Physics-Steerable Video World Models for Embodied Scenario Generation

Nahid Alam
Oreon Labs
nahid.m.alam@gmail.com

Abstract

Embodied agents need world models that can generate physically varied scenarios on demand, without retraining. Recent work shows that physical plausibility is linearly encoded in a frozen VideoMAE encoder at a consistent depth region called the Physics Emergence Zone (PEZ) [3], and that Concept Activation Vectors (CAVs) can steer this representation at inference time. We extend this foundation toward a fully controllable video world model with three new primitives: pixel-space scenario output by coupling PEZ-layer steering to VideoMAE's MAE decoder, compositional control of physics and motion independently (CAV angle $86.6°$*, interaction norm* $\leq$ 0.42*), and temporal scheduling of physical violations across the video. A key finding is that physics is encoded uniformly across all temporal positions in VideoMAE, meaning fine-grained temporal control requires architectural changes beyond joint space-time attention. Together, these primitives enable stress-testing of embodied agents against out-of-distribution physical scenarios with no new data or retraining.*

1. Introduction

Embodied AI agents — autonomous robots, drones, and navigation systems — must reason about physical plausibility to act safely in the world. A robot that cannot distinguish a physically possible outcome from an impossible one will fail at manipulation, grasping, and long-horizon planning. World models [1, 2] are increasingly proposed as the substrate for this physical reasoning, yet they remain largely opaque: one cannot specify "generate a scenario where physics breaks at second 8 with leftward object motion" without retraining or collecting new data.

Recent interpretability work [3] identified a *Physics Emergence Zone* (PEZ) in VideoMAE [5] — a consistent depth region where physical plausibility is linearly encoded and can be steered via Concept Activation Vectors (CAVs) at inference time with no weight updates. However, this leaves three gaps for embodied AI: there is no pixel-space output, no way to compose multiple physical attributes independently, and no temporal resolution over when a violation occurs.

Contributions. We close these gaps with three new primitives:

- **Pixel-space steering pipeline**: coupling PEZ-layer CAV interventions to VideoMAE's MAE decoder produces steered frame reconstructions without any weight updates.
- **Compositional multi-concept control**: independent physics and motion CAVs compose near-linearly ($86.6°$ mutual orthogonality), enabling scenario specification along multiple axes.
- **Temporal scheduling**: applying steering selectively to chosen tubelet positions reveals that physics is uniformly encoded across all temporal positions — a key architectural finding for embodied world model design.

2. Method

PEZ identification. VideoMAE [5] is a 12-layer ViT-Base encoder (D=768, N=1568 tubelet tokens from 16 frames at 224×224). A logistic regression probe on mean-pooled layer representations trained on IntPhys [4] labels peaks at layer l^*=5 (70.1% accuracy, 5-fold, PCA-64), defining the PEZ [3].

CAV extraction. The physics CAV is the normalised probe weight: $\mathbf{v}_{l^*} = \mathbf{w}_{l^*} / \|\mathbf{w}_{l^*}\|_2$, directed toward *impossible* (higher score = more implausible). A motion-direction CAV $\mathbf{v}_{\text{motion}}$ is extracted identically from synthetic ball-trajectory videos with controlled lateral motion.

Steering primitives. At inference time we add scaled CAVs to all patch tokens at the PEZ layer with no weight modifications:

$$\tilde{\mathbf{H}}_{l^*}(\mathbf{x})_i = \mathbf{H}_{l^*}(\mathbf{x})_i + \alpha_p \, \mathbf{v}_{\text{phys}} + \alpha_m \, \mathbf{v}_{\text{motion}}. \quad (1)$$

Setting α_m=0 gives physics-only steering; α_p=0 gives motion-only; both non-zero gives compositional control. Temporal selectivity is achieved by masking the additive term to tokens at chosen tubelet indices $t \in \{0, \ldots, 7\}$, with optional ramp schedules $w_i = t_i/(T/\tau - 1)$.

Pixel-space output. Steered encoder outputs are passed to VideoMAE's frozen MAE decoder. Decoder predictions for masked patches (75% mask ratio) are un-normalised and assembled with original visible patches to produce a full steered video.

3. Results

All experiments use the IntPhys [4] dev split (216 train / 72 val / 72 test, stratified) on a single NVIDIA L40S GPU. **Pixel-space steering (Exp. 1):** steered reconstructions (α=10, 75% mask ratio) differ from unsteered by mean frame-level ℓ_1=0.0021 with PSNR change < 0.01 dB (Table 1). **Compositional steering (Exp. 2):** the angle between $\mathbf{v}_{\text{phys}}$ and $\mathbf{v}_{\text{motion}}$ is 86.6°; physics control (P(impossible)=1.00) is fully preserved across all α_m, with interaction norm $\|\boldsymbol{\varepsilon}\|_2 \leq 0.42$ (Fig. 1). **Temporal localisation (Exp. 3):** any single tubelet, named window, or ramp schedule achieves **100% of the global flip rate** (0.78) — physics is uniformly encoded across all temporal positions (Fig. 2).

Table 1. Pixel-space steering results.

PSNR baseline (dB)	PSNR steered (dB)	Frame diff
18.6	18.6	0.0021

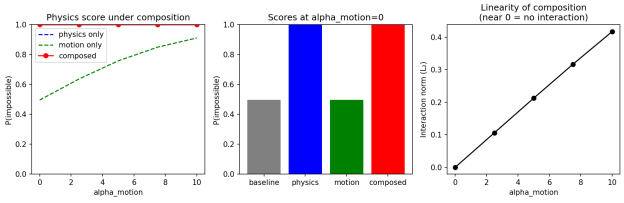

Figure 1. Compositional steering results. *Left*: physics score as α_m varies — P(impossible) stays at 1.00 regardless of motion strength. *Centre*: per-condition scores at α_m=0. *Right*: interaction norm $\|\boldsymbol{\varepsilon}\|_2 \leq 0.42$ confirms near-linear composition of orthogonal CAVs.

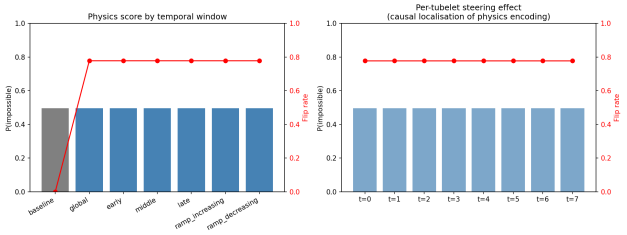

Figure 2. Temporal steering results. *Left*: flip rate for named windows and ramp schedules — all achieve 100% of global. *Right*: per-tubelet flip rate confirms physics is encoded uniformly across all temporal positions in VideoMAE.

4. Broader Impact

Our results carry three implications for embodied AI practitioners. First, physics CAV steering produces adversarial scenarios on demand from any input with no retraining — sweeping α_p and α_m spans a full spectrum of physical plausibility and motion direction from a single frozen checkpoint. Second, near-linear CAV composition (Eq. 1) means scenario properties are independent axes, extending naturally to additional concepts (lighting, object identity) analogously to structured simulation parameter sweeps. Third, physics is uniformly encoded across all tubelet positions in VideoMAE — joint space-time attention prevents precise temporal localisation of violations, so applications requiring fine-grained temporal causal control will need factorized temporal attention or causal masking.

5. Conclusion

We presented physics CAV steering as a practical primitive for embodied scenario generation: a frozen pretrained video world model can be steered at inference time to produce physically adversarial, compositionally specified, and temporally scheduled video scenarios with no weight updates or new data. The finding that physics is uniformly encoded across temporal positions in VideoMAE highlights a concrete architectural gap for the embodied AI community — precise temporal causal control requires moving beyond joint space-time attention. We hope these primitives serve as a foundation for richer physics-aware evaluation and planning tools in embodied systems.

References

[1] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024. 1

[2] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 1

[3] Sonia Joseph, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake Richards, and Mike Rabbat. Interpreting physics in video world models. 2026. 1

[4] Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Veronique Izard, and Emmanuel Dupoux. IntPhys: A framework and benchmark for visual intuition physics. *arXiv preprint arXiv:1803.07616*, 2019. 1, 2

[5] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1