robbe LT Goris, and Eero P Simoncelli. Perceptual straightening of natural videos. *Nature Neuroscience*, 22(6):984–991, 2019. (Cited on pages 2, 3, 5, 18, and 18.)

[28] Olivier J Hénaff, Yoon Bai, Julie A Charlton, Ian Nauhaus, Eero P Simoncelli, and Robbe LT Goris. Primary visual cortex straightens natural video trajectories. In *Nature Communications*, 2021. (Cited on pages 2, 2, 3, 5, and 18.)

[29] Anne Harrington, Vasha DuTell, Ayush Tewari, Mark Hamilton, Simon Stent, Ruth Rosenholtz, and William T. Freeman. Exploring perceptual straightness in learned visual representations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=4cOfD2qL6T. (Cited on page 2 and 2.)

[30] Anne W. Zonneveld, Pascal Mettes, and Iris Groen. Straightening of natural visual sequences in video DNNs: the role of locality and temporal coherence. In *Cognitive Computational Neuroscience (CCN)*, Amsterdam, The Netherlands, 2025. URL https://2025.ccneuro.org/poster/?id=EJMZ9Os8jG. (Cited on pages 2, 2, and 3.)

[31] Keyon Vafa, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. What has a foundation model found? inductive bias reveals world models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=i9npQatSev. (Cited on pages 2 and 11.)

[32] Christian Internò, Jumpei Yamaguchi, Loren Amdahl-Culleton, Markus Olhofer, David Klindt, and Barbara Hammer. The observer effect in world models: Invasive adaptation corrupts latent physics. *arXiv preprint arXiv:2602.12218*, 2026. (Cited on pages 2 and 11.)

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification. (Cited on pages 2, 3, 5, 7, 7, and 18.)

[34] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104. (Cited on pages 2, 5, 7, 7, 18, and 18.)

[35] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent ANNs. In *Advances in Neural Information Processing Systems*, 2019. (Cited on pages 2, 5, 7, 7, and 18.)

[36] Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. In *Advances in Neural Information Processing Systems*, 2020. (Cited on pages 2, 3, 5, 7, 7, and 19.)

[37] Daniel Kaiser, Rico Stecher, and Katja Doerschner. Eeg decoding reveals neural predictions for naturalistic material behaviors. *Journal of Neuroscience*, 43(29):5406–5413, 2023. (Cited on pages 2, 5, and 21.)

[38] OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774. (Cited on pages 2, 2, and 7.)

[39] Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805. (Cited on pages 2, 2, 7, and 7.)

[40] Sand AI. MAGI-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025. (Cited on pages 2 and 8.)

[41] Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines, 2022. URL https://arxiv.org/abs/2106.08261. (Cited on page 2.)

[42] Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. IntPhys: A benchmark for visual intuitive physics reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5016–5025, 2022. (Cited on page 2.)

[43] Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. 2024. URL https://arxiv.org/abs/2311.09048. (Cited on page 2.)

[44] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. (Cited on pages 2, 7, 7, and 8.)

[45] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. (Cited on pages 2, 6, and 7.)

[46] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. (Cited on pages 2, 7, 7, 7, and 8.)

[47] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=AhccnBXSne. (Cited on pages 2, 2, 7, and 8.)

[48] Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-t1: Test-time scaling for video generation. *arXiv preprint arXiv:2503.18942*, 2025. (Cited on pages 2 and 8.)

[49] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin

Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388. (Cited on pages 2, 7, and 9.)

[50] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of EMNLP*, 2024. (Cited on page 2.)

[51] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. (Cited on page 2.)

[52] Tengjiao Yin, Jinglei Shi, Heng Guo, and Xi Wang. Vigor: Video geometry-oriented reward for temporal generative alignment, 2026. URL https://arxiv.org/abs/2603.16271. (Cited on page 2.)

[53] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. (Cited on page 2.)

[54] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *Proceedings of ICLR*, 2024. (Cited on page 2.)

[55] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. (Cited on pages 2 and 10.)

[56] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024. (Cited on pages 2 and 10.)

[57] Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Jp988ELppQ. (Cited on pages 2 and 10.)

[58] Yuta Oshima, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Inference-time text-to-video alignment with diffusion latent beam search. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=c9EAmyYPOv. (Cited on page 2.)

[59] Lorenzo Baraldi, Davide Bucciarelli, Zifan Zeng, Chongzhe Zhang, Qunli Zhang, Marcella Cornia, Lorenzo Baraldi, Feng Liu, Zheng Hu, and Rita Cucchiara. Verifier matters: Enhancing inference-time scaling for video diffusion models. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA, 2025. URL https://bmva-archive.org.uk/bmvc/2025/assets/papers/Paper_1006/paper.pdf. (Cited on page 2.)

[60] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. (Cited on page 2.)

[61] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. (Cited on page 2.)

[62] Christian Internò, Robert Geirhos, Markus Olhofer, Sunny Liu, Barbara Hammer, and David Klindt. Ai-generated video detection via perceptual straightening. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 20672–20705. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1d9a43752c2819e03967c5c1b708169c-Paper-Conference.pdf. (Cited on pages 3 and 5.)

[63] Wenhan Chen, Sezer Karaoglu, and Theo Gevers. Grab-3d: Detecting ai-generated videos from 3d geometric temporal consistency, 2025. URL https://arxiv.org/abs/2512.13665. (Cited on page 3.)

[64] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. URL http://www.blender.org. (Cited on pages 3 and 7.)

[65] Andrew Melnik, Michal Ljubljanac, Cong Lu, Qi Yan, Weiming Ren, and Helge Ritter. Video diffusion models: A survey, 2024. URL https://arxiv.org/abs/2405.03150. (Cited on page 3.)

[66] Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B. Issa, and James J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6):974–983, 2019. (Cited on page 5.)

[67] David J. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9 (2):181–197, 1992. (Cited on page 5.)

[68] Matteo Carandini and David J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012. (Cited on page 5.)

[69] Akiko Ikkai, Andrew W McCollough, and Edward K Vogel. Contralateral delay activity provides a neural measure of the number of representations in visual working memory. *Journal of neurophysiology*, 103(4):1963–1968, 2010. (Cited on page 5.)

[70] Andrew W McCollough, Maro G Machizawa, and Edward K Vogel. Electrophysiological measures of maintaining representations in visual working memory. *Cortex*, 43(1):77–94, 2007. (Cited on page 5.)

[71] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Fx2SbBgcte. (Cited on page 7 and 7.)

[72] NVIDIA team. Cosmos world foundation model platform for physical ai, 2025. URL https://arxiv.org/abs/2501.03575. (Cited on page 7.)

[73] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report, 2023. URL https://arxiv.org/abs/2308.06571. (Cited on page 7 and 7.)

[74] Genmo Team. Mochi 1: A new SOTA in open-source video generation models. https://github.com/genmoai/mochi, 2024. (Cited on page 7.)

[75] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2024. URL https://arxiv.org/abs/2501.00103. (Cited on page 7.)

[76] Epic Games. Epic content license agreement, 2026. URL https://www.unrealengine.com/eula/content. Accessed: 2026-04-22. (Cited on page 7.)

[77] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019. URL https://arxiv.org/abs/1812.01717. (Cited on page 9.)

[78] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. (Cited on page 9.)

[79] Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fréchet video motion distance: A metric for evaluating motion consistency in videos, 2024. URL https://arxiv.org/abs/2407.16124. (Cited on page 9.)

[80] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. (Cited on page 9.)

[81] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012. (Cited on page 10.)

[82] Halely Balaban and Roy Luria. The number of objects determines visual working memory capacity allocation for complex items. *NeuroImage*, 119:54–62, 2015. (Cited on page 10.)

[83] Piyush Nitin Bagad and Andrew Zisserman. Chirality in action: Time-aware video representation learning by latent straightening. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=hchAR53gA0. (Cited on page 11.)

[84] Judson P. Jones and Larry A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987. (Cited on page 19.)

[85] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90. (Cited on page 19.)

Appendix Contents

A Perceptual straightening analysis

At each backbone's best layer (Table 1), we quantify the plausible-vs-violated difference as a paired Cohen's d together with a paired t-test (Table 3). All four backbones produce highly significant differences ($p < 10^{-8}$), with effect sizes ranging from $d = 0.22$ (VOneNet V1) to $d = 0.44$ (CORnet-S IT). Note that these are *turning-angle* effect sizes (straightening [27]).

Table 3: Perceptual-straightening test at each backbone's best layer. $\bar{\phi}^+/\bar{\phi}^-$: mean turning angle on plausible/violated videos. d: paired Cohen's d. Paired t-test over 650 video pairs.

Backbone	Layer	$\bar{\phi}^+$	$\bar{\phi}^-$	Δ	d	p
DINOv2	L12	1.742	1.807	+0.065	0.34	2.2×10^{-17}
DINOv3	L18	1.778	1.836	+0.058	0.34	1.7×10^{-17}
CORnet-S	V1	1.591	1.708	+0.117	0.44	2.0×10^{-26}
VOneNet	V1	2.027	2.042	+0.016	0.22	1.9×10^{-8}

The direction of the straightening effect is universal (violated $>$ plausible, $p < 10^{-8}$, all 57 layer and backbone combinations). The magnitude varies with depth and architecture (Fig. 12, Table 4).

DINOv2 [33] d rises from 0.14 at layer 1 to 0.37, matching the primate V1→IT profile reported by Hénaff et al. [27, 28]. A self-supervised ViT trained only on multi-view invariance spontaneously acquires the same depth-wise gradient observed in biological vision.

DINOv3 [34] d peaks at layer 8 ($d = 0.42$) and descends slightly thereafter. The register tokens introduced in DINOv3 [34] reorganise late-layer representations to reduce attention artefacts, partially attenuating the geometric-straightness signal.

CORnet-S [35] V1 produces the strongest effect of any layer of any backbone ($d = 0.58$), declining through V2 (0.49), V4 (0.49), and IT (0.44). The layer architecturally designated to mimic primate V1 is also the layer with the sharpest plausibility signal—consistent with the IntPhys2 [2] detection results where CORnet-S V1 is also the strongest CORnet-S layer (61.1%).

VOneNet [36] The fixed Gabor front-end produces a weaker but consistent effect ($d \approx 0.20$), essentially unchanged through its ResNet-50 layers [84, 85]. Because the front-end is not trained, the straightening signal emerges at the level of oriented-edge responses tuned to natural-image statistics and does not require further representational depth.

Table 4: Peak Cohen's d for perceptual straightening per backbone.

Backbone	Peak layer	d	p	Profile
DINOv2	L12	0.37	$< 10^{-17}$	Monotonic rise
DINOv3	L18	0.42	$< 10^{-21}$	Mid-depth peak
CORnet-S	IT	**0.58**	$< 10^{-30}$	Decline from V1
VOneNet	V1	0.22	$< 10^{-8}$	Flat

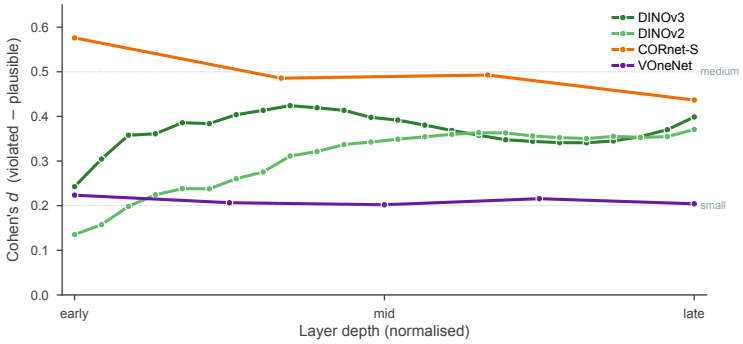

Figure 12: **Cohen's d of the perceptual-straightening effect across normalised layer depth.** Each point is the paired Cohen's d between turning angles of violated and plausible videos, averaged over 650 LikePhys pairs, at one layer of one backbone. All 57 combinations give $d > 0$ ($p < 10^{-8}$). DINOv2 rises monotonically; CORnet-S peaks at V1 (d=0.58); DINOv3 peaks mid-depth; VOneNet is flat.

B Layer sweep analysis

Backbone weights are taken from https://github.com/facebookresearch/dinov2 (DINOv2/v3), https://github.com/dicarlolab/CORnet (CORnet-S), and https://github.com/dicarlolab/vonenet (VOneNet).

LikePhys. Fig. 13 (top) shows per-scenario accuracy vs. layer for each of the five kinematic signals across the four backbones. Two observations motivate the paper's layer choices. First, for the ViT backbones, *angle consistency* peaks at a deep-but-not-final layer (DINOv2 L12, DINOv3 L18) because the very last blocks reorganise features for downstream tasks and lose some of the trajectory regularity. Second, for CORnet-S, *speed variation* grows with depth and peaks at IT, consistent with the recurrent module building temporal sensitivity. VOneNet's best signal is acceleration at the fixed Gabor front-end (V1), which picks up on local curvature changes directly.

IntPhys2. Fig. 13 (bottom) shows per-layer best-signal accuracy on IntPhys2, where the best signal at each layer is the strongest of the five temporal signals in Φ_{temp} (Eq. 1). The trend is the opposite of LikePhys: *mid-level features outperform late features*. DINOv3 L12 (60.5%) beats L18 (58.9%) and L23 (57.5%); CORnet-S V1 (61.1%) beats IT (57.7%). Localised semantic violations—an object vanishing behind an occluder, a ball passing through a wall—are best captured by the mid-level features of DINOv2/v3 and by the V1-like layers of CORnet-S and VOneNet, where local trajectory structure is most cleanly preserved.

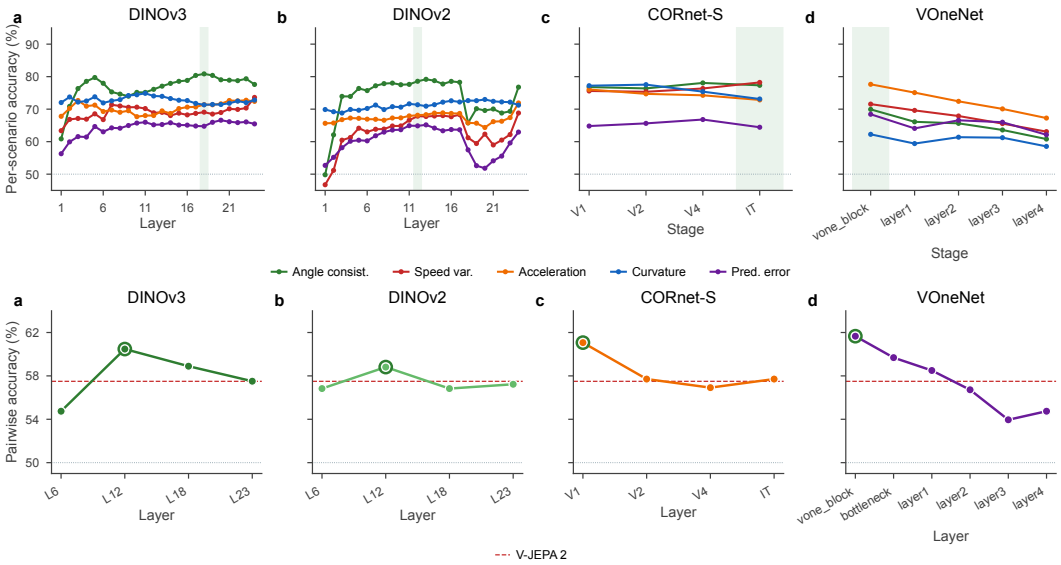

Figure 13: **Layer sweep on LikePhys (top) and IntPhys2 (bottom).** For each backbone, we report per-layer signal accuracy and select the readout layer used in the main text. LikePhys favours deeper layers for ViTs and IT for CORnet-S; IntPhys2 favours mid-level (and V1-like) features.

C Additional VOE results

In Fig. 4, we demonstrate model-brain alignment on object permanence violations in which one object enters occlusion but two exit (the 'Create' scenario), and matched control videos. In Fig. 14, we present a complementary scenario, 'Vanish,' in which two objects enter occlusion but only one exits. For both scenarios, we report the average difference between GEOPHYS signals on valid and invalid videos post-occlusion in Table 5, baseline corrected so that the pre-occlusion difference (when videos are approximately identical) is exactly zero. This controls for minor variations introduced by ADEPT rendering. Finally, Table 6 reports overall GEOPHYS detection performance on additional VOE datasets from neuroscience.

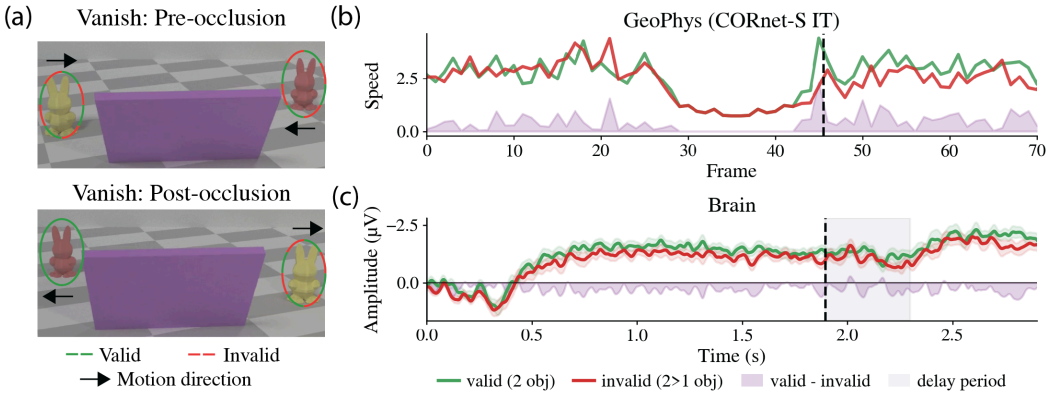

Figure 14: **Detecting object permanence violations in models and brains: Vanish condition** (a) Example VOE stimuli for the Vanish scenario: in valid videos, two objects enter and exit occlusion; in the invalid videos, two objects enter but one exits. (b–c) valid and invalid Vanish signals from GEOPHYS CORnet-S IT (speed) (b) and EEG contralateral delay activity (from [6]; c). Both GEOPHYS and brain signals are elevated for the valid condition after occlusion offset (dashed line).

	CORnet-S IT					*DINOv2*					*DINOv3*					*VOneNet*				
	Spd	Acc	Curv	Ang	Pred	Spd	Acc	Curv	Ang	Pred	Spd	Acc	Curv	Ang	Pred	Spd	Acc	Curv	Ang	Pred
Create $\bar{x}$	-0.25	-0.33	-0.03	+0.03	-0.33	-0.00	-0.01	-0.03	+0.03	-0.01	+0.13	+0.29	+0.02	-0.02	+0.29	+0.00	+0.00	-0.01	+0.01	+0.00
Create σ	0.13	0.32	0.09	0.09	0.32	0.01	0.01	0.08	0.07	0.01	0.17	0.50	0.14	0.13	0.50	0.01	0.02	0.03	0.03	0.02
Vanish $\bar{x}$	+0.16	+0.16	-0.01	+0.01	+0.16	+0.00	+0.01	+0.06	-0.06	+0.01	-0.04	+0.01	+0.06	-0.06	+0.01	+0.00	+0.00	-0.02	+0.02	+0.00
Vanish σ	0.14	0.21	0.05	0.05	0.21	0.01	0.03	0.10	0.10	0.03	0.15	0.28	0.06	0.06	0.28	0.01	0.02	0.03	0.03	0.02

Table 5: Mean valid − invalid GEOPHYS signal post-occlusion (baseline-subtracted). $\bar{x}/\sigma$ across 7 pairs. Spd=Speed, Acc=Acceleration, Curv=Curvature, Ang=Angle Cons., Pred=Pred. Error.

Table 6: **Neural signal alignment detection.** Pairwise accuracy ($n_{\text{correct}}/n_{\text{total}}$, ↑ 0% 100%) on VOE datasets. Perman. = Permanence, Solid. = Solidity, Mat. Viol. = Material Violations.

	Balaban [6]	**Liu** [7]		**Kaiser** [37]
	Perman.	Perman.	Solid.	Mat. Viol.
GeoPhys individual				
DINOv2	**2/2**	4/6	3/6	2/4
DINOv3	1/2	5/6	4/6	0/4
CORnet-S	1/2	2/6	4/6	2/4
VOneNet	**2/2**	3/6	3/6	3/4
GeoPhys ensembles				
Majority	1/2	3/6	4/6	2/4
OR	**2/2**	**6/6**	**6/6**	**4/4**

D Signal breakdown

Table 7 reports the full accuracy and effect size for every backbone × GEOPHYS signal combination on LikePhys. Each backbone specialises in a different signal: CORnet-S on speed variation (d=0.96, the largest single-signal effect size), DINOv2/v3 on angle consistency (d=0.73–0.76), and VOneNet on acceleration (d=0.37). Prediction error provides moderate but consistent signal across all backbones (d=0.25–0.42), never the best but never below chance.

Table 7: Signal breakdown on **LikePhys**: accuracy (%) and Cohen's $|d|$ for every backbone × kinematic signal. **Green**: best signal per backbone.

	Curvature		Speed var.		Acceleration		Angle consist.		Pred. error	
	Acc	$\|d\|$	Acc	$\|d\|$	Acc	$\|d\|$	Acc	$\|d\|$	Acc	$\|d\|$
DINOv2 L12	71	.40	68	.43	68	.31	**79**	**.73**	66	.30
DINOv3 L18	71	.41	69	.55	71	.34	**81**	**.76**	67	.32
CORnet-S IT	73	.51	**78**	**.96**	73	.58	77	.68	70	.42
VOneNet V1	62	.13	72	.46	**78**	**.37**	70	.36	64	.25

E Additional detection analysis

LikePhys: per-physics-domain accuracy. Table 8 groups the 12 LikePhys scenarios by physical domain following [1]. Each backbone specialises: VOneNet dominates rigid-body and optical domains (92.4%, 74.5%), DINOv3 dominates fluid (88.0%), and DINOv2/CORnet-S share continuum (86.5%/86.0%). The majority vote recovers 80–96% across all four domains, compared to 43.5–63.6% for the best VDM baseline (Hunyuan T2V).

IntPhys2: accuracy by difficulty and camera type. Table 9 breaks down IntPhys2 results by the Easy/Medium/Hard sub-splits and Fixed/Moving camera configurations defined in [2]. Unlike VLM

Table 8: **LikePhys: per-physics-domain accuracy (%).** Underline: best VDM baseline (Hunyuan T2V). **Bold**: best GEOPHYS backbone per domain.

	Rigid	Contin.	Fluid	Optical	Overall
Hunyuan T2V	63.6	43.5	51.3	59.0	56.4
DINOv2 L12	86.8	**86.5**	71.3	61.0	78.6
DINOv3 L18	85.0	**86.5**	**88.0**	54.0	80.8
CORnet-S IT	89.8	86.0	65.0	61.0	78.2
VOneNet V1	**92.4**	74.5	57.3	**74.5**	77.6
Majority	94.6	95.5	92.7	80.5	90.9

baselines that degrade from Easy to Hard (Gemini-2.5 Flash: 64.4%→54.5%), GEOPHYS's majority vote is stable across difficulty levels (69–70%). Fixed-camera scenes are slightly easier (71.4% vs. 67.7%), particularly for CORnet-S V1, whose retinotopic structure benefits from spatial stability.

Table 9: **IntPhys2: accuracy (%) by difficulty and camera type.** Published baselines from [2].

	By difficulty			By camera	
	Easy	Med.	Hard	Fixed	Moving
Published baselines					
GPT-4o	57.7	54.8	54.2	57.2	57.7
Gemini-2.5 Fl.	64.4	56.8	54.5	58.7	58.6
V-JEPA 2	54.0	58.5	59.4	54.8	62.4
GEOPHYS *individual*					
DINOv2 L12	50.0	56.2	52.7	53.1	56.2
DINOv3 L12	51.9	53.8	53.6	55.8	51.2
CORnet-S V1	55.8	**62.0**	59.8	**62.7**	54.5
VOneNet V1	52.9	52.0	50.0	50.5	50.0
GEOPHYS *ensembles*					
Majority	69.6	69.5	68.8	71.4	67.7
OR	85.3	92.0	89.9	91.3	91.1
Human	96.2	97.8	95.5	–	–

F PhysicsIQ benchmark protocol

This appendix details the PhysicsIQ benchmark [3] used in Sec. 4.3 to evaluate GEOPHYS as a best-of-N verifier for video generation.

Dataset. PhysicsIQ contains 396 high-quality videos covering 66 real-world physical scenarios across five categories: solid mechanics (38 scenarios), fluid dynamics (15), optics (8), thermodynamics (3), and magnetism (2). Each scenario was filmed from three perspectives (left, center, right) on a static Sony Alpha a6400 with a 16–50 mm lens, at 30 FPS and 3840×2160 resolution. Each scenario was shot *twice* under identical conditions to capture the inherent randomness of real-world physical interactions. The benchmark spans 198 evaluation scenarios (66×3 perspectives), each 8 seconds long.

Conditioning protocol. Each 8-second video is split into a 3-second conditioning window and a held-out 5-second test continuation that serves as ground truth. Image-to-video (I2V) generators receive only the last frame of the conditioning window (the *switch frame*) which is hand-selected per scenario such that the physical event is set up but has not yet occurred (e.g., the first domino is tipped but has not contacted the second). Video-to-video (V2V) generators receive the full 3-second clip as multi-frame conditioning. Generators that accept text receive a human-written description of the conditioning frames that does not give away the continuation; generators that do not (e.g., Stable Video Diffusion) receive only the visual conditioning.

Motion-mask metrics. The benchmark quantifies physical realism via four motion-mask metrics computed between the generated continuation and the real continuation. A binary $h \times w \times t$ *motion-*

mask video is first extracted by thresholding pixel intensity changes across frames; the four metrics summarise this mask in different ways:

- **Spatial IoU** (*where* action happens). Collapse the binary motion mask across time via a max, giving an $h \times w$ binary motion map. Compute IoU against the real motion map:

$$\text{Spatial-IoU} = \frac{|M^{\text{sp}}_{\text{real}} \cap M^{\text{sp}}_{\text{gen}}|}{|M^{\text{sp}}_{\text{real}} \cup M^{\text{sp}}_{\text{gen}}|}. \tag{2}$$

- **Spatiotemporal IoU** (*where and when* action happens). Frame-by-frame IoU on the $h \times w \times t$ binary mask, averaged across t. A model that gets the location right but the timing wrong scores well on Spatial IoU but poorly here.

$$\text{ST-IoU} = \frac{|M_{\text{real}} \cap M_{\text{gen}}|}{|M_{\text{real}} \cup M_{\text{gen}}|}. \tag{3}$$

- **Weighted Spatial IoU** (*where and how much* action happens). Same as Spatial IoU, but collapse time using the per-frame action density. The metric is the pixel-wise minimum over maximum, distinguishing repeated motion (e.g., pendulum) from one-pass motion (e.g., rolling ball):

$$\text{W-IoU} = \frac{\sum_i \min\big(M^{\text{w}}_{\text{real},i},\, M^{\text{w}}_{\text{gen},i}\big)}{\sum_i \max\big(M^{\text{w}}_{\text{real},i},\, M^{\text{w}}_{\text{gen},i}\big)}. \tag{4}$$

- **MSE** (*how* action happens). Pixel-level mean squared error between generated and real frames. Strict appearance similarity. Sensitive to colour and texture hallucinations. Lower is better:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\big(f_{\text{real},i} - f_{\text{gen},i}\big)^2. \tag{5}$$

Aggregate score. The four metrics are combined into a single *PhysicsIQ score* by summing the three IoU metrics and subtracting MSE (with a sign flip since MSE is inverted), then normalising so that the *physical variance* scores 100%. This defines an empirical ceiling: a model achieving 100% would be indistinguishable in motion-mask terms from a real second take. A score of 0% corresponds to no overlap with reality.

G Qualitative best-of-N comparisons

Beyond the aggregate scores of Sec. 4.3, we show what the GEOPHYS selection looks like for each of the five PhysicsIQ categories (solid mechanics, fluid dynamics, optics, thermodynamics, magnetism) beside the ground truth (Fig. 15). Columns left of the dashed line are the shared 3 s real conditioning; columns to its right are the 5 s continuation, sampled evenly; the per-panel Δ gives the PhysicsIQ gain in points.

H Per-signal GEOPHYS BoN performance

We report the per-backbone breakdown of GEOPHYS as a best-of-N verifier on the three matched-comparison generators of Table 2.

Each backbone helps on a different scenario subset. On every generator, every individual backbone already beats the no-verifier baseline by $+5$–$+11$ PhysicsIQ points (Table 10), but no single backbone dominates across all 198 scenarios. The OR ensemble lifts performance by an additional $+4$–$+5$ points over the strongest single backbone, mirroring the complementarity pattern observed for detection (Fig. 5): different backbones flag different violations, and selecting the most-confident backbone per scenario recovers what each misses.

Single-signal gains. On every generator, the strongest single backbone is the one whose detection-winning signal is second-order or near-second-order: DINOv3 L18 angle consistency (which depends on consecutive displacements and is therefore second-order in the trajectory) and VOneNet V1 acceleration each deliver $+8$ to $+10$ PhysicsIQ points, and CORnet-S IT speed variation contributes a comparable $+7$ to $+11$.

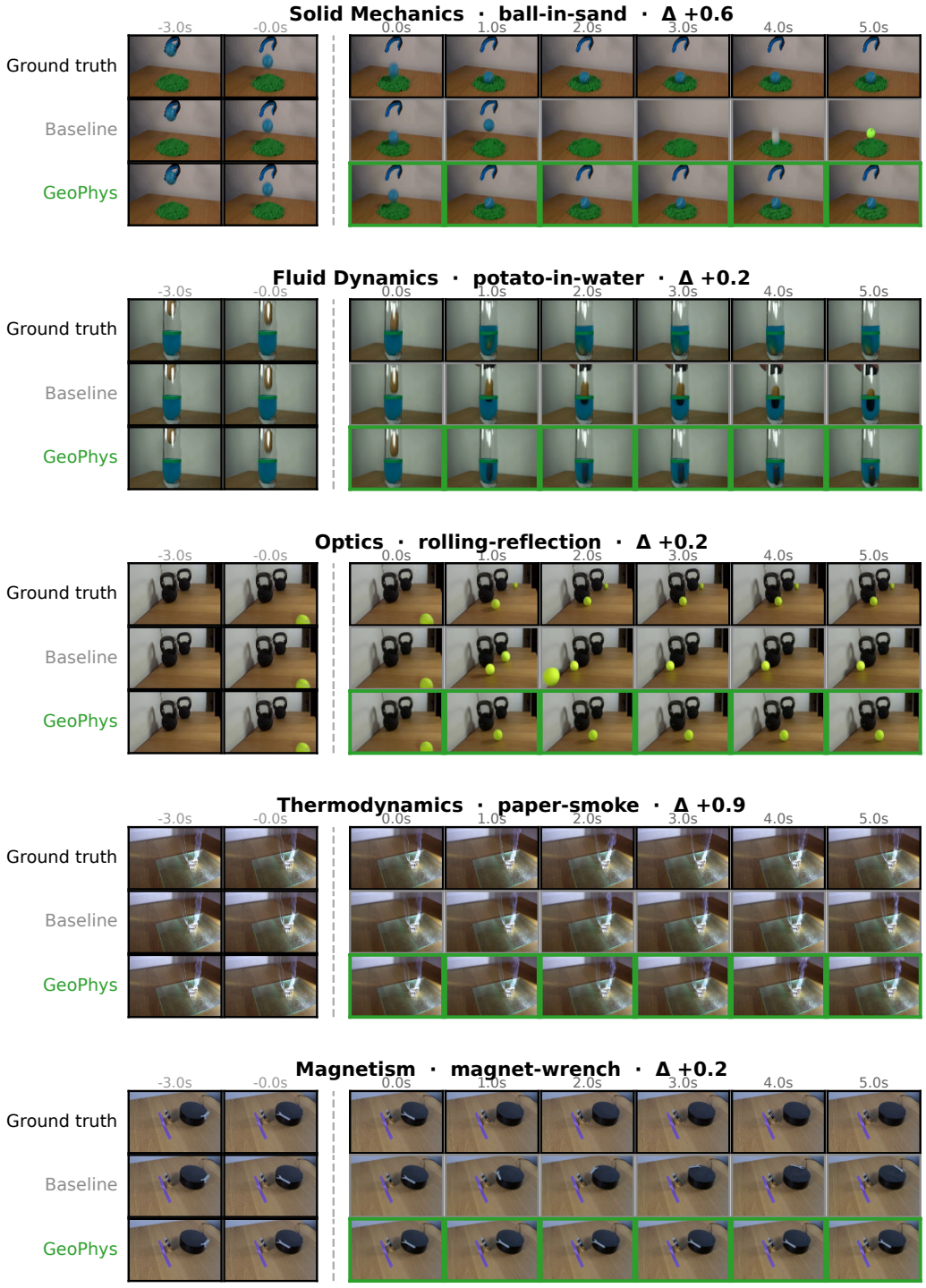

Figure 15: **Qualitative best-of-N selection across physics families** (PhysicsIQ V2V, MAGI-1 24B, N=16). Columns left of the dashed line are the shared real conditioning; columns to its right are the continuations, with the real take (*Ground truth*) for reference.

Table 10: **Per-backbone GEOPHYS BoN performance on the matched candidate pools.** Each row applies the detection-winning backbone–signal pair from Table 7 as the BoN verifier on the same N=16 candidate pool as in Table 2; no signal-search, no PhysicsIQ-specific tuning. The Ensemble row is the OR rule of Sec. 4.2 applied across the four backbones. Δ: gain over the matched no-verifier baseline.

Generator	Backbone (signal)	IoU ↑			MSE ↓	Score ↑	Δ
		Spatial	Spatiotemp.	Weighted Sp.			
MAGI-1 4.5B	Baseline (no verifier)	0.143	0.133	0.070	0.019	18.75	—
	DINOv2 L12 (angle consist.)	0.135	0.178	0.069	0.011	21.81	+3.06
	DINOv3 L18 (angle consist.)	0.130	0.281	0.082	0.008	29.07	+10.32
	CORnet-S IT (speed var.)	0.114	0.266	0.068	0.008	26.39	+7.64
	VOneNet V1 (accel.)	0.132	0.277	0.084	0.008	29.01	+10.26
	Ensemble (OR)	**0.225**	**0.232**	**0.128**	**0.007**	**33.96**	**+15.21**
Wan2.1 14B	Baseline (no verifier)	0.143	0.133	0.070	0.018	18.85	—
	DINOv2 L12 (angle consist.)	0.135	0.178	0.069	0.010	21.95	+3.10
	DINOv3 L18 (angle consist.)	0.130	0.281	0.082	0.007	29.11	+10.26
	CORnet-S IT (speed var.)	0.185	0.215	0.112	0.008	29.81	+10.96
	VOneNet V1 (accel.)	0.223	0.136	0.119	0.009	27.33	+8.48
	Ensemble (OR)	**0.225**	**0.232**	**0.128**	**0.007**	**34.01**	**+15.16**
CogVideoX-5B	Baseline (no verifier)	0.143	0.133	0.070	0.008	19.86	—
	DINOv2 L12 (angle consist.)	0.135	0.178	0.069	0.008	22.04	+2.18
	DINOv3 L18 (angle consist.)	0.130	0.281	0.082	0.007	29.15	+9.29
	CORnet-S IT (speed var.)	0.185	0.215	0.112	0.007	29.87	+10.01
	VOneNet V1 (accel.)	0.223	0.136	0.119	0.007	27.57	+7.71
	Ensemble (OR)	**0.225**	**0.232**	**0.128**	**0.006**	**34.08**	**+14.22**

H.1 Full video-quality metrics

Table 11 gives the complete numbers behind Fig. 10, with a scenario-level standard error on every cell. For the Frechet metrics (FVD, FVMD) we resample the 198 scenarios with replacement and recompute the distance on each resample; for the per-scenario means (LPIPS, VBench-Q) the standard error is $\mathrm{std}/\sqrt{198}$. PhysicsIQ is the reference axis and is reported as a point estimate. The two-family split of the main text is visible directly in the intervals. FVD and LPIPS have standard errors well below the between-selector spread, so their ordering against PhysicsIQ is stable: GEOPHYS attains the lowest FVD and the second-lowest LPIPS among the learnable selectors. FVMD instead has standard errors of 30 to 47, comparable to its full 44-point range, so the selectors overlap and the metric cannot rank them; GEOPHYS at 162.4 ± 47.0 and the baseline at 151.9 ± 37.7 are statistically indistinguishable. VBench-Q varies by 0.4% across selectors, within its ± 0.002 standard error. The Spearman ρ row is taken across the seven selectors; with $n = 7$ these correlations indicate the trend rather than a precise estimate, and the per-cell intervals carry the argument.

Table 11: **Video-quality metrics on PhysicsIQ V2V with scenario-level $\pm$ SE** (MAGI-1 24B, N=16). Point $\pm$ standard error over the 198 scenarios; scenario bootstrap for the Frechet metrics (FVD, FVMD) and $\mathrm{std}/\sqrt{n}$ for the means (LPIPS, VBench-Q). VBench-Q is the mean of VBench's six quality dimensions (breakdown in Table 12). Spearman ρ is vs. PhysicsIQ across selectors.

Selector	PhyIQ ↑	FVD ↓	FVMD ↓	LPIPS ↓	VBench-Q ↑
Baseline	50.1	204.7 ± 20.9	151.9 ± 37.7	0.121 ± 0.007	0.848 ± 0.002
VideoMAE	52.8	203.6 ± 23.2	130.2 ± 29.9	0.118 ± 0.007	0.847 ± 0.002
Qwen2.5-VL	50.5	199.5 ± 20.2	146.2 ± 40.5	0.125 ± 0.007	0.849 ± 0.002
Qwen3-VL	55.9	190.0 ± 18.6	131.0 ± 33.4	0.120 ± 0.007	0.850 ± 0.002
WMReward	62.3	173.9 ± 20.5	138.5 ± 37.2	$\mathbf{0.111 \pm 0.007}$	0.848 ± 0.002
GEOPHYS	**64.5**	$\mathbf{168.4 \pm 17.3}$	162.4 ± 47.0	$\mathbf{0.111 \pm 0.007}$	0.849 ± 0.002
Oracle (UB)	*72.9*	159.8 ± 19.3	117.9 ± 34.2	0.106 ± 0.007	0.848 ± 0.002
Spearman ρ	—	-0.96	-0.32	-0.89	$+0.07$

H.2 VBench dimension breakdown

Table 12: **VBench dimensions on PhysicsIQ V2V** (MAGI-1 24B, N=16), per-dimension mean over the 198 scenarios; the typical scenario-level standard error is in each row label. Six dimensions are saturated and effectively constant across selectors; only dynamic degree varies.

Dimension	Baseline	VideoMAE	Qwen2.5	Qwen3	WMReward	GEOPHYS	Oracle
PhysicsIQ (↑, ref.)	*50.1*	*52.8*	*50.5*	*55.9*	*62.3*	*64.5*	*72.9*
Subject consist. (±.003)	.959	.956	.959	.960	.959	.960	.957
Background consist. (±.001)	.972	.971	.973	.973	.972	.973	.971
Motion smooth. (±.0001)	.9970	.9970	.9970	.9971	.9971	.9972	.9970
Aesthetic (±.006)	.472	.468	.476	.476	.469	.475	.469
Imaging /100 (±.5)	68.9	68.9	69.1	69.4	69.2	69.2	69.3
Temporal flicker (±.0001)	.9981	.9982	.9981	.9982	.9985	.9984	.9983
Dynamic degree (±.018)	.096	.071	.076	.071	.040	.040	.066

Six dimensions are saturated near their ceilings and vary by at most a few tenths of a percent across selectors, so they cannot distinguish the selection strategies. Three of these, motion smoothness, temporal flickering and imaging quality, show a high rank correlation with PhysicsIQ ($\rho > 0.7$), but this is an artefact of ranking values identical to three or four decimal places; their between-selector differences are within one to two standard errors, so the correlation orders essentially equal numbers rather than reflecting a real effect. The only dimension with substantial variation is dynamic degree, which ranges over a factor of two and anti-correlates with PhysicsIQ ($\rho = -0.87$): the more physically plausible selections contain less spurious motion. This matches the FVMD result, where the same suppression of excess motion reads as a divergence from the ground-truth motion distribution. We therefore summarise VBench in the main table by VBench-Q, the mean of the six quality dimensions, and report dynamic degree here, since it measures motion volume rather than perceptual quality.

I BoN compute cost (full)

Table 13 reports the full verifier-path compute cost on the 3,168-video PhysicsIQ candidate pool (N=16 candidates × 198 scenarios). The body's Fig. 11 shows scoring time per video; this table additionally reports parameters, VRAM, and total wall-clock for each method.

Table 13: **Scoring compute cost** on the 3,168-video PhysicsIQ candidate pool. Scoring time and memory are for the verifier path only, on top of the generator. All methods benchmarked on a single H100 GPU.

Method	Verifier backbone	Params	VRAM	per video	total
Foundation-model baselines					
VideoMAE(BoN)	VideoMAE-large	0.3B	1.4 GB	0.47 s	25 min
Qwen2.5-VL(BoN)	Qwen2.5-VL-7B-Inst.	7.0B	16.6 GB	0.28 s	15 min
Qwen3-VL(BoN)	Qwen3-VL-8B-Inst.	8.0B	17.5 GB	0.15 s	8 min
Trained world model					
WMReward [10]	V-JEPA 2 ViT-giant	1.1B	9.3 GB	1.5 s	77 min
GEOPHYS *(frozen, no training)*					
GEOPHYS (1 backbone)	DINOv3 ViT-L	0.3B	1.2 GB	0.25 s	13 min
GEOPHYS (4 backbones)	Ensemble	0.7B	2.0 GB	1.00 s	53 min