Interpreting Physics in Video World Models

Sonia Joseph [1 2] **Quentin Garrido** [1] **Randall Balestriero** [1] **Matthew Kowal** [3] **Thomas Fel** [4] **Shahab Bakhtiari** [2 5]
Blake Richards [2 †] **Mike Rabbat** [1 †]

Abstract

A long-standing question in physical reasoning is whether video models rely on factorized physical state variables, or on task-specific distributed representations. We present the first mechanistic interpretability study of physical variables inside large-scale video encoders, combining layerwise probing, subspace geometry, patch-level decoding, and targeted attention ablations to characterize where and how physical information is organized. Across architectures, we identify a sharp intermediate-depth transition, the Physics Emergence Zone, at which physical variables become linearly accessible. Scalar speed and acceleration are available from early layers, whereas motion direction emerges only at the Physics Emergence Zone, mirroring the V1 to MT motion hierarchy in primate visual cortex. Direction is encoded as a circular high-dimensional population code: dozens of orthogonal probe dimensions must be steered jointly to change the decoded direction, orders of magnitude more than the low-dimensional steering interventions seen in language models. These findings argue against compact physics-engine state variables and support distributed, hierarchically-organized, "brain-like" representations that are nonetheless sufficient for making physical predictions.

1. Introduction

Despite rapid progress in video modeling, it remains unclear whether–and how–video world models represent physical information internally. Prior work has largely addressed this question indirectly, evaluating downstream performance on physics benchmarks while treating the model as a black box (Yi et al., 2020; Garrido et al., 2025; Motamed et al., 2025). As a result, we lack answers to fundamental representational questions: Where in the network physical information is constructed (if at all)? How it is organized across layers and patches? What geometric form does it take?

Understanding this internal organization has implications beyond benchmark accuracy. A model that infers stable physical structure from video could aid scientific modeling in regimes where analytic simulators are incomplete or unavailable, including climate systems, fluid dynamics, and materials science. Models that infer physical variables in a factorized manner—i.e., with distinct representations for quantities such as direction or momentum—could provide interpretable windows into the governing dynamics of the system being modeled rather than black-box predictions alone. These questions closely mirror debates from cognitive science, where accounts of physics representation are often divided between physics engine-based views, which posit compact and reusable latent state variables (Battaglia et al., 2013; Ullman et al., 2017), and heuristic-based views, which emphasize domain-specific rules or perceptual shortcuts without an explicit physics engine (Siegler, 1976; Vasta & Liben, 1996; Davis et al., 2017).

In this paper, we present one of the first mechanistic interpretability analyses of physical variables in video world models. Using layerwise probing, subspace analysis, and targeted ablations, we map where physical information becomes accessible, how it is structured, and what computational substrate supports it.

Physical reasoning in video world models is not formally defined, but broadly refers to the ability to infer object properties, dynamics, and interactions, such as solidity, continuity, and causality (Xue et al., 2023). To this end, we narrow our focus to two complementary diagnostic tasks designed to examine the internal organization supporting physically coherent behavior. First, we analyze a physics task in which models distinguish possible from impossible videos (Riochet et al., 2021; Garrido et al., 2025), capturing coarse-grained physical features including object permanence, shape constancy, and spatiotemporal continuity. Second, to enable fine-grained analysis with ground-truth physical variables, we construct a synthetic toy-ball dataset

[†]Joint last authors. [1]FAIR, Meta Superintelligence Labs [2]Mila & McGill University [3]FAR.AI [4]Kempner Institute, Harvard University [5]Université de Montréal. Correspondence to: Sonia Joseph <soniamollyjoseph@gmail.com>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

Table 1. We consider five distinct possible representations of physics inside video world models. Our findings favor distributed, task-specific representations instead of physics-engine-style assumptions (see Section 8.1 for a discussion).

Physics-engine assumption	**Interpretability prediction**	**Our finding**
Staged derivation	Acceleration is derived from velocity intermediates	Acceleration is decodable at the same stage as velocity without an explicit intermediate velocity state (Sec. 5.2)
Cartesian representation	Motion encoded as (v_x, v_y)	Polar factorization (speed, direction) dominates (Sec. 5.3)
Shared latent physics	Motion variables reused across physical tasks	Direction and intuitive-physics subspaces are nearly orthogonal (Sec. 6.2)
Compact state variables	Motion variables occupy low-dimensional subspaces	Direction spans tens of approximately orthogonal components, steering requires dozens of dimensions (Sec. 7.2)
Object-centric state slots	Motion decodable from specific spatial or temporal patches	Direction becomes spatially redundant across patches post-Physics Emergence Zone (App. C.5)

with precisely controlled velocity and acceleration. Rather than aiming for exhaustive coverage of physical reasoning, we examine in-depth how physical information is structured, transformed, and reused for these two tasks on two state-of-the-art video encoders, V-JEPA 2 (Large, Huge, and Giant) and (Assran et al., 2025) and VideoMAE-v2 G (Wang et al., 2023).

Across all of the models tested, the representation of physics-related information emerges sharply at approximately one-third depth—a transition we call the *Physics Emergence Zone*. The representation of physical variables then peaks in the middle layers, and degrades toward the output. Decomposing motion into finer-grained variables, we find that scalar quantities like speed and acceleration are accessible at the earliest layers, while directional information becomes accessible only at the Physics Emergence Zone. Acceleration does not require an explicit velocity intermediate and can be approximated directly by a single MLP. Direction is a reliable diagnostic of the transition, co-emerging with the ability to distinguish physically impossible videos from physically plausible ones, and with performance on temporal reasoning tasks, such as detecting shuffled videos.

We next examine the relationship between possible–impossible physics judgments and motion direction. Under a physics-engine view with compact, reusable latent states, direction would be expected to support both tasks. Instead, despite their co-emergence, direction and possible–impossible judgments occupy nearly orthogonal representational subspaces, indicating task-specific representations rather than shared latent variables. Both tasks nevertheless rely on a shared circuit-level substrate: attention heads within the Physics Emergence Zone that exhibit unusually local spatiotemporal processing. Suppressing local processing of these heads in the Physics Emergence Zone substantially degrades physics and temporal reasoning performance yet leaves static tasks such as ImageNet classification largely unaffected, showing their critical role in spatiotemporal processing. In addition, we show that motion direction is represented as a high-dimensional population code with circular geometry, reminiscent of population codes in motion neuroscience, requiring coordinated intervention across dozens of dimensions, in contrast to the low-dimensional steering in language models. Overall, our results particularly favor task-specific physical representations over compact, reusable state variables (Tab. 1).

2. Related Work

We situate our work within prior research on video world models, physical reasoning, and interpretability of video transformers.

2.1. Video world models

A *world model* is a learned system whose internal representations capture reusable environmental structure to support prediction, imagination, or planning (Ha & Schmidhuber, 2018). In visual domains, this role is increasingly fulfilled by large-scale unsupervised video models spanning video prediction, robotics, and generative modeling (Ding et al., 2025; Li et al., 2025; Kong et al., 2025).

While recent advances include diffusion-based generators (Ho et al., 2022; Blattmann et al., 2023), their computation is distributed across denoising steps, complicating representation-level analysis. We therefore focus on encoder-based video world models such as VideoMAE-v2 (Wang et al., 2023) and V-JEPA 2 (Assran et al., 2025), whose persistent intermediate representations provide a natural substrate for studying internal physical structure beyond behavioral performance.

2.2. Physical reasoning in video models and cognitive science

Physical reasoning concerns the ability to infer object properties and interactions such as solidity, continuity, and causality (Xue et al., 2023). Despite strong predictive performance, video world models exhibit systematic failures on physical reasoning benchmarks. Models underperform on causal and counterfactual questions in CLEVRER (Yi et al., 2020), show brittleness under violation-of-expectation tests in IntPhys and IntPhys2 (Riochet et al., 2021; Bordes et al., 2025), and fail to generalize in interactive environments such as PHYRE (Bakhtin et al., 2019). Notably, perceptual realism is only weakly correlated with physical correctness (Motamed et al., 2025), raising the question of what internal representations may support physical reasoning.

In parallel, a central debate in cognitive science concerns the representational form of physical reasoning in humans: whether physical judgments rely on compact, reusable latent states—often described as an *intuitive physics engine* (McCloskey, 1983; Battaglia et al., 2013; Ullman et al., 2017)—or instead arise from heuristic, domain-specific reasoning (Siegler, 1976; Vasta & Liben, 1996; Davis et al., 2017). Several reviews frame this distinction as a continuum rather than a dichotomy (Kubricht et al., 2017; Smith et al., 2023). Our work engages this debate at the representational level by analyzing how physical information is organized within video world models.

2.3. Interpretability for video encoders

Most previous interpretability work has focused on text and images, with comparatively limited attention to video due to its higher dimensionality and temporal complexity. Existing video interpretability studies typically rely on proxy tasks or diagnostic benchmarks to assess whether models capture dynamic information, or on visualization-based approaches such as activation maximization and clustering of intermediate representations (Ghodrati et al., 2018; Hadji & Wildes, 2018; Choi et al., 2019; Buch et al., 2022; Kowal et al., 2024). These analyses have revealed strong biases toward static appearance in many video models. However, while prior work can quantify or visualize temporal concepts, it does not characterize the *representational form* used to construct such concepts. In contrast, we perform targeted interpretability analyses to determine where and how physical information is organized within video encoders.

3. Models and Probing Methodology

We study two state-of-the-art video transformer architectures and evaluate physical reasoning using a layer-wise probing methodology.

3.1. Models

Based on the Joint Embedding Predictive Architecture, V-JEPA 2 (Assran et al., 2025) trains an encoder $f_{\boldsymbol{\theta}}$ to map spatiotemporal patches to latent representations, while a predictor $g_{\boldsymbol{\phi}}$ forecasts representations of masked or future patches. This predictive objective encourages temporally structured features over low-level appearance. The architecture extends a vision transformer to video using space–time patches with RoPE embeddings. We analyze the frozen pretrained encoder $f_{\boldsymbol{\theta}}$.

In contrast to V-JEPA 2's latent prediction objective, VideoMAE V2 (Wang et al., 2023) employs masked autoencoding: a large fraction of patches is masked and an encoder-decoder reconstructs the missing pixels, with only the encoder retained after pretraining. This pixel-level reconstruction loss directly incentivizes the preservation of visual detail. We analyze the frozen pretrained encoder and contrast its internal representations with those learned by V-JEPA 2. Appendix D.1 extends this analysis to a total of 14 models across 7 architectural families (latent prediction, pixel reconstruction, diffusion, autoregressive, 3D CNN classification, tracking, and contrastive video-language), with a CNN comparison reported separately in Appendix D.6.

3.2. Probing methodology

To localize where physical information appears in the network, we probe the residual stream at every layer. We primarily use linear probes on mean-pooled space–time patches, which provide a direct readout of what is linearly available in the representation rather than computed by the probe. Because pooling can obscure spatial or temporal structure, we complement these with patch-preserving attentive-mlp probes and interpret both jointly. This setup allows us to first identify where intuitive physics emerges, and then isolate the motion variables and mechanisms that support it. Probe training details are provided in Appendix B; a formal sigmoid-based criterion for declaring a Physics Emergence Zone, together with robustness checks across spatial resolution and frame rate, is given in Appendices D.2 and D.4.

4. The Physics Emergence Zone

To distinguish between physically possible and impossible sequences a model must integrate multiple perceptual cues, capturing sensitivities such as object permanence and spatiotemporal continuity (Baillargeon & DeVos, 1991; Wilcox, 1999; Spelke et al., 1995). Although video transformers can succeed on intuitive physics benchmarks (Garrido et al., 2025), behavioral performance alone does not reveal where the ability to identify impossible physical sequences emerges internally. In this section, we localize the

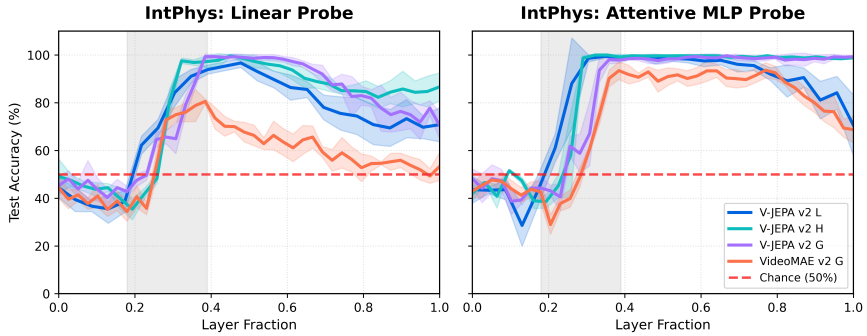


Figure 1. **The Physics Emergence Zone consistently emerges one-third in the model's layers.** We probe performance on V-JEPA 2 (Large, Huge, and Giant) and VideoMAEv2-G across all layers for the possible-vs-impossible physical reasoning task. The shaded area is the emergence zone one-third through the network, where the network starts performing well on the task for linear probes (*left*) and attentive-MLP probes (*right*). For full results, including for the VideoMAE-v2 family, see Appendix C.1.

emergence of this ability across layers and subsequently decompose it into explicit motion variables in Section 5.

4.1. Dataset

To test for the ability to distinguish physically possible vs. impossible sequences, we probe each layer using linear classifiers trained on the IntPhys dataset (Riochet et al., 2021), training the probe on a binary classification task between matched possible and impossible video pairs. Impossible variants violate core physical constraints, including object permanence (objects spontaneously appear or disappear), shape constancy (a cube transforms into a cone), or spatiotemporal continuity (trajectory reversals). Crucially, possible and impossible videos differ only at a single "breakpoint" frame, ensuring that successful classification requires representations that integrate high-level motion dynamics rather than visual cues like texture or color. See Appendix Fig. 5 for dataset examples.

4.2. Distinction between possible and impossible physics emerges one-third through model

Across all V-JEPA 2 model scales (Large, Huge, and Giant), probe accuracies exhibit a remarkably sharp and consistent transition from near chance ($\sim$50%) to high performance ($\sim$85–95%) at approximately one-third of the depth through the encoder (Fig. 1). We refer to this consistent one-third depth transition as the *Physics Emergence Zone*. The location of the Physics Emergence Zone is remarkably consistent across model sizes, indicating a shared computational regime rather than architecture- or scale-specific behavior. VideoMAE-v2-G also exhibits a similar depth-dependent transition; meanwhile, smaller VideoMAE-v2 variants fail to exhibit reliable emergence, potentially due to differences in model capacity, dataset size, or training objective (Appendix C.1).

Breaking down by different types of violations of physics reveals a similar pattern across object permanence, shape constancy, and spatiotemporal continuity (Appendix Fig. 10; see Appendix D.11 for definitions of each violation category), which indicates that the Physics Emergence Zone is not unique to a single type of violation of the laws of physics. Four complementary perspectives on why physical structure emerges at this particular depth, spanning computational staging, neuroscience parallels, attention head diversity, and patch-level transitions, are presented in Appendix D.8.

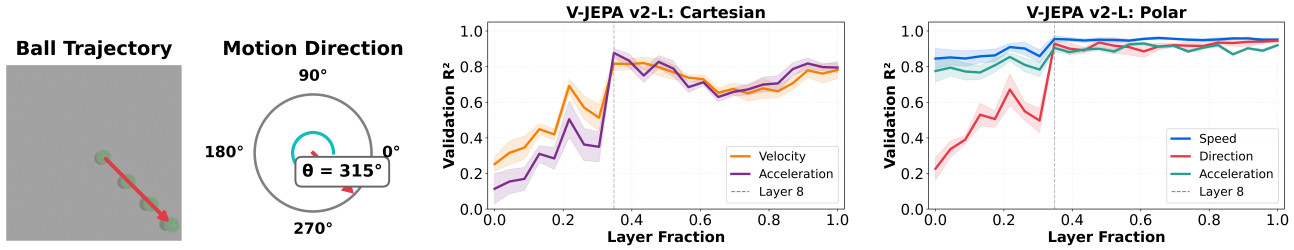
In addition, perhaps counterintuitively, the best possible-vs-impossible physics representations are strongest at intermediate depth: probe accuracy peaks in the middle third of the encoder and degrades toward the output. This indicates that final-layer features do not necessarily preserve the most informative physical structure, consistent with prior findings that intermediate representations can outperform final layers for downstream perception tasks in vision encoders (Bolya et al., 2025). We show that the intermediate representations of the encoder lead to better performance on a downstream intuitive task in Appendix C.1.4.

5. Velocity and Acceleration in the Physics Emergence Zone

Our results demonstrate where the models we are testing can distinguish possible from impossible physics events, but not which physical quantities contribute to that capability. To move beyond coarse behavioral signatures, we turn to explicit physical variables that admit clear ground truth. We focus on Cartesian representations for velocity $\mathbf{v}_t = (v_{x,t}, v_{y,t})$ and acceleration $\mathbf{a}_t = (a_{x,t}, a_{y,t})$, and their polar decompositions, speed $r_t = \|\mathbf{v}_t\|_2$, motion direction θ_t, and acceleration magnitude $\|\mathbf{a}_t\|_2$.

5.1. Synthetic toy ball dataset

We generate synthetic single-ball videos using the Kubric simulator (Greff et al., 2022) with controlled motion param-



(a) Dataset Example *(b)* Cartesian *(c)* Polar

Figure 2. **Motion property encoding across layers.** (a) Example from our synthetic ball rolling dataset showing motion direction θ. (b) Cartesian representations (v_x, a_x) show similar layer-wise emergence patterns, with acceleration available at the same time as velocity. (c) Polar representations (speed, direction, acceleration magnitude) across layer fraction. Direction is only available at the Physics Emergence Zone, while magnitudes are available early.

eters (Fig. 2a). The ball follows straight-line trajectories under either constant velocity or externally induced acceleration, with all other factors held fixed. Ground-truth motion variables are measured in pixels per frame, enabling targeted probing of $\mathbf{v}_t$, $\mathbf{a}_t$, r_t, θ_t, and $\|\mathbf{a}_t\|_2$. Full dataset details are provided in Appendix A.1.2.

5.2. Cartesian representation: acceleration does not rely on velocity

Both Cartesian velocity (v_x, v_y) and acceleration (a_x, a_y) exhibit a transition at the Physics Emergence Zone (Fig. 2b). Notably, acceleration is also decodable with high R^2 from early layers, indicating that acceleration-like signals can be extracted directly from local features without relying on explicit intermediate velocity representations. However, Cartesian variables entangle motion magnitude and direction, making it unclear which aspects of motion are constructed at this stage.

5.3. Polar representation: speed emerges early, while direction emerges at the Physics Emergence Zone

To disentangle magnitude and direction, we reparameterize motion in polar coordinates. Under this decomposition, speed and acceleration magnitude are available from early layers, while directional information becomes reliably decodable only at the Physics Emergence Zone (Fig. 2c). This pattern indicates that the transition is more strongly associated with the emergence of direction rather than of scalar motion quantities. We do a deeper dive into the mechanism behind the globalization of direction in Appendix C.5.

The growing availability of direction in the Physics Emergence Zone persists on the multi-object dataset CLEVRER (Yi et al., 2020), in which object-level probes recover the same Physics Emergence Zone signature across object types and model scales (Appendix Fig. 13b), indicating that direction emergence generalizes beyond single-object motion. For per-object results, see Appendix C.3.

Interestingly, the early availability of speed and the later emergence of direction mirror the motion-processing hierarchy in biological vision: speed-sensitive motion energy appears early, while higher-order pooling gives rise to position-invariant direction selectivity at later stages (Pasternak & Tadin, 2020; Born & Bradley, 2005).

6. What is the Relationship between Possible-vs-Impossible Physics and Direction?

So far, we have shown that representations for possible-vs-impossible physics judgments and motion direction emerge at the same Physics Emergence Zone, but their internal relationship remains unclear. We consider three competing hypotheses about the relationship of the two representations: (i) their co-emergence reflects a generic depth-dependent effect across many tasks, including tasks not related to physical information; (ii) direction is compositionally reused to support possible-vs-impossible judgments, akin to a physics engine, or similarly, both tasks rely on the same underlying latent feature (e.g. spatiotemporal features that are more fine-grained than motion direction); or (iii) the two tasks depend on shared circuit-level computation without any representational overlap in latent space.

In this section, we evaluate each hypothesis. We find that the Physics Emergence Zone is specific to temporally structured tasks, which rules out a generic depth effect. We further show that direction and possible-vs-impossible judgments occupy distinct representational subspaces, which rules out variable reuse and shared underlying latent features. Finally, we identify a shared circuit-level substrate—local spatiotemporal processing in attention heads within the Physics Emergence Zone that supports both tasks, which provides the strongest evidence for the last hypothesis.

6.1. The Physics Emergence Zone is specific to tasks requiring spatiotemporal processing

We first test whether the Physics Emergence Zone reflects a generic depth-dependent pattern, or whether it is selectively associated with tasks that impose specific temporal constraints. To this end, we apply the same layerwise probing analysis to several control tasks, including CLEVRER object counting (Yi et al., 2020), ImageNet classification (Deng et al., 2009), Something-Something-v2 (SSv2) video classification (Goyal et al., 2017), and shuffled versus non-shuffled video discrimination.

Although CLEVRER counting and SSv2 operate on video input, neither task necessarily requires coherent object-level motion or stable direction representations, and can in principle be supported by frame-level information or short-range temporal cues. In contrast, both IntPhys possible–impossible discrimination and shuffled-video detection impose global coherence constraints over time: IntPhys requires maintaining physically consistent object trajectories, while shuffled-video detection requires sensitivity to global temporal order.

Consistent with this distinction, neither CLEVRER counting nor SSv2 exhibits the characteristic one-third emergence signature (Appendix Fig. 12), whereas shuffled-video detection shows a similar emergence pattern. These results suggest that the Physics Emergence Zone is not a generic property of video processing or network depth, but is selectively associated with global spatiotemporal coherence.

6.2. Possible-vs-impossible physics and motion direction do not overlap in latent space

Next, we test whether the possible-vs-impossible and direction tasks share representational space: either the more general possible-vs-impossible physics task compositionally reusing direction, or both tasks relying on the same underlying feature. Both mechanisms would be similar to physics simulators, in which a general and shared set of underlying features generates a variety of physical behavior.

We look at the geometric relationship between their decoding subspaces to find minimal overlap: principal angles between motion and IntPhys subspaces average 69°–83°, with direction closer to IntPhys (69°) than speed (81°). Projection overlap is low—only 7–13% of the IntPhys subspace projects onto direction, and $< 3\%$ onto speed, statistically indistinguishable from random projections (Bjorck & Golub, 1973). For a full account of our method and results, see Appendix C.4. Thus, despite becoming accessible with a probe at the same depth, the two capabilities occupy nearly orthogonal representational subspaces, ruling out representational reuse and shared latent-variable explanations.

6.3. Local attention heads in the Physics Emergence Zone underpin spatiotemporal processing

In the previous section, we established that the internal representations of the possible-vs-impossible physics task and direction are task-specific, without feature reuse between tasks. Still, the two tasks show a shared emergence pattern in the Physics Emergence Zone, which suggests that they may share an underlying computational process– for example, the same attention heads or MLP mechanisms may be responsible for both tasks. Our results from Section 6.1 suggest that the Physics Emergence Zone contains a mechanism that is unique to spatiotemporal processing.

Given that attention heads mediate spatiotemporal processing in transformers across patches, we analyze attention head distance across layers. We find that attention heads outside of the Physics Emergence Zone show relatively homogenous attention profiles. However, uniquely at the Physics Emergence Zone, unusually spatiotemporally local attention heads emerge alongside longer-range heads, resulting in a sharp increase in attention head diversity as measured by distance (Fig. 3). Our method of measuring attention head distance is in Appendix C.6.

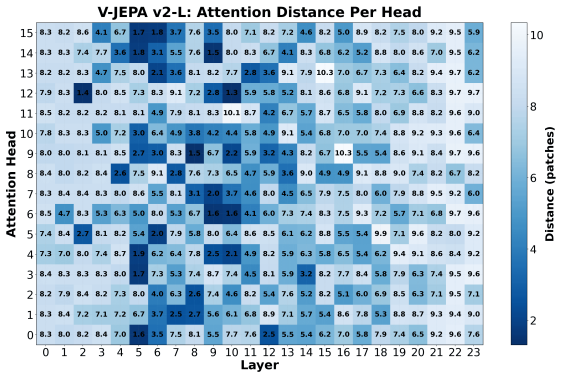
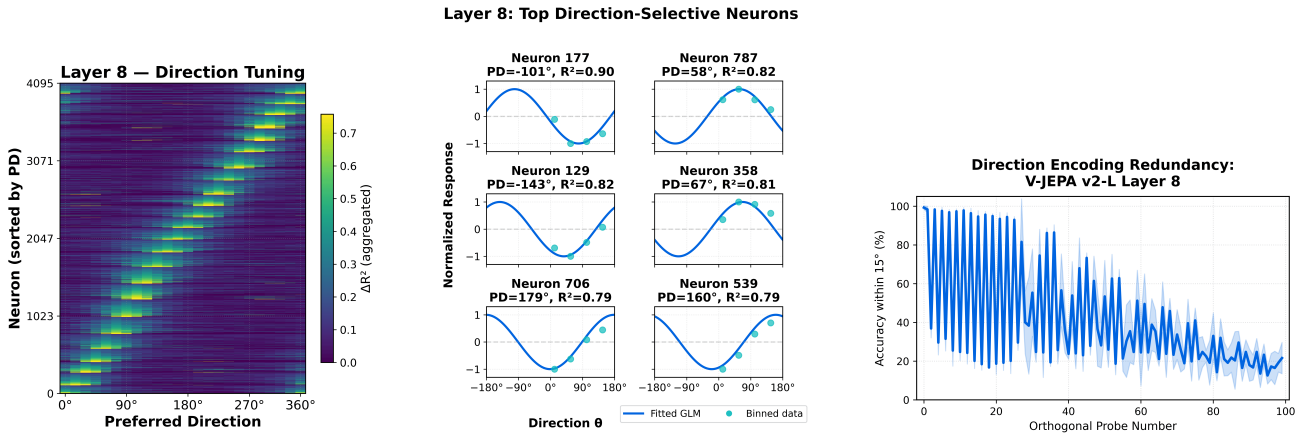


Figure 3. **Spatiotemporally local attention heads crop up uniquely at the Physics Emergence Zone.** Per-head attention locality heatmap showing the coexistence of local and long-range heads at this stage. See Appendix Fig. 19 for a line plot of attention distance.

Table 2. **Local attention suppression at the Physics Emergence Zone degrades performance.** Direction and intuitive physics degrade strongly under spatiotemporal local attention suppression on the Physics Emergence Zone, while ImageNet performance remains largely unchanged. For a full spatiotemporal sweep across a variety of s and t, see Appendix Tab. 4.

Condition	Dir. (R^2)	IntPhys (%)	INet (%)
Base	0.97	78.3	33.7
Spatial (s=7)	0.93	62.2	33.5
Temporal (t=3)	0.83	**51.9**	30.3
Combined (s=3, t=1)	**0.14**	61.7	33.1

We next test whether local attention is functionally responsible for spatiotemporal processing by suppressing local



(a) *(b)* *(c)*

Figure 4. **Direction neurons form a ring-shaped population code with structured redundancy.** (a) At the one-third emergence zone, direction-selective MLP units tile the full angular space and organize into a circular population code. (b) Individual neurons exhibit smooth, sinusoidal tuning to motion direction. (c) Probe accuracy across successive orthogonalizations exhibits a sawtooth pattern, indicating structured redundancy consistent with paired (e.g., sine–cosine) feature encodings.

attention exclusively in the Physics Emergence Zone (Appendix C.6). Our targeted suppression produces severe degradation in both direction decoding and possible-vs-impossible discrimination, while leaving the static task of ImageNet classification largely unaffected (Tab. 2).

Our results identify a shared computational mechanism at the Physics Emergence Zone that supports the possible-vs-impossible physics and motion direction tasks without requiring shared representational subspaces. Our analysis focuses on coarse spatiotemporal reasoning; other compositional physical computations, such as contact dynamics or force-based inference, may rely on additional mechanisms not examined here.

7. Steering the Direction Variable

One gold standard in mechanistic interpretability is the ability to steer representations in latent space to change the model's prediction, establishing the feature's causal impact on the output (Turner et al., 2024; Panickssery et al., 2024; Zou et al., 2025). In the following section, we attempt to steer motion direction to change the decoding. We find a rich and counterintuitive set of behaviors for the representation of direction: circular population geometry, "sawtooth" sin-cosine encoding, and successful steering only along manipulating dozens of orthogonal probe dimensions.

7.1. Direction neurons form a ring-shaped population code at the Physics Emergence Zone

We find that MLP layers (fc1/fc2) at the end of the Physics Emergence Zone selectively encode motion direction (Fig. 4a), with most units strongly direction-tuned (high GLM R^2), and preferred directions tiling 360° (Fig. 4b). At the population level, direction-selective features organize into a unit-circle geometry that is absent in early layers and emerges sharply at the transition (Appendix Fig. 20), consistent with a circular population code in motion neuroscience (Jazayeri & Movshon, 2006). We do not observe an analogous circular organization for speed, indicating a qualitative difference between direction and speed (Appendix Fig. 21). For our full methodology, see Appendix C.7.

However, manipulating only the unit-circle subspace does not effectively steer direction, suggesting that direction is embedded in a higher-dimensional representation beyond a single unit circle. We test this hypothesis in the next section.

7.2. Physics-related variables require many tens of directions to steer

To estimate the dimensionality of motion direction, we iteratively train linear probes, orthogonalize each probe direction, and retrain on the residual representation until performance reaches chance, a method closely related to prior work on iterative nullspace projection and amnesic probing (Ravfogel et al., 2020). We describe our method in detail in Appendix C.11.

We find that speed, direction, and possible-vs-impossible representations are encoded in high-dimensional subspaces. Possible-impossible discrimination requires approximately 20 independent features at the Physics Emergence Zone, while direction decoding requires roughly 40–50 features, increasing to up to 80 near the output layers (Appendix Fig. 22). Interestingly, probe performance exhibits a characteristic sawtooth pattern across successive orthogonalizations (Fig. 4c), consistent with direction being encoded via

approximately sinusoidal feature pairs (e.g., sine–cosine components) – a pattern that we see uniquely in motion direction and not speed (Appendix Fig. 23). A mathematical derivation of the sawtooth pattern from asymmetric capture of paired sine–cosine components, validated by a synthetic experiment matching the observed collapse ratio, is given in Appendix D.5.

We next examine whether the high-dimensional direction representation can be causally controlled along its many feature dimensions. In line with the previous section's results about manipulating the unit circle in the MLP layers, steering along a single feature direction or probe axis produces little to no change in decoded motion direction. In contrast, coordinated interventions across increasing numbers of orthogonal probe directions yield smooth and monotonic reductions in mean angular error (Appendix, Fig. 24). When steering across all probe directions at layer 8, we achieve $< 0.5^\circ$ error to target angles, compared to $> 80^\circ$ for single-probe interventions (Appendix C.12). Effective steering therefore requires manipulating a large fraction of the representational subspace, rather than modifying any individual feature.

Surprisingly, unlike language models, where semantic concepts often align with single or low-rank directions, even complex behaviors such as refusal can often be controlled via a single activation direction or a small set of vectors (Turner et al., 2024; Panickssery et al., 2024; Zou et al., 2025; Arditi et al., 2024), motion direction in video encoders requires coordinated high-dimensional steering.

8. Discussion

Our goal in this work was to characterize the *representational form* through which physical information is made available internally.

8.1. Intuitive physics in cognitive science

A central debate in cognitive science concerns whether intuitive physics relies on compact, reusable latent state variables—akin to a physics engine—or on distributed, task-specific computations (Battaglia et al., 2013; Ullman et al., 2017; Davis et al., 2017). Our results support the latter at the representational level. Despite strong physical behavior, we find no evidence for shared, low-dimensional latent variables: motion direction and possible–impossible judgments occupy nearly orthogonal subspaces, and direction itself requires high-dimensional coordinated steering (Tab. 1). These findings are difficult to reconcile with physics-engine-style representations and instead support distributed, task-specific representations built atop shared spatiotemporal computation.

8.2. Connections to neuroscience

The organization of motion direction we observe closely parallels biological vision. In primate cortex, direction-selective neurons tile angular space and direction is represented as a circular population code rather than an explicit latent variable (Albright, 1984; Jazayeri & Movshon, 2006). Similarly, direction in video world models emerges as a distributed circular geometry. Moreover, the early availability of speed and later emergence of direction mirror the motion-processing hierarchy in cortex, where direction selectivity arises through higher-order pooling (Born & Bradley, 2005; Pasternak & Tadin, 2020). Together, these parallels suggest that video world models may provide an ideal complementary substrate for studying findings in neuroscience.

8.3. Applications to physics simulators

Recent work on physics steering shows that a transformer trained on PDE simulations (not video) can admit low-dimensional activation interventions corresponding to interpretable physical phenomena (McCabe et al., 2025). In contrast, our results show that our video world models encode motion variables in distributed, high-dimensional geometries, suggesting that whether learned models expose compact, interpretable "state variables" is not guaranteed, but depends critically on training domain and objective. As a practical consequence, fine-tuning only the four blocks within the Physics Emergence Zone (16% of V-JEPA 2-L parameters) suffices to reach perfect IntPhys accuracy, while late layers fine-tuned in the same way recover ImageNet performance instead, yielding a clean physics vs. semantics double dissociation (Appendix D.3).

9. Limitations

Our analysis is limited to encoder-based video transformers trained with masked objectives; autoregressive or diffusion video models may exhibit different representational structure. We focus on possible–impossible discrimination and controlled motion variables, which probe core spatiotemporal sensitivity but do not isolate richer physical computations such as contact dynamics, force inference, or long-horizon interaction. Our methods characterize representational accessibility and coarse causal influence rather than a complete circuit-level mechanism, and our synthetic toy-ball dataset may not reflect how physical structure is represented in natural video. Two specific extensions, evaluation on the harder IntPhys2 benchmark (Bordes et al., 2025) and analysis of non-linear motion regimes such as harmonic oscillation and circular orbits, are discussed in Appendices D.9 and D.10.

Scoping our use of "physics". Our use of "physics" follows the violation-of-expectation paradigm from developmental psychology (Baillargeon & DeVos, 1991; Spelke

et al., 1995): the model must distinguish physically possible from impossible scenes, and decode continuous motion variables (speed, direction, acceleration) from controlled stimuli. The Physics Emergence Zone tracks the layers at which this signal becomes accessible; it does not imply that the model "understands" physics in a richer sense, such as counterfactual reasoning, mechanism inference, or generalization to fluid dynamics, mass, or soft-body deformation. Whether deeper notions of physical understanding emerge at the same depth, at different layers, or not at all, is an open question for future work. Extended results spanning 14 models across 7 architectural families, a formal sigmoid-based PEZ criterion, and a layer-targeted fine-tuning experiment are reported in Appendix D.

10. Conclusion

We map where and how physics-relevant information appears inside large-scale video encoders. Across V-JEPA 2 and VideoMAE-v2 G, both possible–impossible discrimination and motion direction emerge at a sharp mid-depth transition—the *Physics Emergence Zone*—after which physics signals peak and then weaken toward the output layers. Motion decomposition reveals a clear asymmetry: scalar magnitudes are available early, while direction becomes linearly accessible only at this transition. Although these abilities co-emerge, direction and possible–impossible judgments occupy nearly orthogonal representational subspaces, while both depend causally on localized spatiotemporal attention. Direction itself is encoded as a high-dimensional, unit circle population code that resists low-dimensional steering. These findings argue against compact, reusable latent physics state and instead support a distributed, task-specific representational regime, in favor of heuristic-based accounts of intuitive physics proposed in cognitive science.

Acknowledgements

Thank you to the JEPA team for the continuous feedback and discussions, including Florian Bordes, Mido Assran, Nicolas Ballas, Amir Bar, and Yann LeCun.

Thank you to the LiNC Lab members for the feedback as well, including Aidan Sirbu, Dane Malenfant, and Colin Bredenberg.

Finally, thank you to Christina Last for the discussions about physics simulators and to Ophira Horwitz for the discussions about incorrect physics in dreams.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Albright, T. D. Direction and orientation selectivity of neurons in visual area mt of the macaque. *Journal of neurophysiology*, 52(6):1106–1130, 1984.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Mojtaba, Komeili, Muckley, M., 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., and Ballas, N. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Baillargeon, R. and DeVos, J. Object permanence in young infants: Further evidence. *Child development*, 62(6): 1227–1246, 1991.

Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L., and Girshick, R. Phyre: A new benchmark for physical reasoning, 2019. URL https://arxiv.org/abs/1908.05656.

Battaglia, P. W., Hamrick, J. B., and Tenenbaum, J. B. Simulation as an engine of physical scene understanding. *Proceedings of the national academy of sciences*, 2013.

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H.-Y. F., Pramod, R., Holdaway, C., Tao, S., Smith, K., Sun, F.-Y., Fei-Fei, L., Kanwisher, N., Tenenbaum, J. B., Yamins, D. L. K., and Fan, J. E. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Bjorck, A. and Golub, G. H. Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594, 1973.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models, 2023. URL https://arxiv.org/abs/2304.08818.

Bolya, D., Huang, P.-Y., Sun, P., Cho, J. H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., Wang, J., Monteiro, M., Xu, H., Dong, S., Ravi, N., Li,

D., Dollár, P., and Feichtenhofer, C. Perception encoder: The best visual embeddings are not at the output of the network, 2025. URL https://arxiv.org/abs/2504.13181.

Bordes, F., Garrido, Q., Kao, J. T., Williams, A., Rabbat, M., and Dupoux, E. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments, 2025. URL https://arxiv.org/abs/2506.09849.

Born, R. T. and Bradley, D. C. Structure and function of visual area mt. *Annu. Rev. Neurosci.*, 28(1):157–189, 2005.

Buch, S., Eyzaguirre, C., Gaidon, A., Wu, J., Fei-Fei, L., and Niebles, J. C. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2917–2927, 2022.

Carreira, J. and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Choi, J., Gao, C., Messou, J. C., and Huang, J.-B. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems*, 32, 2019.

Davis, E., Marcus, G., and Frazier-Logue, N. Commonsense reasoning about containers using radically incomplete information. *Artificial intelligence*, 248:46–84, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ding, J., Zhang, Y., Shang, Y., Feng, J., Zhang, Y., Zong, Z., Yuan, Y., Su, H., Li, N., Piao, J., Deng, Y., Sukiennik, N., Gao, C., Xu, F., and Li, Y. Understanding world or predicting future? a comprehensive survey of world models, 2025. URL https://arxiv.org/abs/2411.14499.

Feichtenhofer, C., Fan, H., Malik, J., and He, K. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Garrido, Q., Ballas, N., Assran, M., Bardes, A., Najman, L., Rabbat, M., Dupoux, E., and LeCun, Y. Intuitive physics understanding emerges from self-supervised pretraining on natural videos, 2025. URL https://arxiv.org/abs/2502.11831.

Ghodrati, A., Gavves, E., and Snoek, C. G. Video time: Properties, encoders and evaluation. *arXiv preprint arXiv:1807.06980*, 2018.

Goyal, R., Kahou, S. E., Michalski, V., Materzyńska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., and Memisevic, R. The "something something" video database for learning and evaluating visual common sense, 2017. URL https://arxiv.org/abs/1706.04261.

Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D. J., Gnanapragasam, D., Golemo, F., Herrmann, C., et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3749–3761, 2022.

Ha, D. and Schmidhuber, J. World models. 2018. doi: 10.5281/ZENODO.1207631. URL https://zenodo.org/record/1207631.

Hadji, I. and Wildes, R. P. A new large scale dynamic texture dataset with application to convnet understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 320–335, 2018.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models, 2022. URL https://arxiv.org/abs/2204.03458.

Jazayeri, M. and Movshon, J. A. Optimal representation of sensory information by neural populations. *Nature neuroscience*, 9(5):690–696, 2006.

Kong, L., Yang, W., Mei, J., Liu, Y., Liang, A., Zhu, D., Lu, D., Yin, W., Hu, X., Jia, M., Deng, J., Zhang, K., Wu, Y., Yan, T., Gao, S., Wang, S., Li, L., Pan, L., Liu, Y., Zhu, J., Ooi, W. T., Hoi, S. C. H., and Liu, Z. 3d and 4d world modeling: A survey, 2025. URL https://arxiv.org/abs/2509.07996.

Kowal, M., Dave, A., Ambrus, R., Gaidon, A., Derpanis, K. G., and Tokmakov, P. Understanding video transformers via universal concept discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10946–10956, 2024.

Kubricht, J. R., Holyoak, K. J., and Lu, H. Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, 21(10):749–759, 2017.

Li, X., He, X., Zhang, L., Wu, M., Li, X., and Liu, Y. A comprehensive survey on world models for embodied ai, 2025. URL https://arxiv.org/abs/2510.16732.

McCabe, M., Mukhopadhyay, P., Marwah, T., Blancard, B. R.-S., Rozet, F., Diaconu, C., Meyer, L., Wong, K. W. K., Sotoudeh, H., Bietti, A., Espejo, I., Fear, R., Golkar, S., Hehir, T., Hirashima, K., Krawezik, G., Lanusse, F., Morel, R., Ohana, R., Parker, L., Pettee, M., Shen, J., Cho, K., Cranmer, M., and Ho, S. Walrus: A cross-domain foundation model for continuum dynamics, 2025. URL https://arxiv.org/abs/2511.15684.

McCloskey, M. Intuitive physics. *Scientific american*, 1983.

Motamed, S., Culp, L., Swersky, K., Jaini, P., and Geirhos, R. Do generative video models understand physical principles?, 2025. URL https://arxiv.org/abs/2501.09038.

NVIDIA, Agarwal, N., et al. Cosmos world foundation model platform for physical AI, 2025.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering llama 2 via contrastive activation addition, 2024. URL https://arxiv.org/abs/2312.06681.

Pasternak, T. and Tadin, D. Linking neuronal direction selectivity to perceptual decisions about visual motion. *Annual Review of Vision Science*, 6(1):335–362, 2020.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. Null it out: Guarding protected attributes by iterative nullspace projection. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Riochet, R., Castro, M. Y., Bernard, M., Lerer, A., Fergus, R., Izard, V., and Dupoux, E. Intphys 2019: A benchmark for visual intuitive physics understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (9):5016–5025, 2021.

Siegler, R. S. Three aspects of cognitive development. *Cognitive psychology*, 8(4):481–520, 1976.

Smith, K. A., Battaglia, P. W., and Tenenbaum, J. B. Integrating heuristic and simulation-based reasoning in intuitive physics. 2023. doi: 10.31234/osf.io/bckes. URL https://doi.org/10.31234/osf.io/bckes.

Spelke, E. S., Kestenbaum, R., Simons, D. J., and Wein, D. Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British journal of developmental psychology*, 13(2):113–142, 1995.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Ullman, T. D., Spelke, E., Battaglia, P., and Tenenbaum, J. B. Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, 2017.

Van Hoorick, B., Tokmakov, P., Stent, S., Li, J., and Vondrick, C. Tracking through containers and occluders in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Vasta, R. and Liben, L. S. The water-level task: An intriguing puzzle. *Current Directions in Psychological Science*, 5(6):171–177, 1996. doi: 10.1111/1467-8721.ep11512379.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., and Qiao, Y. Videomae v2: Scaling video masked autoencoders with dual masking, 2023. URL https://arxiv.org/abs/2303.16727.

Wang, Y., Li, K., Li, X., Yu, J., He, Y., Wang, C., Chen, G., Pei, B., Yan, Z., Zheng, R., Xu, J., Wang, Z., Shi, Y., Jiang, T., Li, S., Zhang, H., Huang, Y., Qiao, Y., Wang, Y., and Wang, L. Internvideo2: Scaling foundation models for multimodal video understanding, 2024. URL https://arxiv.org/abs/2403.15377.

Wilcox, T. Object individuation: Infants' use of shape, size, pattern, and color. *Cognition*, 72(2):125–166, 1999.

Xue, C., Pinto, V., Gamage, C., Nikonova, E., Zhang, P., and Renz, J. Phy-q as a measure for physical reasoning intelligence. *Nature Machine Intelligence*, 5(1):83–93, 2023.

Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al. CogVideoX: Text-to-video diffusion models with an expert transformer, 2024.

Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., and Tenenbaum, J. B. Clevrer: Collision events for video representation and reasoning, 2020. URL https://arxiv.org/abs/1910.01442.

Yuan, J., Zhang, X., Friedrich, F., Beltran-Velez, N., Hall, M., Askari-Hemmat, R., Han, X., Ballas, N., Drozdzal, M., and Romero-Soriano, A. Improving the physics of video generation with vjepa-2 reward signal, 2025. URL https://arxiv.org/abs/2510.21840.

Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., and You, Y. Open-Sora: Democratizing efficient video production for all. https://github.com/hpcaitech/Open-Sora, 2024.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A. Experimental Details

A.1. Dataset Details

A.1.1. Intuitive Physics

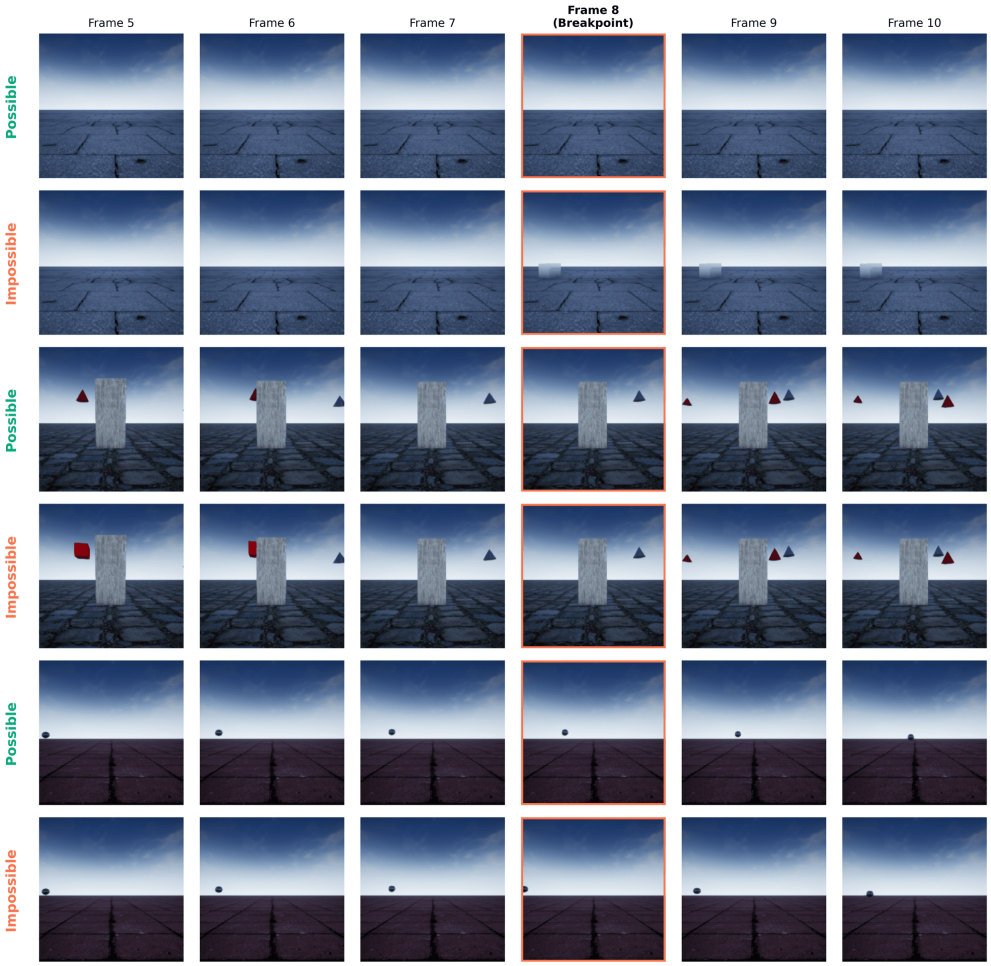


Figure 5. A possible and impossible example from each fold of IntPhys. Each pair differs only at a single "breakpoint" frame and instantiates one of three violation-of-expectation categories from developmental psychology (Baillargeon & DeVos, 1991; Spelke et al., 1995): **(1) object permanence** (an object that should remain present spontaneously disappears), **(2) shape constancy** (an object's geometry changes spontaneously, e.g., a cube becomes a cone), and **(3) spatiotemporal continuity** (an object in motion teleports to a non-adjacent location or its trajectory abruptly reverses). See Appendix D.11 for further discussion.

A.1.2. Synthetic Ball Dataset

We generate two controlled synthetic motion datasets using the Kubric physics simulator (Greff et al., 2022) to study representations of constant-velocity and uniformly accelerated motion. All videos depict a single sphere moving along straight-line trajectories with known ground-truth dynamics.

Velocity dataset. The velocity dataset contains 392 videos (8 directions $\times$ 7 speeds $\times$ 7 start positions), each 16 frames long at 24 fps (0.67 s), rendered at 256×256 resolution. Motion directions are $\theta \in \{0°, 45°, \ldots, 315°\}$, with speed magnitudes $v \in \{1, \ldots, 7\}$ m/s. Start positions are sampled uniformly from $[-2, 2]^2$ m for each (θ, v) pair.

Scenes are generated using Kubric (Greff et al., 2022), with physics simulated in PyBullet 3.2.5 and rendering performed in Blender 2.93. Each scene contains a sphere (radius 0.3 m, mass 1.0 kg), a static 8×8 m floor plane, a fixed overhead perspective camera at (0,0,10) m, and a single directional light.

All friction terms (lateral, rolling, spinning) and restitution are set to zero. The sphere is initialized with nonzero velocity v in direction θ, and no external forces are applied, yielding uniform linear motion throughout the sequence ($\Delta v = 0$).

Acceleration dataset. The acceleration dataset consists of 280 videos (8 directions $\times$ 5 accelerations $\times$ 7 start positions) with identical temporal and spatial resolution. Motion directions follow the same θ set as above, with acceleration magnitudes $a \in \{2, 4, 6, 8, 10\}$ m/s^2. Start positions are sampled uniformly from $[-2, 2]^2$ m for each (θ, a) pair.

The sphere is initialized at rest and subjected to a constant external force $\mathbf{F} = m\mathbf{a}$ applied at each timestep in direction θ. Physics is simulated at 240 Hz (10 substeps per rendered frame). Under frictionless conditions, measured accelerations match target values within 10^{-4} m/s^2, verified via finite differences of recorded velocities. Rendering, scene configuration, and annotations are identical to the velocity dataset.

B. Probe Training Details

For all experiments, we trained linear probes of the form $f(\boldsymbol{h}_\ell) = \boldsymbol{W}\boldsymbol{h}_\ell + \boldsymbol{b}$ on spatiotemporally pooled activations from each layer $\ell \in \{0, \ldots, n-1\}$ of the frozen n-layer encoder. We performed a hyperparameter sweep over 20 configurations, with learning rates $\{10^{-4}, 3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}, 5 \times 10^{-3}\}$ and weight decay $\{0.01, 0.1, 0.4, 0.8\}$, selecting the best model based on validation performance. We used 5-fold grouped cross-validation and report results as mean $\pm$ standard deviation across folds.

C. Additional Experiments

C.1. Possible-vs-impossible physics task

C.1.1. Full results for linear probe

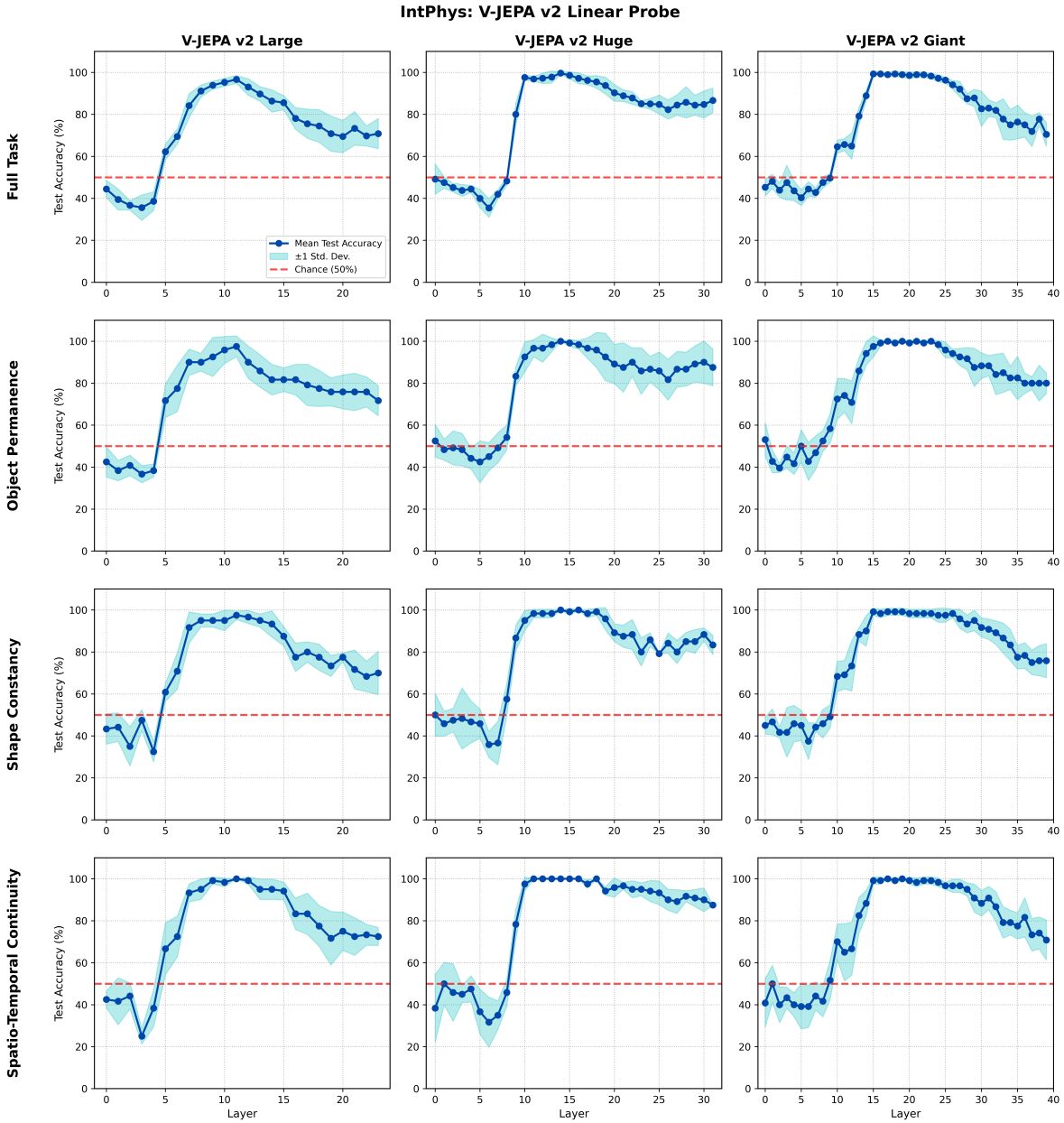


Figure 6. Full results for the linear probe on all sizes of V-JEPA 2.

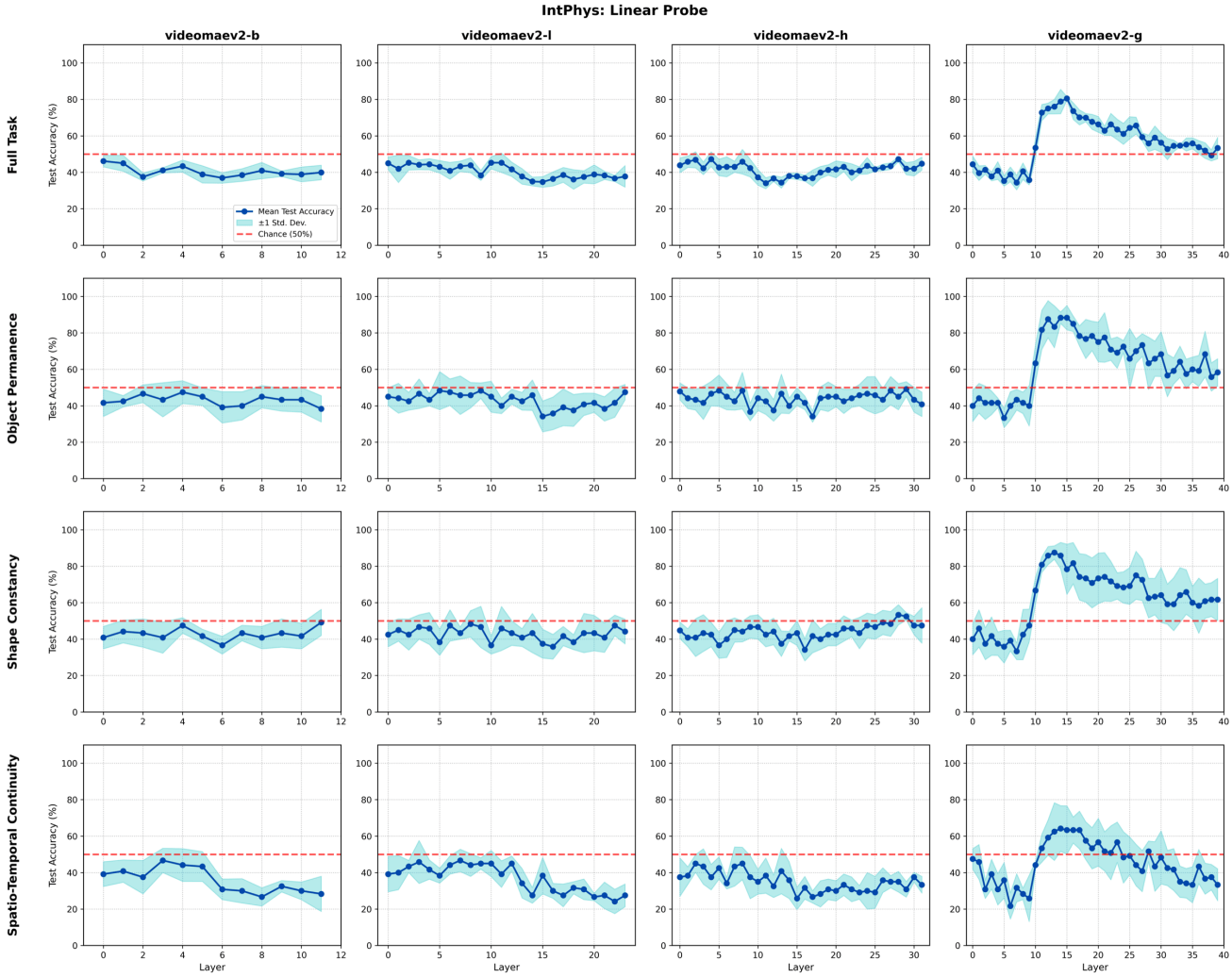


Figure 7. Full results for the linear probe on all sizes of VideoMAE-v2.

C.1.2. Full results for attentive-MLP probe

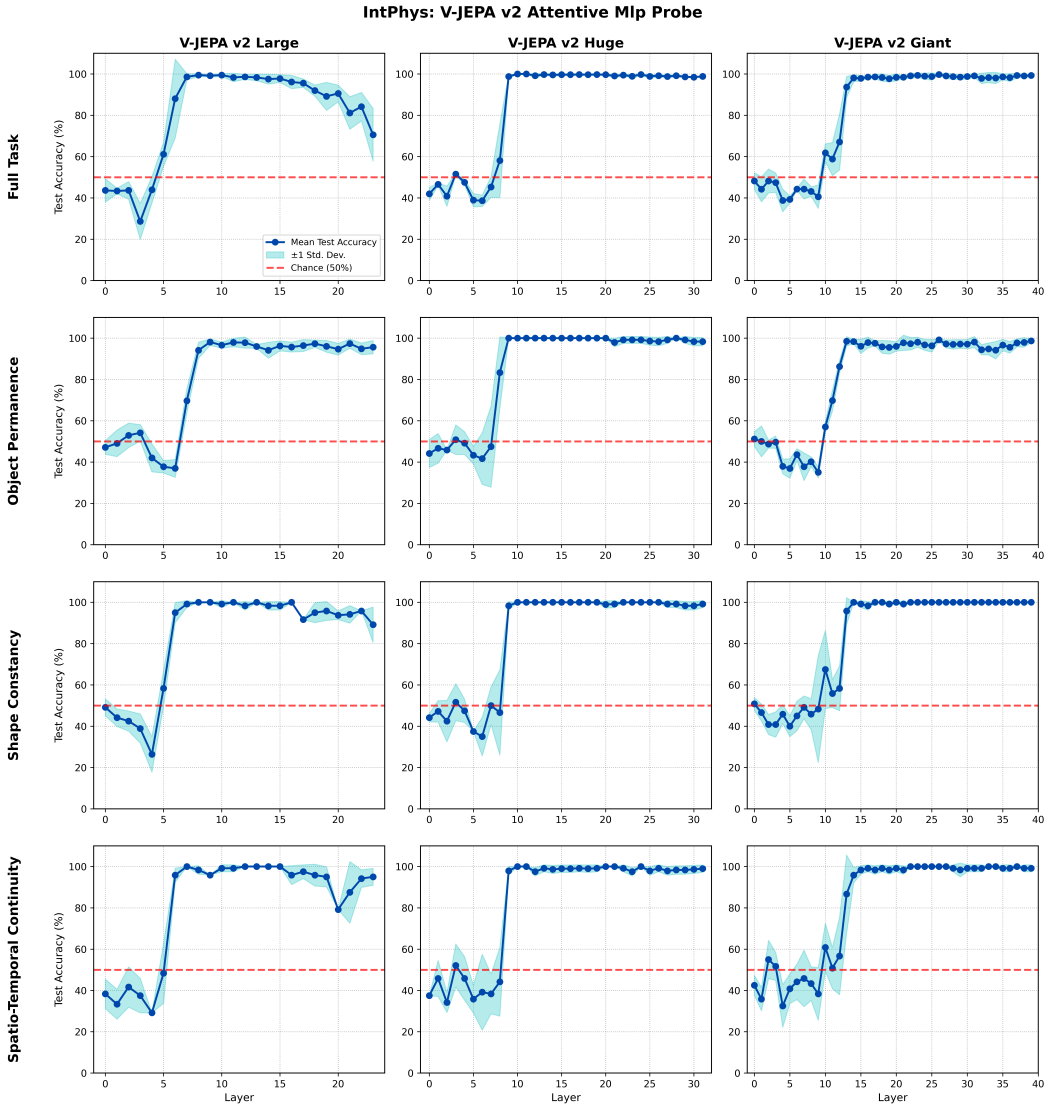


Figure 8. Full results for the attentive-mlp probe on all sizes of V-JEPA 2.

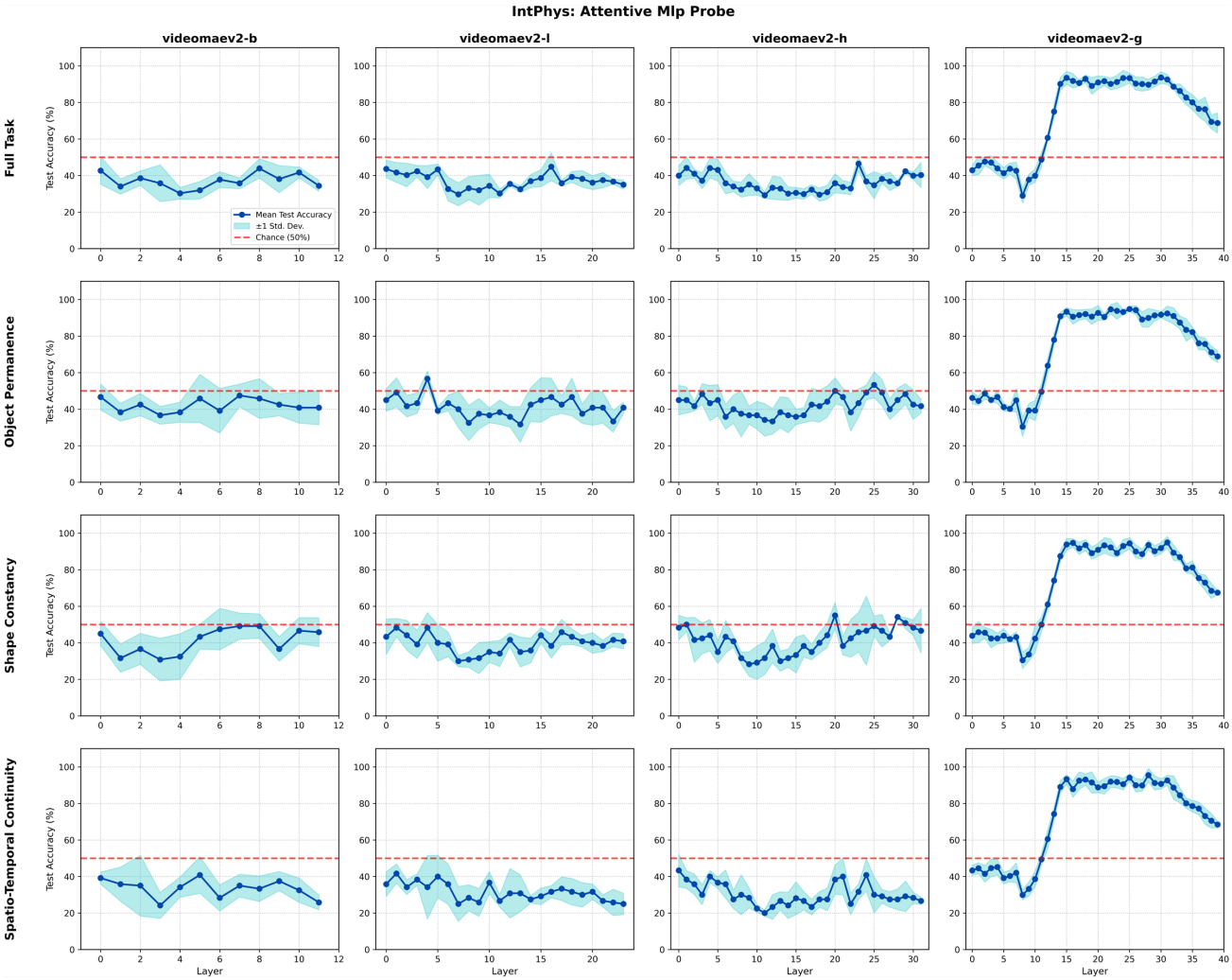


Figure 9. Full results for attentive-mlp probe on all sizes of VideoMAE-v2.

C.1.3. Consistency across possible-vs-impossible physics sub-tasks

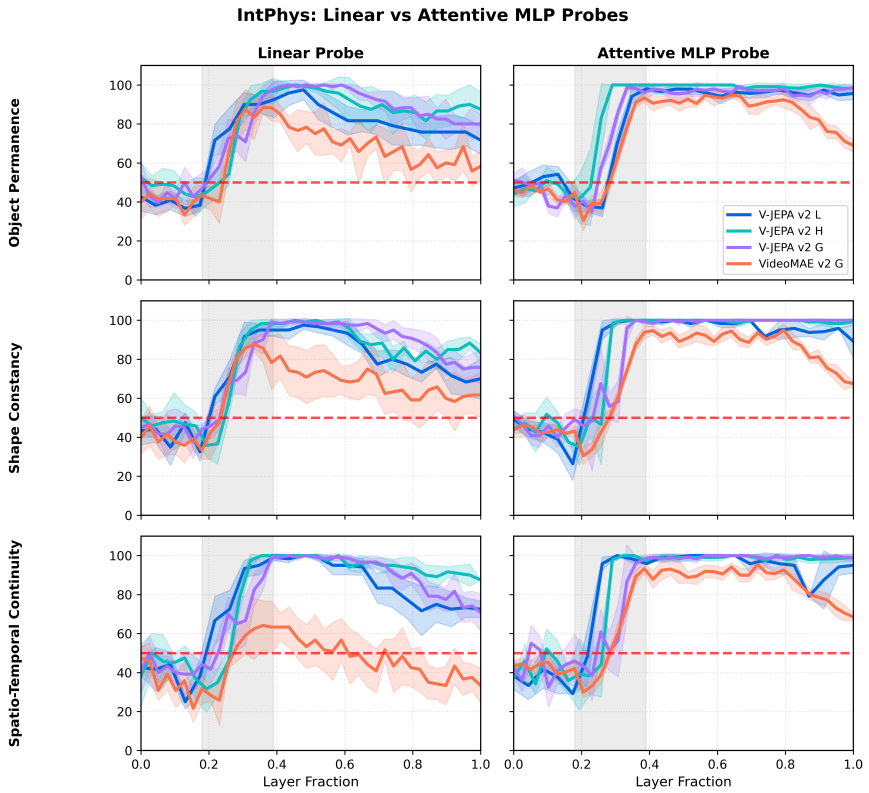


Figure 10. Intuitive physics results broken down by subtasks of object permanence, shape constancy, and spatiotemporal continuity.

In Section 4.2, we show a distinct emergence pattern one-third through the network for the possible vs. impossible physics task. The IntPhys (Riochet et al., 2021) dataset is further divided into three sub-tasks capturing object permanence, shape constancy, and spatiotemporal continuity. Therefore, a natural next question is whether each sub-task has a unique emergence signature.

Probe performance by subtask reveals the same one-third emergence pattern across all three principles: object permanence, shape constancy, and spatiotemporal continuity (Figure 10). This universality suggests that the models may have the same underlying processing for possible vs. impossible tasks.

C.1.4. The middle layer possible-vs-impossible physics representations generalize to better performance on a downstream intuitive physics task

In Section 5.3, we show that perhaps counterintuitively strongest representations for the possible vs. impossible intuitive physics task are at the center of the network across model sizes and architectures. Our findings echo past results in which intermediate instead of final representations were seen to be useful on certain tasks for contrastively vision-language models (Bolya et al., 2025). Similarly, using intermediate instead of end representations may be useful in tasks like using V-JEPA-2 representations to improve the physics plausibility of video generative models (Yuan et al., 2025).

To confirm this hypothesis, we trained from scratch a V-JEPA 2 large predictor on the representations of every layer on the violation-of-expectation framework taken from (Garrido et al., 2025). This downstream task measures the plausibility of the next-frame, and it is a more real-world task than our previous set-up.

We confirm that the middle one third of the network yields the best predictivity on the task (Appendix Figure 11). This suggests that for future tasks involving training on physical representations, the center of the network may be more optimal than the end of the network, where performance degrades.

Interestingly, we also notice that even early layers that give poor performance on only the encoder suddenly perform well

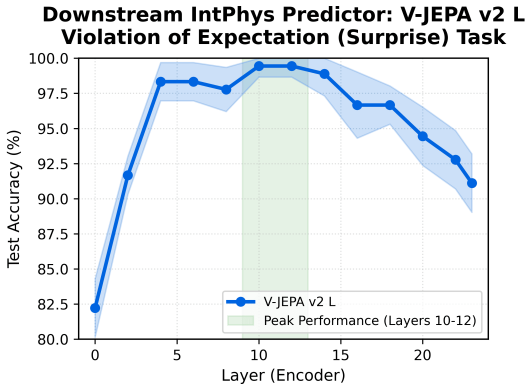


Figure 11. Downstream performance of the V-JEPA 2 Large predictor on the violation-of-expectation intuitive physics set-up. Representations from the middle layers of the encoder provide the best performance compared to final layers.

(Layer 0). This is likely because the predictor itself is learning from the representation.

Future interpretability work will need to investigate why this is the case; we hypothesize it may be related to feature object binding, in which velocity information is most "bound" to the corresponding object. At the end of the network, information is catered to the optimization objective of predicting the next frame in latent space, not necessarily to preserving object-level information.

C.2. Task-Specificity of Physics Emergence Zone

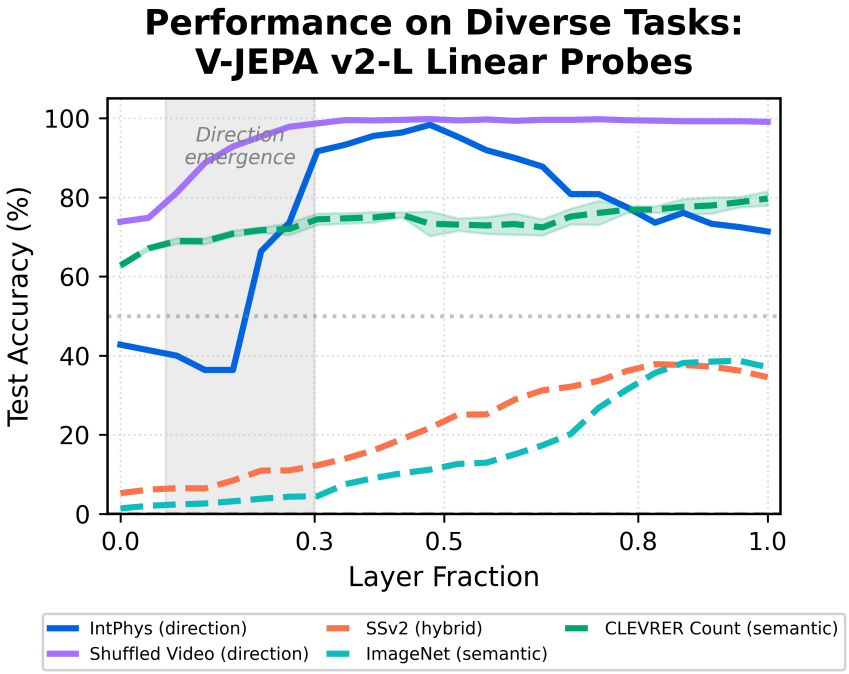
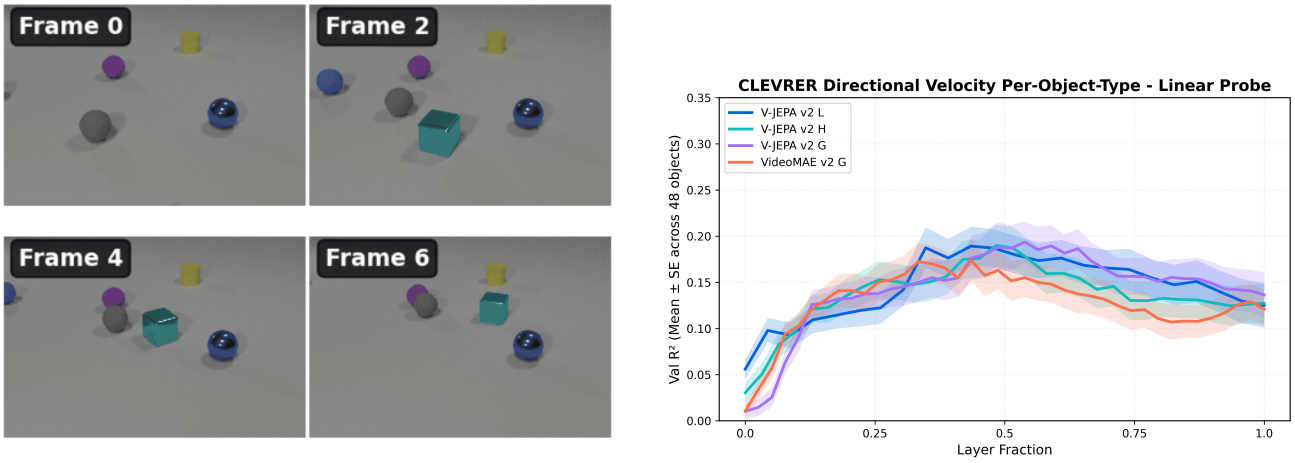


Figure 12. Layer-wise probe performance on control tasks. Object counting, image classification, and standard video classification do not exhibit the one-third emergence signature, indicating it is not a generic depth effect. Only shuffled video detection shows a similar pattern, consistent with relying on temporal order.

C.3. Direction validation on multi-object scenes

In Section 5.3, we validated the emergence of direction for multi-object scenes like the CLEVRER dataset to discover that the Physics Emergence Zone signature was consistent across model architectures (Fig. 13) (Yi et al., 2020). This Appendix

section contains additional results, including a per-object breakdown for accuracy.



(a) *(b)*

Figure 13. CLEVRER dataset: (a) Sample video frames. The first six frames of the CLEVRER video dataset, which shows objects of various colors, shapes, and textures moving across the screen. We train per-object probes on each of the objects to detect its direction. (b) Per-object direction results.

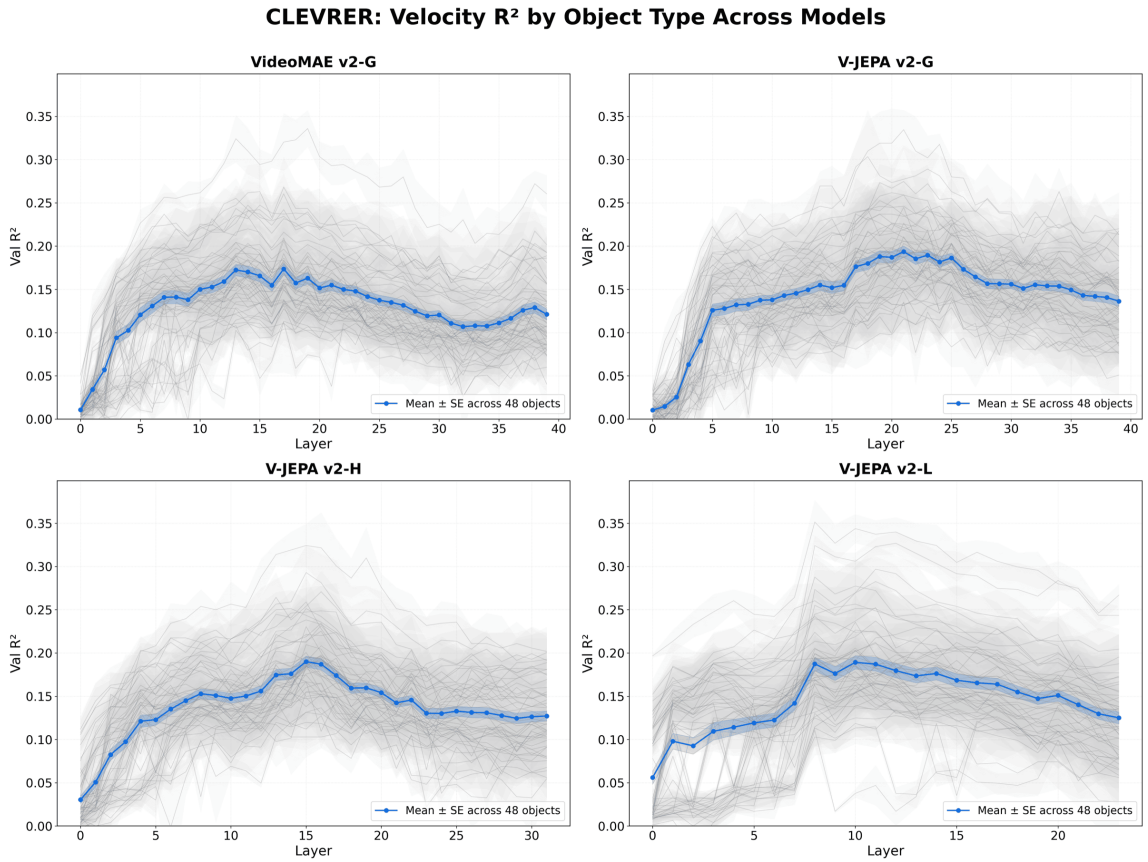


Figure 14. Full results for linear probes for all 48 objects of CLEVRER. Each gray line is an object and the blue is the average across all objects. Error bars are from k-fold cross validation.

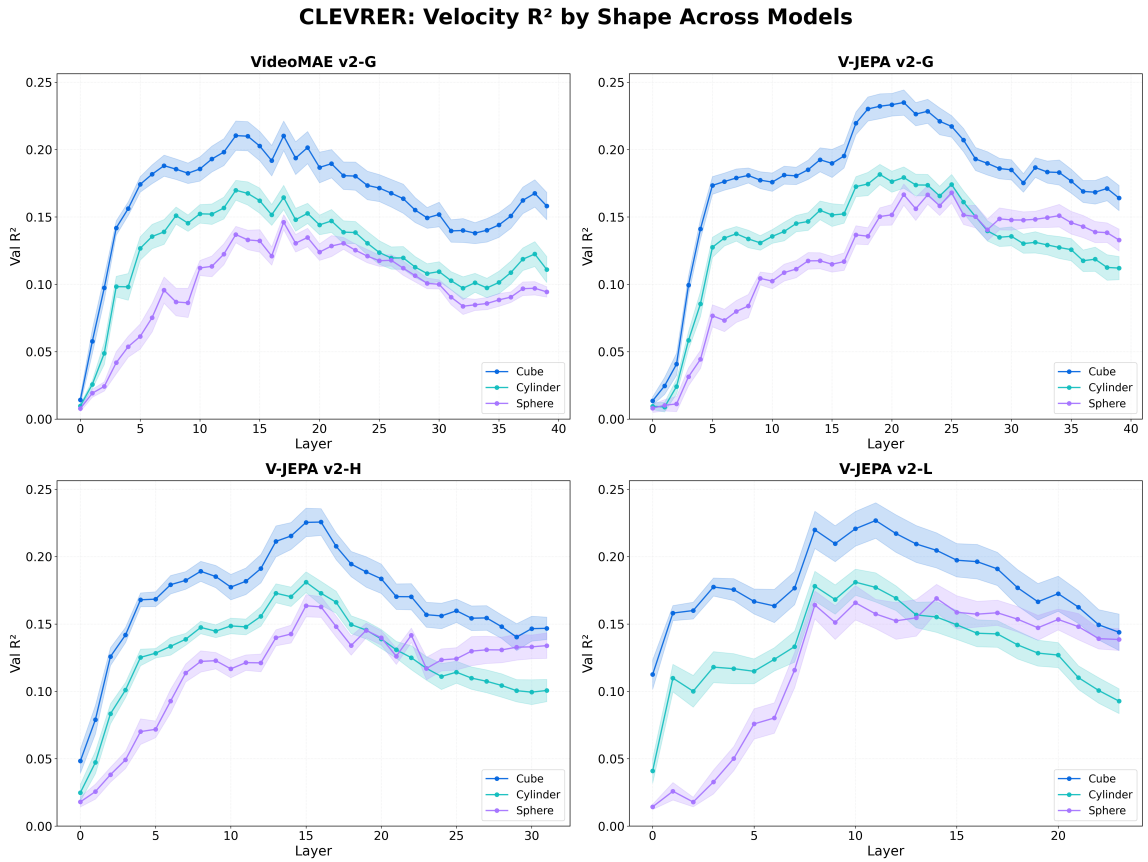


Figure 15. Results for CLEVRER velocity per-object linear probe R² grouped by cube, cylinder, and sphere shape across all model types.

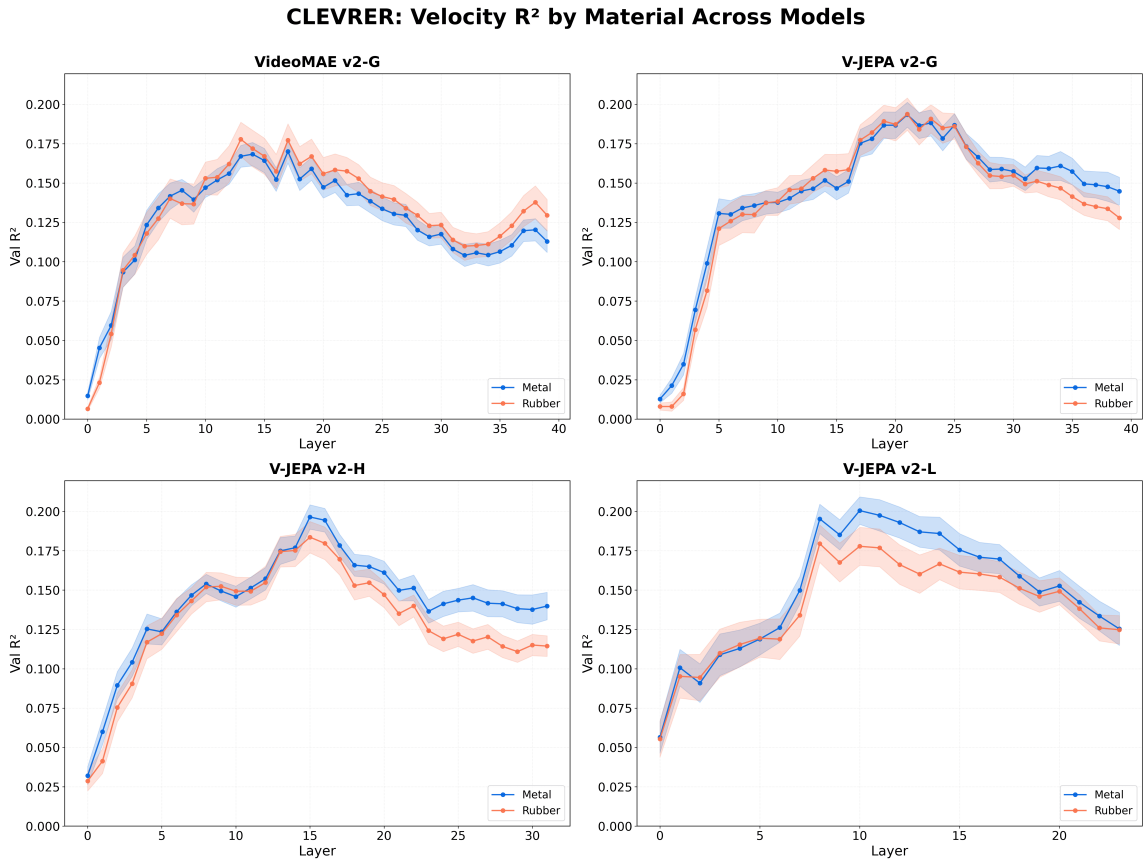
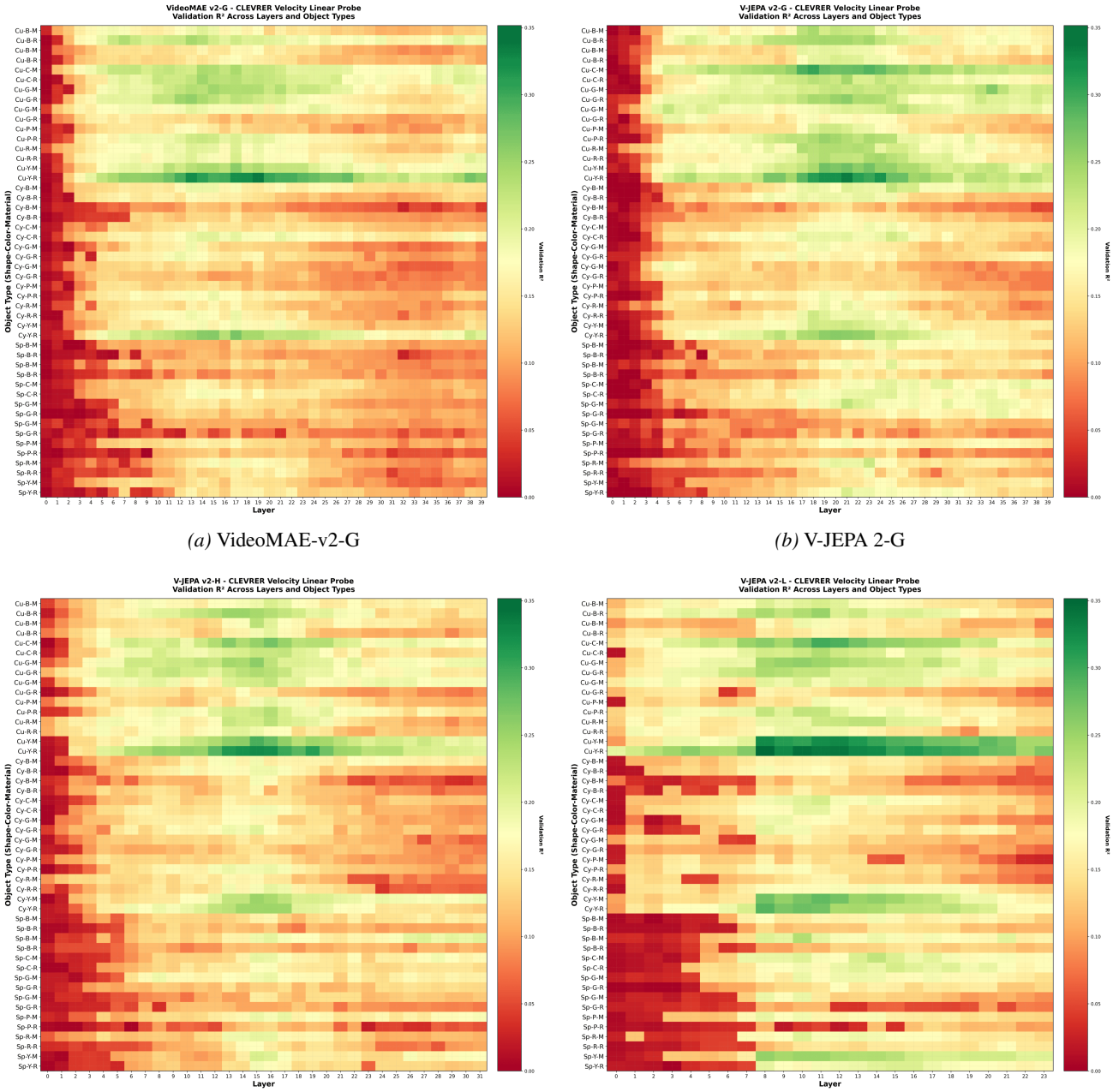


Figure 16. Results for CLEVRER velocity per-object linear probe R² grouped by metal and rubber material across all model types.



(a) VideoMAE-v2-G

(b) V-JEPA 2-G

(c) V-JEPA 2-H

(d) V-JEPA 2-L

Figure 17. CLEVRER velocity linear probe R² heatmaps across layers and object types for all models.

C.4. Subspace Overlap Analysis

In Section 6.2, we examined the relationship between the direction and intuitive physics subtasks to determine whether there is significant representational reuse or shared feature space.

To quantify the relationship between motion encoding and intuitive physics reasoning, we measure the overlap between the subspaces spanned by direction probes, speed probes, and IntPhys probes across layers.

Method. For each layer ℓ, we collect the weight vectors from trained probes. For direction (velocity θ) probes, we use circular regression weights $\mathbf{W}_{\text{dir}} \in \mathbb{R}^{2\times d}$ (sine and cosine components) from each spatial position, yielding a matrix of probe weights. For speed and IntPhys probes, we collect the linear regression weights $\mathbf{w} \in \mathbb{R}^d$. We construct orthonormal bases $\mathbf{Q}_A, \mathbf{Q}_B$ for each subspace via QR decomposition of the stacked weight matrices. We compute three metrics to characterize subspace relationships: *Principal angles.* The principal angles $\theta_1, \ldots, \theta_k$ between subspaces A and B are computed via the singular value decomposition of $\mathbf{Q}_A^\top \mathbf{Q}_B$, where $\cos(\theta_i) = \sigma_i$ (Bjorck & Golub, 1973). We report the mean principal angle $\bar{\theta} = \frac{1}{k}\sum_{i=1}^{k} \theta_i$.

Projection overlap. The fraction of subspace B captured by subspace A is:

$$\text{Overlap}(A \leftarrow B) = \frac{\|\mathbf{Q}_A^\top \mathbf{Q}_B\|_F^2}{\dim(B)} \tag{1}$$

This measures how much of B's variance lies within A.

Grassmann distance. The geodesic distance on the Grassmann manifold:

$$d_G(A, B) = \sqrt{\sum_{i=1}^{k} \theta_i^2} \tag{2}$$

Random Baseline. To assess whether observed overlaps reflect meaningful structure or merely geometric coincidence, we compare against the expected overlap for random subspaces. For two random subspaces A and B with dimensions k_A and k_B in an ambient space of dimension d, the expected projection overlap is (Vershynin, 2018):

$$\mathbb{E}[\text{Overlap}(A \leftarrow B)] = \frac{k_A}{d} \tag{3}$$

Intuitively, a random k_A-dimensional subspace captures a fraction k_A/d of any random direction's variance. For V-JEPA 2-L with embedding dimension $d = 1024$:

- **Direction subspace** ($k = 66$–136): expected random overlap = 6.4–13.3%
- **Speed subspace** ($k = 21$–29): expected random overlap = 2.1–2.8%
- **IntPhys subspace** ($k = 7$–15): expected random overlap = 0.7–1.5%

Results. Table 3 shows the subspace overlap between motion encoding (direction and speed) and IntPhys reasoning across layers. The direction subspace is high-dimensional (66–136 dimensions), speed is lower-dimensional (21–29 dimensions), while IntPhys is compact (7–15 dimensions). Direction and IntPhys subspaces maintain mean principal angles of $69°$–$75°$, with direction→IntPhys overlap of 7–13%. Speed and IntPhys subspaces are more orthogonal ($80°$–$83°$), with overlap below 3%. Critically, these observed overlaps match the random baselines: direction→IntPhys overlap (7–13%) aligns with the expected 6–13% for random subspaces of equivalent dimension, and speed→IntPhys overlap (2–3%) matches the expected 2–3%. This correspondence with chance-level overlap indicates that the motion and IntPhys subspaces share no more structure than would be expected from arbitrary subspaces of the same dimensionality. Despite both capabilities becoming decodable at similar depths in the network, they occupy nearly orthogonal representational subspaces, ruling out representational reuse and shared latent-variable explanations.

Table 3. Subspace overlap between motion encoding (direction/speed) and IntPhys probes (V-JEPA v2-L, d=1024). Overlap$_{A\to B}$ measures the fraction of subspace B captured by subspace A.

	Direction vs IntPhys					Speed vs IntPhys				
Layer	Dir Dim	IP Dim	Angle (°)	Dir→IP Overlap	IP→Dir Overlap	Spd Dim	IP Dim	Angle (°)	Spd→IP Overlap	IP→Spd Overlap
0	30	1	79.1	0.036	0.001	25	1	81.8	0.020	0.001
1	30	1	79.6	0.033	0.001	24	1	82.4	0.017	0.001
2	14	1	81.7	0.021	0.001	25	1	80.3	0.028	0.001
3	114	1	70.6	0.110	0.001	17	1	82.8	0.016	0.001
4	94	1	72.8	0.088	0.001	19	1	82.0	0.019	0.001
5	62	7	75.7	0.064	0.007	16	7	83.5	0.015	0.007
6	96	5	71.4	0.104	0.005	20	5	80.9	0.027	0.007
7	78	11	74.6	0.073	0.010	26	11	81.9	0.022	0.009
8	136	15	69.1	0.129	0.014	28	15	80.7	0.030	0.016
9	122	13	70.7	0.112	0.012	21	13	82.4	0.021	0.013
10	66	12	75.2	0.069	0.012	29	12	81.6	0.025	0.010
11	120	10	70.2	0.117	0.010	26	10	81.2	0.027	0.010
12	122	14	69.9	0.121	0.014	22	14	82.4	0.021	0.013
13	128	13	70.0	0.119	0.012	20	13	82.2	0.023	0.015
14	100	9	71.4	0.103	0.009	21	9	82.6	0.018	0.008
15	122	9	70.7	0.111	0.008	26	9	81.4	0.025	0.008
16	128	14	69.7	0.122	0.013	23	14	81.9	0.023	0.014
17	128	11	68.9	0.132	0.011	21	11	81.9	0.023	0.012
18	128	13	69.0	0.131	0.013	24	13	81.9	0.022	0.012
19	128	7	69.0	0.130	0.007	26	7	81.1	0.026	0.007
20	400	6	51.6	0.386	0.006	25	6	81.3	0.026	0.006
21	400	19	51.5	0.388	0.018	26	19	80.9	0.030	0.022
22	400	6	51.6	0.386	0.006	30	6	79.6	0.034	0.007
23	400	32	51.5	0.389	0.031	31	32	81.3	0.030	0.031

C.5. Direction moves to per-patch encoding

In Section 5.3, we showed that direction uniquely arises at the Physics Emergence Zone compared to other physics-related information like scalar speed and scalar acceleration. The one-third emergence zone appears consistently across tasks and datasets, suggesting that it reflects a structural change in the representation rather than a task-specific artifact. We therefore ask what changes in the representation itself when physical information, specifically direction, becomes decodable.

From a patch-level perspective, we find that direction-related signals are already present in early layers, but remain tightly coupled to specific spatial locations. Direction information is fragmented across patches, such that no single patch contains sufficient information to support reliable decoding or spatial generalization. Mean-pooled probes can nevertheless achieve modest performance by linearly combining these fragmented signals across the frame (Figures 18a).

Around the one-third emergence zone, this structure changes sharply. Direction information becomes broadly distributed across patches, with a marked increase in redundancy and spatial spread. At this stage, individual patches begin to carry sufficient information to support reliable direction decoding on their own. This transition explains why per-patch probe performance rises abruptly at the emergence zone, while mean-pooled performance improves more gradually (Figures 18a, b).

This representational shift also enables spatial generalization (Figure 18b). Probes trained on direction information from one region of the frame begin to generalize to unseen regions only after the emergence zone, indicating that direction is no longer tied to specific spatial coordinates. Instead, it becomes encoded in a globally accessible form that survives pooling operations and supports position-invariant decoding.

Together, these results indicate that the emergence zone corresponds to a transition from local, retinotopic direction signals to a globally distributed representation. This local-to-global shift is reminiscent of the V1→MT hierarchy in biological vision, where early motion signals are spatially localized and later representations pool over space to yield position-invariant direction selectivity.

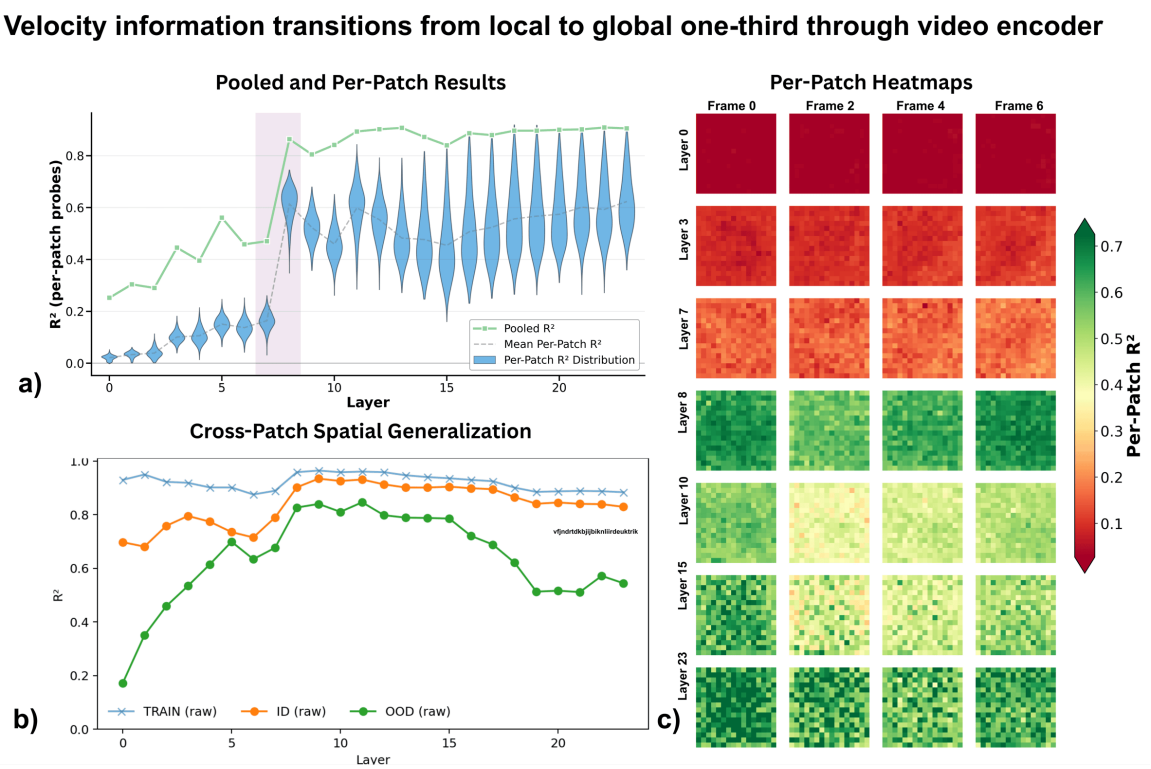


Figure 18. Direction representations transition from local to global at the one-third emergence zone. Results are shown for the synthetic velocity dataset using V-JEPA 2-L. (a) Per-patch linear probe performance across layers. Mean-pooled probes achieve moderate performance earlier by aggregating weak signals across patches, whereas per-patch probes become reliable only at the emergence zone. (b) Spatial generalization performance for probes trained on one half of the frame and evaluated on the other half. Generalization improves sharply at the emergence zone. (c) Per-patch decoding heatmaps illustrate a sharp transition between Layers 7 and 8, where direction becomes decodable from individual patches.

C.6. Attention Distance Analysis

In Section 6.3, we looked at the shared attention head spatiotemporal processing impacted the possible-vs-impossible physics task. We analyze how ablating local attention in V-JEPA v2 affects physics encoding across four tasks: direction prediction (R^2), IntPhys accuracy, per-patch direction decoding (R^2), and ImageNet classification accuracy. V-JEPA v2 uses tubelet embedding with temporal stride 2, so 16 input frames are encoded into 8 temporal tokens. The model processes $T \times N = 8 \times 196 = 1568$ tokens, where each tubelet covers 2 consecutive frames and $N = 196$ is the spatial patch count (14×14 grid from 224×224 images with 16×16 patches).

We define spatial distance between patches as Euclidean distance in patch coordinates: $d_s(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$, measured in patch units (maximum $\sim$18 on the diagonal). Temporal distance is the tubelet difference: $d_t(i,j) = |t_i - t_j|$, ranging from 0 to 7 tubelets.

To test whether local attention is causally important for physics encoding, we ablate attention by masking weights to nearby tokens and renormalizing the remainder. For spatial threshold s, we zero all attention where $d_s(q,k) \leq s$; for temporal threshold t, we zero attention where $d_t(q,k) \leq t$. We evaluate three regimes: (1) spatial-only ablation with $s \in \{1,3,5,7,9,11,13\}$ patches; (2) temporal-only ablation with $t \in \{1,2,3,4,5,6\}$ tubelets; and (3) combined spatiotemporal ablation with paired thresholds.

Large performance drops after ablation indicate reliance on local attention at that scale. Spatial ablation minimally affects direction R^2 but degrades per-patch localization; temporal ablation strongly impacts both direction and IntPhys; combined ablation destroys direction encoding entirely.

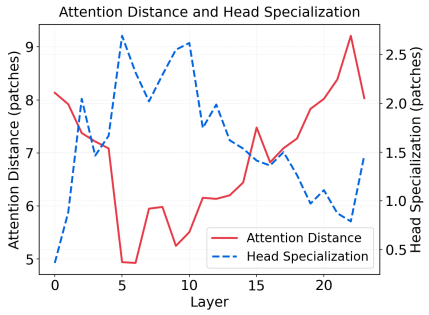


Figure 19. In the physics emergence zone, the average attention head distance drops, and head specialization spikes as local heads emerge among the longer distance heads.

Table 4. Effect of ablating local attention at varying spatial (s, in patches) and temporal (t, in tubelets) thresholds. Spatial ablation minimally affects direction R^2 but degrades per-patch localization; temporal ablation strongly impacts both direction and IntPhys; combined spatiotemporal ablation destroys direction encoding entirely while IntPhys and ImageNet degrade more gradually.

Condition	**Params**	**Direction** R^2	**IntPhys** Acc	**Per-patch** R^2	**ImageNet** Acc
Baseline	s=0, t=0	0.97	78.3	0.72	33.7
Spatial-only (temporal preserved)					
	s=1	0.96	71.4	0.65	33.8
	s=3	0.95	67.2	0.53	33.3
	s=5	0.95	63.9	0.43	33.5
	s=7	0.93	62.2	0.30	33.5
	s=9	0.92	60.6	0.14	33.8
	s=11	0.91	61.1	<0	33.7
	s=13	0.88	60.8	<0	33.9
Temporal-only (spatial preserved)					
	t=1	0.94	76.4	0.64	33.3
	t=2	0.85	60.6	0.48	33.0
	t=3	0.83	51.9	0.41	30.3
	t=4	0.82	50.6	0.36	28.0
	t=5	0.81	50.8	0.29	27.0
	t=6	0.80	50.8	0.24	25.6
Spatiotemporal (both knocked out)					
	s=3, t=1	0.14	61.7	<0	33.1
	s=5, t=2	<0	60.3	<0	31.8
	s=7, t=3	<0	56.4	<0	29.7
	s=9, t=4	<0	56.7	<0	27.3
	s=11, t=5	<0	58.1	<0	19.5
	s=13, t=6	<0	50.8	<0	11.2

C.7. Direction Tuning Analysis

In Section 7.1, we identified that the Physics Emergence Zone has circular population geometry for direction tuning in its MLP neurons. To characterize the direction selectivity of individual neurons within V-JEPA 2, we fit generalized linear models (GLMs) to predict each neuron's activation from the stimulus direction. For each neuron i at each spatiotemporal position (patch $\times$ frame), we model the activation y as a linear function of direction θ:

$$y = \beta_0 + \beta_{\cos}\cos(\theta) + \beta_{\sin}\sin(\theta) + \epsilon \tag{4}$$

where $\theta \in [-\pi, \pi]$ is the motion direction in radians. This sinusoidal basis captures smooth, circular tuning curves commonly observed in biological direction-selective neurons.

We evaluate each neuron's direction tuning strength using cross-validated ΔR^2, computed via k-fold cross-validation ($k = 5$). For each fold, we fit the GLM on training samples and compute the coefficient of determination on held-out validation samples. The cross-validated ΔR^2 quantifies how much variance in the neuron's response is explained by direction, while guarding against overfitting. We regularize the GLM with ridge regression ($\alpha = 10^{-3}$) to ensure numerical stability.

To extract each neuron's preferred direction (PD), we fit the GLM on the full dataset and compute:

$$\text{PD}_i = \arctan 2(\beta_{\sin}, \beta_{\cos}) \tag{5}$$

This yields the direction that maximally activates each neuron. We also compute the direction tuning gain as $\sqrt{\beta_{\cos}^2 + \beta_{\sin}^2}$, which reflects the amplitude of the neuron's direction modulation.

To visualize the population-level direction tuning, we construct heatmaps with neurons on the vertical axis and preferred direction (binned into 24 bins spanning $[-\pi, \pi]$) on the horizontal axis. For each neuron, we assign its cross-validated ΔR^2 to its corresponding preferred direction bin. When multiple spatiotemporal positions for the same neuron fall into the same bin, we aggregate using the maximum ΔR^2 across positions. Neurons are sorted vertically by their peak preferred direction bin, then by tuning strength within each bin. The resulting heatmap reveals whether the network contains a diverse population of direction-tuned neurons spanning all directions, or whether direction tuning is sparse or concentrated at particular angles.

We further find that the circular population code only emerges at the Physics Emergence Zone, and that it is not present at the early layers, with disorganized direction tuning in Layer 0 (Appendix Fig 20)

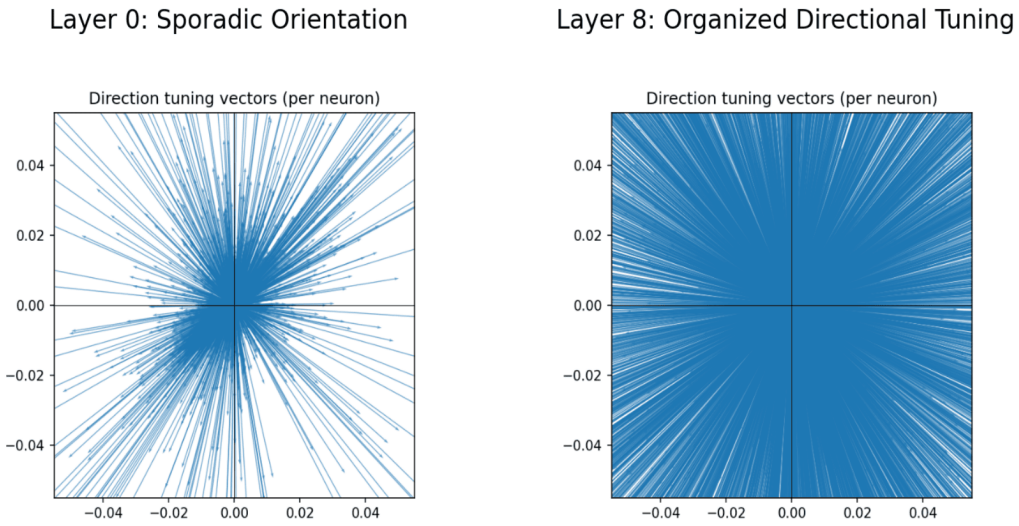


Figure 20. Direction tuning vectors show that direction tuning is sporadic and disorganized at Layer 0, but has emerged in neat organization by Layer 8 of the Physics Emergence Zone

C.8. Speed Tuning Analysis

To characterize speed selectivity, we fit a separate GLM for each neuron predicting activation from stimulus speed. We use a quadratic model to capture neurons with preferred speeds at intermediate values:

$$y = \beta_0 + \beta_r \cdot r + \beta_{r^2} \cdot r^2 + \epsilon \tag{6}$$

where $r \geq 0$ is the speed magnitude. The quadratic term allows the model to capture neurons that respond maximally at a particular speed rather than monotonically increasing or decreasing with speed.

As with direction tuning, we evaluate speed tuning strength using cross-validated ΔR^2 with k-fold cross-validation ($k = 5$) and ridge regularization ($\alpha = 10^{-3}$). To extract each neuron's preferred speed, we compute the vertex of the fitted parabola:

$$r_i^* = -\frac{\beta_r}{2\beta_{r^2}} \tag{7}$$

This preferred speed is only well-defined when $\beta_{r^2} < 0$ (i.e., the parabola opens downward, indicating a true peak). We clip r^* to lie within the observed speed range $[r_{\min}, r_{\max}]$. The speed tuning gain is defined as $|\beta_r|$, reflecting the neuron's sensitivity to speed changes.

Population heatmaps for speed tuning follow the same construction as for direction: neurons are binned by their preferred speed (24 bins spanning the 1st to 99th percentile of observed preferred speeds), with cross-validated ΔR^2 aggregated via maximum within each bin. Neurons are sorted by their peak preferred speed bin. These heatmaps reveal whether the network contains a continuum of speed-tuned neurons or whether speed encoding is concentrated at particular values.

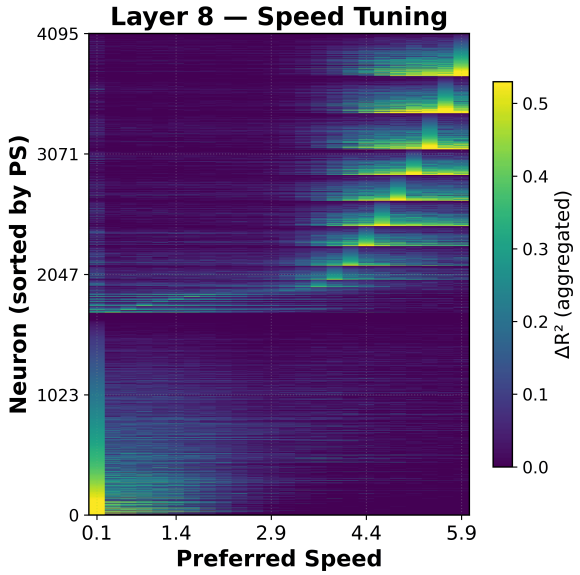


Figure 21. Heatmap of preferred speed for V-JEPA 2 L.

C.9. High feature dimensionality of physics-related representations

In Section 7.2, we discuss the high feature dimensionality of physics-related information.

C.10. Speed representations do not have sawtooth pattern

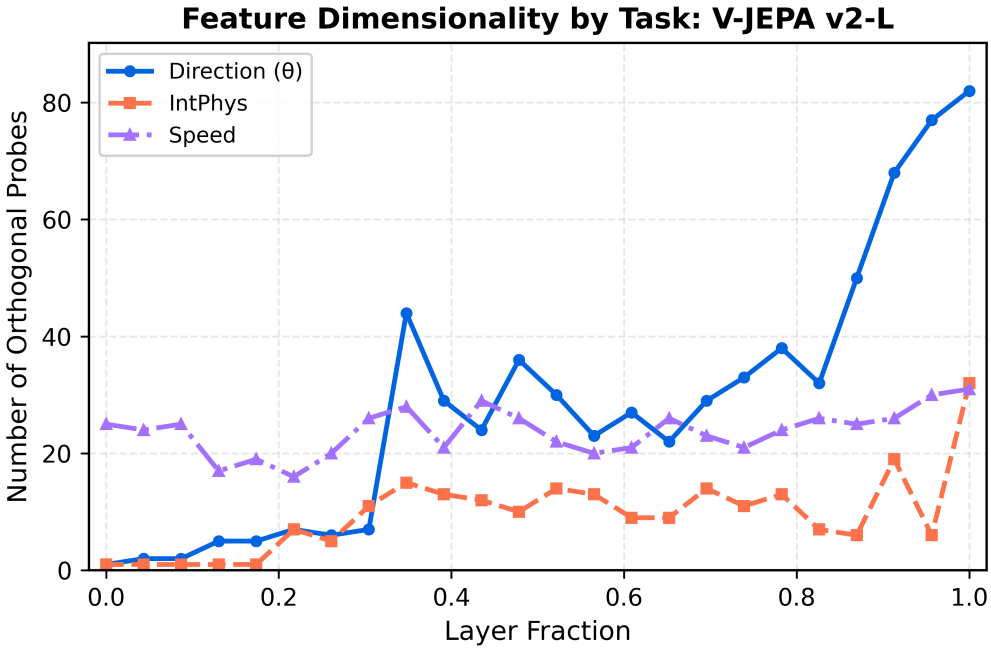


Figure 22. Estimated feature dimensionality of decoded variables across tasks, measured by the number of orthogonal linear probes trainable before performance approaches chance (Direction: $R^2 < 0.3$; Speed: $R^2 < 0.1$; IntPhys: accuracy $< 55\%$). Physical variables require tens of independent feature dimensions for representation.

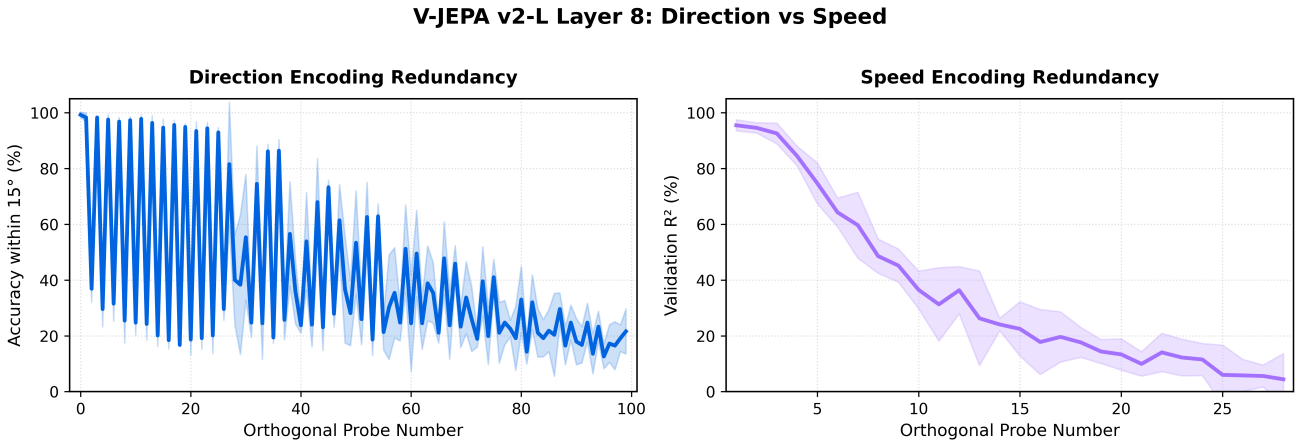


Figure 23. Direction shows a jagged sawtooth pattern, while speed does not.

C.11. Orthogonal Probe Sequence Method

To measure the dimensionality of motion and physics representations, we train sequences of linear probes on progressively orthogonalized activation subspaces. This procedure quantifies how many independent directions in activation space encode a given variable.

Procedure. For a layer with activations $\mathbf{X} \in \mathbb{R}^{N \times d}$ (where N is the number of samples and $d = 1024$ is the embedding dimension), we iteratively:

1. Train a linear probe P_k on the current activations $\mathbf{X}^{(k)}$ to predict the target variable (direction θ, speed, or IntPhys label).
2. Extract the probe weights $\mathbf{W}_k$ and compute an orthonormal basis via QR decomposition.
3. Project out the learned direction: $\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} - \mathbf{X}^{(k)}\mathbf{Q}_k\mathbf{Q}_k^\top$, where $\mathbf{Q}_k$ is the orthonormal basis of $\mathbf{W}_k$.
4. Repeat until probe performance falls below threshold.

The number of probes trained before reaching chance-level performance indicates the effective dimensionality of the representation for that variable.

Probe Architecture. All probes are single-layer linear models:

- **Direction (θ):** Circular regression with outputs $(\sin\theta, \cos\theta)$, trained with MSE loss.
- **Speed:** Linear regression with scalar output, trained with MSE loss.
- **IntPhys:** Binary logistic regression (possible vs. impossible), trained with cross-entropy loss.

Training Details. Probes are trained using Adam optimizer with learning rate $\eta = 10^{-3}$ and weight decay $\lambda = 10^{-4}$ for 100 epochs (direction) or 50 epochs (speed, IntPhys). We use an 80/20 train/test split with a fixed random seed for reproducibility.

Stopping Criteria. The probe sequence terminates when performance approaches chance level:

- **Direction:** $R^2 < 0.1$ or circular MAE $> 80°$ (chance $\approx 90°$)
- **Speed:** $R^2 < 0.05$ or MAE $> 90\%$ of random baseline
- **IntPhys:** Accuracy $< 55\%$ or AUC < 0.55 (chance $= 50\%$)

Subspace Dimensionality. The total number of probes K trained before stopping defines the subspace dimensionality. For direction probes (with 2D output for $\sin / \cos$), the effective dimensionality is $2K$. For speed and IntPhys probes (1D output), the dimensionality equals K. Across layers 0–23, we find direction subspaces of dimension 14–136, speed subspaces of dimension 16–31, and IntPhys subspaces of dimension 1–15 (Table 3).

C.12. Steering

In Section 7.2, we show that internal representations of physics variables are surprisingly high-dimensional. Here we verify that the identified direction subspace causally controls direction encoding through activation steering experiments with proper held-out evaluation.

Subspace Construction. From the orthogonal probe sequence (Appendix C.11), we obtain K probes with weights $\mathbf{W}_k \in \mathbb{R}^{2 \times d}$ predicting $[\sin\theta, \cos\theta]$. We construct an orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times 2K}$ for the direction subspace by stacking the probe weight matrices and applying QR decomposition:

$$\mathbf{V}, _ = \text{QR}\left([\mathbf{W}_1^\top, \mathbf{W}_2^\top, \ldots, \mathbf{W}_K^\top]\right) \tag{8}$$

Steering Procedure. Given activations $\mathbf{x} \in \mathbb{R}^d$ and target angle θ^*, we:

1. **Project** activations onto the direction subspace: $\mathbf{c} = \mathbf{V}^\top\mathbf{x}, \quad \mathbf{x}_\perp = \mathbf{x} - \mathbf{V}\mathbf{c}$
2. **Solve** for target coordinates $\mathbf{c}^*$ via least squares such that all probes predict θ^*
3. **Reconstruct** steered activations: $\mathbf{x}^* = \mathbf{V}\mathbf{c}^* + \mathbf{x}_\perp$

Generalization Experiment. To verify that steering affects the true direction representation—not just the specific probes used for steering—we design a held-out evaluation protocol:

1. **Split** the dataset into disjoint train (70%, 240 videos) and test (30%, 103 videos) sets
2. **Train steering probes** on train set activations using the orthogonal probe sequence (25 probes until $R^2 < 0.1$)
3. **Train a held-out evaluation probe** on test set activations only ($R^2 = 0.99$)
4. **Apply steering** (constructed from train-set probes) to test set activations
5. **Evaluate** whether the held-out probe reads the target direction from steered activations

This protocol ensures the evaluation probe has never seen the steering probes or the activations used to construct the steering subspace, providing a true test of generalization.

Results. We steer videos with 8 discrete motion directions $(0, 45, 90, \ldots, 315)$ toward a single target angle $\theta^* = 90$. This requires angular shifts ranging from 0 (for videos already at 90) to 180 (for videos at 270), with an average shift of approximately 90. At layer 8 of V-JEPA 2-L:

- **Baseline** (no steering): The held-out probe reads the true direction with MAE $= 4.9$ to ground truth, but MAE $= 82.9$ to the target angle (as expected given the average angular distance).
- **After steering with 20 probes**: The held-out probe reads MAE $= 11.9$ to the target, demonstrating that steering *generalizes* to a probe trained on entirely different data—a 71 improvement over baseline.

Figure 24 shows that effective steering requires coordinated intervention across many orthogonal directions: steering with only 1–5 probes yields modest improvement (MAE > 50), while steering with $\sim$20 probes achieves MAE ≈ 12. Importantly, as MAE to the target decreases, MAE to the true labels increases correspondingly, confirming that we are genuinely shifting the representation toward the target direction rather than simply injecting the target into a separate subspace.

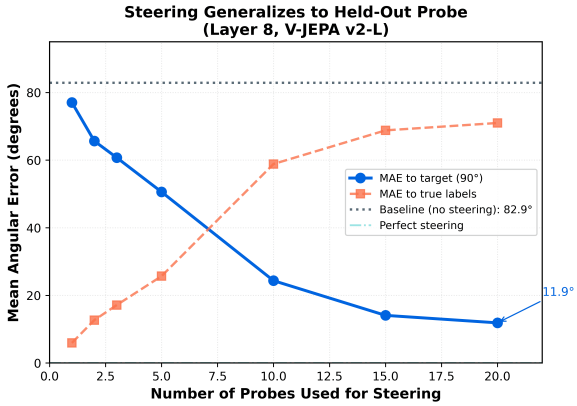


Figure 24. Steering generalizes to held-out data and probes. We train orthogonal probes on the train set (70%) and a separate evaluation probe on the test set (30%). Videos span 8 motion directions (0–315), all steered toward target $\theta^* = 90$. After steering test activations using N train-set probes, the held-out probe (trained only on test data) reads the target direction with MAE decreasing from 82.9 (baseline) to 11.9 with 20 probes. Layer 8, V-JEPA 2-L.

D. Extended Experiments and Analyses

This appendix presents additional experiments and analyses that extend the main paper's findings to a broader set of architectures, formalize the definition of the Physics Emergence Zone (PEZ), and demonstrate that PEZ-localized layers carry the bulk of physics-relevant capacity.

D.1. Extended model coverage: 14 models across 7 architectural families

To assess how broadly our findings generalize, we extend the original model set (V-JEPA 2 and VideoMAE-v2) to 14 models spanning 7 architectural families: latent prediction (V-JEPA 2 L/H/G), pixel reconstruction (VideoMAE-v2 B/L/H/G), diffusion (CogVideoX-2b (Yang et al., 2024), Open-Sora (Zheng et al., 2024)), autoregressive (Cosmos AR-4B (NVIDIA et al., 2025)), 3D CNN classification (SlowFast (Feichtenhofer et al., 2019), I3D (Carreira & Zisserman, 2017)), tracking

(TCOW (Van Hoorick et al., 2023)), and contrastive video-language (InternVideo (Wang et al., 2024)). Models span 28M to 4B parameters.

Table 5 summarizes IntPhys accuracy (possible-vs-impossible discrimination), motion direction R^2 (direction regression on the synthetic ball dataset), and Physion accuracy (Bear et al., 2021) for each model. PEZ presence is determined by the formal sigmoid criterion of Appendix D.2.

Table 5. Extended model coverage. Direction is encoded above chance in every architecture tested (peak $R^2 \geq 0.43$), but only a subset satisfy the formal sigmoid PEZ criterion (✓ markers). Physics plausibility (IntPhys) emerges as a clean PEZ pattern only in models trained with self-supervised predictive objectives (latent prediction, pixel reconstruction at scale, diffusion); classification, contrastive, autoregressive, and tracking objectives encode direction without producing a Physics Emergence Zone for plausibility. Each cell reports the ✓/× PEZ judgment together with the peak value across layers (accuracy for IntPhys and Physion, R^2 for direction).

Model	Objective	Params	IntPhys PEZ	Direction PEZ	Physion PEZ
V-JEPA 2-L	latent prediction	300M	✓ 97%	✓ .96	✓ 75%
V-JEPA 2-H	latent prediction	600M	✓ 100%	✓ .89	✓ 73%
V-JEPA 2-G	latent prediction	1B	✓ 100%	✓ .87	✓ 74%
VideoMAE-v2-G	pixel reconstruction	1B	✓ 84%	✓ .95	× 68%
VideoMAE-v2-B	pixel reconstruction	86M	× chance	✓ .78	× 73%
VideoMAE-v2-L	pixel reconstruction	300M	× chance	✓ .80	× 71%
VideoMAE-v2-H	pixel reconstruction	600M	× chance	✓ .85	× 76%
CogVideoX-2b	diffusion	1.7B	✓ 74%	✓ .88	× 64%
Open-Sora	diffusion	760M	✓ 71%	✓ .69	× 71%
Cosmos AR-4B	autoregressive	4B	× 66%	× .66	× 61%
SlowFast	3D CNN classification	33M	× chance	✓ .83	× 67%
I3D	3D CNN classification	28M	× chance	✓ .88	× 68%
TCOW	tracking	100M	× chance	× .43	× 64%
InternVideo	contrastive	1B	× chance	✓ .65	× 61%

Two patterns are visible. First, motion direction is universal: all 14 models encode direction above chance ($R^2 \geq 0.43$), regardless of objective. Second, physics plausibility is selective: only large self-supervised predictive models (latent prediction, pixel reconstruction at scale, diffusion) develop an IntPhys PEZ. Classification, contrastive, autoregressive, and tracking objectives do not, even when they encode direction comparably well. Within the reconstruction family, only VideoMAE-v2-G clears the IntPhys bar, suggesting a capacity threshold for plausibility emergence under the reconstruction objective. Encoder PEZ depth concentrates at $28\% \pm 4\%$; diffusion models show PEZ at 34–45% depth.

D.2. Formal definition of the Physics Emergence Zone

In the main text, the Physics Emergence Zone is identified visually as the layer at which a sharp accessibility transition occurs. To make this definition quantitative, we fit a four-parameter sigmoid to each layer-wise accuracy or R^2 curve and require:

1. Sigmoid fit quality: $R^2 > 0.9$ (curve is well described by a sigmoid),
2. Inflection depth: inflection point at $\leq 50\%$ of network depth (the transition is in the early-to-middle portion of the network, not at the output),
3. Magnitude: peak accuracy ≥ 15 percentage points above chance, or peak regression $R^2 \geq 0.1$.

A model satisfies the PEZ criterion for a given task if all three conditions hold. Applied to V-JEPA 2-L on IntPhys, sigmoid fits achieve $R^2 > 0.97$; the inflection sits at 25% of network depth with peak 95% accuracy. Across the 14 models in Appendix D.1, the encoder family shows IntPhys PEZ at $28\% \pm 4\%$ depth (mean $\pm$ std over models satisfying the criterion); diffusion models satisfy the criterion at 34–45% depth. The criterion is what underlies the ✓/× entries in Table 5.

D.3. Layer-targeted fine-tuning: PEZ layers carry the physics capacity

If the Physics Emergence Zone identifies the layers in which physics-relevant structure resides, it should be possible to specialize a model for an intuitive-physics task by fine-tuning only those layers. We freeze V-JEPA 2-L and unfreeze 4 of its

24 transformer blocks, then fine-tune on IntPhys with all other parameters held fixed. We compare four layer groups: PEZ layers (blocks 7–10), late layers (blocks 20–23), random layers (blocks 4, 5, 12, 18), and early layers (blocks 0–3). For each group we additionally evaluate CLEVRER direction regression (Yi et al., 2020) (per-object R^2 on the highest-R^2 object type, `cube_yellow_rubber`) and ImageNet classification accuracy.

Table 6. Surgical fine-tuning of 16% of parameters (4 of 24 transformer blocks). The Physics Emergence Zone layers (blocks 7–10) achieve perfect IntPhys accuracy with zero variance across folds, while sacrificing static-image performance; late layers (blocks 20–23) recover ImageNet accuracy but plateau on IntPhys. The double dissociation indicates PEZ layers are physics-specialized and late layers are general-purpose.

Layer Group	Layers Trained	IntPhys Acc (%)	CLEVRER Direction R^2 ($\times 100$)	ImageNet Acc (%)
PEZ layers	7–10	$\mathbf{100.0 \pm 0.0}$	$\mathbf{97.4 \pm 2.6}$	27.8 ± 13.6
Late layers	20–23	85.8 ± 3.3	94.1 ± 2.4	$\mathbf{68.7 \pm 15.8}$
Random control	4, 5, 12, 18	98.1 ± 2.6	92.0 ± 6.6	7.8 ± 5.7
Early layers	0–3	57.5 ± 5.1	85.2 ± 8.0	7.6 ± 4.5

PEZ layers reach 100% IntPhys with zero across-fold variance, despite training only 16% of model parameters. Late layers (also 16%) plateau at 85.8% IntPhys but recover most ImageNet performance (68.7% vs. 27.8% for PEZ layers), giving a clean double dissociation: PEZ layers specialize in physics, late layers in general visual semantics. Random and early-layer baselines show that this separation is not a generic property of partial fine-tuning. The result is consistent with the appendix experiment showing PEZ-layer features outperform final-layer features on downstream physics tasks (Appendix C.1.4), and with concurrent observations by Bolya et al. (2025) that intermediate layers carry task-relevant information often discarded by later layers.

ImageNet variance for PEZ layers is elevated due to strided fold sampling; means are valid but the standard deviation reflects sampling noise rather than instability of the underlying performance.

D.4. Robustness of the Physics Emergence Zone to input degradation

We test whether the PEZ depth depends on input statistics by varying spatial resolution and frame count for V-JEPA 2-L on IntPhys, applying the formal sigmoid criterion of Appendix D.2.

Table 7. PEZ depth is invariant under spatial and temporal degradation of the input. Peak accuracy degrades modestly as resolution drops, and is robust to temporal subsampling. PEZ depth (location of the sigmoid inflection point) varies by at most 4 percentage points across all conditions.

Condition	PEZ Depth	Peak Acc	Sigmoid Fit R^2
Baseline (256×256, 16 frames)	25%	95%	0.995
128×128, 16 frames	25%	69%	0.967
64×64, 16 frames	29%	81%	0.993
256×256, 8 frames	26%	92%	0.995
256×256, 4 frames	25%	93%	0.997

PEZ depth lies in 25–29% across all conditions (mean $26\% \pm 2\%$), with sigmoid fit $R^2 > 0.96$ throughout. Peak accuracy drops from 95% to 69% as spatial resolution falls (consistent with reduced visible motion cues at low resolution), but is essentially unchanged under temporal subsampling down to 4 unique frames. The PEZ therefore reflects computational staging within the architecture rather than an artifact of input statistics.

D.5. Mathematical model of the sawtooth orthogonalization pattern

The main text reports a sawtooth oscillation in iterative null-space probe accuracy when removing direction information (recovery bumps after removals). We provide a mechanism. Let S_m and C_m denote the signal strength of $\sin\theta$ and $\cos\theta$ encoding components after m orthogonalization steps. Suppose at step m the linear probe captures fraction α_m of the available $\sin$ component and β_m of the $\cos$ component, with the asymmetry arising because individual heads encode horizontal and vertical motion preferentially. Then:

$$S_{m+1} = (1 - \alpha_m)\, S_m, \tag{9}$$

$$C_{m+1} = (1 - \beta_m)\, C_m. \tag{10}$$

A circular probe must recover both components jointly, so circular R^2 scales with the slower of the two recursions. When capture is highly asymmetric ($\alpha_m \gg \beta_m$), the $\sin$ component is removed quickly while the $\cos$ component survives; subsequent probes can pick up the residual $\cos$ signal, producing the observed recovery bump. A scalar (e.g., speed) encoding has only a single recursion, $S_{m+1} = (1 - \alpha_m)\, S_m$, with no second component to recover, and therefore no oscillation.

We confirm this mechanism with a synthetic experiment. We construct a paired sin–cos encoding with $R = 60$ redundant copies of $(\sin\theta, \cos\theta)$, and a scalar speed encoding with $R = 28$ redundant copies of a single magnitude. Iterative null-space probing collapses the paired encoding only after 40 orthogonalization steps, while the scalar encoding collapses after 8. The $5{:}1$ ratio (40 vs. 8) closely matches the relative depths observed in real V-JEPA 2 representations. The fine-grained oscillation pattern in real models likely depends on additional structure in the head-level encoding beyond what this minimal model captures.

D.6. CNN comparison: SlowFast and I3D show direction without IntPhys

The 3D CNNs SlowFast (Feichtenhofer et al., 2019) and I3D (Carreira & Zisserman, 2017) provide a useful contrast to transformer encoders. Both are trained on action-classification objectives. We evaluate them on motion direction regression and on IntPhys, applying the same probing protocol used for the encoder family.

SlowFast and I3D both encode motion direction strongly: peak direction R^2 is 0.83 and 0.88 respectively, comparable to V-JEPA 2-L (0.96). However, both remain at chance on IntPhys, with no sigmoid fit satisfying the PEZ criterion. This sharpens the dissociation observed in Appendix D.1: motion direction is encoded across architectural families and training objectives, whereas physics plausibility specifically tracks the use of self-supervised predictive objectives. Architecture alone (transformer vs. CNN) is not the controlling factor; the InternVideo contrastive transformer also shows direction emergence ($R^2 = 0.65$) without an IntPhys PEZ, reinforcing that training objective is the relevant axis.

D.7. Physion benchmark and visual confounds

To evaluate PEZ on more naturalistic stimuli, we add the Physion benchmark (Bear et al., 2021), which contains 1,200 photorealistic trials of physical-prediction scenarios (collisions, draping, rolling, containment, etc.). Per-model Physion accuracies are included in Table 5.

Two caveats accompany the Physion results. First, naturalistic physics benchmarks are known to contain spurious cues such as texture, color, and scene layout that inflate baselines independent of physical reasoning, as the Physion authors themselves note (Bear et al. (2021), Sec. 3). Empirically, models that score at chance on IntPhys nonetheless reach 61–76% on Physion, consistent with this pattern. Second, IntPhys is the only benchmark we evaluate in which paired stimuli differ *only* in physical plausibility, isolating physics from visual confounds; our primary claims are therefore anchored on IntPhys, with Physion serving as a complementary naturalistic check rather than a primary measure. Designing benchmarks that disentangle physics from visual confounds in naturalistic video remains an open problem.

D.8. Why physical structure emerges at one-third depth

The PEZ depth (25–45% across model families; see Appendix D.2) is an empirical observation. Four complementary perspectives help interpret why this particular depth, rather than earlier or later layers, supports the emergence of physical structure.

Computational staging. Early layers build local features (edges, motion energy) over restricted receptive fields. By approximately one-third depth, progressive attention has integrated sufficient spatial and temporal context to support globally coherent motion inference. The PEZ marks the layer at which this aggregation has produced a representation rich enough for both direction decoding and physical-plausibility judgments.

Neuroscience parallel. In primate visual cortex, speed-sensitive neurons appear early in the dorsal stream (V1), while direction selectivity requires higher-order pooling in MT/V5 (Albright, 1984; Born & Bradley, 2005). Our hierarchy mirrors this scalar-before-vector progression: scalar speed is decodable from early layers, while motion direction emerges only at the PEZ.

Attention head diversity. Section 6.3 (Fig. 3) shows a sharp increase in attention head diversity at one-third depth:

spatiotemporally local heads and longer-range heads emerge in the same blocks. The PEZ coincides with the layer at which the network first supports both fine-grained motion tracking and broader temporal integration.

Patch-level transition. Direction encoding transitions from locally encoded (a small number of patches carry direction information) to globally distributed (direction is decodable from many patches) precisely at PEZ depth (Appendix C.5). This transition is consistent with motion information being integrated across the visual field at this stage.

These perspectives are not mutually exclusive: each highlights a different way in which the architecture's processing pipeline transitions at one-third depth.

D.9. IntPhys2 and the next-generation evaluation ceiling

IntPhys is currently the only intuitive-physics benchmark whose paired stimuli differ only in physical plausibility, isolating physics from visual confounds. A more recent benchmark, IntPhys2 (Bordes et al., 2025), extends the violation-of-expectation paradigm to harder scenarios. At present most models, including V-JEPA 2 at all scales, perform near chance on IntPhys2; reliable PEZ analysis is therefore not yet feasible on this benchmark. We view IntPhys2 as a target for future evaluation as model capability increases: understanding where the present ceiling resides, and whether closing it requires architectural changes or training-objective changes, is an open question that the framework developed here is well-positioned to investigate.

D.10. Beyond linear motion: harmonic and circular trajectories

The synthetic toy-ball dataset isolates direction by restricting motion to linear trajectories with constant or uniformly accelerated velocity, providing precise ground-truth control. CLEVRER (Yi et al., 2020) extends this with post-collision angular deflections across 48 object types; V-JEPA 2 shows PEZ consistently across all of these object types, and similarly across the diverse scenarios in Physion (Bear et al., 2021). Together these results suggest the PEZ phenomenon generalizes beyond strict linear trajectories. Explicitly controlled periodic motion, such as harmonic oscillation and circular orbits, would test the framework on a qualitatively different class of dynamics; we leave such experiments to future work that the present probing methodology directly enables.

D.11. Violation types probed in IntPhys

The IntPhys probe-training stimuli used in the main paper consist of matched possible/impossible video pairs that differ only at a single "breakpoint" frame (see Fig. 5 for representative examples). Following the violation-of-expectation paradigm in developmental psychology (Baillargeon & DeVos, 1991; Spelke et al., 1995), the impossible variants instantiate three core violation categories:

Object permanence. An object that should remain present spontaneously disappears across the breakpoint frame.

Shape constancy. An object's geometry changes spontaneously across the breakpoint frame (for example, a cube becomes a cone).

Spatiotemporal continuity. An object in motion teleports to a non-adjacent location, or its trajectory abruptly reverses, across the breakpoint frame.

These categories correspond to the three IntPhys subtasks whose layer-wise emergence is broken down in Fig. 10; the Physics Emergence Zone is observed for all three violation types, indicating that PEZ is not an artifact of any single category.