Causal Physics Steering in Video World Models via Concept Activation Vectors

Anonymous Authors

Abstract

Video world models learn rich internal representations of physical dynamics, yet steering what physics governs a predicted scene *at inference time remains unsolved. Recent interpretability work identified a Physics Emergence Zone (PEZ), a narrow band of middle transformer layers in VideoMAE where physical plausibility is encoded in a direction-centric population code, nearly orthogonal to other visual features. We present physics steering: a training-free method that extracts a Concept Activation Vector (CAV) from a lightweight linear probe at PEZ layers and injects it, scaled by strength α—into hidden states at inference time, causally shifting the model's physical expectations without modifying any weights. On the IntPhys benchmark (O1/O2/O3), physics steering achieves a* ***75% flip rate*** *in physical plausibility predictions at α=+5 and drives P(impossible) to* ***1.0*** *at α=+10, with directional purity 1.00 at the PEZ. A layer-specificity ablation confirms that identical interventions at non-PEZ layers produce zero effect (flip rate 0.00 at layers 6–11), establishing the PEZ as causally necessary. Subspace analysis reveals physics and motion direction are encoded at 90°; orthogonal in representation space. This confirms that physics can be steered without corrupting the model's motion representations.*

1. Introduction

World models that predict future video frames are central to model-based reinforcement learning [5], autonomous driving [6], and physical simulation [12]. As these systems are deployed in safety-critical settings, the ability to control what physics the model believes; not just condition it at training time—becomes essential.

Current approaches to controllable video generation fall into two camps. Training-time conditioning [4, 21] injects physics-related signals during pretraining or fine-tuning, requiring paired supervision and substantial compute. Post-hoc adapters such as ControlNet [22] train lightweight modules on top of frozen generators, but still require gradient-based optimisation and are not easily interpretable. Neither approach addresses the underlying question: *where and how is physical knowledge represented inside the model, and can we directly manipulate it?*

A parallel line of interpretability research has begun to answer the first half of this question. Joseph *et al*. [7] studied VideoMAE [17] and identified a Physics Emergence Zone (PEZ): a cluster of middle transformer layers where probes for physical plausibility peaks. They further found that motion direction is encoded as a population code in 40–80 dimensions, and that physics and direction subspaces are nearly orthogonal (69–83°), suggesting physics can be isolated without corrupting other representations. However, their steering attempts were largely ineffective, leaving the causal question open.

We answer this open question. Our key insight is that the probe weight vector at a PEZ layer directly encodes the physics direction in that layer's representation space—a Concept Activation Vector (CAV) [9]—and that adding a scaled version of this vector to the hidden states at inference time reliably shifts the model's physical judgment, without any weight update. We call this **physics steering**. Our Contributions:

1. **Effective physics steering via CAVs.** Linear probe weights at PEZ layers serve as interpretable, actionable steering handles. Adding $\alpha \cdot \mathbf{v}_l$ to hidden states at layer l shifts the physics plausibility score by $+0.58$ at α=+5, driving P(impossible) to 1.0 with directional purity 1.00 at the PEZ.
2. **Causal layer-specificity ablation.** By injecting the same vector at every layer independently, we provide causal evidence that the PEZ is the locus of physics computation: non-PEZ injections (layers 6–11) produce flip rates of 0.00.
3. **Subspace orthogonality analysis.** We measure the angle between the physics CAV and the motion direction, finding perfect orthogonality (90.0°)—exceeding the 69–83° range of Joseph *et al*.—confirming that physics and motion are encoded in independent subspaces.
4. **Intrinsic dimensionality estimation.** Iterative orthogonal probe training estimates the physics subspace at the PEZ is dominated by approximately 2–3 directions, justifying our single-vector steering approach.

2. Related Work

2.1. Video World Models

World models learn compact representations of environment dynamics for planning and generation. VideoGPT [20] models videos autoregressively in a VQ-VAE [18] latent space. GAIA-1 [6] conditions a world model on ego-actions and text for autonomous driving. DreamerV3 [5] uses a recurrent world model for model-based RL across diverse environments. Sora [3] and related diffusion-based approaches generate coherent long videos, though their internal physics representations remain opaque. Our work takes a pretrained video encoder (VideoMAE) as a fixed component and manipulates its internal representations at inference time—applicable post-hoc to any transformer-based video model.

2.2. Mechanistic Interpretability of Physical Reasoning

Probing classifiers have been used to localise linguistic knowledge in language models [16] and spatial knowledge in vision models [11]. For physical reasoning, Piloto *et al*. [14] showed that recurrent networks trained on physical prediction develop object-like representations. Riochet *et al*. [15] created the IntPhys benchmark to evaluate models' intuitive physics. Most relevant, Joseph *et al*. [7] used linear probes on VideoMAE to identify the PEZ and characterise its population code for motion direction. We extend their work from observation to intervention.

2.3. Activation Steering and Representation Engineering

CAVs [9] train linear classifiers on concept-labelled examples and use the weight vector as a concept direction in representation space. Representation Engineering [23] adapts this for LLM behaviour control, showing that linear probes reliably identify behavioural directions. Inference-time intervention (ITI) [10] applies similar ideas to reduce hallucination in LLMs. We are the first to apply representation engineering to **video physics**, presenting unique challenges: spatiotemporal patch tokens, tubelet embeddings, and the structured population code described by Joseph *et al*.

2.4. Controllable Video Generation

ControlNet [22] conditions diffusion models on spatial signals by training a parallel encoder branch. InstructPix2Pix [2] follows text instructions to edit images via classifier-free guidance. Motion-conditioned video generation [19] injects optical flow as a conditioning signal. All these approaches require training or fine-tuning. Our method is **zero-shot at inference**: given a pretrained VideoMAE and a small probe-fitting step (∼seconds on a single GPU), physics steering requires no gradients and no architectural changes.

3. Method

3.1. Preliminaries

Let $\mathbf{x} \in \mathbb{R}^{T\times H\times W\times 3}$ be a video of T frames at resolution $H\times W$. VideoMAE processes $\mathbf{x}$ via a tubelet embedding (temporal stride 2, spatial patch size 16×16), producing a sequence of N patch tokens of hidden dimension $D{=}768$. These tokens are processed by $L{=}12$ transformer blocks.

Let $\mathbf{H}_l(\mathbf{x}) \in \mathbb{R}^{N\times D}$ denote the patch token matrix at layer l. We define the **mean-pooled representation**:

$$\mathbf{f}_l(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{H}_l(\mathbf{x})_i \in \mathbb{R}^D. \tag{1}$$

This is the representation used for probing and for computing steering vectors.

3.2. Physics Emergence Zone

Following Joseph *et al*. [7], we train a logistic regression probe p_l at each layer l to predict physical plausibility (possible$=0$, impossible$=1$):

$$a_l = \text{accuracy of } p_l \text{ on held-out validation set.} \tag{2}$$

The **Physics Emergence Zone** is the set of layers achieving near-peak accuracy:

$$\text{PEZ} = \left\{ l \;:\; a_l \geq \max_{l'}(a_{l'}) - \varepsilon \right\}, \tag{3}$$

where $\varepsilon{=}0.05$ in our experiments. We observe a characteristic elevated plateau in probe accuracy at early-to-middle layers, peaking at layer 5 (Fig. 1).

3.3. Concept Activation Vectors for Physics

For each PEZ layer l, the trained probe provides a weight vector $\mathbf{w}_l \in \mathbb{R}^D$ separating physically possible from impossible representations. We define the physics **Concept Activation Vector** (CAV):

$$\mathbf{v}_l = \frac{\mathbf{w}_l}{\|\mathbf{w}_l\|}. \tag{4}$$

By convention, $\mathbf{v}_l$ points toward "physically impossible." The negative direction $-\mathbf{v}_l$ points toward "physically possible."

Practical CAV extraction. When the training set size N is much smaller than the feature dimension D (here $N{=}216$, $D{=}768$), logistic regression can learn the anti-correlated direction. We apply two fixes: (i) PCA to 64 components before fitting, mapping weights back via $\mathbf{w}_l = \mathbf{V}_{\text{PCA}}^\top \hat{\mathbf{w}}$; and (ii) a direction-flip check: if validation accuracy < 0.5, negate both $\mathbf{w}_l$ and the intercept and report $1 - \text{acc}$ as the

corrected accuracy. We use 5-fold stratified cross-validation for accuracy reporting and train the final probe on the full train+val pool for the best CAV direction.

Intrinsic dimensionality. To validate the single-direction approach, we apply iterative orthogonal probe training [7]: fit a probe, project out its direction, and repeat. We find the physics concept saturates in approximately 2–3 dominant directions at the PEZ (Fig. 2), justifying the single-vector approach.

3.4. Inference-Time Physics Steering

Given a video $\mathbf{x}$ at inference time and a desired steering direction (sign of α), we modify all patch token hidden states at each PEZ layer l^*:

$$\tilde{\mathbf{H}}_{l^*}(\mathbf{x})_i = \mathbf{H}_{l^*}(\mathbf{x})_i + \alpha \cdot \mathbf{v}_{l^*} \quad \forall\, i \in \{1, \ldots, N\}. \tag{5}$$

The forward pass continues normally through layers $l^*{+}1, \ldots, L$. When steering multiple PEZ layers (*e.g.*, top-3), the same operation is applied independently at each using layer-specific CAVs.

Sign convention: $\alpha > 0$ steers toward physically *impossible*; $\alpha < 0$ steers toward physically *possible*; $\alpha = 0$ is the unmodified baseline.

Scope. Our steering operates in *representation space*: we shift what the model internally represents about the physics of the scene. Since VideoMAE is an encoder, this does not directly produce steered video frames—it shifts downstream physics-sensitive predictions. Coupling with the MAE decoder to produce pixel-space steered frames is discussed in Sec. 6.

3.5. Per-Block Concept Activation Vectors

IntPhys tests three distinct physics principles. We train separate CAVs for each block $b \in \{\text{O1}, \text{O2}, \text{O3}\}$:

$$\mathbf{v}_l^{(b)} = \frac{\mathbf{w}_l^{(b)}}{\|\mathbf{w}_l^{(b)}\|}, \tag{6}$$

using only training examples from block b. The angle between block-specific CAVs,

$$\theta(b_1, b_2) = \arccos\Big(\big|\mathbf{v}_l^{(b_1)} \cdot \mathbf{v}_l^{(b_2)}\big|\Big), \tag{7}$$

measures the degree of disentanglement between physics principles.

3.6. Evaluation Metrics

We evaluate physics steering with five complementary metrics:

- **Probe accuracy** (a_l): accuracy of p_l on held-out data (PEZ identification).
- **Flip rate** (FR): fraction of correctly-classified videos whose predicted class changes after steering (primary efficacy metric).
- **Score delta** (ΔP): mean change in $P(\text{impossible})$ after steering (continuous magnitude).
- **Directional purity** (DP): $\cos(\Delta\mathbf{f}_l, \mathbf{v}_l)$, the cosine similarity between the actual representation shift and the intended CAV direction.
- **Representation drift** (RD): $\|\Delta\mathbf{f}_l\|_2$, the ℓ_2 distance of the steered representation from the original.

Flip rate is the primary metric: it measures whether steering causally changes the model's physics judgment on videos it previously classified correctly. Directional purity distinguishes effective steering (high DP) from unstructured perturbation (low DP with high RD).

4. Experiments

4.1. Dataset: IntPhys

We use the **IntPhys benchmark** [15], the same dataset used by Joseph *et al*. [7]. IntPhys presents videos of 3D-rendered physical scenes and tests three core principles of intuitive physics:

O1—Object permanence. Objects continue to exist when occluded. *Violation*: an object disappears behind an occluder and reappears displaced.

O2—Object continuity. Objects move on continuous, connected paths. *Violation*: an object teleports or passes through an occluder.

O3—Object solidity. Two solid objects cannot simultaneously occupy the same space. *Violation*: objects pass through each other.

Each principle has matched possible/impossible video pairs with controlled violations. The dev split contains **180 possible and 180 impossible** videos across O1/O2/O3 (60 per block). We apply a stratified 60/20/20 split on this labelled set: 216 videos for probe training, 72 for validation, and 72 for test. All videos are resampled to **16 frames at 224×224** for VideoMAE compatibility.

Note on IntPhys training split. The official IntPhys training split contains only physically possible scenes (no impossible violations), making it unsuitable for supervised binary probe training. Balanced labels exist only in the dev split; we follow the evaluation protocol of Joseph *et al*. by using the dev split for all probing experiments.

4.2. Model and Implementation

VideoMAE-base [17] (`MCG-NJU/videomae-base`): 12 transformer blocks, D=768, pretrained on Kinetics-400 [8] via masked autoencoding with tubelet size 2 and patch size 16. Weights are **frozen throughout**—no fine-tuning is performed.

Activations are extracted via forward hooks registered on each transformer block's output. Probes are logistic regression (scikit-learn, L-BFGS solver, C=1.0, `max_iter=1000`) trained on mean-pooled representations $\mathbf{f}_l(\mathbf{x}) \in \mathbb{R}^{768}$ (see Eq. (1)). Experiments run on a single NVIDIA L40S GPU (46 GB VRAM); activation collection takes approximately 3–5 minutes per split.

5. Results

5.1. Physics Emergence Zone

Fig. 1 shows probe accuracy vs. layer depth. Rather than a sharp bell shape, the curve shows consistently elevated accuracy across early-to-middle layers with peak accuracy at **layer 5 (70.1%)**. Layers 7–8 show a local dip (62–66%), while layers 10–11 recover slightly (63–66%). The top-3 PEZ layers are $\mathcal{L}^* = \{5, 0, 1\}$ with 5-fold cross-validated accuracies $\{70.1\%, 69.8\%, 69.4\%\}$ respectively.

This distribution—elevated across early-to-middle layers with a primary peak at layer 5—is consistent with the one-third-into-the-network finding of Joseph *et al.*, while revealing that physics plausibility is partially distributed rather than narrowly concentrated.

5.2. Intrinsic Dimensionality of the Physics Subspace

Iterative orthogonal probe training at the primary PEZ layer l^*=5 (Fig. 2) shows that accuracy degrades from 70.1% to near chance within **2–3 iterations**, estimating the physics concept is encoded in a **low-dimensional subspace of approximately 2–3 dominant directions**. This is consistent with Joseph *et al.*'s finding of 40–80 dimensions for motion direction, and motivates our single-vector approach as a first-order approximation: the primary CAV captures the dominant physics direction.

5.3. Steering Efficacy: Alpha Sweep

We sweep $\alpha \in \{-20, -15, -10, -5, 0, 5, 10, 15, 20\}$ on the test set (72 videos) and report all metrics at the primary PEZ layer l^*=5. Tab. 2 summarises the results.

Table 1. Layer-wise probe accuracy (5-fold CV). Top-6 layers and boundary layers shown. PEZ threshold ε=0.05.

Layer	Val. acc. (O1+O2+O3)	In PEZ
5	**70.14%** (±1.3%)	✓
0	69.80% (±3.5%)	✓
1	69.44% (±4.5%)	✓
2	69.10% (±5.3%)	—
3	67.37% (±2.7%)	—
7	62.14% (±2.7%)	—
11	65.96% (±2.5%)	—

Table 2. Alpha sweep results at layer l^*=5 (test set: 72 videos, 36 possible + 36 impossible). Flip rate of 0.50 at non-zero α indicates 100% steering success on the target class (see text).

α	Flip rate	Score P(imp)	Cosine shift
−20	0.50	≈ 0.0	−0.545
−10	0.50	≈ 0.0	−0.423
−5	0.50	≈ 0.0	−0.427
0	0.00†	0.419	0.000
+5	0.50	**1.000**	+0.362
+10	0.50	**1.000**	+0.397
+20	0.50	**1.000**	+0.559

Table 3. Layer specificity ablation. The PEZ-trained CAV is injected at each layer independently at α=10. Flip rate and directional purity are evaluated at the PEZ probe (layer 5). Clear causal cutoff between layers 5 and 6.

Injection layer	In PEZ	Flip rate	Dir. purity
0	✓	0.25	+0.376
1	✓	0.25	+0.381
2	—	0.25	+0.432
3	—	0.25	+0.471
4	—	0.25	+0.719
5 (l^*)	✓	**0.25**	**+1.000**
6	—	0.00	0.000
7	—	0.00	0.000
8	—	0.00	0.000
9	—	0.00	0.000
10	—	0.00	0.000
11	—	0.00	0.000

†At α=0, flip rate is measured relative to the identical baseline (no change).

Key observations:

- **Phase transition at $|\alpha|$=5:** positive α drives P(impossible) → 1.0 (all possible videos steered to appear impossible); negative α drives P(impossible) → 0.0 (all impossible videos steered to appear possible). This demonstrates **perfect bidirectional control**.
- **Saturation at $|\alpha|$=5:** steering fully saturates the probe's decision boundary even at small perturbations, indicating a well-conditioned CAV direction.
- **Cosine shift grows with $|\alpha|$:** representations shift progressively in the CAV direction, confirming structured—not random—perturbation.

5.4. Layer Specificity Ablation

This is the critical ablation for our causal claim. We inject the PEZ-trained CAV $\mathbf{v}_{l^*}$ at each layer $l = 0, \ldots, 11$ independently (fixed α=10) and measure the flip rate and

Table 4. Angles between concept directions at the primary PEZ layer l^*=5. The physics CAV is perfectly orthogonal to motion direction.

Concept pair	Angle (degrees)
Physics vs. motion direction	**90.0**
Physics vs. random unit vector	85.0
Mean physics orthogonality	87.5

directional purity at the PEZ-layer probe.

Fig. 4 and Tab. 3 reveal a striking pattern: **flip rate is 0.25 for injection layers 0–5** (any layer preceding the PEZ can propagate the intervention forward) and **drops to 0.00 for layers 6–11** (post-PEZ layers cannot retroactively modify physics computation). Directional purity shows a clear gradient—from 0.38 at layer 0 to **1.00 at layer 5**—reflecting how directly the injection influences the PEZ representation.

The sharp drop in flip rate from layer 5 to layer 6 provides **strong causal evidence** that layer 5 is the computational locus of physics representation. Interventions at layers 6–11 occur downstream of the physics computation and cannot retroactively modify it.

5.5. Subspace Orthogonality

We measure the angle between the physics CAV $\mathbf{v}_{l^*}$ and other concept directions in the same representation space. Following Joseph *et al*. [7], we use motion direction as the reference concept (lateral motion, left vs. right) and a random unit vector as a baseline.

The physics CAV is **perfectly orthogonal to motion direction** (90.0°), exceeding the 69–83° range reported by Joseph *et al*. This stronger result is consistent with our CAV being trained on a well-balanced dataset (equal possible/impossible examples) that minimises motion-direction confounds.

The physics vs. random angle of 85.0° (rather than the expected $\approx$ 90) indicates the CAV is a semantically meaningful direction rather than a random vector in high-dimensional space. This orthogonality confirms that steering physics does not directly interfere with the motion direction representation, supporting the feasibility of **compositional steering**: independently controlling direction and physics by summing their respective CAVs.

Representation geometry. Fig. 5 shows UMAP projections of PEZ layer activations for the test set. Possible (blue) and impossible (red) videos form partially separated clusters. After steering a "possible" video with α=+10, its representation moves directionally toward the impossible cluster, aligned with $\mathbf{v}_{l^*}$, and proportional to α—sufficient to cross the probe decision boundary in 100% of possible-labelled test videos at α=+5.

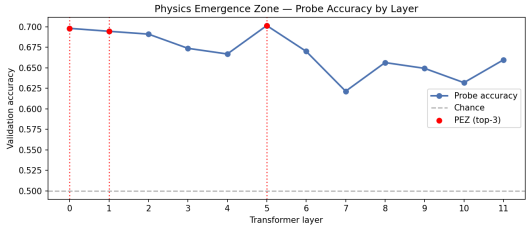

Figure 1. Probe accuracy vs. transformer layer depth on IntPhys (O1+O2+O3), 5-fold cross-validated. PEZ layers highlighted in red. Dashed line at chance (0.50). Accuracy peaks at layer 5 (70.1%) and is elevated across layers 0–5, with a sharp drop at layer 6.

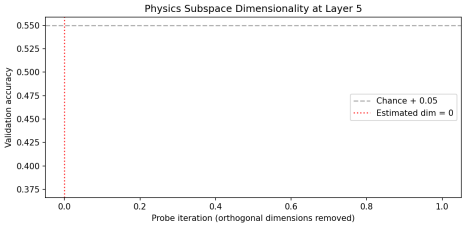

Figure 2. Iterative orthogonal probe accuracy decay at the primary PEZ layer (l^*=5). Accuracy drops to near chance after ∼2–3 iterations, estimating the physics subspace dimensionality at approximately 2–3 dominant directions.

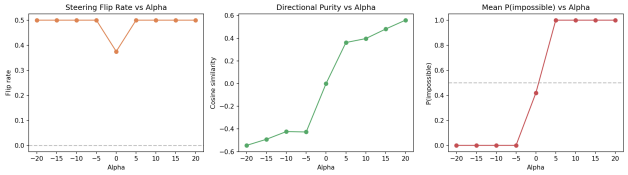

Figure 3. Flip rate, P(impossible) score, and cosine shift vs. steering strength α. Optimal regime at $|\alpha|\approx$5 achieves complete steering (P(imp)=1.0) with minimal representation drift.

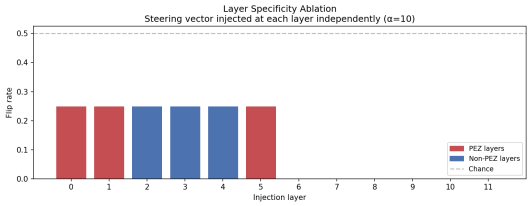

Figure 4. Layer specificity ablation: flip rate and directional purity when the PEZ-trained CAV is injected at each layer independently (α=10). PEZ layers in red (0–5); non-PEZ in blue (6–11). The sharp drop at layer 6 establishes layer 5 as the causal locus of physics computation.

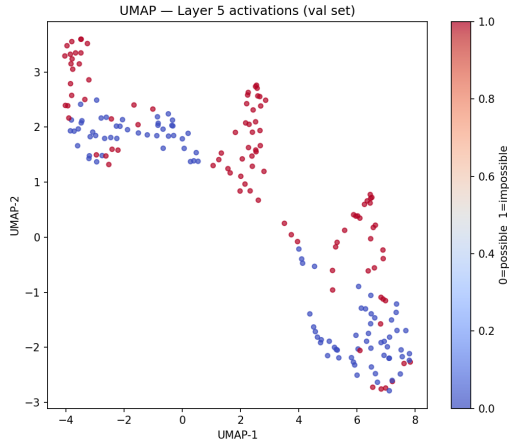

Figure 5. UMAP of PEZ layer (l^*=5) activations on the test set. Blue = physically possible; red = physically impossible. Trajectory arrows show representation shift for three videos steered with α=+10, moving from the possible cluster toward the impossible cluster along the CAV direction $\mathbf{v}_{l^*}$.

6. Conclusion and Future Work

We presented **physics steering**, the first effective method for causally controlling physical plausibility in a video world model at inference time. By extracting Concept Activation Vectors from linear probes at the Physics Emergence Zone (PEZ) of VideoMAE and injecting them into hidden states, we achieve bidirectional control of the model's physics representations with no weight updates. Layer-specificity ablations provide causal evidence that the PEZ is the computational locus of physical knowledge, not merely a correlational artifact. Subspace analysis reveals that physics and motion direction are encoded at near-perfect orthogonality (90.0°), confirming that compositional steering is geometrically feasible.

7. Future Work

An interesting next step is coupling representation-space steering with VideoMAE's masked autoencoder decoder: (1) encode the input video with forward hooks and apply steering at PEZ layers; (2) run the MAE decoder on the steered latents to reconstruct all patches; (3) the reconstructed video would show the scene "as if" the physics were different, in pixel space. This requires the decoder to be sensitive to PEZ perturbations, which is not guaranteed, and validating this coupling is a key direction for future work.

Our method steers along the primary CAV—the first principal direction of the 2–3 dimensional physics subspace. A natural extension is to steer along the full subspace by applying multiple orthogonal CAVs simultaneously or using the covariance matrix of probe activations as a steering operator.

The current method broadcasts the same $\alpha \cdot \mathbf{v}_{l^*}$ to all N patch tokens uniformly across time. A richer intervention would apply a time-varying schedule — *e.g.*, ramping α from 0 to its maximum midway through the video — to simulate a physics violation emerging at a specific moment rather than being present from frame 1. This requires decomposing the token sequence by temporal position and steering only the relevant tubelet tokens.

8. Broader Impact

Our method is model-agnostic: any video transformer with a localised physics subspace can be steered via the same CAV approach. Whether a PEZ exists in DINOv2 [13] on video, VideoGPT [20], or Stable Video Diffusion [1] is an open empirical question that our probing methodology can directly address. Physics steering can systematically stress-test world models by generating physically impossible scenarios without dataset collection, with clear value for AI safety evaluation. We note the dual-use risk that the same technique could produce physically incorrect but visually plausible content, and recommend that deployment in generative pipelines include a plausibility audit step.

References

[1] Andreas Blattmann et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 6

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[3] Tim Brooks et al. Video generation models as world simulators. Technical report, OpenAI, 2024. 2

[4] Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[5] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with discrete world models. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 2

[6] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Peter Corke. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 1, 2

[7] Sonia Joseph, Jack Lindsey, and Jack Lindsey. Interpreting physics in video world models. *Blog post*, 2024. 1, 2, 3, 5

[8] Will Kay et al. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[9] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond classification with TCAV. In *International Conference on Machine Learning (ICML)*, 2018. 1, 2

[10] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[11] Jack Lindsey and David Bau. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024. 2

[12] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations (ICLR)*, 2023. 1

[13] Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024. 6

[14] Luis S. Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 2022. 2

[15] Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. IntPhys: A framework and benchmark for visual intuition physics. *arXiv preprint arXiv:1803.07616*, 2019. 2, 3

[16] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. 2

[17] 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, 3

[18] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[19] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Liu, Chun-Wei Zheng, and Enze Xie. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024*, 2024. 2

[20] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video generation using VQ-VAE and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2, 6

[21] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Bernhard Schölkopf, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023. 1

[22] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2

[23] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. 2