MOST OF THE GEOMETRY CREDITED TO VIDEO FOUNDATION MODELS IS A GROUND PLANE: SIX PROBE CHOICES, AND THE SMALL EFFECT THAT SURVIVES THEM

Anonymous authors
Paper under double-blind review

ABSTRACT

A video foundation model is credited with 3D structure when depth decodes from its features, and a naive probe on driving footage reads R^2 near 0.85. A quadratic in the patch's own row and column reaches $R^2 = 0.844$ on the same data; regress it out, and the probe that read 0.848 adds $+0.122$ over it. Most of that number is a ground plane. Image position is one of six choices a depth probe makes, with motion, noise timestep, block, temporal window and patch grid; none is consistently reported, and in an exploratory audit on ten open checkpoints each moves the verdict, by $1.6\times$ to about $50\times$ or a change of sign. We give a configuration that closes all six and test what survives. On 200 real driving clips, every choice controlled but the block, which is fixed the usual way, four of four predictive encoders carry depth beyond position and flow and zero of six generators do; selecting the block on held-out splits as well keeps the encoder count and lets one generator cross at $+0.0188$, 2.5 times below the weakest encoder. The classes separate by magnitude, not by kind: an effect worth 0.05–0.12 of variance, an order of magnitude below the headline numbers it replaces. It survives a dense optical-flow control, which narrows the margin by a third, and a VAE control, after which encoders keep 70% of their advantage. Representational convergence, often read as corroboration, points the other way: the converging models that align patch for patch share a subspace that transfers a depth probe between them at ceiling, yet position-residualized depth decodes from it in $0/5$ models, so what independently trained video models agree on is where things sit in the frame. Harness, per-experiment results, preregistration and decision ledger are released.

1 INTRODUCTION

Video foundation models are credited with 3D structure because depth decodes from their frozen features; on KITTI footage a naive linear probe reads R^2 near 0.85. Shallow readouts recover depth, pointmaps and camera pose from these models, and Huang et al. (2026) read them as showing that generation models "exhibit a strong understanding of 3D objects and scenes" without 3D training; El Banani et al. (2024) make it definitional, positing that 3D awareness means that representations encode scene structure. Neither reports the readout with position or camera motion regressed out.

Whether that structure is there decides three uses. As world models the same features are the state a planner searches: V-JEPA 2-AC plans by minimizing representation distance to a goal on a robot arm (Assran et al., 2025). As generators they are scored on physical plausibility, where the best of eight models reaches 29.5% of a 100% ceiling and realism does not predict the score (Motamed et al., 2026). As backbones their priors are repurposed for depth (Hu et al., 2025), as image generators' were (Ke et al., 2024). Each use is premised on depth that is not image position.

Yet most of that readout is a ground plane. On the same footage row and column alone, as a quadratic, reach $R^2 = 0.844$, and the naive number does not separate the models it is used to separate. Three of ours read 0.782 to 0.848 naive on the same KITTI frames. Once the position basis is removed they span $+0.122$ to -0.098, the last of them predicting depth *worse* than the ground plane alone

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.

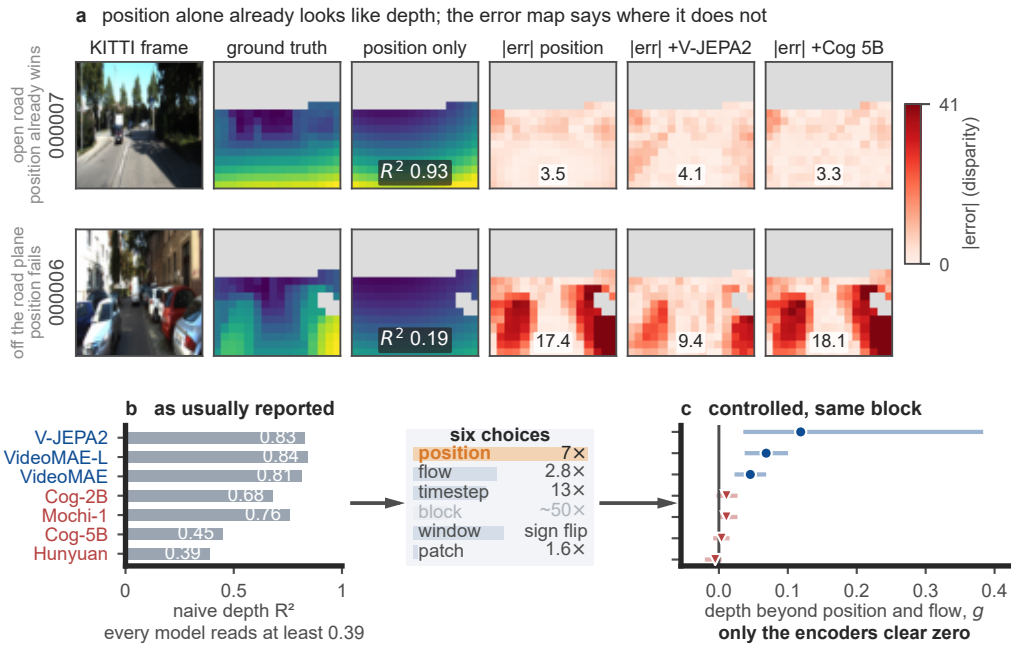


Figure 1: **What the probe reads, how far each choice moves it, and what survives. (a)** KITTI frames, disparity on the probe's 16×16 grid, the quadratic image-position prediction, and three designs' absolute error on one scale (40 released scenes, no flow control; all in Figure 7). On open road position alone reaches $R^2 = 0.93$; off the road plane it falls to 0.19, and V-JEPA2 roughly halves the error where CogVideoX-5B does not. **(b)** Seven checkpoints with no control (V-JEPA2-H has none). **(c)** The same features and scenes with every choice controlled but the block; bars are floor-subtracted g with 95% bootstrap intervals. The middle panel gives each choice's fold change from Table 1, each on its own model and data; the block is greyed because (c) keeps it.

(Appendix G, Figure 1). Position is the first of six such choices; the other five are the motion control, the noise timestep, the block read, the temporal window and the patch grid. Each moves the verdict by $1.6\times$ to about $50\times$, or flips its sign, none is consistently reported (Table 1), and on the same models and data one setting reads depth in every checkpoint while another leaves it in the encoders alone (Figure 1b,c). We do not recompute published results; on models we control, the reported number is one evaluation of $\Delta(m; \pi)$, a map over the six choices (Eq. 5).

We measure that map on ten open checkpoints: a 0.4B predictive encoder and five latent video-diffusion DiTs of 2B–13B over four families run through a five-rung preregistered ladder on KITTI with its controls, and four more at the verdict. The probe is closed-form ridge throughout, so every effect is the model's rather than probe engineering's (§3). One quantity carries the positive half: $\mathrm{CKA_c}$, the mean linear CKA between a model's scene-pooled features and the four models that converge on a shared subspace, on 40 unlabeled KITTI clips. One forward pass per clip, no labels, no training, no fitted parameters. It separates the pool into two occupied regimes, and the cluster clears the rungs the outsiders fail (Figure 2); its one out-of-sample test failed (§6.3).

Contributions. **(1) An audit.** We measure the six choices with model and data held (§4), retiring five probing defaults (Appendix R); no published probe we know of controls more than one.

(2) A protocol. We state what to report instead and what closing each choice costs (§4.2); it leaves the ranking intact and decides whether there is a claim.

(3) A verdict. On 200 real driving clips with every choice controlled but the block, four of four predictive encoders keep depth that position and flow do not explain, and no generator does. Selecting the block out of fold as well, one generator crosses, 2.5 times below the weakest encoder (§5). The effect is worth 0.05–0.12 of variance, the size of the structure-from-motion claim.

(4) What convergence selects. Converging models that align patch for patch share a subspace where depth probes transfer at ceiling and position-residualized depth does not decode: image layout (§6).

2 Related work

Probing geometry in frozen features, and the controls that discipline it. Depth, pointmaps and correspondence are recoverable from video (Huang et al., 2026) and image (El Banani et al., 2024) foundation models, latent image diffusion (Chen et al., 2023; Tang et al., 2023) and feed-forward reconstruction (Wang et al., 2024; 2025); Bratulić et al. (2025) ask our question of another model class. Each fixes one probe configuration; we vary it. Geometry Forcing (Wu et al., 2026a) installs alignment with a designated teacher where we measure what already exists between independently trained models. Linear probes (Alain & Bengio, 2017) are the standard readout and their critics supply our discipline: control tasks separate probe capacity from encoded structure (Hewitt & Liang, 2019), and decodability licenses no claim about use (Belinkov, 2022).

Representational similarity, and how video models are scored. CKA (Kornblith et al., 2019) is the instrument, alongside canonical-correlation methods (Raghu et al., 2017; Morcos et al., 2018) and RSA (Kriegeskorte et al., 2008). Because it can be driven by a few dominant directions and can disagree with functional equivalence (Ding et al., 2021; Davari et al., 2023), we pair two kernels ($r{=}{+}0.983$), shuffle nulls and a stitching-style check (Lenc & Vedaldi, 2015; Bansal et al., 2021). Convergence is argued to grow with scale (Huh et al., 2024) and to leave alignable spaces across model zoos (Dravid et al., 2023; Achara et al., 2026); here it is measured, and non-monotone in scale. Our pool spans latent video diffusion (Rombach et al., 2022; Ho et al., 2022; Blattmann et al., 2023), DiT and MMDiT backbones (Peebles & Xie, 2023; Esser et al., 2024) and predictive encoders (Bardes et al., 2024; Assran et al., 2025). Video generators and world models are usually scored on their output—control (Ha & Schmidhuber, 2018; Hafner et al., 2025), interactive generation (Bruce et al., 2024; Yang et al., 2024), realism, plausibility and rendered geometry (Huang et al., 2024; Duan et al., 2025; Kang et al., 2025; Bansal et al., 2025; Wu et al., 2026b)—and we read the representation. Loss-side scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) do not constrain which competences appear (Wei et al., 2022; McKenzie et al., 2023), and our scale paradox (§6.1) is an inverse-scaling instance with a measured mechanism.

3 Setup

Models and features. One predictive encoder and five latent video-diffusion DiTs, four architectures, 0.4B–13B, carry the ladder and the convergence pool: V-JEPA2 (Assran et al., 2025), Wan2.2-TI2V-5B (Team Wan et al., 2025), CogVideoX-2B and -5B (Yang et al., 2025), HunyuanVideo-13B (Kong et al., 2024) and Mochi-1 (Genmo Team, 2024); §4 and §5 add VideoMAE and VideoMAE-large (Tong et al., 2022), V-JEPA2 ViT-H and LTX-Video (HaCohen et al., 2024), ten checkpoints in all, four predictive encoders and six generators. Diffusion features are noised latents at a fixed timestep for the ladder and a clean latent for the controlled probes; blocks are fixed per model by the L1 sweep (Appendix C) and selected out of fold in §5.2. The ladder is five preregistered rungs per model, L1 linear presence, L2 causal use, L3 lawfulness under a symmetry, L4 compositional generalization and L5 behaviour under cue conflict (§6.2).

Data and probe. KITTI-2015 scene-flow frames (Menze & Geiger, 2015; Geiger et al., 2012) with disparity ground truth: 200 scenes for L1, 100 for L3, and 40 with real 21-frame motion clips for every cross-model analysis. Synthetic clips drive L4 (bump arrangements) and L5 (cue conflict). Probes are scene-disjoint, 6 seeds, on 16×16 patches (Wan2.2 and LTX-Video 8×8).

Preregistration and controls. All pass rules, controls and kill criteria were locked before data and are released with the harness; deviations are append-only and exploratory results are marked †. Every rung is tested against a random-feature control, an untrained-init control where applicable and a shuffled-label permutation (Hewitt & Liang, 2019), with bootstrap 95% CIs and Benjamini–Hochberg correction (Benjamini & Hochberg, 1995) across blocks and models.

The two quantities this paper reports. With $d \in \mathbb{R}^N$ the depth target over the N valid patches of the scenes scored, $P \in \mathbb{R}^{N\times 5}$ the quadratic image-position basis $[\,r, c, r^2, c^2, rc\,]$, $f \in \mathbb{R}^{N\times q}$ the motion term ($q = 1$ for the frame-difference proxy, $q = 3$ for a dense field's mean u, v and magnitude), $Z_m \in \mathbb{R}^{N\times 50}$ model m's frozen features whitened to 50 dimensions (Eq. 11) and m_0 the same architecture at random initialization, read at the same block:

$$\Delta(m) = R^2(d \mid P, f, Z_m) - R^2(d \mid P, f)\,, \qquad g(m) = \Delta(m) - \Delta(m_0), \tag{1}$$

$$\mathrm{CKA_c}(m) = \frac{1}{|C \setminus \{m\}|} \sum_{c\in C\setminus\{m\}} \mathrm{CKA}(A_m, A_c)\,. \tag{2}$$

The naive readout is $R^2(d \mid Z_m)$ alone, and g is the increment over the untrained floor $\Delta(m_0)$. Both regressions share one ridge estimator on $X \in \mathbb{R}^{N\times(J+1)}$, the controls alone ($J = 5 + q$) or the controls with Z_m appended ($J = 55 + q$) plus an intercept, fit on training scenes, scored on held-out ones:

$$\hat\beta \;=\; \left(X^\top X + \lambda D\right)^{-1} X^\top d, \qquad R^2 \;=\; 1 - \frac{\|d - X\hat\beta\|^2}{\|d - \bar d\|^2}, \tag{3}$$

with $D = \mathrm{diag}(1, \ldots, 1, 0)$ so the intercept is unpenalised, $\lambda = 1$ for every model, dataset and arm, and $\bar d$ the held-out mean; nothing in the probe is tuned per model. In Eq. 2, C is the set of four models that converge (§6.1), so each member is compared with the other three, and $A_m \in \mathbb{R}^{S\times k_m}$ is model m's features averaged over patches per clip, for $S = 40$ clips at its native width k_m, column-centred. The similarity is linear CKA (Kornblith et al., 2019); for two such matrices A and B,

$$\mathrm{CKA}(A, B) \;=\; \frac{\|A^\top B\|_F^2}{\|A^\top A\|_F\,\|B^\top B\|_F} \;=\; \frac{\langle K_A, K_B\rangle}{\sqrt{\langle K_A, K_A\rangle\,\langle K_B, K_B\rangle}}, \qquad K_A = AA^\top, \tag{4}$$

with $K_A \in \mathbb{R}^{S\times S}$.

4 Six Choices Decide the Verdict, and Image Position Decides Most of It

On KITTI image position alone reaches $R^2 = 0.844$ (Figure 1a), so most of the audit runs on motion-parallax stimuli, where 67% of depth survives residualization, and each choice is read on the data where it is largest (Table 1). The audit is exploratory, run after the preregistered ladder of §6.

One map, six choices. A depth probe fixes six choices before it reads anything, so a paper reports one evaluation of

$$\Delta(m; \pi), \qquad \pi \;=\; (\,P,\; f,\; \tau,\; \ell,\; w,\; \gamma\,), \tag{5}$$

where P is the image-position basis regressed out and f the motion term. The other four fix how the features are read: τ the noise level a generator is read at, ℓ the block read (a transformer layer), w the temporal window and where the depth target sits in it, and γ the patch grid. Table 1 moves one choice at a time with the other five held; the probes we build on control at most one.

Position-independent depth exists, and it lies outside what the models share. With the position basis removed, V-JEPA2 decodes depth† at $R^2_{\text{resid}} = \mathbf{0.517}$, where $R^2_{\text{resid}} = R^2\big(d - [P, \mathbf{1}]\,\hat w_P \mid Z_m\big)$ scores the depth left by a ridge fit $\hat w_P$ on position over the training scenes (Eq. 12; 6 seeds, $[0.509, 0.523]$). VideoMAE reaches $\mathbf{0.948}$ $[0.945, 0.952]$ at a block chosen without the motion test, and VideoMAE-large repeats it. Random-initialized controls read -0.025 and -0.017, and freezing the middle frame drops the readout to 0.076 and -0.009, so it is learned and motion-derived. It decodes at $+0.172$ from the directions the converged models do *not* share and at -0.074 from those they do (§6.4), which is why the residual gate of §6.4, reading only the shared directions, missed it.

4.1 The Audit, Choice by Choice

Table 1 and Figure 1 price the six. *Position* removes the most, $7\times$, and *motion* a further $2.8\times$. The *timestep* falsified our own paradigm law: at the ladder's noised timestep all four generators read zero beyond position and flow. At a clean latent CogVideoX-5B recovers $\mathbf{0.153}$ against an encoder

Choice	Uncontrolled	Controlled	Swing	n	Data
Image position	0.848 naive depth	$+0.122$ residualized	$7\times$	8	KITTI
Optical flow	0.534 (position only)	0.188 (position+flow)	$2.8\times$	$3^{\ddagger}$	synthetic
Noise timestep	0.326 (clean latent)	0.025 (our fixed t)	$13\times$	1	synthetic
Network block	0.147 (best block)	0.003 (worst probed)	$\sim 50\times$	3	synthetic
Temporal window	$+0.295$ (2.1 s, centred)	-0.082 (8.5 s span)	sign flip	1	ScanNet
Patch resolution	$+0.0479$ (8×8)	$+0.0298$ (16×16)	$1.6\times$	1	KITTI

Table 1: Six choices condition every verdict here and in the probing literature it builds on. Values† are Δ (Eq. 1) with the stated control in place, on the model and data where each choice is largest (V-JEPA2 in rows 1 and 5, VideoMAE in row 2, CogVideoX-2B in rows 3, 4 and 6; synthetic is motion-parallax stimuli). n is how many models each magnitude rests on: position is controlled in every row of Table 12, and three of the six rest on one model. The block's $\sim 50\times$ is the largest of the three models swept ($1.6\times$ to $54.6\times$). †Exploratory, outside the ladder's preregistration. ‡Variable-gain stimuli, where flow and depth decouple, controlling the stimulus's own flow magnitude rather than a dense field (that is §5.3); V-JEPA2 and CogVideoX-2B move $1.5\times$ and $2.8\times$, and the VideoMAE run predates the loading repair (Appendix T).

reference of 0.199, and CogVideoX-2B, at the ladder's block, stays at 0.014 (Appendix S); row 3 of Table 1 is its swing with position alone controlled. It reaches the ladder too: over five noise levels Wan2.2 stays lawful at $5/5$ blocks until pure noise while LTX-Video falls from $9/9$ to $0/9$, so one committed level is safe for one model and wrong for the other (Appendix L). The *block* swings one measurement about $50\times$ (Table 10). The *window* is why V-JEPA2, reading 64 frames to VideoMAE's 16, read -0.054 on ScanNet (Dai et al., 2017) where VideoMAE cleared: four seconds of camera motion against a one-frame target. The readout moves monotonically from -0.082 at 8.5 s to $+\mathbf{0.295}$ at 2.1 s centred, and a generator swept the same way reaches $+0.0773$ centred, a quarter of the swing (Appendix G). The *patch grid* moves CogVideoX-2B from $+0.0298$ at 16×16 to $+0.0479$ at 8×8, $1.6\times$. That 8×8 grid, which does not match the 16×16 position basis, had kept Wan2.2 out of every controlled probe until §5 reads it at its own grid.

4.2 What to Report Instead, and What It Changes

The corner π^* we would ask a probing paper to report closes each choice against the nuisance its row measures, in the order the audit prices them. Three are free and belong before any number is quoted: (i) regress out P, five extra columns in X (Eq. 3), the largest removal (row 1); (ii) read diffusion features at a clean latent (row 3); (iii) centre the depth target in the window and keep each model's native patch grid (rows 5 and 6). Two cost something. A dense flow field costs one forward pass per clip (row 2). Choosing the block on one set of splits to report it on another costs a probe fit per block per split, though one forward pass yields every block (row 4), the largest swing and the one our own generator count turned on. Report g rather than Δ (Eq. 1).

Adopting π^* does not re-rank our pool: naive and controlled readings of seven checkpoints agree at Spearman $+0.93$, 2 of 21 pairs reversed. It decides whether a claim exists: naively every model reads $R^2 \geq 0.389$; with the block fixed, only the encoders' intervals exclude zero (Appendix Q).†

Where the instrument is blind. At L1 it does not discriminate: an untrained V-JEPA2 decodes depth at 0.758, so linear decodability is an architectural prior (Appendix F). Where depth is mostly a ground plane little is left to ask with: position alone takes 72–84% of KITTI's depth variance, and the effect surviving there is 0.05–0.12 against the 0.199 that V-JEPA2 keeps beyond position and flow on the motion-parallax stimuli. And the window is inert at short spans: from 0.43 to 1.73 s with the target at the window start, a generator stays within -0.0124 to $+0.0122$, so V-JEPA2's sign flip at 8.5 s is invisible to a 13-frame probe.

5 As the Choices Close, a Small Effect Survives, and the Classes Separate by Magnitude

Table 12 gives every model. The 200-scene dissociation is a preregistered comparison on a post-hoc instrument (PREREG A.6; metric fixed after the ladder ran); §5.3 and the indoor runs are exploratory.

5.1 Four of Four Encoders, and a Generator Count That Depends on the Layer

With the block where the L1 sweep put it, motion controlled by the frame-difference proxy and generators read at a clean latent, four of four encoders clear their floors on 200 KITTI scenes ($+0.046$ to $+0.119$, bootstrap intervals excluding zero). None of the four 16×16 generators does (-0.005 to $+0.011$), and Wan2.2 and LTX-Video at their native 8×8 read $+0.0092$ and $+0.0112$, six-seed point estimates ($+0.0097$ and $+0.0078$ under a dense flow field, exploratory): zero of six. Measured in the 60-scene run at the same fixed block, V-JEPA2 and VideoMAE-large reach 0.091 and 0.083 over random-initialized floors of 0.044 and 0.047: gaps of 0.047 and 0.036, and both floors clear the detection criterion alone. Frame count does not split the classes: VideoMAE reads 16 frames, inside the generators' 13–17, and clears at $+0.046$. A partial counterpart: models predicting in representation space pass violation-of-expectation tests of intuitive physics where pixel-space predictors sit near chance (Garrido et al., 2025); our generators predict VAE latents, one step from pixels.

The unit of the test is the family. The four encoders are two architectures and the six generators five, so the unit is the family ϕ, scored by the mean $\bar{g}_\phi$ of its checkpoints: $+0.1086$ and $+0.0571$ for the encoder families $\mathcal{E}$ against -0.0053 to $+0.0112$ for the generator families $\mathcal{G}$, without overlap. With $\mathcal{F} = \mathcal{E} \cup \mathcal{G}$ and $\bar{g}_\mathcal{A}$ the mean over a set $\mathcal{A}$ of families, the separation margin and the exact one-sided permutation test over every choice of two families as encoders are

$$\mathcal{M} = \min_{\phi \in \mathcal{E}} \bar{g}_\phi - \max_{\phi \in \mathcal{G}} \bar{g}_\phi, \tag{6}$$

$$p = \binom{7}{2}^{-1} \Big| \big\{ \mathcal{A} \subset \mathcal{F} : \ |\mathcal{A}| = 2, \ \bar{g}_\mathcal{A} - \bar{g}_{\mathcal{F} \setminus \mathcal{A}} \ \geq \ \bar{g}_\mathcal{E} - \bar{g}_\mathcal{G} \big\} \Big|. \tag{7}$$

Here $\mathcal{M} = +0.0459$, and the test gives $1/21$, $p = 0.048$, the smallest p two families against five can produce. It rules out a labeling coincidence, nothing more; the size is $\mathcal{M}$ (Appendix H). Over checkpoints instead of families $\mathcal{M} = +0.0344$, the margin the flow control of §5.3 narrows.

5.2 Each Model at Its Own Best Block

Each block is chosen on three 70/30 splits of the 200 scenes and scored on three others. With Δ_s the increment of Eq. 1 on the split drawn by seed s, $\mathcal{S}_{\text{sel}} = \{1, 7, 13\}$ and $\mathcal{S}_{\text{eval}} = \{42, 100, 2026\}$,

$$\hat{\ell}(m) = \arg\max_{\ell} \ \frac{1}{3} \sum_{s \in \mathcal{S}_{\text{sel}}} \Delta_s(m, \ell), \qquad g^\star(m) = \frac{1}{3} \sum_{s \in \mathcal{S}_{\text{eval}}} \Big[\Delta_s\big(m, \hat{\ell}\big) - \Delta_s\big(m_0, \hat{\ell}\big) \Big], \tag{8}$$

with the floor m_0 read at the chosen block. Scored this way at its committed block, every model lands inside its Table 12 interval, all eight of them. Every encoder still clears at its best block ($+0.0475$ to $+0.1789$), and one generator does: Mochi-1 at its deepest block, $+0.0188$ $[+0.0031, \ +0.0327]$. That is 2.5 times below the weakest encoder and 9.5 times below the strongest, and three of four generators clear at no block probed. The categorical count, zero of six generators, does not survive this step (Appendix S). The interval on a selected maximum is uncorrected for the blocks it was chosen from, so its coverage is below nominal. $\mathcal{S}_{\text{sel}}$ and $\mathcal{S}_{\text{eval}}$ split the same 200 scenes differently rather than into disjoint sets, a weaker separation than disjoint scenes would give. Wan2.2 and LTX-Video, on a different pipeline, keep their Table 17 values.

5.3 Four Ways the Surviving Effect Could Be Wrong

Four attacks, pass rules fixed before they ran, all exploratory (Appendix I). Higher-order position: a position basis as wide as the model's features absorbs six percent of V-JEPA2's increment. Linear readout: a nonlinear readout, regularized gradient boosting (Friedman, 2001), of the class through which 3D understanding is read out of generators elsewhere (Huang et al., 2026), leaves the split in place. Motion: a dense optical-flow field in place of the frame-difference proxy narrows the encoder–generator margin by a third without the ranges overlapping. The VAE is part of the explanation and not all of it: the bare CogVideoX latent reads at the generator ceiling, while encoders on VAE round-tripped frames keep 72% (VideoMAE) and 70% (V-JEPA2) of their gap, on one VAE.

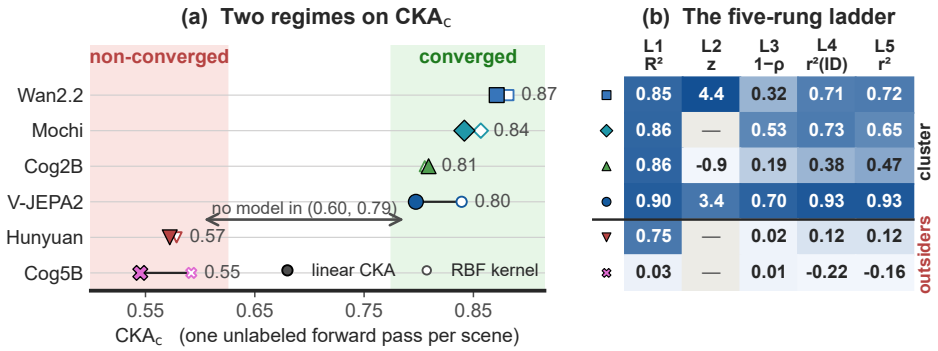


Figure 2: **Within the pool, which models carry geometry is predictable from $\mathrm{CKA_c}$. (a)** $\mathrm{CKA_c}$ splits the pool into two occupied regimes on 40 scenes (outsider max 0.59, cluster min 0.80); filled = linear CKA, open = RBF kernel. The *ordering* is stable but not the *location*: on independent 20-scene halves the scale shifts by ≈ 0.16 (Appendix T), so $\mathrm{CKA_c}$ ranks and gives no portable threshold. **(b)** The ladder as an evidence matrix, rows as in (a): L1 held-out R^2, L2 steering z (Eq. 14), L3 $1-\rho$ (Eq. 9), L4 in-distribution r^2 and L5 r^2 of the geometry cue (Appendix D); grey = not run. Hunyuan-13B decodes depth at 0.75 and fails every higher rung, which decodability misses.

5.4 Indoors, and a Generative Test That Failed

On ScanNet the same probe puts CogVideoX-2B at $+0.012$ against an untrained floor of -0.014, while VideoMAE and VideoMAE-large hold at $+0.121$ and $+0.236$ (Appendix P). Two NYU Depth v2 (Silberman et al., 2012) protocols do not count, one failing a preregistered feature-health gate that a DPT-Large anchor passes (Ranftl et al., 2021; 2022) and one reversing CogVideoX-2B on warp-induced parallax, where its untrained floor is already 0.159. A generative test of the same claim failed. Ablating an extracted SfM basis inside CogVideoX-5B during generation lowers the output's depth fidelity by 0.291, and a matched appearance basis by 0.305. The test does not separate them, and we record a failed attempt (Appendix P).

6 The Pool Does Converge, and What It Converges On Is Layout

Independently trained video models reportedly converge on a shared representation, read as corroboration that they share real structure. On this pool they do, and what they share is layout.

6.1 Two Occupied Regimes, and Nothing in Between

The CKA matrix (Appendix Figure 6a) comes down to four numbers: cluster pairs 0.72–0.89, outsider-to-cluster 0.49–0.60 against random-feature CKA of 0.32–0.46, the two CogVideoX scales 0.577, and Hunyuan–Cog5B 0.407, at that level. Collapsed to $\mathrm{CKA_c}$ (Figure 2a), the pool occupies two disjoint intervals with no model between them, on linear and RBF-kernel CKA alike ($r = +0.983$ element-wise). At n=6 the per-rung rank correlation is not significant on its own ($\rho = +0.43$ on L1, $+0.60$ on L4; $p = 0.40$ and 0.21). The prediction claim therefore rests on the evidence matrix, where on every rung the outsiders can be scored on both sit below all four cluster members (Figure 2b), and on the enumeration test of Appendix F. The partition survives independent halves, leave-one-out, pseudo-model controls, RBF-kernel CKA and the enumeration of all ten n=5 labelings; the one criterion that disagrees groups Hunyuan with the cluster (Table 6). Architecture does not predict membership: three different DiTs and an encoder converge while two copies of one DiT split.

The two outsiders fail differently, and scale predicts neither. One is decorrelated flat, the other over-smooth flat (Appendix Figure 6b), no more similar to each other than to the cluster, so "failing" is not one condition. CogVideoX-5B shares CogVideoX-2B's architecture, exits the cluster and loses linear depth ($0.86 \rightarrow 0.03$), and its untrained floor equals its trained value. Training *failed to build* linearly accessible geometry rather than destroying it, where the same control on Wan shows training building it. Parameter count is blind in both directions here; $\mathrm{CKA_c}$ is not (Appendix F).

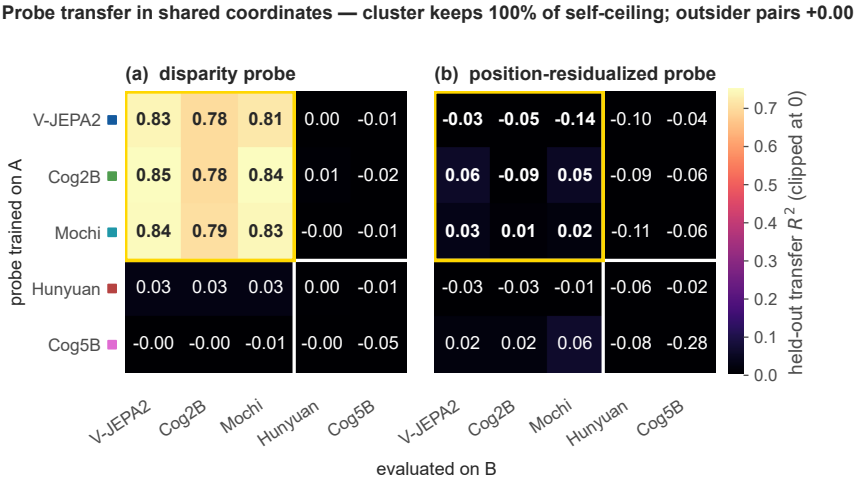


Figure 3: **The shared subspace transfers between models, and what it carries is layout.** Left: a ridge depth probe trained in model A's CCA-aligned shared coordinates, evaluated on model B's aligned features on held-out scenes; cluster pairs transfer at the self-transfer ceiling and outsider pairs at ≈ 0. Right: the same protocol on position-*residualized* depth, where no pair transfers.

6.2 The Cluster Clears the Rungs the Outsiders Do Not

Each rung of Figure 2b had its own falsifier. *L1*, linear presence, discriminates nothing: depth decodes in five of six models and an untrained V-JEPA2 reaches 0.758. *L2* splits causal use: steering the depth direction shifts the decoded value monotonically, scored as z against random directions, while the same push through denoising never moves the *generated* video past 3σ, so the code is internal and distributed. *L3* scores reflection lawfulness (Lenc & Vedaldi, 2015), a symmetry none of these architectures builds in (Cohen & Welling, 2016). With σ the horizontal flip of a clip, M the mirror permutation of the patch grid, Π a random permutation of it and δ the mean per-patch cosine distance,

$$\rho \;=\; \frac{\delta(F(\sigma x),\; MF(x))}{\delta(F(\sigma x),\; \Pi F(x))}, \tag{9}$$

is 1 at chance and lower when the model obeys the law. The gradient is model-specific, with both outsiders at chance. *L4*, composition, loses nothing out of distribution wherever a read-out exists. *L5*'s preregistered reading, "appearance-bound" like texture bias (Geirhos et al., 2019) because on textureless conflict stimuli every model tracked the shading cue, fell to a texture control: with a trackable parallax cue three of the four cluster models follow geometry at every probed block, and on real KITTI footage the V-JEPA2 oracle tracks it at 0.951 against a painted shading field. Mochi gains at one block of seven and Hunyuan a fifth as much, so the surviving reading is narrower: five of six models, all but CogVideoX-5B, track a trackable cue (Appendix D).

6.3 The Held-Out Forecast: $\mathrm{CKA_c}$ Called Mochi-1, the Locality Screen Did Not

We preregistered a cheaper predictor, the locality drop $\Lambda = c(1) - c(6)$ of the autocorrelation in Appendix Figure 6b (Eq. 13), and used it to forecast the then-unmeasured Mochi-1 ladder as ABSENT. $\mathrm{CKA_c}$ called it present, and the measured ladder returned $R^2 = 0.855$. The screen is dead (Appendix S), and the forecast that survived cost zero fitted parameters: Mochi has the *lowest* locality drop in the cluster and full ability, so locality marks the failure mechanism and $\mathrm{CKA_c}$ makes the ability call. The second forecast was lost. LTX-Video, never probed here before, scores $\mathrm{CKA_c} = 0.826$, inside the cluster band. The forecast *able* was filed before a depth label went near the model, and the model passed one of the three rungs it commits to, $1/3$ (Appendix J); by the preregistration's rule $\mathrm{CKA_c}$ ranks within the pool it was built on and does not port. One miss belongs to the rung, LTX being lawful at $9/9$ blocks two steps cleaner on the noise schedule (§4.1): a smaller error than the score records, and still one, since a portable criterion must survive its protocol.

6.4 What the Cluster Shares, and the Gate It Fails

What the converged models share is image layout (Figure 3). Patch-level CCA finds 7–10 shared dimensions between a cluster member and the other members, and 0–1 between an outsider and anyone. Depth decodes from them as well as from the full features and not from the ten to thirteen whitened directions outside them, and the localization is causal under ablation (Appendix M). For the transfer test, whiten models A and B to 20 dimensions each on the training scenes (Eq. 11), let $\kappa_1 \geq \cdots \geq \kappa_{20}$ be the canonical correlations of their paired patches, and let U_A and U_B hold their first $\max(1, \nu_{AB})$ canonical coordinates. With $\hat{\beta}_A$ the ridge fit (Eq. 3) of U_A on the training scenes,

$$\nu_{AB} = \big|\{i : \kappa_i > 0.5\}\big|, \qquad T(A \to B) = R^2\Big(d^{\mathrm{te}},\ U_B^{\mathrm{te}} \hat{\beta}_A\Big), \\ T_{\mathrm{rel}} = \frac{\mathrm{mean}_{A \neq B}\ T(A \to B)}{\mathrm{mean}_B\ T(B \to B)}, \tag{10}$$

so every transfer is cross-model *and* scene-disjoint, and $T(B \to B)$, where all 20 directions align, is B's own ceiling. Over the six ordered pairs of V-JEPA2, CogVideoX-2B and Mochi-1 ($\nu_{AB} = 7$–9; Wan2.2's 8×8 grid does not align) the probe transfers at $T_{\mathrm{rel}} = 0.818/0.816$, 100% of the self-transfer ceiling, as model stitching (Lenc & Vedaldi, 2015; Bansal et al., 2021) makes precise. The bootstrap 95% interval on T_{rel} is $[0.966,\ 1.045]$, and cluster–outsider transfer sits at 0.004. The shared subspace decodes image position ($0.93/0.90$) better than depth (0.81), its depth content is low-frequency, and over the quadratic position basis it adds $+0.023$. Position-*residualized* depth decodes from the shared subspace in $0/5$ models, Wan2.2 and LTX-Video at their native 8×8 included, and the residualized probe transfers nowhere (-0.006; permutation control -0.31). The claim that the shared subspace *is* the geometry fails its gate and is rescoped to layout (Appendix S).

7 Limitations

Four predictions falsified, two corrections, one rescope. Four predictions were falsified by the controls and held-out forecasts run against them: a paradigm law, the appearance-bias reading of L5, the locality screen, and the claim that $\mathrm{CKA_c}$ ports past its pool. Two measurements were corrected: patch-random L1 splits leaked within-scene structure, so cross-model comparisons are scene-disjoint and the ladder's patch-random L1 column is labelled; and the generator count was read at a block chosen by naive decodability, so "zero of six" is the committed protocol's count and not the models' (§5.2). One surviving claim was rescoped by its own gate, the shared subspace being layout rather than geometry. Each, with what falsified and replaced it, is a row in Appendix S.

Six checkpoints cannot support a scaling claim, the regimes are read from a pool we assembled, and the mechanism behind the encoder–generator split stays open. Confirmatory numbers are KITTI driving scenes; the ladder lacks L2 for three of the six models and its L4 stimuli are synthetic; $\mathrm{CKA_c}$'s one out-of-sample test failed and it hands a practitioner no portable threshold; the VAE control covers one VAE, its round trip a lower bound; steering stays under 3σ across four protocols; and the earlier VideoMAE numbers predate a silent attention-bias loading defect (all in Appendix T).

8 Conclusion

Most of the geometry credited to video foundation models is a ground plane, and five further probe choices move what is left by up to a change of sign (Table 1). Control every choice but the block on real driving video and a small, real effect remains: all four encoders clear their floors and none of the six generators does at the block the L1 sweep picks. Select the block out of fold as well and the encoders hold while one generator crosses, 2.5 times below the weakest, so the classes separate by magnitude. What convergence selects is image layout. Four predictions of ours were falsified by their own controls and forecasts, and everything needed to make the choices differently is released.

AI use statement

In this work, we used generative AI tools for the following tasks with required disclosure: helping develop the conceptual framework (the six-choice audit and the falsification ladder); proposing and

refining hypotheses and candidate directions; designing and giving feedback on methodology and experiments, among them the position-residualized decoding gate, the cross-model probe-transfer protocol, the untrained floors and the NYU feature-health gate; implementing methods and operating the experiment runs (the extraction harness, the per-model adapters, the probe and analysis code); cleaning and reformatting data sets (the loaders that turn KITTI, NYU Depth v2 and ScanNet into patch grids); generating synthetic data sets, in the limited sense that the tools wrote the code that renders the synthetic L4 and L5 stimuli, while no data point was produced by a generative model itself; and interpreting results. We have not used generative AI tools for translation, and formulating mathematical claims, providing ingredients for proofs, writing proofs, and qualitative or thematic data analysis are not applicable to this work. Among the tasks whose disclosure is recommended rather than required, the tools drafted and edited sections of the text, which the authors revised; searched for and summarized related work; wrote and edited the figure scripts; and suggested section structure and titles.

We have reviewed all AI-assisted work. The governing quantity, the hypotheses and the preregistered decision rules were selected and locked by the authors before the corresponding data existed, and every later deviation is logged in the released preregistration. All experiments were executed on compute the authors control. Every number in the text, tables and figures was checked by the authors against the committed per-experiment result files, a released checking script re-derives the headline ones from those files, and every citation's metadata was fetched from DBLP, Semantic Scholar or the official proceedings page rather than written from memory. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

All probes are closed-form ridge on cached activations. Every number in the figures, the results tables and the calibration battery regenerates from a committed result JSON via the released scripts. Two exceptions are marked where they appear, and we list them here as well. The v1–v3 rows of Table 25 exist only in the run log (the v4 sweep is the committed one), and the ScanNet clean-timestep setting is fixed in the experiment script rather than recorded in its output JSON. All ten checkpoints are official public releases, and KITTI-2015, NYU Depth v2 and ScanNet are public (ScanNet under its terms of use: we release per-clip metrics, never frames). The anonymized archive submitted as supplementary material contains the extraction harness and per-model adapters (one activation hook each), every experiment script with its committed result JSON, the figure scripts, and a manifest naming the file behind each figure and table in this paper. It also holds the preregistration with its locked pass rules and kill criteria, the decision ledger with a memo for every experimental gate, the retraction ledger behind Appendix S, and the checking script with the manuscript source it checks. Two things are left out for size and are regenerated by the scripts that need them: the cached per-patch feature arrays, a gigabyte across the pool, which the extraction scripts rebuild from the public checkpoints, and the raw frames, which come from the public datasets. Every figure redraws from the committed JSONs alone, with two exceptions. The schematic of Figure 10 was drafted by an image-generation model from the released specification. Figure 1(a) and Figures 7 and 9 also read the cached features and the KITTI frames; the numbers their captions quote are committed as JSON and checked like every other.

References

Akshit Achara, Tatiana Gaintseva, Matéo Mahaut, Pritish Chakraborty, Viktor Stenby Johansson, Melih Barsbey, Emanuele Rodolà, and Donato Crisostomi. Multi-way representation alignment. In *International Conference on Machine Learning (ICML)*, 2026.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research (TMLR)*, 2024.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Jelena Bratulić, Sudhanshu Mittal, Thomas Brox, and Christian Rupprecht. On geometric understanding and learned priors in feed-forward 3D reconstruction models. *arXiv preprint arXiv:2512.11508*, 2025.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal M. P. Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott E. Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024.

Yida Chen, Fernanda Viégas, and Martin Wattenberg. Beyond surface statistics: Scene representations in a latent diffusion model. In *NeurIPS Workshop on Diffusion Models*, 2023.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning (ICML)*, pp. 2990–2999, 2016.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2443, 2017. doi: 10.1109/CVPR.2017.261.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Amil Dravid, Yossi Gandelsman, Alexei A. Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1934–1943, 2023.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas J. Guibas, Justin Johnson, and Varun Jampani. Probing the 3D awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21795–21806, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, pp. 12606–12633, 2024.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis (SCIA)*, Lecture Notes in Computer Science, pp. 363–370, 2003. doi: 10.1007/3-540-45103-X_50.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361, 2012. doi: 10.1109/CVPR.2012.6248074.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

Genmo Team. Mochi 1. `https://github.com/genmoai/mochi`, 2024.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. DepthCrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2005–2015, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024.

Zixuan Huang, Xiang Li, Zhaoyang Lv, and James M. Rehg. How much 3D do video foundation models encode? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 384–394, 2026.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, pp. 20617–20642, 2024.

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 *International Conference on Machine Learning (ICML)*, pp. 28991–29017, 2025.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 991–999, 2015.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research (TMLR)*, 2023.

Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3061–3070, 2015.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5732–5741, 2018.

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, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6076–6085, 2017.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12179–12188, 2021.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(3):1623–1637, 2022. doi: 10.1109/TPAMI.2020.3019967.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision (ECCV)*, pp. 746–760, 2012. doi: 10.1007/978-3-642-33715-4_54.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Team Wan, Ang Wang, Baole Ai, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research (TMLR)*, 2022.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3D representation for consistent world modeling. In *International Conference on Learning Representations (ICLR)*, 2026a.

Jiaxin Wu, Yihao Pi, Yinling Zhang, Yuheng Li, and Xueyan Zou. Quantitative video world model evaluation for geometric-consistency. *arXiv preprint arXiv:2605.15185*, 2026b.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations (ICLR)*, 2024.

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)*, 2025.

Symbol	Meaning
$\mathrm{CKA_c}$	mean linear CKA of a model's patch-averaged clip features to the convergence-cluster members other than itself (Eq. 2), on 40 unlabeled KITTI clips. One forward pass per clip, no labels and no training. The governing quantity.
$c(h)$, Λ	spatial autocorrelation of per-patch features at grid distance h (Eq. 13); the locality drop is $\Lambda = c(1) - c(6)$.
R^2_{pos}	depth variance explained by a quadratic image-position basis $[r, c, r^2, c^2, rc]$ alone. The size of the layout prior in a dataset.
Δ	$R^2(\text{depth} \mid \text{pos}, \text{flow}, Z) - R^2(\text{depth} \mid \text{pos}, \text{flow})$: what the features add over image position and optical flow together. The controlled structure-from-motion metric.
$\Delta_{\min}$	the smallest Δ across seeds, reported beside the mean so a positive verdict cannot rest on one lucky split.
Z	the frozen feature block being probed (top-50 PCA-whitened, Eq. 11).
ν_{AB}	number of CCA directions with canonical correlation above 0.5 between two models' features (Eq. 10): the shared-subspace dimension.
T_{rel}	mean cross-model transfer over mean self-transfer (Eq. 10).
$\mathcal{M}$	separation margin, weakest encoder minus strongest generator, over families or checkpoints (Eq. 6).
$g^\star$	g at the block chosen on separate splits (Eq. 8).
R^2_{resid}	R^2 of the depth that position leaves (Eq. 12).
z	L2 steering effect against random directions (Eq. 14).
L1–L5	the five preregistered ladder rungs: linear presence, causal use, lawfulness under a symmetry, compositional generalization, and behavior under cue conflict.
†	exploratory: run outside the ladder's preregistration.

Table 2: Symbols, fixed once and used throughout.

A Notation

B Estimands and test statistics

The main text defines the increment and its floor (Eq. 1), $\mathrm{CKA_c}$ (Eqs. 2 and 4), the probe (Eq. 3), the choice vector (Eq. 5), the family test (Eqs. 6 and 7), the block chosen on separate splits (Eq. 8), L3 lawfulness (Eq. 9) and the transfer ratio (Eq. 10). The remaining quantities are defined here, each as the released code computes it.

Whitening. With $F_m \in \mathbb{R}^{N \times k_m}$ a block's per-patch features, column-centred, and $F_m = U\Sigma V^\top$ its thin SVD, the features are projected on the first n_{w} principal directions and scaled to unit variance:

$$Z_m \;=\; F_m\, V_{:,1:n_{\mathrm{w}}}\, \mathrm{diag}\!\left(\frac{\Sigma_{11}}{\sqrt{N}}, \ldots, \frac{\Sigma_{n_{\mathrm{w}} n_{\mathrm{w}}}}{\sqrt{N}}\right)^{-1}. \tag{11}$$

For Δ, $n_{\mathrm{w}} = 50$ and the fit uses every patch of the scenes scored, without labels; for the CCA of Eq. 10, $n_{\mathrm{w}} = 20$ and the fit uses the training scenes only, its mean and directions then applied to the held-out scenes.

Residualized target. With $X_c = [P, \mathbf{1}]$ and $\hat{w}_c$ the ridge fit ($\lambda = 1$) of d on X_c over the training scenes,

$$R^2_{\text{resid}}(m) \;=\; R^2(d - X_c\hat{w}_c \mid Z_m)\,, \tag{12}$$

scored on the held-out scenes by Eq. 3, the depth of both splits residualized with the same $\hat{w}_c$. Its denominator is the variance position leaves, where the denominator of Δ is the variance of d.

Spatial locality. With $\hat{z}_i(x)$ the unit-normalized feature of patch i in clip x at the block of best depth decoding, p_i its grid cell and $\mathcal{P}_h = \{(i, i') : \|p_i - p_{i'}\|_1 = h\}$ the ordered patch pairs at grid distance h,

$$c(h) \;=\; \frac{1}{S\,|\mathcal{P}_h|} \sum_{x} \sum_{(i,i') \in \mathcal{P}_h} \hat{z}_i(x)^\top \hat{z}_{i'}(x), \qquad \Lambda \;=\; c(1) - c(6), \tag{13}$$

over the $S = 40$ clips of $\mathrm{CKA_c}$; Λ is the locality drop of §6.3.

Causal use (L2). With w the unit direction of the depth probe at the steered block and $\bar{D}(\alpha; u)$ the mean depth the read-out block's probe decodes after the steered block's features are pushed along the unit direction u by α standard deviations of their projection on w,

$$\eta(u) \;=\; \frac{\bar{D}(+2; u) - \bar{D}(-2; u)}{4}, \qquad z \;=\; \frac{\eta(w) - \mathrm{mean}_r\ \eta(u_r)}{\mathrm{sd}_r\ \eta(u_r)}, \tag{14}$$

over random unit directions u_r, 20 for V-JEPA2 and 10 for each generator. z is the L2 column of Figure 2b.

Swing. The swing column of Table 1 is the ratio of the uncontrolled to the controlled value, or a sign flip where the two differ in sign.

C Model and probe specification

Model	Architecture	Params	Feature source	Grid	$\mathrm{CKA_c}$ (role)
V-JEPA2	encoder (ViT-L)	0.4B	encoder block 16	16×16	0.80 (cluster)
Wan2.2-TI2V	diffusion DiT	5B	DiT block 0	8×8	0.87 (cluster)
CogVideoX-2B	DiT (expert-adaLN)	2B	DiT block 14	16×16	0.81 (cluster)
Mochi-1	DiT (AsymmDiT)	10B	DiT block 18	16×16	0.84 (cluster)
HunyuanVideo	DiT (MMDiT)	13B	best of sweep (flat)	16×16	0.57 (outsider)
CogVideoX-5B	DiT (expert-adaLN)	5B	all 9 layers probed	16×16	0.55 (outsider)
VideoMAE	masked-video encoder	0.09B	encoder block 7	$14 \to 16$	—
VideoMAE-large	masked-video encoder	0.3B	encoder block 18	$14 \to 16$	—
V-JEPA2-ViT-H	encoder (ViT-H)	0.6B	encoder block 20	16×16	—
LTX-Video	diffusion DiT	2B	DiT block 7	8×8	0.83 (held out)

Table 3: The ten checkpoints, the block each probe reads (fixed by the L1 sweep; diffusion features from in-distribution noised latents at a fixed timestep), the per-patch grid, and the governing quantity $\mathrm{CKA_c}$ (mean linear CKA to the other cluster members, computed from the committed CKA matrix). CogVideoX-2B and -5B share one architecture and differ only in scale — the pair behind the scale paradox (§6.1). The four checkpoints below the rule enter only in §4, §5 and, for LTX-Video, the held-out forecast of §6.3: they are not in the convergence pool and are not counted in the n=6 statistics; the three encoders carry no $\mathrm{CKA_c}$, and LTX-Video's was scored after the pool was locked (Appendix J). Ten checkpoints are measured in total.

What constrains the rest of the paper Two rows carry constraints the later sections inherit. Wan2.2 and LTX-Video are the two models on an 8×8 grid, so they are absent from the patch-level analyses that need matching patches: the CCA subspace and the transfer matrix. Wan2.2 is also the model with the highest $\mathrm{CKA_c}$ and the strongest causal mediation, so for most of this project the checkpoint our convergence story ranked first was the one the geometry story could not test. The controlled probes turned out not to need matching patches, and both are measured in them at their native grid (Appendix J); what the exclusion was hiding is the sixth choice. The four checkpoints below the rule were added after the ladder closed: the three encoders to ask whether the encoder result depends on V-JEPA2 in particular, and LTX-Video as the held-out forecast (§6.3). The encoders carry no $\mathrm{CKA_c}$ because $\mathrm{CKA_c}$ is defined against the six-model pool, and none of the four is counted in any n=6 statistic.

D The ladder, per model and per rung

L5, layer by layer. The texture control, the only changed variable, falsified the appearance-bound reading of L5: with a trackable parallax cue three of the four cluster models follow geometry at every probed layer (V-JEPA2 $R^2 = 0.93$ at shading correlation 0.21), Hunyuan gains a fifth as much, Cog5B has no signal either way (Fig. 11), and on real KITTI footage the V-JEPA2 oracle tracks real geometry at 0.951 against a painted shading field (Fig. 12). Mochi does not follow the pattern, gaining at one layer of seven and losing at four.

The ladder rungs in full. *L1*, linear presence, discriminates nothing: depth decodes in five of six models, an untrained V-JEPA2 reaches 0.758, and no ordering here predicts anything later. *L2*, causal use, splits: steering the depth direction shifts the decoded value monotonically (V-JEPA2 $z = 3.4$, Wan $z = 4.4$, with $0/20$ and $0/10$ random-direction exceedances), while the same push through the full denoise loop never moves the *generated* video past 3σ under four protocols (Table 25), so the causal geometric code is internal and distributed. *L4*, composition, degrades by ≤ 0 wherever a read-out exists, and the outsiders have no base to generalize from.

Model	L1 R^2 (layer)	L2 z	L3 resid.	L4 ID→OOD	L5 R^2
V-JEPA2	0.87 (L16) / 0.90*	3.4	0.30 [0.30, 0.31]	0.93 → 0.93	0.93
Wan2.2	0.82 (L0) / 0.85*	4.4	0.68 (0.55 DiT)	0.71 → 0.73	0.72
CogVideoX-2B	0.86 (L14)	−0.9	0.805 (0.80 DiT)	0.38 → 0.51	0.47
Mochi-1	0.855 (L18)	—	0.475	0.73 → 0.73	0.645
HunyuanVideo	0.75 (flat)	—	0.985	0.12 → 0.13	0.12
CogVideoX-5B	0.03 (< 0 at 8/9)	—	0.986	−0.22 → −0.18	floor
LTX-Video (held out)	0.835 (L7)	—	0.993 (0.022 clean)	−0.04 → +0.09	—
controls	−0.16 to −0.01	0/20, 0/10	1.0 chance	ratio 1.01	—

Table 4: Every rung, every model. L1 is the patch-random sweep the ladder was filed with; scene-disjoint splits on the 40 cross-model scenes read 0.79–0.83 for the three cluster models that have them (Table 21), and no L1 verdict changes. * the 40-scene motion-clip protocol used by the cross-model analyses; the two protocols agree on every verdict. L2 (—) is unrun for Mochi and undefined for the outsiders, which lack the L1 direction the intervention steers. L3 is the residual over the random-permutation ceiling (lower = lawful); "DiT-only" applies the flip in latent space, bypassing the VAE. Every L3 is read at schedule index 350; §4.1 shows the verdict depends on that choice. LTX-Video, the held-out model of §6.3, is listed for comparison and is not in the n=6 pool; its L3 is given at the committed index and at a clean latent, and its L4 in-distribution value is negative, so the probe does not fit. L4 reports the compositional split (2–3 bumps → 5–6 bumps) at each model's best layer; MAE and normalized MAE agree with R^2 throughout. L5 is motion-geometry R^2 in the textured condition.

What the columns do and do not separate Read the columns before the rows. L1 separates almost nothing: five of six models clear it, and an untrained V-JEPA2 reaches 0.758 on the same probe (Appendix F), so linear decodability of depth is close to free. The separation appears at L3 and L4, where the outsiders sit at chance and the cluster spreads over a wide model-specific range. Mochi's row is the one that falsified our preregistered screen: it has the lowest locality drop in the cluster and it clears every rung it was run on. The dashes are not missing data in the ordinary sense. L2 steers the L1 direction, so a model that has no L1 direction cannot be given an L2 score at all.

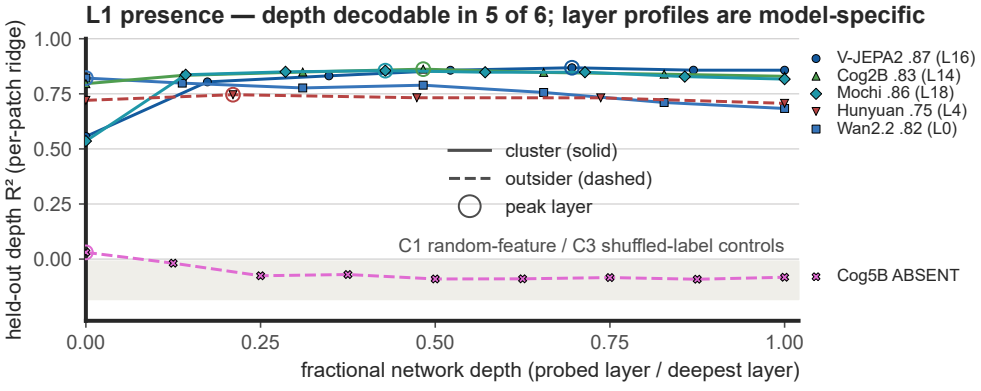


Figure 4: L1 per-layer depth R^2 on fractional network depth; controls band at bottom; Cog5B ABSENT at every layer.

What the curves show that the table cannot The L1 column of Table 4 is one number per model. These are the profiles behind those numbers, and they have no common shape. V-JEPA2 climbs from 0.798 at block 6 to 0.904 at block 18 and then flattens; the DiTs peak in the middle and fall away; CogVideoX-5B never leaves the control band at any depth. The band itself is why the rest of the figure can be trusted: the random-feature and shuffled-label controls sit at -0.041 and -0.051 and stay flat across the network, so no curve here is a probe fitting noise. The practical consequence is the one the main text records as a limitation. Because the profiles are model-specific, the layer each probe reads has to be fixed by a labelled sweep, and scoring a new checkpoint with $\mathrm{CKA_c}$ needs that sweep or a heuristic to stand in for it.

E THE LADDER: PREREGISTERED RULES, AND EVERY DEVIATION FROM THEM

The pass rules below are quoted from the locked preregistration. The third column is what was actually run. A paper that reports only the second column is reporting a plan; a paper that reports only the third is reporting a result with its rule chosen afterwards.

The shape of the deviations The deviations have a direction worth naming. Every narrowing removed a target, a residual or a stimulus family; none moved a threshold or a pass rule. That asymmetry is deliberate, because thresholds are the part a reader cannot check after the fact, while a narrowed scope is visible in the results table. The cost is that three rungs measure less than their names promise. L3 tests reflection equivariance where the plan said lawfulness without qualification, L5 covers one illusion family of six, and L1 covers depth alone of four geometric targets. Claims in the main text are scoped to the executed column, not the preregistered one.

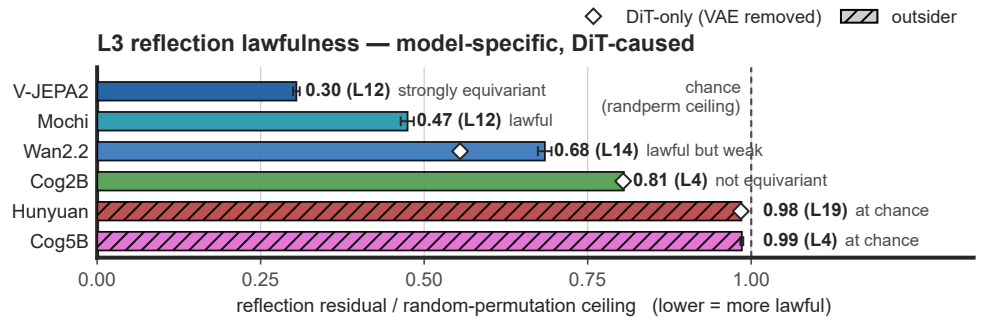


Figure 5: L3 reflection lawfulness with bootstrap CIs; diamonds: VAE bypassed (DiT-only): lawfulness survives, and improves for Wan.

Reading the lawfulness metric, and the control that keeps it honest The quantity plotted is ρ of Eq. 9, so 1.0 is chance and lower is more lawful, and a layer counts as lawful only when the bootstrap upper bound on ρ falls below 1 *and* the equivariance residual is smaller than the invariance residual $\delta(F(\sigma x), F(x))$: the first says the match is mirror-specific rather than generic smoothness, the second that the model is not simply ignoring the flip. Two models sit at chance (≥ 0.985) and cannot be ranked at all; the rest spread from V-JEPA2 at 0.30 to CogVideoX-2B at 0.805. The diamonds are the control that matters for the L3 row of Table 5. Applying the flip in latent space with the VAE bypassed leaves Cog2B unchanged (0.804 against 0.805) and makes Wan *more* lawful (0.55 against 0.68), so the effect belongs to the transformer and not to the autoencoder wrapped around it. What the figure does not show is equivariance in the strong sense. None of these architectures has reflection symmetry built in, and the metric asks only whether a flipped input lands closer to the equivariant prediction than a permutation does.

F CALIBRATION BATTERY, CHECK BY CHECK

The two failure modes, measured. Patch-feature cosine similarity against Manhattan distance separates three regimes (Fig. 6b): structured decay, decorrelated flat (Hunyuan) and over-smooth

Rung	Preregistered rule	Executed, and how it differs
L1	Presence. Linear probe $R^2 > 0.5$ at some layer AND above the random-feature and untrained-init controls, bootstrap CI excluding the control, Benjamini–Hochberg across layers and models. Targets: per-patch depth, camera pose, occupancy, surface normal.	Ridge on per-patch depth, KITTI disparity ground truth, 200 scenes. Of the four preregistered targets only depth is carried through the ladder; the others did not go past pilot runs. The split was tightened mid-project from patch-random to scene-disjoint after we measured $\approx +0.07$ of within-scene leakage, which lowered cluster values from 0.85–0.90 to 0.78–0.83. Every cross-model comparison uses the scene-disjoint numbers; the L1 column of Table 4 is the patch-random sweep and is labelled so (Appendix S).
L2	Causality. Steering along the L1 direction changes decoded geometry, with the effect's CI excluding a matched random-direction control.	Two arms. Internal: steer at one layer, read depth at a later one (z against 10–20 random directions). Behavioral: push the same direction through the full denoise loop, decode to RGB, and probe the generated video with an independent oracle. The behavioral arm is stricter than the preregistered rule and fails for both generators it can be run on (Appendix O); we report both. Mochi's L2 is unrun, and the outsiders have no L2 because it needs the L1 direction they lack.
L3	Lawfulness. Four residuals — multi-view consistency, epipolar Sampson, rigid-motion equivariance, occlusion ordering — each significantly below the controls, with the equivariance residual growing sub-linearly in transform magnitude.	One of the four: reflection equivariance, a symmetry built into none of these architectures. Run with the VAE in the loop and again in latent space with the VAE bypassed, because a lawfulness result that lives in the autoencoder is not a result about the generator.
L4	Compositional generalization. Read, not pass/fail: relative degradation of the L1 probe from in-distribution to novel geometric configurations, with effect size and CI.	Trained on 2–3-bump arrangements, tested on unseen 5–6-bump ones. Reported as R^2 ID→OOD. Two controls the plan did not ask for: the compositional-split target-variance ratio (1.01, so the generalization is not a variance artifact) and agreement with MAE and normalized MAE.
L5	Geometry-versus-appearance conflict. A geometry-following index over six illusion families (textured-flat, mirror, Ames room, impossible object, forced perspective, trompe-l'oeil), each with a matched real-3D control.	One family, the textured-flat case, as motion-parallax stimuli with shading decoupled from geometry; the index is replaced by the motion-geometry R^2 and the correlation with the shading cue. The texture control (only the texture changes) falsified the reading we filed for this rung. Validated once on real KITTI footage with painted shading, for V-JEPA2 only.

Table 5: Every rung, its locked rule, and the distance between the rule and the run. Three of the five were narrowed (L1's target set, L3's residual set, L5's stimulus families) and two were tightened beyond the plan (L2's behavioral arm, L4's variance controls). The narrowings are why the paper's class-level claims are scoped to what was measured rather than to the rung's name.

flat (Cog5B; Mochi, which nonetheless converges and passes). Every model beats a per-sample spatial-shuffle null at $p < 0.001$, but the margins rank them, Cog5B's $z = 7.4$ against Cog2B's $z = 294$. Both failure modes destroy spatial discrimination and are no more similar to each other than to the cluster, so "failing" is not one condition.

What the two partials cost The eleven passes are not where this table earns its place. A reader who counts them concludes the cluster is a discovered object; what the data support is weaker and more specific. The algorithmic clustering is not significant against its own permutation null, and the four validity criteria disagree about how many clusters there are, with the gap statistic preferring k=2 and putting Hunyuan inside. So the partition is one we drew, which then survived exact enumeration, leave-one-out, independent halves, pseudo-models and a second kernel. That is a defensible object at n=6 and it is not a discovered cluster. The two informative rows carry a separate warning: an

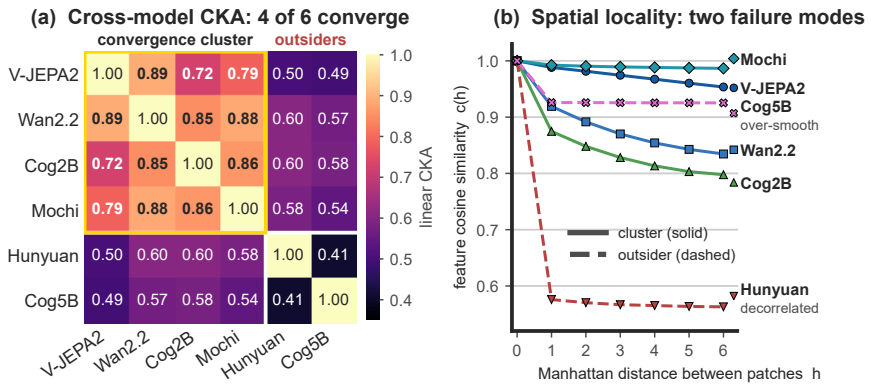


Figure 6: **Two occupied regimes, on two instruments. (a)** Cross-model linear CKA on 40 KITTI scenes: four of six models form a convergence cluster (outlined); the two outsiders sit below every cluster pair and lowest of all to each other. **(b)** Spatial autocorrelation $c(h)$ of per-patch features at grid distance h (Eq. 13) gives three regimes: structured decay, decorrelated-flat (Hunyuan, 0.576) and over-smooth-flat (Cog5B 0.926; Mochi 0.992, which converges and passes the ladder anyway).

untrained encoder reaches L1 0.76, so any paper whose evidence is a linear depth probe has measured something an architecture prior already provides.

G THE CHOICES, MEASURED

Each subsection isolates one choice in the probe with everything else held fixed. Five are here; the sixth, patch resolution, is in Appendix J, where the model it was hiding is also measured. All of it is exploratory in the sense of §4: run under the audit plan after the preregistered ladder closed, outside the ladder's preregistration.

G.1 IMAGE POSITION

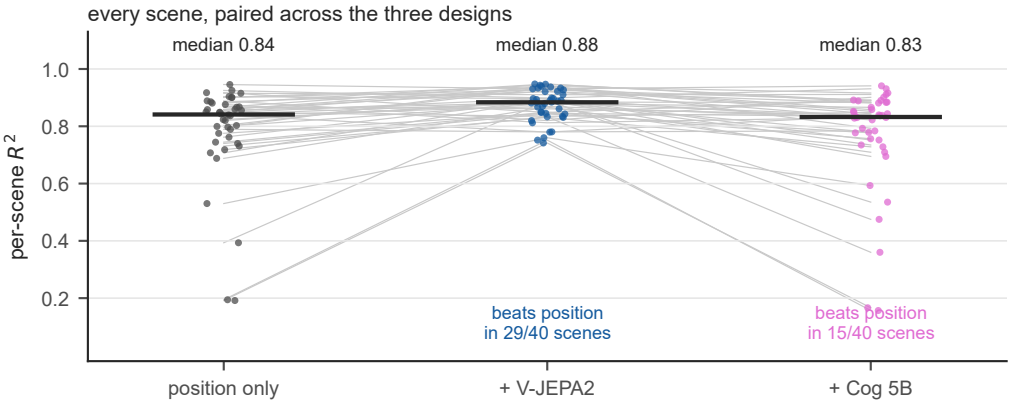


Figure 7: **Every scene, paired across the three designs.** All 40 released scenes, one line each, on the same no-flow-control setting as Figure 1. The encoder beats position in $29/40$ scenes and the generator in $15/40$. Eleven are scenes where the encoder is *worse*, which a pooled mean would hide.

What to take from it Two things in this table are separable. The first is that CogVideoX-2B reads 0.782 naive on KITTI and -0.098 residualized: not a small positive, a negative, meaning its features predict depth worse than the position prior alone once that prior is removed. The second is the gap between the two stimulus sets. A probe run on driving footage has 16% of the depth variance

Check	Objection it answers	Result	Verdict
Negative controls	shuffled and random features must not show locality	real/shuffled locality ratio $43\times$–$846\times$	pass
Permutation test	locality could be a spatial-sampling artifact	1000 permutations per model, $z = 7$–294, all $p < 0.001$	pass
Cross-metric rank	the ordering could be metric-specific	Spearman $\rho = 0.6$ to 0.9 against L1	pass
Cluster validity	the partition could be one we drew	hand partition silhouette 0.325, $p = 0.073$; algorithmic best-k null $p = 0.51$	partial
Multi-criterion k	the number of clusters could be assumed	silhouette and Davies–Bouldin choose k=3, Calinski–Harabasz k=4, gap k=2	partial
Held-out halves	the result could be one sample	two independent 20-scene halves: $p = 0.071$ and 0.029	pass
Leave-one-out	one model could carry the partition	silhouette positive in 6/6 folds, mean 0.342	pass
Pseudo-model control	any two representations cluster if you let them	adding random-Gaussian and channel-shuffled pseudo-models, n=8, $p = 0.006$; random Gaussian ranks last	pass
Exact enumeration	the split could be one of many equally good ones	all ten n=5 labelings: ours ranks 1/10 on the summed separation and 1/10 on the minimum; per-rung pairwise dominance 6/6	pass
Bootstrap CIs	point estimates could be noise	every locality drop's 95% CI excludes zero (6/6)	pass
Kernel robustness	linear CKA could be the cluster	RBF CKA agrees element-wise at $r = 0.983$; RBF partition $p = 0.057$	pass
Untrained encoder	an architecture prior could decode depth	random-init V-JEPA2 reaches L1 0.758 and fails every rung above	informative
Untrained DiTs	the same, for generators	Cog5B untrained $0.031 \approx$ trained; Wan untrained 0.48 vs trained 0.85	informative
Random-feature probe	the probe could fit anything	Gaussian features at 256/1024/3072 dims: test R^2 -0.009 to -0.146, ABSENT at every width	pass
External anchor	our probe could be weak or strong in absolute terms	DPT-Large on the same grids: affine $R^2 = 0.829$, Spearman 0.935	pass

Table 6: Fifteen checks, each written against a specific objection rather than as a generic robustness sweep. Eleven pass outright. Two are marked partial and are the honest limits of n=6: the algorithmic clustering is not significant against its own permutation null ($p = 0.51$), so the partition stands as a hand partition with a $p = 0.073$ silhouette; and the four cluster-validity criteria do not agree on k, with the gap statistic preferring k=2, which groups Hunyuan with the cluster. Two are marked informative because they are controls, not tests. An untrained encoder already reaches L1 0.76, which is why L1 discriminates nothing and the rungs above it carry the paper.

Stimulus	R^2_{pos}	var. left	Model	naive R^2	position-residualized
KITTI	0.844	0.156	V-JEPA2	0.848	$+0.122$
			CogVideoX-2B	0.782	-0.098
			Mochi-1	0.843	$+0.016$
motion	0.330	0.670	V-JEPA2	0.673	$+0.476$
			CogVideoX-2B	0.288	-0.050
			Mochi-1	0.358	-0.043

Table 7: What a quadratic image-position basis already explains, and what survives removing it. On KITTI the prior takes $R^2 = 0.844$ of the depth variance, so a naive probe reporting 0.848 (V-JEPA2) is reporting little else, and the same features residualized give $+0.122$. On motion-parallax stimuli the prior takes 0.330, leaving 67% of the variance, and V-JEPA2 reaches $+0.476$. How much room the control leaves is a property of the dataset a probe runs on, not of the model it probes.

left to explain after position; on motion-parallax stimuli it has 67%. Neither number is about the models. Both decide what a probe run on that data can detect, and only the second leaves enough room to separate models cleanly. That is why the sharp dissociation lives on synthetic motion and the real-video effects in Appendix H are an order of magnitude smaller.

G.2 OPTICAL FLOW

Model	position control only	+ flow control	shrinkage
V-JEPA2	$+0.294$	$+0.199$	-0.095
VideoMAE	$+0.534$	$+0.188$	-0.347
CogVideoX-2B	$+0.019$	$+0.007$	-0.012

Table 8: Adding the flow control to the position control on the same variable-gain motion-parallax stimuli, 6 seeds each, both columns Δ (Eq. 1). VideoMAE loses 0.347 of its 0.534, a factor of $2.8\times$, and V-JEPA2 0.095 of its 0.294, $1.5\times$: that part of what survived position control was recoverable from the flow magnitude alone. The VideoMAE row predates the loading repair of Appendix T. The flow control is what separates a model that has factored camera motion out of the scene from one that has encoded the motion itself.

What to take from it The shrinkage column is the price of the control, and it is charged where the uncontrolled result was largest. VideoMAE loses two thirds of its increment and V-JEPA2 a third; CogVideoX-2B, near zero already, loses 0.012. That asymmetry is the reason we treat the flow control as necessary rather than conservative: without it the encoders' increments would read 1.5 to 2.8 times larger than they are. Mochi-1, HunyuanVideo and CogVideoX-5B were not run on these stimuli; their position-residualized R^2 on the constant-gain stimuli is already inside the seed noise (-0.007, -0.025, $+0.005$), and the flow control lowered every increment it was run on.

G.3 DIFFUSION NOISE TIMESTEP

Latent read at	appearance	optical flow	**depth**
no noise added	$+0.366$	$+0.669$	$\mathbf{+0.326}$
index 999 (least noise)	$+0.320$	$+0.667$	$\mathbf{+0.322}$
index 700	$+0.005$	$+0.384$	$\mathbf{+0.172}$
index 350 **(our setting)**	-0.003	$+0.123$	$\mathbf{+0.025}$
index 100 (most noise)	-0.006	-0.063	$\mathbf{-0.068}$

Table 9: CogVideoX-2B at DiT layer 14 on the synthetic motion-parallax stimuli: one model, one probe, five points on the noise schedule. Indices address the descending 1000-step schedule, so a higher index means less noise; index 350 is the fixed in-distribution setting every ladder measurement in this paper uses. The controlled depth increment moves from $+0.326$ to $+0.025$, a $13\times$ swing, and our operating point sits near the bottom of it. Our own control recorded that verdict as false.

What to take from it Read the appearance and flow columns before the depth one. At the noisiest setting all three collapse together (appearance -0.006, flow -0.063, depth -0.068), which is a pipeline that has stopped delivering content rather than a model without geometry. At our operating setting the pipeline is still delivering: flow decodes at $+0.123$ while depth sits at $+0.025$. That combination is what makes the reading interpretable, and it is what our original diffusion-zero claim failed to check. A negative is only a negative when the same extraction still recovers something else.

The probe that reports the verdict cannot see this choice. Swept over the same timestep axis and over layers, an *uncontrolled* depth probe on 200 KITTI scenes barely moves. Wan2.2 spans 0.683–0.825 across 28 timestep$\times$layer cells with a widest row of $1.20\times$, and CogVideoX-2B spans 0.753–0.862 with a widest row of $1.08\times$. Against the $13\times$ the controlled metric moves over that same axis, the naive probe is flat. A choice that changes the answer by an order of magnitude leaves no trace in the measurement most papers report, which is a sufficient explanation for why the choice goes undocumented.

The layout verdict does not depend on it. Re-running the residual gate (§6.4) with CogVideoX-2B's features swapped for clean-latent ones and everything else identical moves R^2_{shared} by at most 0.041: V-JEPA2 $+0.071 \rightarrow +0.041$; CogVideoX-2B $-0.013 \rightarrow +0.027$; Mochi-1 $+0.113 \rightarrow +0.087$. The pass count is $0/3$ under both timesteps, so the sharpest objection to §6.4, that its layout finding is an artifact of reading diffusion features under noise, is answered by measurement rather than by argument.

G.4 NETWORK LAYER

Model (clean latent)	swept layer, and Δ beneath it						max
CogVideoX-2B	L4	L9	L14	L19	L24	L29	
Δ	$+0.003$	$+0.003$	$+0.014$	$+0.057$	$+0.147$	$+0.146$	$+0.147$
HunyuanVideo	L2	L5	L8	L11	L14	L18	
Δ	$+0.015$	$+0.016$	$+0.013$	$+0.013$	$+0.016$	$+0.020$	$+0.020$
Mochi-1	L6	L14	L22	L30	L38	L46	
Δ	$+0.002$	$+0.001$	$+0.000$	$+0.000$	$+0.001$	$+0.004$	$+0.004$

Table 10: The controlled metric swept over depth in three diffusion generators on the synthetic motion-parallax stimuli, all read at a clean latent so the timestep choice is closed first. CogVideoX-2B moves $55\times$ across its own layers, from $+0.003$ at L9 to $+0.147$ at L24, and the L14 our ladder fixes reads $+0.014$, near the bottom of its own range. Hunyuan and Mochi stay flat and near zero everywhere, so their negatives are not layer artifacts. The like-for-like encoder reference on these same stimuli is V-JEPA2's $+0.199$ (Table 8), which no layer of any of the three reaches. These values are not comparable with Table 12: that table is 200 real KITTI scenes, where the position control leaves 28% of depth variance, against 68% here, so the two stimulus sets put the same metric on different scales.

What to take from it The three rows answer different questions and only one of them is about our own choices. CogVideoX-2B's spread makes the layer a real degree of freedom: a paper reporting its best layer and a paper reporting its middle layer would disagree by an order of magnitude on the same checkpoint, and both would be reporting honestly. Hunyuan and Mochi are the stronger result. Their negatives hold at every probed layer, so those two cannot be rescued by layer selection, and the generator side of the dissociation does not rest on where we happened to read. What no row shows is a generator reaching encoder range: the largest value anywhere in the sweep is 0.147, against 0.199 for an encoder on the same stimuli.

G.5 TEMPORAL WINDOW

Stride	target frame	span (s)	naive R^2	Δ	$\Delta_{\min}$	seeds>0
4	first	8.53	-0.018	$\mathbf{-0.082}$	-0.156	$1/6$
2 (baseline)	first	4.27	$+0.013$	$\mathbf{-0.054}$	-0.214	$2/6$
2	centred	4.27	$+0.206$	$\mathbf{+0.139}$	-0.022	$5/6$
1	first	2.13	$+0.210$	$\mathbf{+0.136}$	$+0.048$	$6/6$
1	centred	2.13	$+0.365$	$\mathbf{+0.295}$	$+0.203$	$6/6$

Table 11: V-JEPA2 on ScanNet: 104 clips over 13 rooms, splits grouped by room, 6 seeds. Only the temporal window changes; the model, layer, probe, data and target are the configuration of §4.1. V-JEPA2 reads 64 frames, so stride sets how much camera motion the window spans while the depth target describes a single frame. The verdict moves from -0.082 (1/6 seeds) at 8.5 s to $+0.295$ (6/6, minimum $+0.203$) at 2.1 s with the target centred, which is above VideoMAE-large's $+0.236$ on the same rooms. Centring alone lifts the baseline configuration from -0.054 to $+0.139$.

What to take from it Two variables move here and they are separable. Stride sets how much camera motion the window spans, and the increment rises monotonically as the span shrinks. Centring the target is a second, independent effect: at fixed stride 2 it moves the verdict from -0.054 to $+0.139$, which is already a sign change without touching the span. Both point at the same mismatch. The probe integrates over an interval, the target describes one instant in it, and the further the target sits from the interval's centre the more of the model's representation is about geometry the target does not describe. Nothing here is specific to V-JEPA2 except the 64-frame window that makes the mismatch large enough to flip a sign.

Why this one cost us a false negative. The other four choices were closed by controls we had planned. This one arrived as a failed replication: VideoMAE cleared the indoor probe and V-JEPA2 did not, in the same run, on the same rooms. We recorded the failure, named the window as the

untested explanation, and ran the sweep, which overturned our own negative. The general form is the temporal extent a probe integrates over relative to the frame its target describes, and we have found it unreported in the probing literature we build on. Whether it moves a diffusion verdict is no longer an argument. Sweeping the same window on CogVideoX-2B over the same ScanNet clips, with 13 frames at strides 1 to 4, spans 0.43 to 1.73 s: read the way we read V-JEPA2, with the target at the window start, it is flat at -0.0124 to $+0.0122$, so the mismatch that flipped V-JEPA2's sign does not arise at 13 frames. Centring the target is a different matter: it rises monotonically with span, $+0.0087$ to $+0.0773$, clear of the uncentred band from stride 2 up, over an untrained floor of -0.0059. The window is therefore not inert for a generator. Its floor-subtracted best, $+0.0832$, stays under the weakest encoder's ScanNet value ($+0.1212$), so the count in Table 12 is unchanged, and the 200-scene protocol already centres its target. Preregistered in Section G with this decision rule fixed before the data; exploratory in the Section A sense.

H THE CONTROLLED VERDICT, EVERY MODEL AND EVERY DATASET

The split, tested as a split. §5.1 gives the family test and why its p is the floor of the design. Counting checkpoints instead of architectures gives $1/210$, $p = 0.005$, and we report the family number because checkpoints of one architecture are not exchangeable. Here is every cell behind the range and the count.

Family	Model	Δ trained	floor (untrained)	gap	gap 95% CI	> 0?	gap, real flow†
encoder	V-JEPA2	$+0.076$	-0.042	$\mathbf{+0.119}$	$[+0.035, +0.384]$	yes	$+0.111$
encoder	V-JEPA2-ViT-H	$+0.063$	-0.036	$\mathbf{+0.099}$	$[+0.026, +0.337]$	yes	n/r
encoder	VideoMAE-large	$+0.096$	$+0.028$	$\mathbf{+0.069}$	$[+0.037, +0.100]$	yes	$+0.061$
encoder	VideoMAE	$+0.068$	$+0.022$	$\mathbf{+0.046}$	$[+0.023, +0.069]$	yes	$+0.037$
generator	CogVideoX-2B	$+0.030$	$+0.019$	$\mathbf{+0.011}$	$[-0.004, +0.027]$	no	$+0.013$
generator	Mochi-1	$+0.018$	$+0.007$	$\mathbf{+0.011}$	$[-0.001, +0.027]$	no	$+0.012$
generator	CogVideoX-5B	$+0.017$	$+0.013$	$\mathbf{+0.004}$	$[-0.008, +0.015]$	no	$+0.005$
generator	HunyuanVideo	$+0.012$	$+0.018$	$\mathbf{-0.005}$	$[-0.020, +0.004]$	no	-0.005

Table 12: Real KITTI driving video, 200 scenes, every choice controlled but the block, which the L1 sweep fixes: position and flow regressed out, every generator read at a clean latent, a fixed 16×16 grid, the depth target centred in the 21-frame window (the setting §4.1 selects), bootstrap over scenes. Floor and gap are $\Delta(m_0)$ and $g(m)$ of Eq. 1, the floor run through the identical pipeline. Four of four encoders clear zero and zero of four generators do; Wan2.2 and LTX-Video (8×8) are in Table 17, six-seed point estimates inside the generators' range. The floor cuts both ways: CogVideoX-2B's raw $+0.030$ sits on an untrained $+0.019$, while V-JEPA2 and ViT-H have negative floors, so with floors clipped at zero the encoder range is $+0.046$ to $+0.076$ rather than $+0.119$ at the top; the generators and the verdict do not move. †Exploratory: the same features re-read with a dense optical-flow field in place of the frame-difference proxy (§5.3); ViT-H has no untrained builder and is not re-read.

What to take from it Read the floor column first. Every generator's untrained floor sits within 0.011 of its trained value, so its raw increment is not evidence of anything learned, and the largest raw increment (CogVideoX-2B at $+0.030$) is the clearest case: its untrained twin returns $+0.019$ of the same. The encoders' floors are also a warning about method. V-JEPA2's floor is -0.042 with an interval spanning 0.36, which is unstable, and its gap interval inherits that width ($[+0.035, +0.384]$). We report the gap because it is the quantity a learning claim needs, but the two VideoMAE rows, whose floors are small and positive, carry the encoder result more safely than the two V-JEPA2 rows do.

What to take from it Column four is the point of the table: position and optical flow alone reach 0.755–0.762 on this data without any model features at all. Every model in the table is being asked to improve on that. The last column is small for all of them because the baseline is already high, and the naive column misleads for the same reason – though not in the way we first assumed. We expected the control to reorder the models and it mostly does not: rank correlation between the naive column and the controlled one is $+0.857$, and only three of twenty-one pairs reverse. What the control changes is the magnitude. The naive column spreads the seven models over 0.453 and reads every one of

Model	layer	naive R^2	$R^2_{\text{pos,flow}}$	$R^2_{\text{pos,flow},z}$	Δ
V-JEPA2	16	$+0.828$	0.755	0.847	$+0.091$
VideoMAE-large	18	$+0.842$	0.761	0.844	$+0.083$
VideoMAE	7	$+0.815$	0.761	0.817	$+0.056$
CogVideoX-2B	24	$+0.681$	0.762	0.785	$+0.024$
CogVideoX-5B	20	$+0.449$	0.762	0.776	$+0.019$
Mochi-1	24	$+0.760$	0.761	0.772	$+0.015$
HunyuanVideo	4	$+0.389$	0.761	0.767	$+0.011$

Table 13: The same KITTI protocol on 60 scenes, decomposed. (This is the 60-scene run; the bootstrap table above is the 200-scene one.) Column three is what an uncontrolled probe would report; column four is what image position and optical flow explain without any features at all; the last column is the difference the features make. The ordering by naive R^2 and the ordering by Δ disagree: Mochi-1 is fourth on the naive number (0.760) and near the floor once controlled ($+0.015$), while V-JEPA2 keeps $+0.091$.

them as carrying depth; the controlled column puts four of the seven within 0.012 of each other, and with the untrained floor subtracted three of seven clear zero. Hunyuan's naive 0.389 is below what position and flow reach alone, so its features subtract. The ordering was never the problem. The two orderings disagree concretely: Mochi-1 ranks fourth by naive R^2 and second-to-last by increment, the largest fall in the table. A paper that reports the naive column is ranking models by how well their features happen to align with a ground-plane prior.

Model	naive R^2	Δ	$\Delta_{\min}$	seeds> 0	untrained floor
VideoMAE-large	$+0.318$	$\mathbf{+0.236}$	$+0.201$	$6/6$	-0.145
VideoMAE	$+0.205$	$\mathbf{+0.121}$	-0.011	$5/6$	-0.150
CogVideoX-2B	$+0.082$	$\mathbf{+0.012}$	-0.141	$5/6$	-0.014
V-JEPA2	$+0.013$	$\mathbf{-0.054}$	-0.214	$2/6$	-1.691

Table 14: ScanNet: real indoor RGB-D video, 104 clips over 13 rooms, splits grouped by room so no room appears on both sides, 6 seeds, layers fixed in advance from the KITTI runs. Image position explains only $R^2 = 0.083$ of ScanNet depth against 0.733 on the 60 KITTI scenes of Table 13, so the control leaves far more room here and the margins are correspondingly larger. CogVideoX-2B lands at $+0.012$ against an untrained floor of -0.014, which is how the NYU contradiction ($+0.394$, warped parallax) resolves as a warping artifact. V-JEPA2's row is the failing window configuration; Table 11 corrects it. Its untrained floor of -1.691 is why we do not floor-subtract on this dataset: doing so would manufacture a spurious $+1.6$.

What to take from it Indoors the control is weak, so the margins are large and the failure modes are visible. The untrained encoder floors run from -0.145 to -1.691, far below zero, because a random-init encoder on grouped room-disjoint splits does worse than predicting the mean. That is why we do not floor-subtract here: the same arithmetic that gives a clean gap on KITTI would report $+1.6$ for V-JEPA2 on ScanNet, which would be an artifact of an unstable floor and not a result. The row that settles a dispute is CogVideoX-2B's. On NYU, where parallax is induced by warping each image by its own disparity, the same model scored $+0.394$; on real indoor video with real camera motion it scores $+0.012$.

I THE FOUR ATTACKS, IN FULL

§5.3 states the verdicts; this appendix keeps the numbers and the reasoning behind them. Each attack was appended to the preregistration with its pass rule before any of its data existed, the first two in Section F of the released file, the dense flow field in Section E and the VAE control in Section D, and all four are exploratory in the sense of Section A.

Attack 1: higher-order image position. Position in this paper is five terms, so if what the encoders add were a smoother function of where a patch sits, a richer basis would absorb it. We swept the basis to 50 random Fourier features (Rahimi & Recht, 2007) of (row, column), the width the model's whitened features get, giving the position basis and the model matched capacity. V-JEPA2's increment

goes from $+0.1310$ to $+0.1226$, keeping 94% of itself, while the largest generator reads $+0.0389$. A position basis with the model's own capacity absorbs six percent of the effect.

Attack 2: a nonlinear readout. Huang et al. (2026) read generators with a shallow transformer and conclude they understand 3D; we read linearly and find nothing beyond position. Both can hold if the information is present but not linearly accessible, and until now this paper could not tell the two apart, our own earlier nonlinear probe having returned $R^2 = -16.85$, a failed probe rather than a null. A regularized gradient-boosted readout, early-stopped on a validation split cut from the training scenes only, raises V-JEPA2 from $+0.1191$ to $+0.1331$ while the generators reach at most $+0.0483$, a margin of $+0.0848$. The dissociation is a property of the representation, not of the readout class we chose. It does not settle whether a transformer readout on their data and targets would agree; that needs their pipeline, not ours.

Real optical flow in place of the frame-difference proxy. The flow regressor in every row of Table 12 is the mean absolute grayscale frame difference per patch. It has no direction and it confounds motion with texture, so every "beyond position and flow" in this paper was, literally, beyond position and frame difference. We re-read the same features in the same extraction pass with a dense optical-flow field pooled to the grid (mean u, mean v, magnitude; Farnebäck (2003), no learned weights), so the two columns are paired. The field is not the proxy renamed: their correlation is 0.54. The encoders give up 0.007 to 0.009 of their gap and the generators gain up to $+0.002$: the weakest encoder reads $+0.0367$ and the strongest generator $+0.0132$, a margin of $+0.0235$ against the $+0.0344$ the proxy gave. The proxy was carrying about a third of the margin and none of the generators' null. The ranges still do not overlap, so the dissociation survives its own control, and "beyond position and flow" is now true as written. The proxy column reproduces every published gap to within 0.002, which is the check that the pipeline did not move between the two readings. The separation also holds seed by seed: the weakest of the eighteen encoder seed values is $+0.0257$ and the strongest of the thirty-six generator seed values is $+0.0177$.

Model	gap, proxy	gap, real flow	change
V-JEPA2	$+0.1188$	$+0.1112$	-0.0076
VideoMAE-large	$+0.0685$	$+0.0615$	-0.0070
VideoMAE	$+0.0456$	$+0.0367$	-0.0089
CogVideoX-2B	$+0.0109$	$+0.0132$	$+0.0023$
Mochi-1	$+0.0109$	$+0.0119$	$+0.0010$
CogVideoX-5B	$+0.0039$	$+0.0048$	$+0.0009$
HunyuanVideo	-0.0054	-0.0046	$+0.0008$
Wan2.2 (8×8)	$+0.0095$	$+0.0097$	$+0.0002$
LTX-Video (8×8)	$+0.0091$	$+0.0078$	-0.0013

Table 15: Floor-subtracted gap under the published frame-difference proxy and under a dense Farnebäck flow field (mean u, v, magnitude), both computed on one extraction pass over the same features, 200 scenes, six seeds. The proxy column reproduces Table 12 to within 0.002. ViT-H has no untrained builder and is not re-read.

The VAE control. Every generator is read behind a lossy video VAE and no encoder is, so "generators do not carry it" has a rival reading: features behind a VAE have lost the parallax that structure-from-motion needs. Two arms, both preregistered with their rules. Probing CogVideoX-2B's VAE latent itself, with no transformer in the path, gives a floor-subtracted gap of $+0.0121$. That latent has 16 channels, so the comparison is made at 16 whitened dimensions, where the encoders read $+0.0237$ to $+0.0318$ and the generators $+0.0077$ to $+0.0119$: the bare latent sits at the generator ceiling and at half the weakest encoder, which is no clean call under the rule. The second arm runs an encoder on frames that have been through the VAE (encode, decode, pixel MSE 0.030 and 0.017, so the frames really changed). VideoMAE keeps $+0.0328$ of its $+0.0456$ and V-JEPA2 $+0.0828$ of its $+0.1187$: 72% and 70% of the published advantage survive a round trip that is itself a lower bound on the damage, since a decoded frame keeps full resolution while a DiT reads the latent directly. So the latent the generators read carries about half of what the weakest encoder extracts at matched dimension, their transformers add nothing on top of it, and encoders on the VAE's decoded pixels keep most of their advantage. The cross-class claim stands with that qualifier attached. One VAE was tested; Wan's and LTX's were not.

Arm	frames	Δ trained	floor	gap	pixel MSE
latent, CogVideoX-2B VAE	13	$+0.0122$	$+0.0001$	$+0.0121$	—
round-trip, VideoMAE	16	$+0.0562$	$+0.0234$	$+0.0328$	0.030
round-trip, V-JEPA2	64	$+0.0740$	-0.0088	$+0.0828$	0.017

Table 16: PREREG Section D. The latent arm probes the VAE latent with no transformer in the path; its latent has 16 channels, so it is whitened to 16 dims and read against reference rows recomputed at 16 dims (encoders $+0.0237$ to $+0.0318$, generators $+0.0077$ to $+0.0119$; rule D.5 K-1). The round-trip arm runs the encoder on frames the VAE has encoded and decoded, a lower bound on the loss; non-zero pixel MSE is the guard that the frames changed. Floors are the same architecture at random initialisation on the same inputs.

J WAN2.2, THE PATCH GRID, AND A HELD-OUT FORECAST

The held-out forecast, block by block. The main text states the score and the rule it triggers; here is the run. We put the criterion on record again and lost. Of the three things *able* commits to, L1 returned $R^2 = 0.835$ against controls of -0.616 and -0.254, L3 returned chance at every one of nine blocks, and L4 returned a probe that does not fit its own training split, with in-distribution R^2 of -0.039 at the best block, so the out-of-distribution half had nothing to degrade from and the rung's own no-degradation read would have passed it on a drop it never had. Two scoring rules were fixed after the numbers were seen, are declared below, and both move the verdict against the forecast. One of the two failures belongs to the rung, not the model. Every DiT's L3 here is read at one schedule index; two steps cleaner, LTX is lawful at $9/9$ blocks (§4.1). The L4 failure has no such cause. So the criterion did not misjudge whether LTX is lawful, it misjudged whether the rung as run would say so. The two results below, Wan2.2 at its native grid and the patch-grid swing, were added after the audit closed; both are exploratory in the sense of §4.

The pipeline is the published one. Before reading anything into a new row, the same script was pointed at CogVideoX-2B at 16×16: it returns $R^2_{\text{pos}} = 0.7326$, $R^2_{\text{pos,flow}} = 0.7624$, naive 0.6809 and $\Delta = +0.0237$ at n=60, matching the 60-scene run on every field, and $+0.0298$ at n=200, matching the 200-scene one. The grid is the only free parameter.

Model	grid	n	Δ	$\Delta_{\min}$	gap over floor
CogVideoX-2B	16×16	200	$+0.0298$	$+0.0169$	—
CogVideoX-2B	8×8	200	$+0.0479$	$+0.0303$	—
Wan2.2	8×8	200	$+0.0273$	$+0.0136$	$+0.0092$
untrained floor	8×8	200	$+0.0181$	$+0.0023$	—
LTX-Video	8×8	200	$+0.0307$	$+0.0156$	$+0.0112$
untrained floor	8×8	200	$+0.0195$	$+0.0057$	—
CogVideoX-2B	16×16	60	$+0.0237$	$+0.0061$	—
CogVideoX-2B	8×8	60	$+0.0552$	$+0.0283$	—
Wan2.2	8×8	60	$+0.0305$	$+0.0022$	$+0.0064$
untrained floor	8×8	60	$+0.0241$	$+0.0134$	—

Table 17: Wan2.2 and LTX-Video measured at their native grid, with the control that makes them readable. Patch resolution moves the metric on a fixed model and fixed data by $1.6\times$ at n=200 and $2.3\times$ at n=60, so Wan's row is comparable to the 8×8 control and not to the 16×16 rows. Wan's gap over its own untrained floor is $+0.0092$, inside the generators' range and an order of magnitude below the encoders'. Unlike the bootstrap table, these are six-seed point estimates without intervals.

The held-out forecast, in the order it happened. LTX-Video was extracted with no depth label touched, passed a structural gate (64 patches, patch-variation ratio 0.174, locality drop $+0.0211$), and scored CKA_c over nine blocks: 0.497, 0.613, 0.674, 0.777, 0.782, 0.758, 0.826, 0.708, 0.605. The maximum, 0.8259 at block 20, is inside the cluster band, so the locked rule predicted *able*. That prediction was committed, and only then was L1 run: R^2 across the same nine blocks spans 0.747–0.835, best at block 7, against a random-feature control of -0.616. L3 then returned 0.9927 at best (chance; every block's mirror residual above its identity residual) and L4 an in-distribution R^2 below zero at every block, so the forecast is scored $1/3$ and wrong. Two scoring rules were fixed

after the numbers were seen and are declared here. L3 uses the flag the pool was scored with, which also requires the mirror residual to beat the identity residual; the literal reading of the preregistration (ratio below one) would pass LTX at 0.996 and would also flip CogVideoX-2B at 0.804, a published row. L4 requires the in-distribution probe to beat the mean predictor, because a probe that fits nothing shows no drop: at block 0 every drop is under 0.15 on a probe with $R^2_{\text{ID}} = -0.039$. Both rules move the verdict against the forecast; both readings are in the JSON.

K THE RESIDUAL GATE AT 8×8

phase19c's $0/3$ was zero of the three 16×16 cluster models; Wan2.2 was excluded on the same inherited grid argument that Appendix J lifts. The gate was rerun at 8×8 with phase19c's scoring imported unchanged (whitening, CCA, all four thresholds), pooling the 16×16 maps 2×2 and reading Wan2.2 and LTX-Video at their native grid. Each model's shared subspace is defined against every other model in the set, so adding a model changes the reference for every row. Position alone reads $R^2 = 0.834$ at this grid.

Model	R^2_{resid} shared	private	full	k_{sh}	passes
Control: the original three, pooled to 8×8					
V-JEPA2	-0.215	-0.290	-0.015	12	no
CogVideoX-2B	$+0.029$	-0.220	-0.158	10	no
Mochi-1	$+0.092$	-0.108	$+0.011$	9	no
Extended: the three plus the two 8×8 *models*					
V-JEPA2	$+0.022$	-0.317	-0.015	11	no
CogVideoX-2B	$+0.018$	-0.208	-0.158	11	no
Mochi-1	$+0.104$	-0.109	$+0.011$	9	no
Wan2.2	-0.033	-0.347	-0.149	10	no
LTX-Video	-0.077	-0.108	-0.126	3	no
Same, with LTX-Video re-extracted at block 20 (its CKA_c *argmax)*					
V-JEPA2	$+0.031$	-0.319	-0.015	11	no
CogVideoX-2B	$+0.024$	-0.212	-0.158	11	no
Mochi-1	$+0.105$	-0.110	$+0.011$	9	no
Wan2.2	-0.020	-0.355	-0.149	10	no
LTX-Video (block 20)	-0.069	-0.106	-0.115	2	no

Table 18: The residual gate at 8×8. The control block is the check that pooling did not move a verdict: all three original models still fail, and no criterion flips. The extended block is the paper's $0/5$. LTX-Video shares $k = 3$ of 20 whitened directions with the pool at block 7 and $k = 2$ at block 20, against 9–11 for the others: at the patch level it is the least converged model in the set, whichever layer is read, while its scene-pooled CKA_c places it in the cluster.

L L3 AS A FUNCTION OF THE NOISE LEVEL

Every DiT's lawfulness rung was read with the latent noised to schedule index 350 of 1000. The dial was turned on the two 8×8 models with nothing else changed: the same 100 scenes, the same blocks, the latent flip, the random-permutation baseline and its bootstrap. "clean" adds no noise and passes timestep 0. τ is the flow-matching noise fraction read off each scheduler, not the index.

M THE SHARED SUBSPACE, STEP BY STEP

What bounds the layout component. The gate decides it: position- *residualized* depth fails to decode from the shared subspace in $0/5$ models, at 16×16 for the three cluster models and, once Wan2.2 and LTX-Video were re-admitted at their native 8×8, reading -0.03 and -0.08; pooling the original three to 8×8 flips no criterion (Appendix K). What the subspace encodes instead is the patch's own row and column, near ceiling, while the private complement returns neither (Table 22, Figure 9). The word *convergence* earns its place: the four cluster members decode patch row at 0.938–0.964 where the two outsiders decode it at chance (exploratory). The position-residualized probe

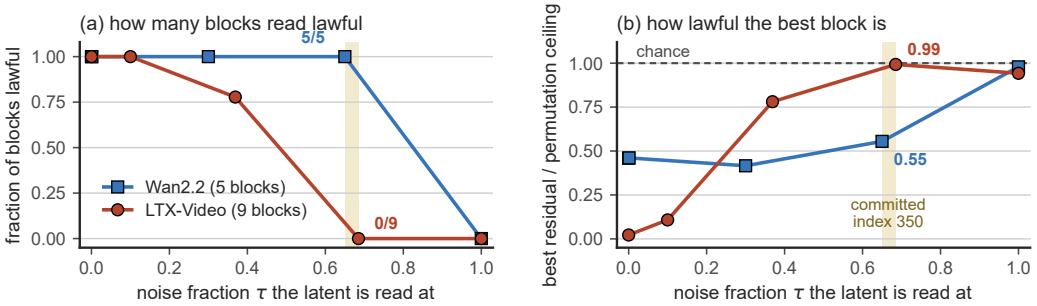


Figure 8: **A lawfulness verdict at one noise level is a verdict about that level.** **(a)** Fraction of blocks that read lawful against τ. **(b)** The best block's residual over the permutation ceiling; the dashed line is chance. The band is the committed protocol, schedule index 350. Wan2.2 is flat until pure noise; LTX-Video collapses between $\tau = 0.37$ and 0.685, and the band sits on the far side of it.

Model	index	τ	lawful	best	ρ	mirror	identity
Wan2.2	clean	0.000	5/5	14	0.459	0.0951	0.1833
Wan2.2	999	0.001	5/5	14	0.461	0.0949	0.1824
Wan2.2	700	0.300	5/5	4	0.416	0.0746	0.1664
Wan2.2	**350**	**0.650**	**5/5**	**4**	**0.555**	**0.0819**	**0.1369**
Wan2.2	0	1.000	0/5	14	0.979	0.0429	0.0384
LTX-Video	clean	0.000	9/9	0	0.022	0.0001	0.0018
LTX-Video	999	0.100	9/9	7	0.108	0.0010	0.0063
LTX-Video	700	0.369	7/9	17	0.781	0.0169	0.0190
LTX-Video	**350**	**0.685**	**0/9**	**17**	**0.993**	**0.0515**	**0.0491**
LTX-Video	0	1.000	0/9	14	0.942	0.0585	0.0502

Table 19: L3 against the noise level the latent is read at, everything else fixed (100 scenes, latent flip, random-permutation baseline, bootstrap). "lawful" counts blocks whose mirror residual sits below both the permutation ceiling and the identity residual; "best" is the most lawful block and the three ratios are read there. Bold is the committed protocol.

transfers nowhere (-0.006; permutation control -0.31), confirming the verdict from an independent protocol. The shared subspace is a low-frequency layout code that on driving data carries the scene's depth profile. Five measurements in sequence, each one narrowing what the cross-model convergent subspace can be.

1. How many directions two models actually share. CCA on top-20 PCA-whitened per-patch features, counting directions with canonical correlation ≥ 0.5; the directions that clear it run 0.92–0.97, and random 20-dimensional features give ≈ 0.04, so the threshold is not what manufactures them. The three cluster members share 7–9 of 20 directions with each other; every pair involving an outsider shares 0 or 1, and the two outsiders share 0 with each other. So this is not one cluster against a rival cluster. It is one cluster and two models that are isolated from everything, including from one another. Wan is excluded from this analysis: its 8×8 grid does not align patch-for-patch with the others.

What to take from it The first four columns establish that the shared subspace holds whatever the full representation could decode, and its orthogonal complement holds none of it. The two ablation columns rule out the obvious deflationary reading. Removing k random directions leaves 0.49–0.58, so the collapse under shared-direction removal is not about how many dimensions were taken; removing the k least-shared directions changes nothing, so it is not about which end of the spectrum they came from. The last column is where the section turns. Everything the subspace decodes, a quadratic function of image position had already explained, to within 0.03. The subspace is real, cross-model and causal, and it is not geometry.

LTX-Video, ρ per block					
block	clean ($\tau = 0.00$)	999 ($\tau = 0.10$)	700 ($\tau = 0.37$)	350 ($\tau = 0.69$)	0 ($\tau = 1.00$)
0	0.022	0.116	0.926	1.093	1.038
3	0.029	0.117	0.822	1.020	1.034
7	0.022	0.108	0.797	1.012	1.038
10	0.037	0.128	0.814	1.008	1.010
14	0.090	0.184	0.796	0.994	0.942
17	0.088	0.169	0.781	0.993	0.948
20	0.068	0.174	0.846	1.016	0.985
24	0.023	0.188	0.879	1.018	1.000
27	0.039	0.234	0.898	0.993	1.004
Wan2.2, ρ per block					
block	clean ($\tau = 0.00$)	999 ($\tau = 0.00$)	700 ($\tau = 0.30$)	350 ($\tau = 0.65$)	0 ($\tau = 1.00$)
4	0.475	0.477	0.416	0.555	0.985
9	0.499	0.500	0.491	0.619	0.990
14	0.459	0.461	0.447	0.595	0.979
19	0.513	0.515	0.485	0.597	0.995
24	0.543	0.545	0.525	0.629	0.988

Table 20: The sweep of §4.1, block by block. Each cell is the mirror residual over the random-permutation ceiling for one block at one noise level; below 1 and below the identity residual is lawful. LTX-Video's collapse is not one block's: every block moves together between $\tau = 0.37$ and 0.685. Wan2.2's five blocks stay under 0.6 until $\tau = 1$.

Model	k_{sh}	full	shared	private	ablate k rand.	ablate least-sh.	beyond pos.
V-JEPA2	10	0.799	**0.830**	-0.065	0.486	0.830	$+0.033$
CogVideoX-2B	8	0.786	**0.766**	-0.006	0.581	0.781	$+0.008$
Mochi-1	7	0.828	**0.832**	-0.029	0.582	0.829	$+0.028$

Table 21: 2–4. Scene-disjoint 70/30 splits throughout. The shared subspace decodes depth as well as the full representation (mean 0.809 against the full features' own value) while the private complement decodes none (mean -0.033). Ablating the k shared directions destroys decoding; ablating k random directions leaves 0.49–0.58; ablating the k *least* shared directions leaves it intact, so the localization is causal and not a dimension-count effect. The last column is the increment over a quadratic image-position basis: mean $+0.023$. Everything the subspace decodes, position already explained.

What to take from it The ordering inside each row is the diagnosis. If the shared subspace were a depth code, depth would decode from it better than image row and column do; instead row and column beat depth in all three models. Within depth the split is just as clean: the low-pass component decodes at 0.71 mean and the high-pass component at 0.16, so what is shared is the smooth large-scale part of the depth map, the part a ground-plane layout already implies. The private rows are the control. If these numbers were an artifact of the decoder or the split, the complement would show some of the same structure. It shows none.

What to take from it The matrix is block-structured rather than graded. The 3×3 cluster block is uniformly near ceiling and everything touching an outsider is near zero, in both directions: probes fitted in an outsider's coordinates do not read cluster models, and probes fitted in the cluster do not read outsiders. There is no intermediate cell anywhere in the table. This is the strongest form the convergence claim takes, because transfer is a harder test than similarity: two representations can score high on CKA while a probe fitted in one is useless in the other. What transfers is also what the previous two tables identified, so the conclusion is narrow. A layout code transfers across architectures at ceiling; the geometry that is not layout transfers nowhere.

N Representational similarity to ground-truth geometry

What to take from it This is the one analysis in the paper that reaches the conclusion from outside the probing framework, and it agrees. Every cluster member correlates with ground-truth scene geometry above its permutation null after appearance is partialled out, so the convergence is not

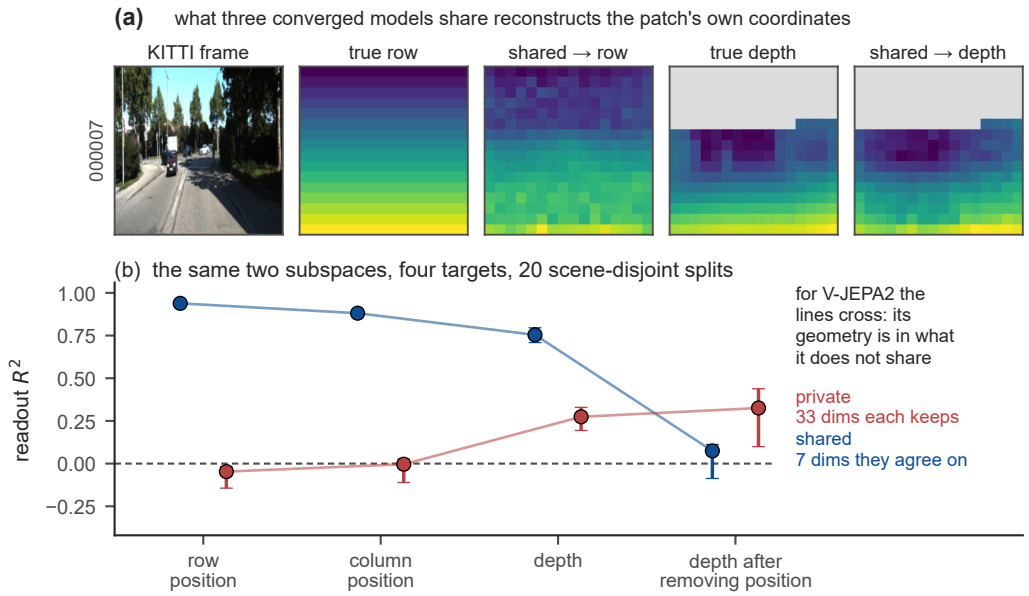


Figure 9: **What three converged models share is a coordinate system.** V-JEPA2's features, whitened to 40 directions, against CogVideoX-2B and Mochi-1 concatenated (Wan2.2's 8×8 grid does not align patch-for-patch): the seven directions with canonical correlation above 0.8 are the shared subspace and the other 33 its private complement. **(a)** The KITTI frame, then row and depth decoded from the shared subspace and painted on the 16×16 grid it is pooled to, each beside its ground truth. The patch's own row comes back almost exactly: the directions these models agree on encode where a patch sits in the frame. Its "depth" is that same gradient. **(b)** Over 20 scene-disjoint splits, points at medians and bars at the 10th and 90th percentiles. The shared subspace reads row 0.94 and column 0.88, and the private complement reads neither (-0.05, -0.00). On raw depth the shared subspace still leads, 0.75 against 0.27, because depth and image position nearly coincide on driving footage. Residualize position out and the order reverses: shared 0.07, private 0.33. That crossing is V-JEPA2's: with CogVideoX-2B or Mochi-1 as host, the private complement reads position-residualized depth at -0.01 and 0.06, no better than the shared subspace (0.06 and 0.06). Same 40 scenes as every subspace number in §6.4; Table 21 cuts at 20 whitened directions and canonical correlation 0.5, leaving a private complement of ten to thirteen directions that reads raw depth at -0.03 where the 33 here read 0.27.

Decoded from the shared subspace	depth	row	column	depth low-pass	depth high-pass
V-JEPA2	0.830	0.933	0.890	0.704	0.158
CogVideoX-2B	0.766	0.931	0.928	0.694	0.121
Mochi-1	0.832	0.939	0.866	0.736	0.202
private complement	−0.065	−0.089	+0.041	−0.053	−0.023
private complement	−0.006	−0.008	+0.024	−0.021	+0.031
private complement	−0.029	−0.017	+0.003	−0.016	+0.050

Table 22: 5. What the shared subspace encodes. It decodes vertical image position (0.934 mean) and horizontal position (0.895) better than depth itself (0.809), and within depth it carries the low-pass component (0.711) and not the high-pass one (0.160). That is a scene-layout code: where things are in the frame, smoothly. The three private rows are the orthogonal complement in each model and decode nothing.

purely an appearance effect. The columns we do not smooth are the second and third. Appearance similarity exceeds geometry similarity for every model in the table, and color exceeds both for every model but Hunyuan. These representations track geometry; they are not mostly about geometry.

fitted in ↓ / read in →	V-JEPA2	Cog2B	Mochi	Hunyuan	Cog5B
V-JEPA2	$+0.830$	$+0.783$	$+0.807$	$+0.002$	-0.015
Cog2B	$+0.845$	$+0.785$	$+0.842$	$+0.009$	-0.022
Mochi	$+0.840$	$+0.791$	$+0.834$	-0.001	-0.015
Hunyuan	$+0.033$	$+0.033$	$+0.033$	$+0.005$	-0.013
Cog5B	-0.000	-0.002	-0.008	-0.002	-0.049

Table 23: A ridge depth probe fitted in one model's CCA-aligned shared coordinates and evaluated on another model's features, scene-disjoint throughout. Within the cluster the mean cross-model transfer is 0.818 against a self-transfer ceiling of 0.816, the mean of each target model's own self-transfer (the diagonal), so the probe transfers at 100% of ceiling. Transfer between the cluster and an outsider averages $+0.004$ over both directions. What transfers this well between independently trained architectures is the layout code of Table 22, not geometry: the same probe fitted on position-residualized targets transfers nowhere (-0.006 mean over the same pairs).

Model	RSA geometry	RSA appearance	RSA color	geometry | appearance	perm. p
V-JEPA2	$+0.193$	$+0.235$	$+0.354$	$\mathbf{+0.125}$	0.000
Wan2.2	$+0.117$	$+0.183$	$+0.304$	$\mathbf{+0.060}$	0.030
Mochi-1	$+0.161$	$+0.182$	$+0.272$	$\mathbf{+0.108}$	0.002
CogVideoX-2B	$+0.073$	$+0.107$	$+0.195$	$\mathbf{+0.040}$	0.031
CogVideoX-5B	$+0.062$	$+0.067$	$+0.073$	$\mathbf{+0.042}$	0.038
HunyuanVideo	$+0.033$	$+0.047$	$+0.040$	$\mathbf{+0.019}$	0.148

Table 24: Representational similarity between each model's per-scene representation and three external similarity structures on the same 40 scenes: ground-truth geometry, raw appearance, and color. The fourth column partials appearance out of the geometry correlation, and 1000-permutation p values test it. Every cluster member survives at $p < 0.05$; Hunyuan does not ($p = 0.148$). The asymmetry we disclose rather than smooth is in columns two and three: cluster-mean appearance RSA ($+0.177$) exceeds cluster-mean geometry RSA ($+0.136$), and color RSA is higher still. Convergence to ground-truth geometry survives an appearance control, but geometry is not what these representations are mostly tracking — which is the same conclusion the subspace chain reaches by a different route.

Hunyuan is the negative control that came for free: it fails the permutation test at $p = 0.148$, the same verdict it receives on every rung above L1.

O L2 BEHAVIORAL PROTOCOLS

Protocol	Inject layer	Window	α sweep	Wan z	Cog2B z
v1 single-step†	L9	step 17	$\{0.5, 1, 2\} \cdot s$	$+0.20$	$+0.20$
v2 multi-step†	L9	steps 5–24	$\{2, 5, 10\} \cdot s$	$+0.34$	$+0.12$
v3 multi-step†	L4 (peak)	steps 5–24	$\{10, 20, 50\} \cdot s$	$+1.51$ mono.	-1.64 mono. (opp.)
v4 full-window	L0	steps 0–49	$\{10, 20, 50\} \cdot s$	$+0.17$	$+0.09$

Table 25: The four behavioral L2 protocols behind §6.2: the depth direction is injected during denoising, the generated video is decoded to RGB, and its depth is probed by an independent V-JEPA2+KITTI oracle (train $R^2 = 0.88$). s is the per-layer activation scale; z is quoted at α=20. **Only the v4 sweep survives as a committed result file** (10 seeds), and its own verdict fields record both the causal and the monotonicity check as failed. The v1–v3 rows are recorded in the run log only, are marked exploratory (†) for that reason, and are reported here because the sweep they document is why we stopped: they are not regenerable from the release. Across v4's three α the largest $|z|$ is 0.28 (Wan) and 0.21 (Cog2B), so no committed protocol comes near 3σ. Maximum injection also breaks generation for *random* directions, putting the null standard deviation at 4.30 (Wan) and 6.81 (Cog2B) and erasing distinguishability, which is why pushing harder buys nothing.

Why the sweep stopped here This table exists to show how hard we pushed before reporting a null. It is not itself a result. Three of the four protocols are marked exploratory because they survive only in the run log, and we include them because a reader judging whether the internal–behavioral

gap is real needs to know the range of the search, which the release does not contain. The v3 row is the informative one: at maximum injection Wan moves monotonically in the intended direction while Cog2B moves monotonically in the opposite one, the signature of a generator being broken rather than steered. The v4 null standard deviations make the same point in numbers. Once random directions disrupt generation as thoroughly as the depth direction does, no injection strength can tell them apart, and a larger sweep would buy a larger null rather than a result.

P THE INDOOR REPLICATIONS IN FULL

The main text gives the verdicts; this appendix keeps the runs behind them.

Two indoor protocols were run on NYU Depth v2 and neither counts, for different reasons; real indoor video does count, and it agrees with KITTI. A causal test of the same claim inside a generator did not separate the hypotheses.

The indoor test of the gate is inadmissible, in either direction. An attempted transfer of the residual gate (§6.4) to NYU Depth v2 (Silberman et al., 2012) is scoped out on a preregistered feature-health gate: a DPT-Large anchor (Ranftl et al., 2021; 2022) on the same grids is healthy (per-scene affine $R^2 = 0.74$) while four of five video models' features fail every target variant, a static-clip extraction artifact. NYU neither supports nor counts against the claim.

One protocol disagreed; real video settles it. A cross-domain replication on NYU reversed the verdict for CogVideoX-2B, which scored $+0.39$ there against $+0.36$ and $+0.43$ for the encoders. NYU has no video, so parallax is *induced* by warping each image by its own disparity. The random-initialized CogVideoX floor is already 0.159 (V-JEPA2: -0.023): that architecture reads the warp. Re-running the identical probe on ScanNet (real indoor video, 104 clips over 13 rooms, splits grouped by room, diffusion latents read clean), CogVideoX-2B falls to $+0.012$ against its own floor of -0.014, so the NYU score was artifact. VideoMAE and VideoMAE-large hold at $+0.121$ and $+0.236$ (the latter $6/6$ seeds, minimum $+0.201$). Only one of the six generators was retested there. Indoor video also leaves the control far more room than driving: image position explains $R^2 = 0.083$ of ScanNet depth against 0.733 on the 60 KITTI scenes of Table 13, so the KITTI margins were small.

The generative test, in full. We attempted to confirm the dissociation causally by ablating an extracted SfM basis inside CogVideoX-5B during generation and measuring the depth fidelity of the output. Ablating the SfM basis drops fidelity by 0.291; ablating a matched appearance basis drops it by 0.305. The two ablations are indistinguishable, so the experiment does not separate them and we record it as a failed attempt rather than as support.

Q WHAT ADOPTING THE PROTOCOL CHANGES

The main text states the configuration and the headline consequence; here is the arithmetic behind it, computed from already-committed results.

Adopting it does not re-rank the pool†. Across the seven checkpoints that have both readings, the naive R^2 and the floor-subtracted controlled gap agree at Spearman $+0.93$ and disagree about 2 of 21 pairs. What changes is the size of the claim and whether there is one: naively every model reads at least $R^2 = 0.389$ and would pass for "depth decodes"; controlled at the fixed block, four of four encoders clear their floors and none of the six generators does, three of them on a positive mean their own interval does not support. The compression is the part worth carrying: V-JEPA2 holds $30.44\times$ the controlled signal of CogVideoX-5B while reading only $1.843\times$ as high naively, though the ratio's denominator is a gap of $+0.004$ whose interval $[-0.008, +0.015]$ covers zero (Table 12), so its size is not stable and the main text reports the order of magnitude instead. A naive number cannot be converted into a controlled one after the fact, which is why the six have to be reported.

R FIVE PROBING DEFAULTS, AND WHERE EACH ONE FAILS

The six choices of §4 are choices inside a recipe; this appendix says what auditing them costs the recipe. Each default below meets a measurement in this paper rather than an argument, and the

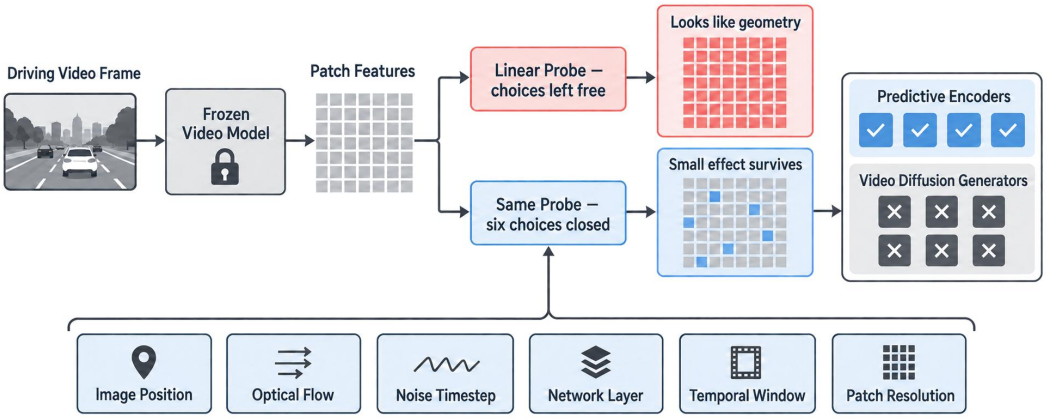


Figure 10: **The same features, read two ways.** One frozen model, one probe, one dataset, read twice. With the six choices left free, a linear readout reports $R^2 = 0.848$ for KITTI depth. With every choice controlled but the block on 200 clips, the same probe leaves $+0.046$ to $+0.119$ in four of four predictive encoders and none of six generators (Tables 12 and 17; per-choice swings in Table 1). The two grids show *how much* survives, not where. Diagram only, no data (Appendix S).

probe is closed-form ridge throughout, so every effect is attributable to the model and not to probe engineering.

Default	What it meets here	Where
Decodability indicates the capability	L1 decodes depth in five of six models and an untrained V-JEPA2 reaches 0.758 on the same probe, so linear presence separates nothing on this pool	§6.2, App. D
Bigger models represent more	CogVideoX-5B shares CogVideoX-2B's architecture, exits the convergence cluster and loses linear depth ($0.86 \rightarrow 0.03$); its untrained floor sits at its trained value, so training failed to build the structure rather than destroying it	§6.1
Spatial locality indicates spatial structure	The locality drop, preregistered as a cheaper predictor, forecast Mochi-1 as ABSENT and the measured ladder returned present at $R^2 = 0.855$; Mochi has the cluster's lowest locality drop and full ability	§6.3, App. S
A depth probe reads depth	A quadratic image-position basis alone reaches $R^2 = 0.844$ on KITTI and V-JEPA2's 0.848 naive decode becomes $+0.122$ once that basis is regressed out; what the converged models share decodes position better than it decodes depth	§6.4, App. M
A rung is a rung	L3's lawfulness verdict is a verdict about the noise level it is read at, and the controlled increment swings $13\times$ over the timestep and about $50\times$ over the layer on fixed models and fixed data	§4.1, App. G

Table 26: Five defaults of representation probing and the measurement each one meets. The values are stated with their controls where the last column points; none of the five is argued.

S HONESTY APPENDIX: THE CLAIMS THAT DID NOT SURVIVE

How Figure 10 was made. Every other figure in this paper is generated by a committed script that reads its result files and is checked against them. Figure 10 is the one exception, and it is a diagram rather than a measurement: it was drafted by an image-generation model from a written specification released with the harness. The specification forbids numerals in the image for the obvious reason, so the panel carries no quantity of its own; the values in its caption are typeset from Tables 1, 12 and 17

Claim as filed	What falsified it	What replaced it
World models follow appearance under conflict	texture control (only texture changed): $0.60/0.55 \rightarrow 0.93/0.21$	cue-trackability law; real-footage validation at 0.95
Paradigm decides where and whether depth appears	where: Cog2B flat/mid at $n{=}3$; whether: at a clean latent CogVideoX-5B reads 0.153 beyond position and flow, where the ladder's noised timestep read zero	profiles are model-specific; every controlled probe reads generators at a clean latent
Locality predicts ability (preregistered screen)	held-out Mochi: predicted ABSENT, measured 0.855	$\mathrm{CKA_c}$ convergence predicts; locality is the failure-mechanism signature
L1 $= 0.86$–0.90 (patch-random)	within-scene leakage $\approx +0.07$	scene-disjoint 0.78–0.83 canonical for every cross-model comparison; the ladder's L1 column stays patch-random and is labelled so
The shared subspace *is* the geometry	residual gate: $0/3$ models (now $0/5$ at 8×8) decode position-residualized depth from it (position alone $R^2 = 0.844$); residual probe does not transfer	rescoped to a shared layout code; the layout probe transfers cluster-wide at the self-ceiling
$\mathrm{CKA_c}$ predicts ability out of sample	held-out LTX-Video: cluster band (0.826), predicted able; L3 chance at every block, L4 in-distribution $R^2 < 0$; scored $1/3$	$\mathrm{CKA_c}$ ranks within the pool it was built on; a rung read at one noise level is a verdict about that level
Zero of six generators carry depth beyond position and flow	PREREG H: each model's block selected on held-out splits rather than by naive decodability — Mochi-1 crosses at $+0.0188$ $[+0.0031, +0.0327]$, encoders unchanged at $+0.0475$–$+0.1789$	a separation of degree: one generator clears, $2.5\times$ below the weakest encoder, and three clear at no block

Table 27: Every claim this project filed and then withdrew, what falsified it, and what replaced it. Four are predictions a control or a held-out forecast falsified, two are measurements corrected — one for leakage, one for a block chosen by naive decodability — and one is a claim rescoped by its own gate: the seven §7 counts. The full ledger of kill criteria, gate memos and deviations ships as supplementary material.

like any other number in the text. The two grids inside it are schematic renderings of effect size, not spatial maps, and nothing in the figure was measured.

How to read the two tables together They answer different questions and are not one list. The first records claims this project filed and then withdrew: four predictions falsified, two measurements corrected, one claim rescoped. Every row was decided by a measurement rather than by argument, and two of them by a forecast filed before the model was probed. The second records objections raised against surviving claims, each with the measurement that discharged it. The asymmetry is the honest summary of the project: seven claims did not survive and eight challenges did not land, the last two being the controls the headline number leans on. The shared subspace appears in both tables, falsified as geometry in the first and rescoped as a layout code in the same row, because the gate that falsified the strong reading left the weaker one standing and transferring at ceiling.

The two figures behind the first row These are the measurements that falsified the claim that world models follow appearance under geometric conflict, and the one that replaced it. Texture is the only variable that changes. With no trackable parallax cue, V-JEPA2 at L12 reads 0.598 on motion-geometry and correlates 0.552 with the shading cue, which is what we first reported. Give it a cue it can track and the same model at the same layer moves to 0.931 and 0.213; the gain holds at all four layers it was probed at (Fig. 11). The original claim was an artifact of stimuli that made geometry untrackable, and its replacement is a law about cue trackability instead of a claim about what models prefer. The second panel is the validation, shown two pages on: over 60 KITTI scenes with a painted shading field, the probe correlates 0.950 with true geometry and -0.152 with the painted cue, a gap

Caveat raised	Discharged by
L3 lawfulness could come from the VAE, not the DiT	flipping in latent space (VAE bypassed): Wan DiT-only 0.55 vs 0.68 with VAE — *more* lawful without it; Cog2B unchanged (0.804 vs 0.805)
L4 generalization R^2 could be a target-variance artifact	compositional-split target-variance ratio 1.01; MAE and nMAE agree with R^2
L5 trackability law measured only on synthetic stimuli	real KITTI footage with painted shading: the probe tracks real geometry at 0.95 and is not fooled by the fake cue
CKA convergence could be circular ("successful models just look alike")	RSA to ground-truth geometry survives partialling out raw-RGB appearance (permutation $p < 0.05$ for every cluster member); the remaining asymmetry — appearance RSA $+0.18$ exceeds geometry RSA $+0.14$ — is disclosed rather than smoothed
Cog5B: did scale-up *destroy* geometry or never build it?	untrained-DiT floor $0.031 \approx$ trained $\Rightarrow$ never built; Wan untrained 0.48 vs trained $0.85 \Rightarrow$ training does build it where it exists
NYU indoor null could count against the shared subspace	feature-health gate: NYU features fail every repaired target in 4/5 models while the DPT anchor on the same frames is healthy ($\lvert\rho\rvert = 0.72$) $\Rightarrow$ extraction artifact (static-clip/domain), so NYU is inadmissible in either direction
"Flow" was mean absolute frame difference, not flow	re-read with a dense optical-flow field (mean u, v, magnitude) on the same features: encoders lose 0.007–0.009, generators gain ≤ 0.002, weakest encoder $+0.037$ vs strongest generator $+0.013 \Rightarrow$ survives; margin narrows from $+0.034$ to $+0.024$ (§5.3, exploratory)
Generators are read behind a VAE, encoders are not	bare CogVideoX latent reads $+0.012$, the generator ceiling at matched 16 dims and half the weakest encoder; encoders on VAE round-tripped frames keep 72% (VideoMAE) and 70% (V-JEPA2) of their gap $\Rightarrow$ the VAE is part of the explanation, not all of it; one VAE tested (§5.3, exploratory)

Table 28: Objections to the surviving claims and the control experiments that discharged them. Unlike the table above, these claims survived their challenge; each row cites the deciding number.

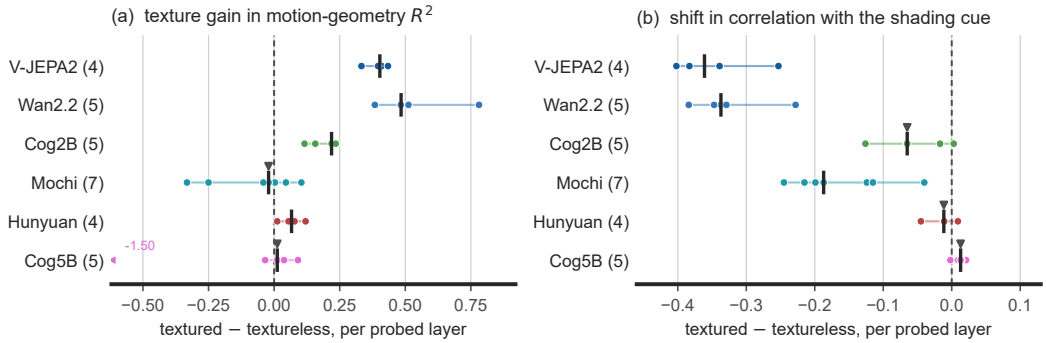


Figure 11: **Texture gating, at every probed layer.** Texture is the only changed variable. One row per model, one dot per probed layer (count in the label), bar at the median. **(a)** Gain in motion-geometry R^2. V-JEPA2, Wan2.2 and Cog2B gain at every layer they were probed at (medians $+0.40$, $+0.48$, $+0.22$). Mochi and CogVideoX-5B change sign across layers (▼): Mochi gains only at the shallowest of seven layers and loses at four of them, so its median is -0.02. Hunyuan gains at every layer but by a fifth of the cluster's median ($+0.07$), which is the sense in which the rescue fails for it. CogVideoX-5B's -1.50 at its deepest layer is drawn at the axis edge with its value rather than dropped. **(b)** The same layers, shift in correlation with the shading cue; negative is movement away from following appearance. An earlier version of this figure read each model at one layer, chosen by the argmax of the same textured endpoint it plots, which biased every slope upward; drawing all layers removes the choice rather than replacing it.

of 1.103 $[1.030, 1.171]$ that does not shrink as the fake cue is strengthened to $\alpha = 0.9$ (1.105). One caveat travels with it: this validation covers V-JEPA2 only.

T LIMITATIONS IN FULL

The main text states these in brief (§7); each is given here with the numbers it rests on.

n=6 is small, and the guards are what that n allows. Six checkpoints cannot support a scaling claim, and the two regimes are read from a pool we assembled. Every guard runnable at this n is in Appendix F: exact enumeration over all ten n=5 labelings, leave-one-out, independent halves, pseudo-model controls, two kernels. The mechanism analysis over six heterogeneous checkpoints with collinear axes rules out simple structural explanations while leaving the mechanism open; two untested readings fit the failure signatures (MMDiT attention trading locality for cross-modal mixing; within-architecture scale-up sliding toward an over-smooth attractor that a nonlinear readout can still decode (Huang et al., 2026)), and separating them needs training-time access.

The ladder lacks L2 for three models, and the domain is driving. Mochi is measured on four of five rungs (its L2 is unrun), and the outsiders have no L2 because L2 needs the L1 direction they lack. Wan is excluded from patch-level subspace analyses (its 8×8 grid). All confirmatory numbers are KITTI driving scenes; the gate's indoor test is inadmissible rather than negative (Appendix P). L4 stimuli are synthetic bump arrangements, and the real-footage L5 validation covers V-JEPA2 only.

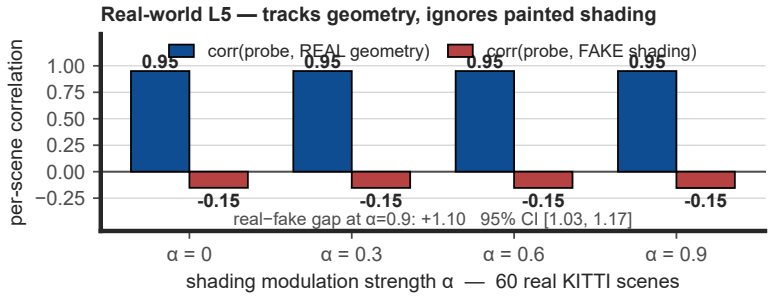


Figure 12: Real-footage L5 on 60 KITTI scenes: per-scene correlation of the depth probe with true geometry and with a painted shading field, swept over the strength α of the fake cue. The two curves do not approach each other.

The real-footage validation, and how far it reaches The last caveat above is what this figure shows and does not show. On 60 real KITTI scenes the probe tracks true geometry at 0.950 and the painted shading field at -0.152, and the separation is flat in the strength of the fake cue: at $\alpha = 0.9$ the gap is still 1.105. So the cue-trackability law is not an artifact of synthetic stimuli. It is also one model at one layer. Nothing here says a generator behaves the same way on real footage, and we did not run one.

Behavioral causality is sub-threshold, and pushing harder is self-defeating. Single-direction steering stays under 3σ in generated video across the four protocols we ran. The maximum-injection protocol breaks generation for random directions too, putting the null standard deviation at 4.30 and 6.81, so a stronger intervention buys no discriminability (Table 25). We report the internal–behavioral gap as a finding and stop there.

Wan2.2's exclusion, and how it was resolved. Through the audit Wan2.2 appeared in no experiment of §4, because its 8×8 patch grid did not match the position basis every controlled probe regresses out. That was not an incidental omission: Wan has the pool's highest $\mathrm{CKA_c}$ (0.87) and its strongest causal mediation ($z = 4.4$), so the model our convergence story ranked first was the one the geometry story could not test. The exclusion turned out to be inherited rather than necessary, because the basis is analytic in (row, column) and the depth target pools to any grid, and Wan is now measured at its native grid (Appendix J): $+0.0273$ over an untrained floor of $+0.0181$, a gap of $+0.0092$. It

is a generator like the others. What the episode leaves behind is the sixth choice the exclusion was hiding, patch resolution, which moves the metric by 1.6–$2.3\times$ on a fixed model.

The second held-out forecast, scored. LTX-Video was the criterion's one out-of-sample test and it failed, $1/3$. The failure has a documented cause on one rung (the noise level, Appendix L) and none on the other (L4's in-distribution probe does not fit at any block). Both are reported. The preregistration's own rule for this outcome is the one used: $\mathrm{CKA_c}$ is described as ranking within a pool, not as a portable classifier, and no claim is upgraded. Cosmos-2.5, the planned second held-out model, never loaded and stays dropped; adding a model after the first failed is the selection the preregistration exists to prevent.

The temporal window is a fifth choice, and we found it by failing. Our first ScanNet configuration integrated V-JEPA2 over ≈ 4 seconds of camera motion against a first-frame depth target and reported no indoor structure-from-motion for it; the window sweep overturns that (-0.082 to $+0.295$, §4), so the indoor result rests on both encoder families. We keep the failing configuration: a probe's temporal extent relative to what its target describes flips signs, is unreported in the literature we build on, and we did not anticipate it either. We then replaced the argument with a measurement. CogVideoX-2B swept the same way is flat when the target sits at the window start, and rises to $+0.0773$ when it is centred, a swing of 0.0897 against V-JEPA2's 0.3767 on the same clips. A 13-frame window does not reproduce the sign flip, but it is not inert either, and we had argued it would be.

A loading defect in one encoder, found by a smoke test, and its reach. Published VideoMAE checkpoints store attention biases in the model's original layout (`q_bias`, `v_bias`), which recent `transformers` releases do not map onto their `query/key/value.bias` parameters: loading reports them missing and silently randomizes them. Our ScanNet runs repair the mapping explicitly and assert that it fired. The earlier VideoMAE numbers in this paper were produced before the defect was known, on a different library version. On the 200 KITTI scenes the repair has since been measured (PREREG H.7): at the block chosen out of fold it raises VideoMAE from $+0.0475$ to $+0.0611$ and VideoMAE-large from $+0.0937$ to $+0.1010$, so the defect understated both and moved no verdict, and the tables keep the unrepaired values, the conservative side of every encoder–generator comparison. The synthetic-stimulus value has not been re-run and carries the caveat; the ScanNet values do not.

$\mathrm{CKA_c}$ ranks; it does not hand over a threshold. Recomputed on the two independent 20-scene halves, every $\mathrm{CKA_c}$ value moves by about 0.16 and two models fall inside the empty band on one half, three on the other, while the rank order holds. A practitioner can order candidates against a reference pool; they cannot carry our cut-off to their data. Two further limits: the layer each probe reads was fixed by a *labelled* depth sweep, so scoring a new model needs that sweep or a heuristic; and the algorithmic clustering that recovers our partition is not significant against its own permutation null ($p = 0.51$), so the partition stands as a hand partition with a $p = 0.073$ silhouette rather than as a discovered cluster.