DOES PHYSICS LIVE IN THE ACTIVATIONS? LOCALIZING PHYSICAL QUANTITIES IN VIDEO DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generation models produce strikingly realistic sequences and are increasingly proposed as world models, yet recent benchmarks reveal pronounced deficits in their physical reasoning. This raises the question of whether these models internalize physical principles or merely reproduce familiar motion patterns. We address this by probing internal representations of video Diffusion Transformers (DiTs) for simulator-derived ground-truth physical quantities spanning kinematic motion and rigid-body dynamics under gravity and contact. We find that these quantities are linearly decodable with high accuracy early in the denoising process, substantially outperforming a baseline decoded directly from the model's own noised latents—indicating that the relevant physical information is actively constructed during denoising, rather than already present in the input.

Additionally, we show that activations at on-object tokens carry the relevant physical information and that quantities defined over multiple frames are readable from single latent frames. Hence, information is sharply localized within the token sequence and is computed globally but stored locally. The probes further show partial extrapolation, transferring unchanged to scene variations and object configurations they were never fit on, so what they read is not simply a correlate of the training scenes. When fitted directly in the raw activation space, the probing directions can serve as steering vectors to change the model's output.

Video generation models (Yang et al., 2025; Team Wan et al., 2025; Zheng et al., 2025) can produce remarkably realistic motion, and show impressive results in vision understanding tasks (Wiedemer et al., 2025; Wang et al., 2026). This makes them natural candidates for general-purpose world models (Brooks et al., 2024; Yang et al., 2024). Yet recent evaluations show that this visual realism does not imply reliable physical behavior: generated videos can look plausible while violating basic physical constraints (K.M et al., 2026; Gu et al., 2026; Guo et al., 2025). This raises an important question that output-level evaluations cannot answer: when a video model produces physically consistent motion, does it do so because it represents the underlying physical quantities, or does it simply reproduce patterns that are common in its training data?

We study this question from the perspective of the model's internal representations as illustrated in Fig. 1. Rather than asking whether a generated video looks physically plausible, we ask whether the model's activations contain explicit information about the physical state of the scene. In particular, following the Linear Representation Hypothesis, we investigate whether quantities such as speed, direction, momentum, and gravitational acceleration are represented in a form that can be linearly decoded from the hidden states of video DiTs such as Wan 2.1 (Team Wan et al., 2025), CogVideoX-1.5 (Yang et al., 2025), and Open-Sora 2.0 (Zheng et al., 2025).

To make this question measurable, we render controlled scenarios with a physics simulator to gather exact ground-truth values for every physical quantity of interest. We encode these videos, forward-diffuse them to a range of noise levels, and record the activations of a single frozen DiT forward pass. Ridge probes then predict the simulator-derived quantities across layers, noise levels, and token subsets. We compare those results against probes applied directly to the model's noised latent

Project page: `https://physics-in-activations.github.io/`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

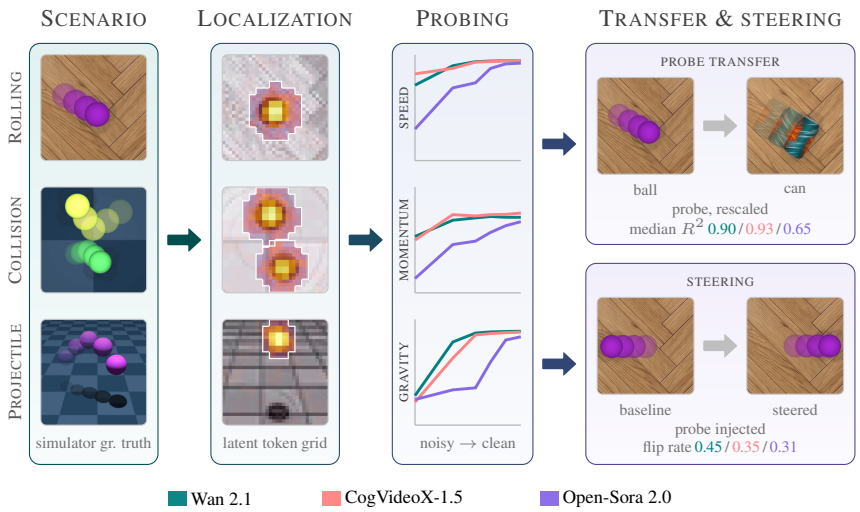


Figure 1: What a frozen video DiT knows about physics. For each scenario (left), the physical signal concentrates on the object's tokens, which the probe pools as a disk-shaped ROI (*localization*; schematic heat, not a measured map). A linear probe recovers speed, total momentum $\boldsymbol{p}_{\text{tot}}$ and gravity from the residual stream along the denoising trajectory (*probing*), and its direction holds up downstream (right): it still reads the quantity once the ball is replaced by a can, while a full-resolution probe injected back flips the rolling direction.

and image conditioning, allowing us to distinguish information already present in the input from information constructed inside the transformer.

In all three models, physical quantities are strongly linearly decodable from the activations and well beyond what the input carries: quantities such as direction become readable long before the input latent reveals it, and gravity, which the latent does not reveal even when clean, is decoded precisely from the activations. Hence, physical information is not simply read from the models' input, but is constructed inside the transformer during denoising. The information is highly localized — a Region of Interest (ROI) covering 1–2% of the tokens around the object beats a global readout over the whole token sequence — and temporally structured, with each token-frame favoring the physical state of its own phase. Speed and gravity are readable from a single token-frame, even though they require information gathered across time. Thus, the transformer aggregates information across the sequence and stores it locally around the object. The probes transfer to variations of background, object shape and count, as well as perspective, suggesting that the decoded signal captures physical structure rather than only scene-specific appearance. Finally, we ask whether model-internal representations are causally active in generation. For the rolling direction, raw-activation probes, when injected back, can change the generated motion.

Our contributions can be summarized as follows:

- We establish that physical quantities are linearly decodable from video DiT activations.
- We find that those properties are computed across the generated video, but stored locally around the relevant objects.
- We show that representations encoding physical quantities transfer across scene variations and are causally active in generation.

1 Related Work

Physics benchmarks A growing body of work quantifies the physical correctness of video models. Intuitive-physics benchmarks test plausibility judgments in vision models (Riochet et al., 2022; Bordes et al., 2025). For generative models, one line rates Text-to-Video (T2V) outputs through perceptual metrics (K.M et al., 2026), binary judgments (Gu et al., 2026) or human ratings of physi-

cal adherence (Guo et al., 2025), while another matches generations against tracked ground truth, in simulation (Jain & Wu, 2026) and in real footage (Tragoudaras et al., 2026; Motamed et al., 2026), or trains DiTs on synthetic 2D mechanics to test sequence completion (Kang et al., 2025). Others compare outcome distributions of stochastic events against analytical references (Sadeghi et al., 2026), or assess symbolic reasoning through overlays of physical quantities (Yao et al., 2026). The picture is consistent: generations look plausible but fail under closer inspection (Guo et al., 2025; Yao et al., 2026; Gu et al., 2026; Tragoudaras et al., 2026; Motamed et al., 2026; Sadeghi et al., 2026). While some benchmarks find proprietary models outperforming open-source ones (Yao et al., 2026), others report mixed rankings (Gu et al., 2026; Jain & Wu, 2026; Guo et al., 2025).

Physics probing in video models While the previously mentioned works evaluate models solely through their outputs, none examine internal representations. A parallel line of work addresses this gap through the lens of mechanistic interpretability, probing video models' internal representations for physical quantities or plausibility. Joseph et al. (2026) find that quantities such as speed and acceleration become linearly accessible at roughly one-third network depth in Video Joint Embedding Predictive Architecture (V-JEPA) and Video Masked Autoencoder (V-MAE). Ballout et al. (2025) show that the vision encoder in Multimodal Large Language Models (MLLMs) forms representations from which physical plausibility is decodable, but this capability does not transfer through vision–language projection. Punzo et al. (2026) compare plausibility representations across video foundation models, spanning encoder architectures and DiTs, finding V-JEPA to perform best. In contrast, Esmati et al. (2026) report that DiTs outperform encoder-based architectures on a plausibility readout. Additionally, they show that causal influence and linear decodability are dissociated across network depth. Li et al. (2026) instead ask why physics *fails*, tracing T2V failures to attention heads whose Rotary Position Embedding (RoPE) decay locks motion in early. Tang et al. (2026) exploit probes during video generation in DiTs as a ranking signal to prune implausible video candidates early in the diffusion process.

The majority of prior work asks *is it plausible?* We instead probe video DiTs directly against ground-truth continuous physical quantities, reframing the question as *does it know?* The works most aligned with this goal either operate on video encoders rather than generative models (Joseph et al., 2026), or primarily assess physical plausibility, with only a limited side experiment probing quantitative targets in smaller models and simple 2D animations (Esmati et al., 2026).

Steering in video models Activation steering — adding a direction to a model's internal state at inference time to change its output — is well established for Large Language Models (LLMs) and is beginning to be adapted to video DiTs: mostly for safety, with a single supervised direction (Kundu et al., 2026) or closed-loop control in a contrastive subspace (Hong et al., 2026), but also for motion control through test-time latent optimization (Xiao et al., 2024), for generation quality by renormalizing rare large-magnitude activations (Cheng et al., 2026), and for multi-event generation by fusing prompts along the denoising trajectory (Liao et al., 2026).

2 Setup

2.1 Synthetic Data Generation

Real-world video is difficult to annotate with exact physical labels. We therefore use the Genesis physics simulator (Genesis AI Team, 2026) to generate ground-truth videos with exact analytic or simulator-derived annotations for every physical observable of interest.

We consider three scenarios covering complementary classes of observables: **(i) Rolling without slipping** for kinematics (speed, direction); **(ii) Elastic collision** of multiple balls for interaction and a conserved quantity, total momentum, and pre-/post-collision phases providing time-varying targets; and **(iii) Projectile motion** for gravitational acceleration and launch angle, neither of which is detectable from any single frame. We render 200 videos per scenario, varying appearance and initial conditions. Scenario details, generation parameters and the released dataset are described in Appendix A.1.

2.2 Models and Activation Capture

We use the largest open-weight Image-to-Video (I2V) variants of three State-of-the-Art (SOTA) architectures: Wan 2.1 (14B) (Team Wan et al., 2025), whose plain architecture, unlike its Mixture of Experts (MoE)-based successor, avoids expert-routing confounds; CogVideoX-1.5 (5B) (Yang et al., 2025); and Open-Sora 2.0 (11B) (Zheng et al., 2025) in its Hunyuan-Variational Autoencoder (VAE) release (Kong et al., 2024), which keeps compression rate and VAE architecture comparable across models, see Appendix A.2.

All three models are DiTs that denoise compressed video latents. A VAE maps a clip $\boldsymbol{x} \in \mathbb{R}^{F\times H\times W\times C}$, with F frames of spatial resolution $H \times W$ and C channels, to a latent representation $\boldsymbol{z} \in \mathbb{R}^{F'\times H'\times W'\times C'}$, with roughly $8\times$ spatial and $4\times$ temporal compression. The decoder reverses this process after denoising. The latent is split into $p_t \times p_h \times p_w$ spatio-temporal patches, flattened row-major into n_{tok} tokens, and linearly embedded in the residual stream $\boldsymbol{r}_l \in \mathbb{R}^{n_{\text{tok}}\times d}$ processed by L DiT blocks $l \in \{0, \dots, L-1\}$. Because attention spans the full sequence, every token is contextualized by the entire clip. Each token index s has a fixed latent position (f, h, w), a mapping our localization experiments read in reverse. We call the tokens with frame index f a *token-frame*, arranged as a $H'/p_h \times W'/p_w$ *token grid*.

Activation Capture The models are trained to make single-step predictions at arbitrary noise levels. Rather than sample from pure noise, we encode a labeled ground-truth video $\boldsymbol{x}_{\text{gt}}$ as $\boldsymbol{z}_0 = \mathcal{E}(\boldsymbol{x}_{\text{gt}})$, diffuse it to a chosen noise level t to obtain $\boldsymbol{z}_t$, and run the denoising pass from there. We record each layer's activations $\boldsymbol{h}_l(t) \in \mathbb{R}^{n_{\text{tok}}\times d}$. This keeps inference within the training distribution while maintaining exact physical labels. Videos use each model's native resolution and frame count, and the resulting activation–label pairs train the probes. During capture, we condition on the first ground-truth frame and use an empty prompt to avoid a mismatch between ground-truth video and text conditioning.

3 Probing Physical Observables

3.1 Method

Estimator Probes are ridge regressions on standardized features, with both the standardization statistics and the penalty λ fit on training folds only. We report R^2 under k-fold cross-validation ($k = 5$) with folds assigned at the video level, so no frame of a held-out video is seen during fitting. For vector-valued targets (direction, total momentum) we average the per-component R^2 uniformly. Probes are trained on the residual stream $\boldsymbol{r}_l$. Noise levels are reported as the Signal-to-Noise Ratio (SNR) of the noised latent in dB, which, unlike a denoising step or diffusion timestep, is comparable across models with different noise schedules. A permutation null, refitting every probe on targets shuffled across videos, leaves R^2 at chance throughout (mean -0.056; Appendix B.1), validating the probing protocol.

Pooling and localization Time-varying targets are probed from spatially pooled activations, $\bar{\boldsymbol{h}}_{\text{sp}} \in \mathbb{R}^{F'\times d}$, retaining the temporal latent-frame axis so that each token-frame serves as a regression sample; targets constant over the clip are instead probed from the spatio-temporal mean, $\bar{\boldsymbol{h}}_{\text{st}} \in \mathbb{R}^{d}$ (Appendix B.1). Orthogonally, we vary which tokens are pooled: either all tokens (the *global* readout) or only a disk-shaped ROI centered on the object's simulator-derived position projected onto the latent grid. The disk's radius is set relative to the projected object radius in each latent frame (default factor 1.5). Both token sets are mean-pooled, so the two readouts differ only in which tokens enter the mean, not in feature dimension or estimator. Unless stated otherwise we report the ROI readout.

Input baseline The model's input already carries physical signal, from the noised latent $\boldsymbol{z}_t$ (Section 2.2) and from the clean conditioning image. As a reference we therefore probe the input itself: the baseline $\boldsymbol{z}_{tc}$ concatenates $\boldsymbol{z}_t$ with the conditioning latent and is pooled exactly like the activations (Appendix B.1).

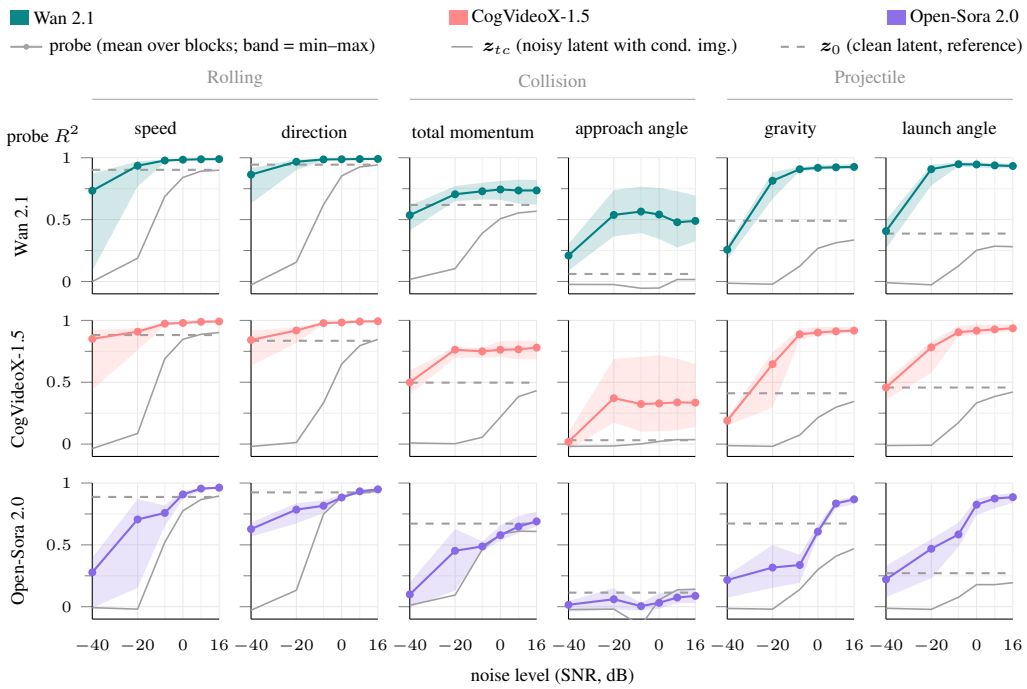


Figure 2: Probe decodability (R^2, $n = 200$ videos, 5-fold video-level cross-validation) from DiT activations at the ROI, across noise levels (SNR of the noised latent). Solid colored lines are the mean over transformer blocks, bands their min–max spread (10 blocks uniformly across network depth). Gray: the latent references, ROI masked the same way — solid z_{tc} (z_t with the clean conditioning anchor), dashed z_0 (clean latent). Collision targets use a multi-object readout (Appendix B.1).

3.2 Physical Quantities Are Linearly Decodable

Figure 2 reports, for every target and model, what a linear probe recovers from the DiT's activations against what a similar probe recovers from the model's own input, along the denoising trajectory. Every probe here reads the ROI token set, and the baseline z_{tc} is masked the same way, so the two differ only in what they are computed from. Three findings hold across the grid.

At the maximum measured noise level (-40 dB), the input provided to the model is uninformative, leaving the matched baseline at chance performance across all evaluated quantities. Despite this severe degradation, the model's residual stream is already capable of precisely resolving the base kinematics, while also gaining performance in the other evaluation scenarios. This indicates that DiTs actively construct physical representations over the course of the denoising trajectory, rather than passively inheriting them from the noisy input.

As the input becomes cleaner, the physical targets split into two distinct groups, determined by the amount of information inherently carried by the clean latent. The first group consists of targets, such as rolling kinematics, where the clean latent ultimately holds the requisite information. For these properties, the model's performance margin over the baseline correspondingly shrinks as the denoising process advances. The second group comprises quantities like gravity, launch angle, and approach angle, where the clean latent carries minimal informative data. Here, the model maintains a substantial predictive margin: gravity and launch angle stay $+0.40$ to $+0.69$ above the baseline in every model even at the cleanest measured input, and approach angle $+0.47$ and $+0.30$ for Wan and CogVideoX — though Open-Sora does not recover that target at all (-0.05).

Beyond the differences across physical quantities, the emergence of this information depends heavily on both the network depth and the denoising timeline. Across all evaluated models, decodability remains remarkably consistent throughout the network's depth: physical properties become readable within the initial transformer blocks and persist to the final output layers with peaks in early-mid layers (see Appendix B.2.1). However, the models diverge significantly along the denoising trajec-

tory. Wan and CogVideoX build the evaluated quantities early, reaching peak performance while the input is still heavily noised. In contrast, Open-Sora requires a much cleaner input, with decodability only rising near the midpoint of the diffusion process. Table B.1 extends this to more quantities and reports the global readout alongside the ROI readout.

3.3 Spatio-temporal Localization

Having established that physical quantities can be read from the model's activations, we next ask where this physical information is stored within the token sequence. Unless stated otherwise, we use a fixed mid-depth layer (Layer 21) and a 0 dB denoising level, where all models already provide reliable physical readouts. We examine three complementary aspects: spatial localization across token positions, temporal localization across token-frames, and temporal aggregation, i.e. where information from several frames is combined.

Spatial localization We first ask whether physical quantities are represented at the location of the object that carries them. We compare five token sets at fixed probe, feature dimension and sample count: the single token closest to the object center (ROI), an on-object disk of $1.5\times$ the object size, two size-matched off-object decoys (one pinned at the object's reflected first-frame position, one tracking its reflected trajectory; Appendix B.1), and the global readout over the whole token grid. We evaluate these conditions on the single-object rolling and projectile scenarios, where the off-object regions cannot overlap with another physical object.

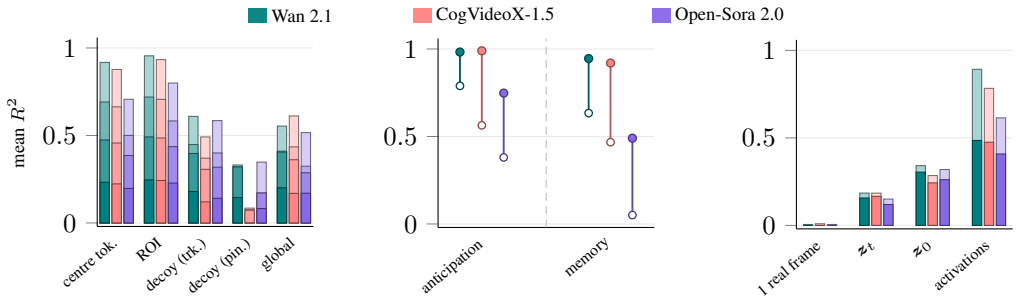
The results in Fig. 3a show a clear spatial concentration of the signal. The ROI consistently outperforms the global readout, despite using only roughly 1–2$\%$ of the tokens, and it outperforms the single center token by a smaller margin (0.04–0.09 in mean R^2). Thus, the information is highly localized, but extends beyond the center token, and the radius sweep (Fig. B.2) shows this benefit saturating around the $1.5\times$ used here.

The tracked decoy, which moves like the object but on the opposite side of the grid, retains some decodability for directional quantities, so trajectory-correlated information might leak through tracking. The on-object region nonetheless remains substantially more informative, and pinning the decoy at its initial location removes even this trajectory information: decodability drops sharply, particularly for CogVideoX. The readout is therefore tied to the object's location, not to the number of pooled tokens or to its trajectory.

Temporal localization For quantities that vary within a clip, we ask whether a token-frame represents the physical state at its own point in time. The elastic collision gives each object exactly two speeds, before and after contact. We assign every token-frame to its phase and fit one probe for each speed from its own phase or from the other one: the post-collision speed from pre-collision frames (*anticipation*), and the pre-collision speed from post-collision frames (*memory*).

As shown in Fig. 3b, the strongest readout consistently comes from a token-frame's own phase. Information nevertheless extends across the collision: pre-collision token-frames retain substantial information about the post-collision speed, while post-collision token-frames retain less information about the preceding speed across all models. Relative to the own-frame score for the *same* target, anticipation retains 80%, 57%, and 51% of the information for Wan, CogVideoX, and Open-Sora, respectively, while the reverse direction retains 67%, 51%, and 10%. The two phases are nearly independent ($r = 0.22$), ruling out the possibility that anticipation merely reflects correlation. Thus, temporal localization is graded rather than strict: token-frames preferentially encode the physical state of their own phase, while information about other phases remains partially accessible through the model's representations.

Temporal aggregation In the previous experiment, we asked whether a token-frame primarily represents the physical state at its own point in time. Here, we go one step further and consider quantities for which there is no single-frame state to recover: speed and gravitational acceleration are the first and second derivatives of position and must be computed from multiple frames. While we can see in Fig. 2 that these quantities are readable from the full sequence, here we want to evaluate if DiTs aggregate them across time.



(a) SPATIAL LOCALIZATION — which tokens. The ROI beats the global readout. Each bar stacks the four targets of the two single-object scenarios — from saturated to pale, rolling speed & direction, projectile gravity and launch angle — at a quarter of their R^2 each, so the bar height is their mean.

(b) TEMPORAL LOCALIZATION — which token-frame. Filled dots show the target read from its own frame; hollow dots show the *same* target read from the other phase of the collision. Own-frame readouts are highest for every model, and anticipation consistently outperforms memory.

(c) TEMPORAL AGGREGATION — where the quantity is aggregated: the same ROI probed at a single frame, stacking rolling speed (saturated) and projectile gravity (pale) at half their R^2 each. Neither is readable from a single real frame and only partly from the latents but highly from the activations.

Figure 3: Localization of the ROI readout: *spatially* a and *temporally* b along with *temporal aggregation* c. Bars are probe R^2 under video-level k-fold cross-validation.

To test this, we probe four representations of a single frame: the original pixel frame, the noised latent ($\boldsymbol{z}_t$), the clean latent ($\boldsymbol{z}_0$), and the corresponding DiT activation, as presented in Fig. 3c. The pixel frame is at chance, as expected, since a single position sample contains no information about its temporal derivatives. The clean latent already contains some information because the VAE temporally compresses roughly four real frames into one latent frame — eight for CogVideoX-1.5, whose DiT pairs latent frames ($p_t = 2$) — making speed partially recoverable before the DiT. The DiT activations, however, recover substantially more, indicating that the transformer aggregates information beyond what is available in the local latent frame alone. This effect is particularly clear for gravity: the clean latent carries little information ($R^2 = 0.07$–0.12), while the corresponding activations reach $R^2 = 0.41$–0.81. Hence, the observed temporal aggregation is not simply a consequence of the VAE compression: the DiT gathers information across frames and makes the resulting quantity locally readable.

3.4 GENERALIZATION ACROSS OBJECTS AND SCENES

To distinguish probes that read physical quantities from those that exploit scene-specific visual regularities, we transfer probes to controlled Out-of-Distribution (OOD) scene variations, see Fig. A.1. In rolling, we swap the wooden floor for a synthetic checkerboard (background), the ball for a soda can (shape), or the top-down camera for an oblique side view (perspective). In collision, we increase the number of balls from two to five (count).

A scene change can alter a readout's scale and offset without changing its direction. We therefore calibrate each transferred probe using only a gain a and offset b, fitted on ten target-scene videos so that $\boldsymbol{q} = a\,\boldsymbol{w} + b$. The probe weights and readout direction $\boldsymbol{w}/\|\boldsymbol{w}\|$ remain fixed. Please note that ten videos cannot guarantee the optimal affine calibration; hence, the appendix gives the uncalibrated scores and the calibration ceiling (Appendix B.2.2). This approach separates transfer of the learned di-

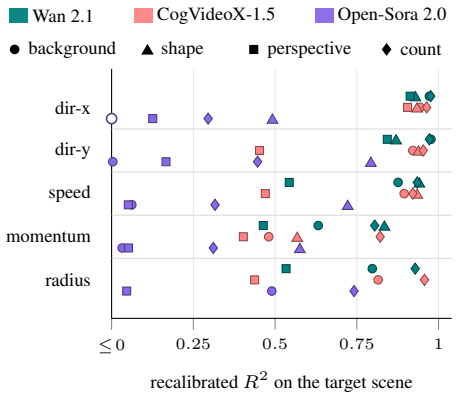


Figure 4: A ROI probe fitted on one scene and evaluated on another after recalibration. Hollow marks on the left edge are off-scale values below 0.

rection from scene-specific calibration. For each model, we use the layer selected by pooled source-scene cross-validation and evaluate at 0 dB SNR.

Figure 4 reveals a pronounced separation between the models. Open-Sora transfers poorly, with a median $R^2 = 0.295$ compared with 0.842 in-scene, whereas Wan and CogVideoX-1.5 retain high scores, with medians of 0.876 and 0.905, compared with 0.981 and 0.977 in-scene, respectively. Viewpoint changes pose the greatest challenge, while changes in object count and shape — both without background changes — extrapolate most reliably. Across targets, directional quantities are generally the most robust, whereas appearance-related quantities perform worst: radius and momentum, whose mass term depends on size in our scenes.

Additional results in Table B.2 show that global probes transfer through the shape change (median recalibrated $R^2 = 0.773$) but collapse under every other shift (background 0.032, perspective 0.095, count -0.006). The shape swap is the only one that leaves both the background and the number of objects intact, so it is the only one under which a global mean sees nearly the same scene. Localization therefore not only supports within-scene decodability but also enables transfer across scenes.

3.5 Steering

Decoding a quantity from the activations shows that the representation *correlates* with the physical state, not that the model uses it to produce its output. Steering tests the latter: if pushing the activations along a probe direction reliably changes the generated motion, that direction is causal.

Protocol We steer the rolling ball's direction (left vs. right), the simplest target for which we obtain a usable steering direction (see Section 4). At block l, we take the steering direction $\boldsymbol{d}_l \in \mathbb{R}^{n_{\text{tok}} \times d}$ directly from the raw activation-space probe, keeping one vector per token rather than pooling. We then push each token activation in $\boldsymbol{h}_l$ a fraction α toward its corresponding row of $\boldsymbol{d}_l$ using spherical linear interpolation (SLERP).

We apply the intervention only during the high-noise phase of generation, from pure noise until the noise-level ratio reaches approximately -20 dB (see Appendix C). To isolate the effect of steering, we always apply the rightward direction and evaluate only seeds that roll left without intervention. Since the seed is fixed, any rightward motion after steering is a flip caused by the intervention; we report the fraction of such flips as the *direction flip ratio*.

Controls Two comparisons bracket the probe direction. The *leftward* probe direction is a negative control: it pushes toward the motion the seeds already have, so it should flip nothing. A rightward *prompt* — the same seed, with the text conditioning stating that the ball rolls right — measures how far the conditioning channel alone moves the motion.

Results The negative control behaves as intended: the leftward direction flips none of the seeds to the right on any backbone. Prompting is at best partially effective, flipping 30% of seeds for Wan but only 1.7% and 4% for CogVideoX and Open-Sora. The rightward probe direction, in contrast, steers every backbone well above the control and above the prompt (Fig. 5): the direction read from the activations is causally efficacious in all three models.

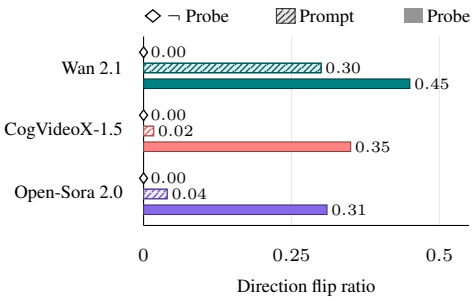


Figure 5: Direction flip ratio on left-rolling seeds, per backbone: leftward probe direction (¬ Probe, negative control, marked at 0), rightward prompt and rightward probe direction. All directions are full-resolution. Configurations in Table C.1.

4 Discussion and Limitations

Discussion The DiTs *add information beyond their input*: averaged over depth, their activations outperform the noised-latent baseline. Physical quantities are readable even at high noise, indicating that dynamics are fixed before detailed appearance.

The resulting state is *computed globally but stored locally*. Small object-centered ROIs score higher and transfer far better than the global readout, while distant tokens carry little of the relevant information. This unsupervised object-centric representation has practical consequences. Only a small portion of tokens suffices when the object's position is known (the simulator provides it here; real video would require tracking). For multi-object scenes, a global readout can recover only system-level quantities such as total momentum, whereas per-object quantities require a per-object readout. Each object also supplies a separate training sample.

The direction survives a scene change; its scale does not. Probes do not transfer one-to-one, because the shift changes the readout's gain and offset, but its direction is shared across the scenes we tested. That direction is also causally active: moving activations along a direction probe changes the generated motion in every model.

Wan and CogVideoX behave similarly throughout. Open-Sora matches them in-distribution at low noise on five of six quantities, but fails the two tests that go beyond in-distribution readout: it barely transfers, and its probe direction is the least effective for steering. In-distribution decodability therefore does not distinguish world-model candidates; transfer against a target-scene ceiling and steering with the probe's own direction do.

Decodability alone does not establish physical understanding and benchmark failures attest to the models' limited capabilities. The models are not trained specifically to produce physically correct video, so failure in video generation may reflect the conditioning or training mix. Moreover, our probes read physical *states*, not whether the law relating states holds.

Limitations Our *scope* is rigid-body mechanics in uncluttered, few-object scenes, matching the motion classes of Kang et al. (2025). We omit dissipation, rotation, deformable or articulated bodies, multi-body contact, and phenomena beyond mechanics, including fluids, thermodynamics, and optics. A direct extension is to hidden material parameters that, like gravity, are revealed only through motion: friction, restitution, stiffness, and viscosity.

We capture activations from a *forward-noised real video*, not the model's own sampling trajectory. They are therefore off-trajectory, although capture requires no inversion.

Some targets are naturally correlated, making it hard to tell which quantity the model encodes. Momentum magnitude and kinetic energy correlate at $r = 0.85$, and both track total mass ($r = 0.71$ and 0.73). All main targets are otherwise uncorrelated within each scenario. Conversely, targets already *decodable from the clean latent* cannot distinguish a model that denoises well from one that makes physics decodable.

Steering beyond direction. Steering other physical properties remains difficult. High-scoring probes for them are available only from pooled activations, while high-dimensional raw-activation probes fail with limited samples. In our experiments, injecting a single vector uniformly across all tokens did not yield functional steering results, which might be due to the spatio-temporal structure of the DiT token sequence with massive activations (Cheng et al., 2026). Hence, the pooled probe family collapses when injected for steering (Appendix C). Only steering a ROI in the token-grid, following an idea similar to that of Liu et al. (2026), who gate tokens by a soft spatio-temporal mask, might help, but how to apply it during generation—for example, by tracking the object—remains open.

5 Conclusion

We probed three open-weight video DiTs against simulator ground truth to ask whether they internalize physics or merely reproduce familiar motion. *Physics lives in the activations*: speed, direction, momentum and gravity are linearly decodable, the DiTs add information their input does not carry, and that information is computed globally but stored locally, at the object's own tokens. Its readout direction survives controlled scene changes and steers the motion of generated videos — although Open-Sora, despite matching the others in-distribution, transfers poorly. A decodable state alone is not sufficient for physically consistent generation: whether the laws linking these states hold in the models' own representations, and why models with decodable state still fail physics benchmarks, are the questions that follow.

AI USE STATEMENT

In this work, we used generative AI tools to propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, interpret results. We have not used generative AI tools to generate synthetic data sets or to help develop theoretical models or conceptual frameworks. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, create or edit software code, draft parts of a research paper, summarize or analyze existing literature, brainstorm, identify relevant literature, format references, propose a title or keywords for a research paper. We have reviewed all AI-assisted work. All AI-assisted code was reviewed through manual pull requests and verified by our automated test pipelines before use. AI-suggested hypotheses, experimental designs and interpretations were critically assessed by the authors and validated against our experimental results. All references were drawn from the authors' own curated literature library. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, crowdsourced annotation, or personal or identifiable data. All videos are rendered with the Genesis physics simulator from procedurally sampled rigid-body scenes, and their labels come directly from the simulator. The three video models we study (Wan 2.1, CogVideoX-1.5 and Open-Sora 2.0) are publicly released open-weight checkpoints, used frozen and in accordance with their respective licenses. We release our code and the rendered dataset with its simulator labels to support reproducibility; neither contains sensitive content. Our steering experiments show that internal activations can be edited to change generated motion. Activation-level control of generative video is a dual-use capability, but our interventions are limited to the lateral rolling direction of a single object in synthetic scenes and add no capability beyond existing prompt- and motion-control methods.

REPRODUCIBILITY STATEMENT

All three models are open-weight checkpoints, used frozen (Section 2.2); the simulated scenarios and their ground-truth quantities are specified in Section 2.1 and Appendix A.1, and the probing protocol in Section 3.1. Code to reproduce every experiment — simulation, activation capture, probing, steering, and the export of every figure and table — is available at `https://anonymous.4open.science/r/DOES-PHYSICS-LIVE-IN-THE-ACTIVATIONS-35ED/`, and the rendered videos with their simulator labels at `https://osf.io/ey6h5/overview?view_only=a264137a73b0426d80e4987ea23fcaaf`.

REFERENCES

Mohamad Ballout, Serwan Jassim, and Elia Bruni. Pixels to principles: Probing intuitive physics understanding in multimodal language models, 2025. URL https://arxiv.org/abs/2507.16572.

Florian Bordes, Quentin Garrido, Justine T. Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments, 2025. URL https://arxiv.org/abs/2506.09849.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI, February 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

Xianhang Cheng, Yujian Zheng, Zhenyu Xie, Tingting Liao, and Hao Li. Steering video diffusion transformers with massive activations, 2026. URL https://arxiv.org/abs/2603.17825.

Parsa Esmati, Somjit Nath, Katja Hofmann, Derek Nowrouzezahrai, Samira Ebrahimi Kahou, and Majid Mirmehdi. The invisible hand of physics: When video diffusion models know more than they show, 2026. URL https://arxiv.org/abs/2606.05328.

Genesis AI Team. The role of simulation in scalable robotics, Genesis World 1.0, and the path forward. Genesis AI Blog, May 2026. URL https://www.genesis.ai/blog/the-role-of-simulation-in-scalable-robotics-genesis-world-10-and-the-path-forward.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Eric Wang. PhyWorldBench: A comprehensive evaluation of physical realism in text-to-video models. In *International Conference on Learning Representations (ICLR)*, pp. 75130–75164, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/79d86433c2acd12b6fa98553435d226e-Abstract-Conference.html.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2VPhysBench: A first-principles benchmark for physical consistency in text-to-video generation, 2025. URL https://arxiv.org/abs/2505.00337.

Jihoon Hong, Alice Chan, Qiyue Dai, Julian Skifstad, and Glen Chou. Activation steering of video generation models via reduced-order linear optimal control, 2026. URL https://arxiv.org/abs/2606.04775.

Swarnim Jain and Shangzhe Wu. RigidBench: Evaluating rigid-body physics in video generation models, 2026. URL https://arxiv.org/abs/2608.15555.

Sonia Joseph, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake Richards, and Mike Rabbat. Interpreting physics in video world models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2602.07050.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28991–29017, 2025. URL https://proceedings.mlr.press/v267/kang25g.html.

Megha Mariam K.M, Aditya Arun, Zakaria Laskar, and C. V. Jawahar. PhyEduVideo: A benchmark for evaluating text-to-video models for physics education. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8690–8699, March 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/K.M_PhyEduVideo_A_Benchmark_for_Evaluating_Text-to-Video_Models_for_Physics_Education_WACV_2026_paper.html.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. HunyuanVideo: A systematic framework for large video generative models, 2024. URL https://arxiv.org/abs/2412.03603.

Rohit Kundu, Arindam Dutta, Sarosij Bose, Athula Balachandran, and Amit K. Roy-Chowdhury. Pulling the REINS: Training-free safety alignment of video diffusion models via representation steering, 2026. URL https://arxiv.org/abs/2606.17257.

Yueyan Li, Haibo Wang, Caixia Yuan, and Xiaojie Wang. Why Do Video Diffusion Models Violate Physics? Unveiling the Flaws in Attention Mechanisms, September 2026. URL http://arxiv.org/abs/2609.23658. arXiv:2609.23658 [cs.CV].

Ruotong Liao, Guowen Huang, Qing Cheng, Guangyao Zhai, Lei Zhang, Xun Xiao, Thomas Seidl, Daniel Cremers, and Volker Tresp. TunerDiT: Training-free progressive steering of diffusion transformer for multi-event video generation, 2026. URL https://arxiv.org/abs/2605.31590.

Zhixuan Liu, Peter Schaldenbrand, Yijun Li, Long Mai, Aniruddha Mahapatra, Cusuh Ham, Jean Oh, and Jui-Hsien Wang. TokenDial: Continuous attribute control for text-to-video generation in visual dial space. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2026. doi: 10.48550/arXiv.2603.27520. URL https://arxiv.org/abs/2603.27520.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 948–958, March 2026. URL https://openaccess.thecvf.com/content/WACV2026/papers/Motamed_Do_Generative_Video_Models_Understand_Physical_Principles_WACV_2026_paper.pdf.

Samuele Punzo, Niccolò Caselli, Ippokratis Pantelidis, Francesco Massafra, Salvatore Lo Sardo, and Mohammadreza Salehi. How do video foundation models encode intuitive physics? probing across pretraining paradigms, 2026. URL https://arxiv.org/abs/2606.09646.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Veronique Izard, and Emmanuel Dupoux. IntPhys 2019: A benchmark for visual intuitive physics understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5016–5025, September 2022. doi: 10.1109/TPAMI.2021.3083839. URL https://ieeexplore.ieee.org/document/9442261/.

Jonathan Sadeghi, Jenny Seidenschwarz, Jesse Allardice, Sirish Srinivasan, Benjamin Graham, and Jeffrey Hawke. CaliBench: Are the stochastic dynamics of video world models physically calibrated? *Transactions on Machine Learning Research*, September 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=rWE29rkvDz.

Chujun Tang, Lei Zhong, and Fangqiang Ding. Seeking physics in diffusion noise, 2026. URL https://arxiv.org/abs/2603.14294.

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

Antonios Tragoudaras, Chenyu Zhang, Daniil Cherniavskii, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Evaluating newtonian mechanics in video generative models with real physical systems. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2504.02918.

Ruisi Wang, Zhongang Cai, Fanyi Pu, Junxiang Xu, Wanqi Yin, Maijunxian Wang, Ran Ji, Chenyang Gu, Bo Li, Ziqi Huang, Hokin Deng, Dahua Lin, Ziwei Liu, and Lei Yang. Demystifying video reasoning. In *European Conference on Computer Vision (ECCV)*, 2026. URL https://arxiv.org/abs/2603.16870.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners, 2025. URL https://arxiv.org/abs/2509.20328.

Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. In *Advances in Neural Information Processing Systems*, volume 37, pp. 76115–76138. Curran Associates, Inc., 2024. doi: 10.52202/079017-2423. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8b2fc235787852ead92da2268cd9e90c-Abstract-Conference.html.

Sherry Yang, Jacob C. Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Position: Video as the new language for real-world decision making. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56465–56484, 2024. URL https://proceedings.mlr.press/v235/yang24z.html.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations (ICLR)*, pp. 83048–83077, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ce31378e9f41d8907e97dab172b6c559-Abstract-Conference.html.

Runmao Yao, Kairui Hu, Yukang Cao, Ruisi Wang, Shulin Tian, Ziang Cao, Weichen Fan, Ziqi Huang, Yuhao Dong, Hao Li, Zhaoxi Chen, Zhongang Cai, Lei Yang, and Ziwei Liu. Apple-π: Benchmarking thinking with video towards law-grounded physical intelligence, 2026. URL https://arxiv.org/abs/2607.16401.

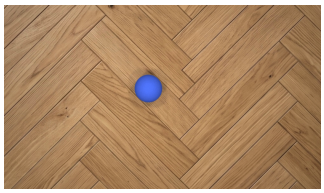
Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei, Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang, Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao, Yuqi Wang, Ziang Wei, and Yang You. Open-Sora 2.0: Training a commercial-level video generation model in $200k, 2025. URL https://arxiv.org/abs/2503.09642.

A Data and models

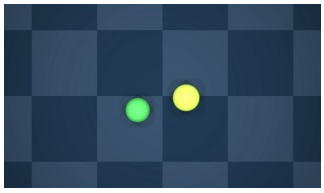
A.1 Simulation scenarios

All three scenarios are defined by a list of rigid objects — each with a mass, a radius, a color, and an initial position and velocity — as well as global scene parameters such as the gravity vector. Mass is tied to the rendered radius by a constant-density law, so that mass becomes visually decodable. Otherwise, absolute momentum and kinetic energy are unidentifiable from the video (non-unique) and therefore out of reach of any probe. The rolling scenario is stepped by the rigid-body solver, while the collision and projectile scenarios are instead rendered following the exact analytical trajectory, because Genesis' rigid contacts are effectively inelastic and hence are not suitable for our use case. Alongside the frames we store the per-frame position, velocity and momentum of every object, together with the scene's sampled parameters, camera pose and seeds. Table A.1 lists what is sampled, and from where. Each description below shows a scenario's example initial frame.

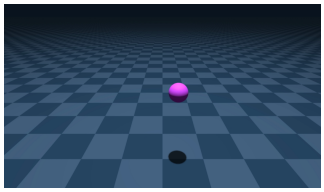
Rolling without slipping A single ball rolls across a horizontal floor without slipping. Hence, speed and heading are conserved for the whole clip and the labels are constant in time. Initial conditions are chosen in such a way that the ball does not leave the frame within the simulated time. The probe targets are the speed $\|\boldsymbol{v}\|$ and it's direction, read out as the unit vector $\hat{\boldsymbol{v}} = \boldsymbol{v}/\|\boldsymbol{v}\|$.



Elastic collision Two balls, *both* in motion, meet in a frictionless elastic collision. They travel at constant speed between collisions. Setups that would either have a negligible impulse, and no meaningful collision, or where any of the balls leave the frame are rejected. The probe targets are the total momentum $\boldsymbol{p}_{\text{tot}} = m_1\boldsymbol{v}_1 + m_2\boldsymbol{v}_2$, kept as an (x, y) vector, and the approach angle θ_{app}, the angle between the two velocity vectors before contact.



Projectile A single ball launches mid-air and flies a parabolic arc under a per-scene gravity magnitude g. The magnitude of g is drawn from a low band so that the ball never reaches the ground within the clip. The launch has x, y and z components, so the path is a parabola in a vertical plane. The probe targets are the gravity magnitude g and the launch angle θ_0, the angle between the horizontal plane and the projectile's velocity at the moment of release.



Scene variants Each scenario has variants that hold its sampler and physics fixed and change exactly one factor, which is what makes them usable as transfer tests (Section 3.4). For rolling, *background* swaps the UV-mapped wood floor for an infinite checker plane; *shape* replaces the ball with a cylindrical can lying on its side, rolling with the same no-slip angular velocity at a fixed radius; *perspective* changes the top-down camera for an oblique side view. For the collision, a *count* variant replaces the two-body construction with a five-object elastic hard-ball billiard on a bounded plane, invisible reflecting walls at the frame edges keeping all five in view. The can is rendered from an analytic solution.

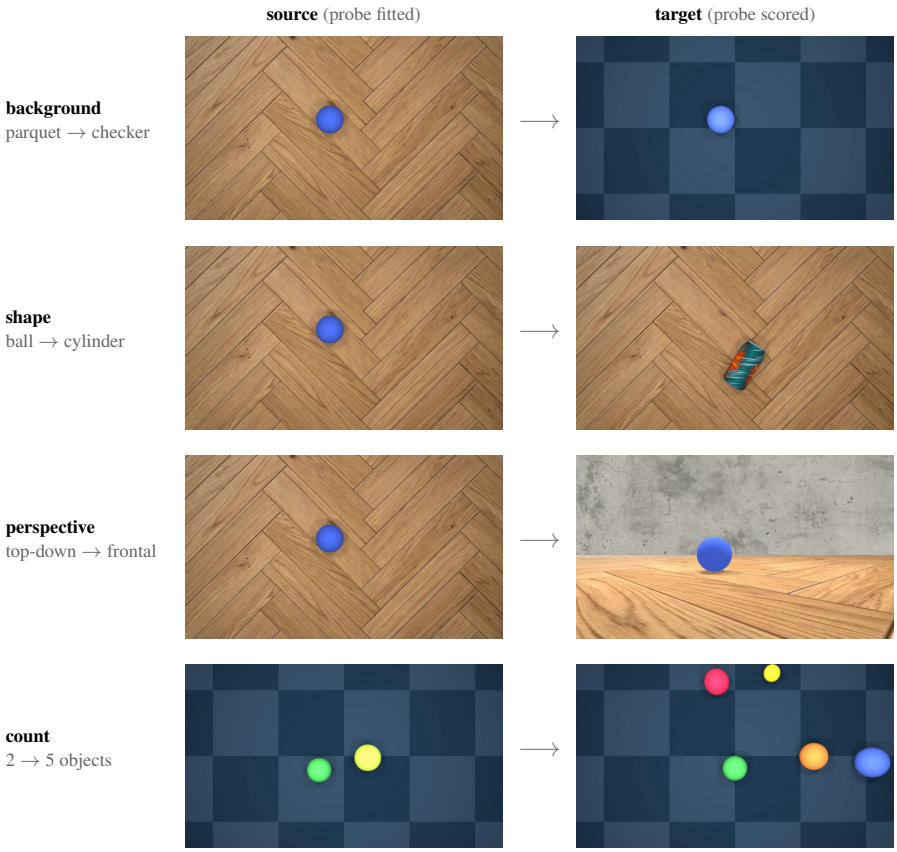


Figure A.1: The four transfer shifts of Section 3.4. Each row fits the probe on the left scene and scores it on the right, and changes exactly one factor between them.

Table A.1: Sampled parameters per scenario. $\mathcal{U}$ denotes a uniform draw; T is the clip duration and m_{ref}, r_{ref} the reference mass and radius of the density law. Radius and colour are appearance nuisances in every scenario.

	Quantity	Symbol	Sampled from
Shared	mass	m	$m_{\mathrm{ref}}\,(r/r_{\mathrm{ref}})^3$, i.e. constant density $\rho = 1000\,\mathrm{kg/m^3}$
	colour	–	uniform over a fixed ten-colour palette
	rendering	–	81 frames at 24 fps ($T = 3.375$ s)
	simulation step	Δt	10 ms
Rolling	radius	r	$\mathcal{U}[0.06, 0.12]$ m
	speed	$\|\boldsymbol{v}\|$	$\mathcal{U}[0.03, 0.12]$ m/s
	heading	φ	$\mathcal{U}[0, 2\pi)$
	start position	(x, y)	$\mathcal{U}[-0.10, 0.10]^2$ m
Collision	total mass	$M = m_1 + m_2$	$\mathcal{U}[0.3, 3.0]\,m_{\mathrm{ref}}$
	mass fraction	m_1/M	$\mathcal{U}[0.20, 0.80]$ (radii follow from the density law)
	speeds	$\|\boldsymbol{v}_i\|$	$\mathcal{U}[0.04, 0.12]$ m/s
	headings	$\varphi_1, \varphi_2 - \varphi_1$	$\mathcal{U}[0, 2\pi), \mathcal{U}[30^\circ, 330^\circ]$
	impact angle	α	$\mathcal{U}[5^\circ, 75^\circ]$ away from head-on, side drawn $\pm$
	contact time	t_c/T	$\mathcal{U}[0.25, 0.65]$
	contact point	$\boldsymbol{p}_c$	$\mathcal{U}[-0.05, 0.05]^2$ m
Projectile	radius	r	$\mathcal{U}[0.08, 0.12]$ m
	gravity	g	$\mathcal{U}[0.14, 0.26]$ m/s^2
	vertical launch	$v_{z,0}$	$\mathcal{U}[0.24, 0.43]$ m/s
	downrange launch	$v_{x,0}$	$\pm\mathcal{U}[0.08, 0.30]$ m/s, sign drawn $50/50$
	lateral launch	$v_{y,0}$	$\pm\,0.07$ m/s
	launch angle	θ_0	derived from $(v_{x,0}, v_{y,0}, v_{z,0})$: 39.6°–79.4°, median 60.3°
	start position	$\boldsymbol{p}_0$	$\mathcal{U}\big([-0.2, 0.2] \times [-0.15, 0.15] \times [0.80, 0.90]\big)$ m

A.2 Models and implementation details

Table A.2 lists the three backbones, the checkpoints we capture from, and the clip geometry each one is run at. All three are the largest open-weight I2V variants of their architectures, run at their own native resolution and frame count rather than a shared one, so that inference stays in distribution (Section 2.2).

Table A.2: Backbones, checkpoints and capture geometry. *Blocks* is the number of transformer blocks the model exposes. Open-Sora's DiT is a Multi-Modal Diffusion Transformer (MM-DiT): its depth is 19 double-stream plus 38 single-stream blocks, indexed consecutively.

Model	Checkpoint	Frames	Resolution	Blocks
Wan 2.1 (14B)	`Wan-AI/Wan2.1-I2V-14B-480P-Diffusers`	81	480×832	40
CogVideoX-1.5 (5B)	`zai-org/CogVideoX1.5-5B-I2V`	81	768×1360	42
Open-Sora 2.0 (11B)	`hpcai-tech/Open-Sora-v2`	81	480×832	57

Latent geometry All three models compress by the same factor overall, which is what makes their token grids comparable. Wan and CogVideoX use a $4\times$ temporal, $8\times$ spatial VAE with a DiT patch size of $1\times2\times2$; CogVideoX-1.5 additionally groups latent frames in pairs ($p_t = 2$). Open-Sora 2.0 is run in its HunyuanVAE configuration.

Noise levels Captures are taken at six noise levels, reported throughout as the SNR of the noised latent: $\sigma \in \{0.990, 0.909, 0.715, 0.500, 0.285, 0.137\}$, i.e. -40, -20, -8, 0, $+8$ and $+16$ dB. The same σ grid is used for every model, which is what makes the noise axis comparable across schedules. Unless stated otherwise, localization results are read at a fixed mid-depth block and 0 dB, where all three models already give reliable readouts.

B PROBING

B.1 PROBING PROTOCOL DETAILS

Time-varying targets Targets that vary within the clip are probed from spatially pooled activations that keep the latent frame sequence intact, $\bar{\boldsymbol{h}}_{\text{sp}} \in \mathbb{R}^{F' \times d}$. Latent frames then enter the regression as individual samples sharing one weight vector, so the probe stays d-dimensional. Since the VAE compresses multiple video frames into a single latent frame, the physical target quantity is averaged over the corresponding time window.

Off-object controls The two decoys of Section 3.3 reflect the object's position through the center of the token grid, and either center a size-matched disk at the reflected position of the first frame (*decoy pinned*) or follow the reflected trajectory (*decoy tracked*); the tracked decoy therefore moves exactly like the object, but mirrored. If the object starts at the center, the decoy disks are pushed outward so that they do not overlap it. We evaluate the decoys on the single-object rolling and projectile scenarios only, where an off-object region cannot overlap another physical object.

Multi-object readout Both collision targets are read against scene-level values from all objects' ROIs, and differ in how the objects are combined: approach angle concatenates the ROIs, since a relative quantity cannot be read from a symmetric summary, while total momentum $\boldsymbol{p}_{\text{tot}}$ averages them.

Input baseline Activations are captured in an I2V setting from a denoising process started at a noised latent $\boldsymbol{z}_t$ (Section 2.2), a compressed and noised version of the simulator's ground-truth video. To measure the information this input already carries, we concatenate $\boldsymbol{z}_t$ with the conditioning latent $\boldsymbol{z}_{\text{img}} \in \mathbb{R}^{1 \times H' \times W' \times C'}$, repeated across the F' latent frames, along the channel dimension, giving the baseline signal $\boldsymbol{z}_{tc} \in \mathbb{R}^{F' \times H' \times W' \times 2C'}$, and train a probe using the same pooling variants — spatial and spatio-temporal, global and ROI — as those on the DiT activations.

Permutation null Because a probe sees far more features than videos, we verify that the protocol itself yields nothing on structureless labels: we refit every probe on targets shuffled across videos, under the same folds, penalty search and pooling, with 50 permutations. This leaves R^2 at chance for every quantity, model, layer and noise level, with a mean of -0.056.

B.2 ADDITIONAL PROBING RESULTS

	Wan 2.1		CogVideoX-1.5		Open-Sora 2.0	
	global	ROI	global	ROI	global	ROI
Rolling						
speed	0.78 (+0.53)	0.98 (+0.14)	0.74 (+0.56)	0.98 (+0.13)	0.64 (+0.49)	0.91 (+0.13)
direction$_x$	0.85 (+0.33)	0.99 (+0.07)	0.80 (+0.84)	0.99 (+0.25)	0.53 (+0.22)	0.87 (-0.01)
direction$_y$	0.85 (+0.81)	0.99 (+0.20)	0.81 (+0.90)	0.98 (+0.42)	0.52 (+0.03)	0.90 (-0.00)
momentum	0.91 (+0.15)	0.93 (+0.45)	0.92 (+0.19)	0.90 (+0.40)	0.83 (+0.14)	0.85 (+0.33)
kinetic energy	0.81 (+0.23)	0.89 (+0.34)	0.85 (+0.24)	0.87 (+0.32)	0.69 (+0.19)	0.78 (+0.25)
mass	0.96 (+0.03)	0.96 (+0.68)	0.93 (+0.21)	0.90 (+0.54)	0.96 (+0.16)	0.94 (+0.37)
radius	0.98 (+0.04)	0.99 (+0.68)	0.94 (+0.20)	0.97 (+0.56)	0.96 (+0.19)	0.95 (+0.37)
Collision						
approach angle	0.33 (+0.34)	0.54 (+0.59)	0.06 (+0.06)	0.33 (+0.31)	-0.04 (-0.02)	0.03 (-0.02)
total momentum	0.82 (+0.37)	0.74 (+0.24)	0.90 (+0.83)	0.76 (+0.54)	0.64 (+0.02)	0.58 (-0.01)
total kinetic energy	0.83 (+0.09)	0.79 (+0.15)	0.93 (+0.14)	0.79 (+0.17)	0.81 (+0.01)	0.73 (+0.19)
total mass	0.97 (+0.06)	0.95 (+0.30)	0.97 (+0.06)	0.91 (+0.21)	0.99 (+0.07)	0.96 (+0.30)
speed (t)	n/a	0.95 (+0.55)	n/a	0.95 (+0.70)	n/a	—
direction$_x$ (t)	n/a	0.95 (+0.42)	n/a	0.95 (+0.98)	n/a	0.42 (+0.22)
direction$_y$ (t)	n/a	0.95 (+0.96)	n/a	0.96 (+1.19)	n/a	—
momentum (t)	n/a	0.84 (+0.56)	n/a	0.81 (+0.52)	n/a	—
kinetic energy (t)	n/a	0.75 (+0.54)	n/a	0.76 (+0.65)	n/a	—
Projectile						
gravity	0.13 (+0.07)	0.92 (+0.65)	0.23 (+0.30)	0.90 (+0.69)	0.17 (+0.16)	0.61 (+0.31)
launch angle	0.61 (+0.53)	0.95 (+0.69)	0.65 (+0.44)	0.92 (+0.58)	0.71 (+0.62)	0.83 (+0.65)
initial speed	0.57 (+0.22)	0.90 (+0.15)	0.68 (+0.34)	0.90 (+0.28)	0.64 (+0.22)	0.80 (+0.05)
radius	0.88 (+0.26)	0.89 (+0.62)	0.73 (+0.33)	0.83 (+0.36)	0.88 (+0.42)	0.89 (+0.44)
speed (t)	0.84 (+0.70)	0.97 (+0.35)	0.83 (+0.87)	0.94 (+0.65)	0.75 (+0.62)	—
direction$_x$ (t)	0.93 (+1.20)	0.99 (+0.33)	0.93 (+1.43)	0.99 (+0.89)	0.72 (+0.35)	—
direction$_y$ (t)	0.16 (+1.16)	0.49 (+1.58)	0.14 (+0.64)	0.20 (+1.10)	-0.18 (+0.21)	—
direction$_z$ (t)	0.93 (+0.36)	0.99 (+0.61)	0.94 (+0.92)	0.99 (+0.58)	0.92 (+0.79)	—

Table B.1: Decodability of physical targets from ROI and global readouts, at $\sigma = 0.500$ (0 dB; Fig. 2 gives the full noise dependence). Each cell is the probe R^2 averaged over 10 transformer blocks across network depth, block 0 excluded, with the gap over that arm's own z_{tc} baseline in parentheses. Rows marked (t) are time-varying: one ridge over all (video, token-frame) pairs rather than a mean of per-frame scores. Per-ball rows are averaged over the two balls, which agree to ≤ 0.03 throughout, and have no comparable global cell (n/a; multi-object).

B.2.1 DECODABILITY ACROSS DEPTH, NOISE AND ROI SIZE

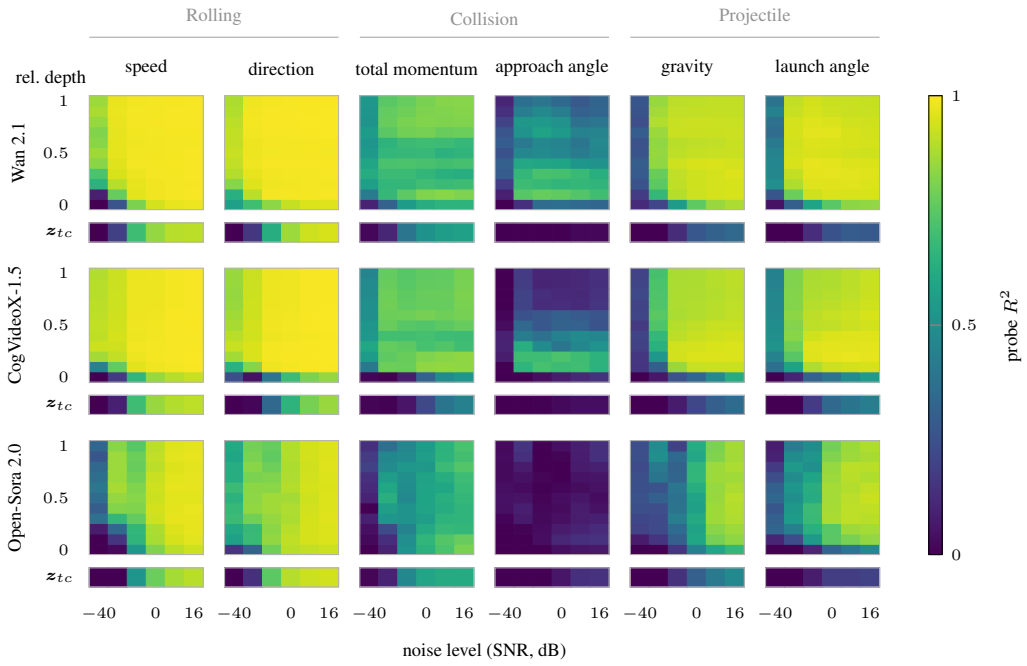


Figure B.1: The depth × noise grid behind Fig. 2. Each cell is one probe (R^2, $n = 200$, 5-fold video-level cross-validation) fitted at one noise level and one transformer block; the main figure's line is the column mean of this grid and its band the column min–max. The y axis is *relative* depth, block index divided by $L - 1$ for a network of L blocks, not a raw block number: the models have 40, 42 and 57 blocks, so only the fraction of the network is comparable. Each panel samples eleven uniformly spaced blocks. The small panel below each grid is that panel's matched z_{tc} baseline on the same colour scale.

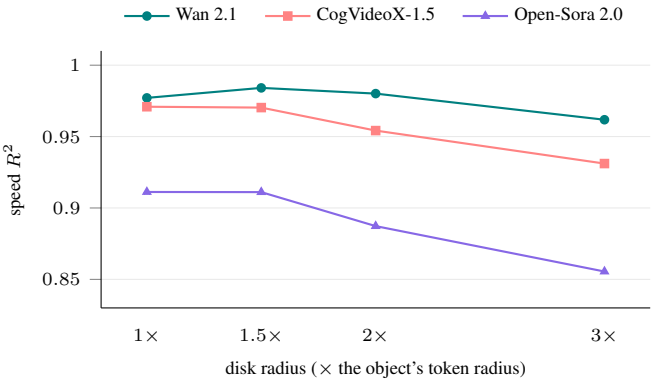


Figure B.2: Disk-radius sweep behind the single ROI bar of Fig. 3a: speed R^2 against the disk radius, as a multiple of the object's own token radius, at one token-frame. The curve is flat between 1× and 1.5× before falling away. The 1.5× radius is chosen as default. Rolling closeup-room, n=200, layer 21, 0 dB, all arms d-dimensional means over a tracked disk.

B.2.2 EXTRAPOLATION EXTENDED RESULTS

Applied unchanged — with the same weights and training-set standardization statistics — the probe appears not to transfer: 26 of 57 evaluations have negative R^2, and the median sits at just 0.16

(Table B.2). The corresponding predictions nevertheless track the targets closely, with a median Pearson correlation of $r = 0.896$. The apparent failure is therefore mostly a calibration failure: the scene shift changes the readout's gain and offset, while preserving much of its ordering of examples. Because R^2 is unbounded below, even a modest gain or offset mismatch can overwhelm an otherwise informative readout.

The table separates these effects. *as-is* evaluates the source probe without adjustment; r measures transfer of its ordering independent of scale; and *recal* fits only a target-scene gain and offset. The resulting scores test whether the learned direction transfers after accounting for a scene-specific change of units, rather than allowing the probe to be refit.

The gap between r and *as-is* is therefore a failure of gain and offset rather than of direction. The two arms separate localization from decodability: the global arm's *tgt CV* is high throughout, so a global readout does carry these quantities within a scene — what it does not do is survive the move to another one, with a median r of 0.327 against the ROI arm's 0.896. Applied unchanged, the global probe fails almost everywhere: 35 of its 42 cells are negative against 22 of 42 for the ROI arm, and the worst reaches -2483.

			Wan 2.1				CogVideoX-1.5				Open-Sora 2.0			
condition	quantity	arm	tgt CV	as-is	r	recal.	tgt CV	as-is	r	recal.	tgt CV	as-is	r	recal.
background	dir-x	ROI	0.99	0.95	0.99	0.97	0.99	0.89	0.97	0.94	0.89	−14.36	0.32	−0.03
		global	0.91	−63.82	0.53	0.24	0.95	−33.11	0.33	0.08	0.84	−1110	0.06	−0.03
	dir-y	ROI	0.99	0.95	0.99	0.98	0.99	0.79	0.96	0.92	0.88	−13.14	0.31	0.00
		global	0.91	−4.83	0.56	0.27	0.95	−149	0.66	0.37	0.80	−155	0.04	−0.04
	speed	ROI	0.99	0.42	0.94	0.88	0.99	0.55	0.95	0.89	0.93	−30.70	0.42	0.06
		global	0.84	−664	0.23	0.04	0.90	−229	0.24	0.02	0.85	−4.14	0.27	−0.28
	momentum	ROI	0.93	0.58	0.80	0.63	0.89	0.11	0.71	0.48	0.93	−185	0.34	0.03
		global	0.91	−288	0.18	0.03	0.93	−1363	0.12	0.01	0.93	−1104	0.19	0.04
	radius	ROI	0.99	−0.98	0.90	0.80	0.98	0.47	0.91	0.82	1.00	−103	0.71	0.49
		global	0.99	−848	0.51	0.21	0.97	−2483	0.16	0.02	1.00	−12.81	0.24	0.03
	median	*ROI*	0.99	0.58	0.94	0.88	0.99	0.55	0.95	0.89	0.93	−30.70	0.34	0.03
		global	0.91	−288	0.51	0.21	0.95	−229	0.24	0.02	0.85	−155	0.19	−0.03
shape	dir-x	ROI	1.00	0.80	0.97	0.93	1.00	0.51	0.97	0.93	0.98	−3.02	0.75	0.49
		global	0.97	0.54	0.95	0.85	0.98	0.53	0.97	0.93	0.97	−0.00	0.74	0.37
	dir-y	ROI	1.00	0.82	0.95	0.87	1.00	0.77	0.97	0.94	0.99	−25.25	0.90	0.79
		global	0.97	0.17	0.96	0.93	0.99	0.81	0.98	0.95	0.99	−32.09	0.81	0.62
	speed	ROI	0.99	−2.27	0.97	0.94	0.99	−3.29	0.97	0.93	0.98	−25.84	0.85	0.72
		global	0.92	−0.66	0.89	0.76	0.97	−3.07	0.92	0.82	0.98	−97.57	0.62	0.33
	momentum	ROI	0.99	0.75	0.92	0.83	0.99	0.28	0.85	0.57	0.98	−1.12	0.77	0.58
		global	0.92	0.37	0.76	0.34	0.97	0.68	0.92	0.79	0.98	−1.96	0.57	0.32
	radius	ROI	–	–	–	–	–	–	–	–	–	–	–	–
		global	–	–	–	–	–	–	–	–	–	–	–	–
	median	*ROI*	1.00	0.77	0.96	0.90	1.00	0.40	0.97	0.93	0.98	−14.13	0.81	0.65
		global	0.94	0.27	0.92	0.80	0.98	0.61	0.94	0.88	0.98	−17.02	0.68	0.35
perspective	dir-x	ROI	0.97	0.92	0.97	0.91	0.97	0.87	0.96	0.91	0.83	−33.13	0.49	0.13
		global	0.87	−2.47	0.76	0.46	0.89	−6.83	0.63	0.26	0.59	−146	0.02	−0.13
	dir-y	ROI	0.96	0.55	0.92	0.84	0.95	0.34	0.71	0.45	0.78	−2.29	0.48	0.17
		global	0.70	−11.76	0.38	0.05	0.80	−7.26	0.15	−0.10	0.57	−146	0.04	−0.20
	speed	ROI	0.91	0.05	0.75	0.54	0.92	−0.38	0.69	0.47	0.55	−21.79	0.27	0.05
		global	0.59	−321	0.58	0.30	0.79	−97.57	0.38	0.11	0.20	−984	0.01	−0.03
	momentum	ROI	0.78	−34.39	0.69	0.46	0.79	−19.48	0.65	0.40	0.66	−140	0.31	0.05
		global	0.73	−7.13	0.62	0.37	0.70	−20.56	0.52	0.25	0.42	−748	0.22	−0.10
	radius	ROI	0.93	−24.90	0.74	0.53	0.91	−15.06	0.68	0.44	0.96	−29.19	0.37	0.05
		global	0.92	−290	0.50	0.24	0.82	−629	0.33	0.09	0.93	−384	0.22	−0.25
	median	*ROI*	0.93	0.05	0.75	0.54	0.92	−0.38	0.69	0.45	0.78	−29.19	0.37	0.05
		global	0.73	−11.76	0.58	0.30	0.80	−20.56	0.38	0.11	0.57	−384	0.04	−0.13
count	dir-x	ROI	0.98	0.98	0.99	0.98	0.98	0.96	0.98	0.96	0.62	−0.04	0.58	0.30
		global	−0.18	−4.67	0.32	−0.01	−0.36	−1.34	0.26	0.00	−0.65	−3.58	0.08	−0.07
	dir-y	ROI	0.98	0.97	0.99	0.97	0.98	0.95	0.98	0.95	0.70	0.16	0.67	0.45
		global	−0.17	−0.91	0.40	−0.19	−0.34	−4.43	0.23	−0.53	−0.14	−2.45	0.24	−0.42
	speed	ROI	0.96	0.92	0.97	0.93	0.96	0.92	0.96	0.92	0.60	−2.07	0.57	0.32
		global	−0.62	−0.72	0.13	−0.08	−0.55	−0.83	0.14	0.00	−0.38	−5.03	0.13	−0.14
	momentum	ROI	0.91	0.36	0.91	0.80	0.90	0.69	0.91	0.82	0.71	−6.18	0.60	0.31
		global	0.29	−14.48	0.36	0.04	0.25	−9.14	0.29	−0.01	0.46	−10.26	0.31	0.03
	radius	ROI	0.98	0.73	0.97	0.93	0.98	0.90	0.98	0.96	0.93	−0.57	0.89	0.74
		global	0.90	−7.59	0.37	0.04	0.69	−12.16	0.28	0.01	0.96	−5.05	0.12	−0.02
	median	*ROI*	0.98	0.92	0.97	0.93	0.98	0.92	0.98	0.95	0.70	−0.57	0.60	0.32
		global	−0.17	−4.67	0.36	−0.01	−0.34	−4.43	0.26	0.00	−0.14	−5.03	0.13	−0.07
	all shifts	*ROI*	0.98	0.73	0.95	0.88	0.98	0.55	0.96	0.91	0.89	−14.36	0.57	0.30
		global	0.90	−7.13	0.51	0.24	0.89	−9.14	0.33	0.08	0.84	−32.09	0.22	−0.03

Table B.2: Per-cell transfer numbers behind Fig. 4. *tgt CV* is the cross-validated score of a probe fitted *and* read on the target scene: the optimal in-distribution score we could achieve. *as-is* applies the source probe unchanged and is unbounded below. r is the Pearson correlation on the target scene, which bounds what the calibration of Section 3.4 could reach. *recal* is what that ten-video calibration actually achieves — the quantity plotted in Fig. 4, given here per cell. *radius* is not evaluable under the object swap: the can's radius is fixed by construction, so the target has no variance — the object-swap summary rows are therefore over four quantities, not five.

C STEERING PROTOCOL DETAILS

Our steering experiments push the activations of a generation along a direction associated with rightward motion and check whether the ball actually changes course. Below we describe the steering

protocol, the text conditioning and configuration used, and what a steering direction is, how it is injected, and how its effect is scored. A direction can be kept at full resolution, with one vector per token, or pooled into a single vector; only the former steers (*Full-resolution and pooled directions* below), and *Per-backbone results* below reports what it achieves on each backbone.

Full-resolution and pooled directions A *full-resolution* direction at block l is a tensor $\boldsymbol{d}_l \in \mathbb{R}^{n_{\text{tok}} \times d}$ on the same token grid as the activations, with one row $\boldsymbol{d}_{l,s} \in \mathbb{R}^d$ per token position $s = 1, \ldots, n_{\text{tok}}$. The direction is the weight vector $\boldsymbol{w}_l$ of a probe fitted on full-resolution activations, reshaped to $n_{\text{tok}} \times d$; its activations come from single-step forward noising, as in the probing experiments: a labeled simulator video is encoded, noised to a chosen level and passed once through the model (Sections 2.2 and 3.1). A *pooled* direction $\bar{\boldsymbol{d}}_l \in \mathbb{R}^d$ is estimated the same way from spatio-temporally mean-pooled activations, as used by the probes of Section 3.1, and so carries no token layout. Pooled directions, however, do not steer. Pushes weak enough to leave the video intact keep the direction flip ratio at noise level, while stronger pushes make the video deteriorate or collapse before the ball flips. A likely reason is that the moving object occupies only a handful of the n_{tok} token positions: most rows $\boldsymbol{d}_{l,s}$ are near zero and a few are large, so their token average, applied identically at every position, retains only a small fraction of the per-token signal ($\approx 3\%$ of $\sum_s \|\boldsymbol{d}_{l,s}\|^2$ for a realized-motion contrast). Raising the push strength to compensate mostly adds a uniform drift to background tokens that carry none of the signal, which the video absorbs as a smear or global pan before the object's own motion changes. All steering results therefore use full-resolution directions.

Injection A push rotates each token's activation $\boldsymbol{h}_{l,s}$ a fraction $\alpha \in [0, 1]$ toward its target $\boldsymbol{d}_{l,s}$ on the unit sphere:

$$\texttt{slerp}: \quad \boldsymbol{h}_{l,s} \leftarrow \frac{\sin\big((1-\alpha)\theta\big)}{\sin\theta}\,\boldsymbol{h}_{l,s} + \frac{\sin(\alpha\theta)}{\sin\theta}\,\boldsymbol{d}_{l,s}, \qquad \theta = \arccos\left(\frac{\boldsymbol{h}_{l,s}\cdot\boldsymbol{d}_{l,s}}{\|\boldsymbol{h}_{l,s}\|\;\|\boldsymbol{d}_{l,s}\|}\right) \tag{1}$$

For a full-resolution direction, each token is rotated toward its own row. For a pooled direction, the target is the same vector at every position, $\boldsymbol{d}_{l,s} \equiv \bar{\boldsymbol{d}}_l$. The push is applied at a single block l, during the first quarter of the denoising trajectory only (Table C.1).

Prompts Steered generations use a *neutral* prompt that describes the scene without naming a direction; the prompt reference uses a *directed* prompt that additionally states that the ball rolls right. The two differ only in the highlighted phrase:

NEUTRAL PROMPT

> *A single continuous shot: a purple sphere with a fine matte rubber-grain texture sits at rest on a warm honey-toned wood-plank floor, seen directly from above, a same wood-plank flooring extending uninterrupted to every edge of the frame, with no walls or horizon in view filling the background. The ball begins rolling smoothly sideways, accelerating horizontally until it exits the **left or right edge** of the frame.*

DIRECTED PROMPT

> *A single continuous shot: a purple sphere with a fine matte rubber-grain texture sits at rest on a warm honey-toned wood-plank floor, seen directly from above, a same wood-plank flooring extending uninterrupted to every edge of the frame, with no walls or horizon in view filling the background. The ball begins rolling smoothly sideways, accelerating horizontally until it exits the **right edge** of the frame.*

Evaluation Every push is evaluated on *left-rolling seeds*, whose unintervened generation produces a ball rolling left, with the direction oriented toward rightward motion. Its *direction flip ratio* is the fraction of these seeds whose steered motion is classified as rightward. The classification is measured, not eyeballed. Each generated clip is passed through a Mask R-CNN instance segmentation (COCO weights) on sampled frames; the object's mask centroid is tracked across those frames, and the resulting track is reduced to a motion class: *left*, *right*, or *unclassified*, where the last covers

ambiguous, vertical, or unreliable tracks due to degenerations or object spawns. A track is classified only if its longest unbroken run of detections covers at least 30% of the clip and at least 70% of frame pairs agree in sign with the net displacement, so a handful of large steps cannot carry a jittery track; one axis counts as dominant at a ratio of 1.5. Every steered clip shares its seed with an unsteered $\alpha = 0$ clip and is compared against that clip rather than against a pooled baseline, which removes seed-to-seed variance and ensures that a seed already rolling in the target direction is never counted as a flip. Two references put the direction flip ratio in context: a *prompt* reference, the flip ratio of an unsteered generation under the directed prompt (Appendix C), run with the same Classifier-Free Guidance (CFG) scale and number of denoising steps as the probe push; and a *reversed* push, the same configuration rotating toward $-\boldsymbol{d}_{l,s}$ instead, which should flip no seeds to the right. The configuration behind each reported result is listed in Table C.1.

Per-backbone results Table C.1 lists, for each backbone, the configuration of the probe push together with the direction flip ratio of the probe direction and of both references, the numbers Fig. 5 plots. All results use full-resolution directions and are evaluated on the same 52 known left-rolling seeds for every backbone, the probe direction and both references. The probe direction exceeds the prompt reference on every backbone, and the reversed push flips none of the seeds to the right.

Table C.1: Steering configuration and results behind Fig. 5. The first block gives the per-backbone settings of the probe push, which the negative control ($\neg$ Probe) reuses with the push reversed. The second block is shared by every generation, including the prompt reference: an unsteered generation under the directed prompt, with the same CFG scale and number of denoising steps as the probe push. The last block gives the direction flip ratio on the 52 left-rolling seeds, plus the fraction of seeds the reversed push leaves classified as rolling left.

	Wan 2.1	CogVideoX-1.5	Open-Sora 2.0
Steered block l	21	3	16
Push strength α	0.275	0.25	0.25
CFG scale w		6.0	
Denoising steps		40	
Steered steps		first quarter (first 10 of 40)	
Direction flip ratio			
Probe	0.45	0.35	0.31
Prompt reference	0.30	0.017	0.04
$\neg$ Probe	0.00	0.00	0.00
$\neg$ Probe, still classified left	0.77	0.50	0.33