
When Does Latent Prediction Error Support Physical Surprise Claims in Video World Models?

SUBMISSION VERSION, NOT CAMERA-READY

This PDF will be updated.

Abstract

Prediction errors from video world models are often treated as physical surprise, but they can also reflect appearance, static content, motion, or ordinary prediction difficulty. We present an evaluation framework that separates what frozen features contain, how they organize appearance and physical differences, and what changes a model’s own prediction error. We study IntPhys2, a synthetic benchmark of possible and impossible events, and exact appearance and motion changes rendered with Kubric, a controllable synthetic video generator. We compare frozen features from a self-supervised latent predictor, a masked pixel reconstruction model, and simple color and motion statistics. Appearance remains easy to recover across these features, while one evaluated self-supervised video model exposes only modest information about whether events are possible or impossible. Looking at the latent geometry, appearance changes become smaller, line up more consistently across scenes, and overlap much less with physical changes in deeper layers. Across three evaluated action-conditioned world models, incorrect actions consistently raise native prediction error, but appearance changes with state held fixed produce either larger or smaller errors than approximately matched opposite actions depending on the model and score. Exact Kubric rerenders produce similar error changes for visual and physical interventions. A model can therefore represent physical information and react to incorrect actions without guaranteeing that its native prediction error measures physical surprise.

1 Introduction

Video world models are expected to capture regularities that make visual futures predictable, so predictive likelihood, reconstruction error, latent distance, and denoising loss are often operationalized as model "surprise" or physical plausibility [Smith et al., 2019, Piloto et al., 2022, Riochet et al., 2018, Dai et al., 2023, Garrido et al., 2025, Bordes et al., 2025, Maes et al., 2026, Yuan et al., 2025]. Related scores also select physically plausible generations [Baldassarre et al., 2025, Yuan et al., 2026].

In a latent-predictive world model, an encoder maps video observations to representations, and a predictor forecasts a future representation from context, optionally conditioned on actions. For predictor p , representation f , observed future x_{t+1} , and distance d , a score

$$S_t = d(p(f(x_{\leq t})), f(x_{t+1}))$$

directly establishes only prediction difficulty in the chosen representation. A large value may reflect physical inconsistency, appearance, static content, motion, latent anisotropy, or ordinary model error; decodable physics labels likewise do not make the native error physically specific. This distinction matters when residuals localize events, compare videos, or reward trajectories.

We study construct validity by separating *information accessibility*, *representation geometry*, and *native-error semantics*. Distinct experiments use IntPhys2 representations, Physion++ local errors, released action-conditioned predictors, and exact Kubric renders [Bordes et al., 2025, Tung et al., 2023, Greff et al., 2022]; no result is pooled across these roles.

A split audit superseded two earlier IntPhys2 probes; our post-hoc five-fold cross-fit holds each of 57 sealed scenes out once. V-JEPA2-G reaches 57.9% [54.2%, 61.8%] on possible/impossible classification across all four appearance conditions, exceeding VideoMAE2-G and PixelSummary by 8.7 pp [4.4, 13.3] and 7.9 pp [4.1, 11.8]. LeWorldModel errors respond to action mismatches, yet color changes preserving simulator state and trajectory can shift them as much or more. Under the same PushT protocol, DINO-WM assigns lower error to hue changes than to opposite actions with approximately matched visible effects.

Our contributions are:

1. A practical evaluation framework that separates what information frozen features contain, how they organize appearance and physical differences, and what changes a model’s own prediction error. The tests prevent success on one question from being used as evidence for a stronger claim.
2. An IntPhys2 representation study showing that videos altered in hue, brightness, or object color move closer to their clean versions across depth in V-JEPA2 and VideoMAE2 even though the appearance condition remains decodable, while V-JEPA2 yields a modest possible versus impossible readout. After accounting for the overall scale and directional bias of the latent space, appearance changes become smaller, show more consistent patterns across different video scenes, and overlap much less with physical changes in deeper layers.
3. Controlled experiments showing that an action-conditioned predictor can react more strongly to an appearance change that preserves simulator state and trajectory than to an incorrect action with approximately matched visible effects. Static content tests find high errors without physical inconsistency, while exact renders find similar errors from visual and physical changes, showing why physical sensitivity alone does not make prediction error a reliable measure of physical surprise.

2 Related work

2.1 Predictive video representation learning

Self-supervised video learning increasingly trains models to predict masked or future visual structure without requiring action labels. Joint-embedding predictive architectures (JEPAs) replace pixel reconstruction with prediction of target representations. JEPAs and hierarchical JEPAs have been proposed within a broader architecture for non-generative predictive world models whose representations are intended to be simultaneously informative and predictable [LeCun, 2022]. V-JEPA learns from masked video regions and V-JEPA2 extends this family toward video understanding, prediction, and planning [Bardes et al., 2024, Assran et al., 2025]. Masked-reconstruction methods such as VideoMAE and VideoMAE V2 learn video features by recovering masked observations [Tong et al., 2022, Wang et al., 2023, Alrasheed et al., 2026]. These families are increasingly used as components of video world models because their representations support temporal recognition and downstream decision making. However, objective-level or downstream-task comparisons do not fix the meaning of a model’s native prediction residual. A large latent discrepancy can reflect generic prediction difficulty, visual novelty, appearance change, ordinary motion, or a physically inconsistent transition. The predictive objective leaves the residual’s physical meaning unspecified. We evaluate that meaning with direct controls. Evidence that a model forecasts representations well or transfers robustly to a downstream task consequently does not establish whether its pointwise score is calibrated across visual conditions or reflects an expected causal transition.

2.2 Action and physics information in latent representations

Recent work gives strong evidence that video latents can encode information relevant to actions and physical structure. Yeom et al. evaluate diverse video and image encoder families with a common inverse-dynamics probe and find that action recoverability is driven primarily by temporal video pre-

training rather than pixel reconstruction fidelity [Yeom et al., 2026]. Nilaksh et al. similarly compare reconstruction and semantic latent spaces under a fixed robotic world-model protocol, finding that visual fidelity alone is insufficient for selecting latents that support planning and policy evaluation [Nilaksh et al., 2026]. Recent work on deterministic-world cloning further improves long-horizon prediction by explicitly regularizing latent geometry, demonstrating that representation structure materially affects rollout error [Xia et al., 2026]. We address the complementary question of what such errors measure and which controls are required before interpreting them as physical surprise. Joseph et al. use layerwise probes, subspace analysis, and attention ablations to show that physics-related information becomes accessible at intermediate depths in V-JEPA2 and VideoMAE2 [Joseph et al., 2026]. At the representation level, Punzo et al. probe frozen video foundation models on IntPhys2 and Minimal Video Pairs, and show that the accessibility of physics-relevant information depends on model family, layer, and temporal readout [Punzo et al., 2026]. PhyLatent further identifies appearance sensitivity, physical-state aliasing, and compressed counterfactual action branches as distinct failure modes of an action-conditioned JEPA latent, then adds training-time constraints to address them [Zeng et al., 2026].

These studies concern what can be decoded from a representation, how its geometry organizes actions and states, or whether it is useful for downstream control. They therefore motivate our setting, but leave open whether the native error of a predictor over such a latent space is itself semantically specific to physical inconsistency. In particular, an inverse-dynamics probe, a policy score, or a state-grounding head can reveal physically useful information without showing why a frozen predictor assigns a large error to a particular observed transition. This distinction matters because a predictor can operate over a physically organized latent space while remaining sensitive to visual factors that are irrelevant to the physical event being evaluated.

2.3 Physical reasoning and violation-of-expectation evaluation

Violation-of-expectation benchmarks assess whether a model distinguishes physically plausible from implausible events. IntPhys introduced controlled paired videos for visual intuitive-physics reasoning, while IntPhys2 increases environmental complexity and provides a more recent standardized evaluation [Riochet et al., 2018, Bordes et al., 2025]. Physion and Physion++ evaluate physical prediction and online inference of different physical properties from video [Bear et al., 2021, Tung et al., 2023]. Garrido et al. show that self-supervised pretraining on natural video can yield intuitive-physics competence, strengthening the case that physical structure can emerge without explicit symbolic supervision [Garrido et al., 2025]. LeWorldModel evaluates an action-conditioned JEPA using prediction discrepancies as a surprise signal for physically implausible events [Maes et al., 2026]. CRONOS instead tests whether generated videos preserve the same physical event under controlled changes to scene, viewpoint, object category, and appearance [Begiristain et al., 2026].

Together, these benchmarks and diagnostics make predictive discrepancy a useful candidate signal, but they usually evaluate a ranking over plausible and implausible videos, an explicit downstream readout, or the fidelity of a generated rollout. They do not jointly isolate static prediction difficulty, appearance-only variation, lawful dynamical change, and physical inconsistency for the same native scalar. In particular, an error increase at an impossible event is not by itself sufficient to show physical specificity if an appearance-only intervention or a nearly static window produces an equal or larger increase.

Our construct-validity evaluation uses matched static, appearance, lawful-motion, and impossible-dynamics controls to identify what causes native latent prediction error to increase. These controls determine when an error supports a narrower physical-surprise interpretation.

3 Construct-validity framework

We separate six gates (Table 1). Failure at one gate limits the inference but does not negate evidence at earlier gates. Accessible physical information can coexist with appearance sensitivity, and action-conditioned error can coexist with poor physical specificity.

For IntPhys2, the final post-hoc five-fold cross-fit keeps all pairs and appearances from each scene together and holds every scene out once. Each fold fits a balanced logistic regression after training-

Table 1: Operational audit for a proposed physical-surprise score. Each gate states the minimum control and the strongest valid inference.

Gate	Required control	Passing inference
Readout	unseen independent groups; fixed-capacity probe	physical information is accessible to the readout
Geometry	unrelated-scale, shuffled-direction, and matched-subspace references	controlled organization beyond global scale and anisotropy
Action	identical target under lawful and inconsistent actions	residual uses action-transition compatibility
Appearance/static	appearance changes with simulator state fixed and frozen frame controls	response is not explained by those alternatives
Specificity	matched rendered effect and lawful/impossible comparison	inconsistency adds signal beyond visual change
Calibration	cross-scene/event/nuisance comparison	score magnitude is comparable over the claimed population

only standardization. Appendix C.2.2–C.2.4 documents the superseded folds, sealed-test status, and uncertainty estimand.

For each task, correctness is averaged within pair and scene before a 10,000-draw fixed-prediction bootstrap over all 57 held-out scene values. Paired model deltas preserve fold, scene, and pair alignment. The intervals condition on fitted probes and omit probe-refitting and fold-assignment variability.

At each frozen layer we ask three questions. *Magnitude*: how far does a perturbation move a pooled, L2-normalized representation relative to the median cross-scene distance between unrelated clean videos? *Direction*: does that perturbation move different scenes coherently, measured by the standard mean resultant length (MRL) [Mardia and Jupp, 1999], beyond a cross-scene shuffled-correspondence null? *Subspace*: what fraction of a held-out possible-to-impossible displacement lies in a $k = 4$ appearance basis fit only on training scenes, relative to a dimension-matched random subspace? Scene is the independent unit; all five layers and 10,000-draw scene bootstrap/null procedures were fixed on corrected development before one completed sealed geometry evaluation. Appendix C.2.4 gives exact estimands, null construction, and sensitivity.

We select LeWorldModel because it uses action-conditioned prediction discrepancy as a physical-surprise signal and releases controllable task checkpoints [Maes et al., 2026]. DINO-WM and JEPA-WM provide released PushT checkpoints over frozen DINOv2 features [Oquab et al., 2024, Zhou et al., 2024, Terver et al., 2026]. We did not evaluate V-JEPA-2/2.1-AC because no released checkpoint matches the TwoRoom or PushT action interface, and we found no original-author PLDM pretrained weights; comparisons are therefore restricted to released, task-matched checkpoints.

For native-error experiments, LeWorldModel reuses target bytes under lawful, zero, and opposite actions. Hue changes only rendered color; simulator state and physical trajectory remain fixed, although PushT color may carry task relevant information. DINO-WM and JEPA-WM reuse the PushT targets, manifest, rendered-effect matches, and bootstrap; DINO-WM scores final-step visual-latent MSE, while JEPA-WM uses its native preprocessing, context, proprioception, and predictor to score final one-step visual embedding MSE. Kubric uses identical 12-frame contexts across exact rerenders and is exploratory at eight scenes.

4 Results

4.1 Motion and static content in local prediction error

We test whether local prediction error follows visible motion or can remain high in nearly static windows. Using V-JEPA2-L, we compare local prediction error over fixed 16-frame windows from 10 Physion++ videos with RGB motion and frozen frame baselines (Appendix A.2–A.3). The analysis is self-contained: its V-JEPA2-L scores are not combined with the representation and rerendering experiments that follow.

Across 273 complete windows, raw prediction error has Pearson correlation $r = 0.596$ with mean consecutive RGB motion, and each video’s highest-error window contains substantial motion. Six of the 10 videos also contain a high-error near-static window with mean RGB difference below

0.01. All six have nonpositive residuals after subtracting the corresponding frozen frame score, with mean residual -0.0350 . The cutoff and candidate selection rule were exploratory. The 6/10 count identifies this failure mode in the inspected videos; population prevalence remains unknown. Local prediction error responds to motion and static content difficulty; interpreting it as an event signal requires motion and frozen frame controls.

4.2 Appearance and physics label accessibility

We test whether frozen video features make appearance condition and the possible versus impossible label linearly accessible. Measuring both determines whether accessible physics information coexists with retained appearance information. Using the 57-scene IntPhys2 sealed population under the post-hoc * five-fold scene cross-fit and the layers originally selected on development data, we evaluate V-JEPA2-G and VideoMAE2-G on identical sample identities and four appearance conditions, while PixelSummary provides frame RGB statistics, color histograms, and frame difference statistics. The population contains 114 matched pairs, 228 source videos, and 912 appearance conditioned samples. Appearance classification predicts clean, hue shifted, brightness shifted, or selectively re-colored input. Physics classification predicts possible versus impossible across all four appearance conditions, while clean-only physics uses a separately trained classifier on clean videos.

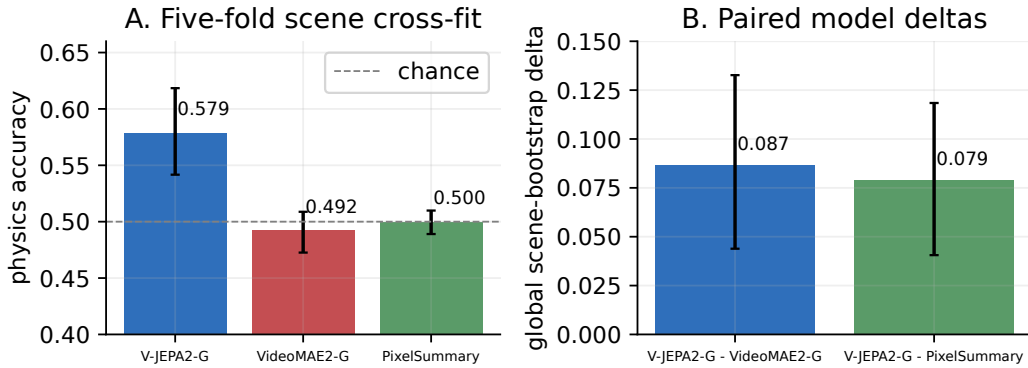


Figure 1: Five-fold IntPhys2 scene cross-fit on the 57-scene sealed population. Every scene is held out once and contributes one aggregated scene value; intervals and paired deltas bootstrap all 57 unique scenes globally while conditioning on fitted probes.

Table 2: IntPhys2 classification readouts and paired V-JEPA2-G comparator differences on the sealed population. The first three rows report mean out-of-fold scene accuracy in percent; rows below the separator report paired accuracy differences in percentage points. Brackets give fixed-prediction 95% global scene-bootstrap intervals.

Feature family or contrast	Appearance classification	Physics classification	Clean-only physics
V-JEPA2-G	97.7% [95.7, 99.3]	57.9% [54.2, 61.8]	54.8% [51.3, 58.8]
VideoMAE2-G	95.3% [92.8, 97.6]	49.2% [47.3, 50.9]	49.6% [46.5, 52.2]
PixelSummary	79.1% [73.7, 84.5]	50.0% [48.9, 51.0]	51.8% [48.7, 54.8]
V-JEPA2-G — VideoMAE2-G	2.4 pp [−0.1, 5.0]	8.7 pp [4.4, 13.3]	5.3 pp [0.4, 10.1]
V-JEPA2-G — PixelSummary	18.6 pp [13.5, 24.1]	7.9 pp [4.1, 11.8]	3.1 pp [−1.3, 8.3]

Figure 1 visualizes the physics readout and paired differences, while Table 2 reports all three classification tasks. Appearance remains recoverable in every feature family. V-JEPA2-G reaches 57.9% [54.2%, 61.8%] on all-appearance physics classification and exceeds VideoMAE2-G by 8.7 percentage points [4.4, 13.3] and PixelSummary by 7.9 percentage points [4.1, 11.8]. The clean-only V-JEPA2-G minus PixelSummary interval includes zero. The results show that modest physics accessibility coexists with highly accessible appearance information under the evaluated probe protocol. Training objective and architecture remain confounded. PixelSummary excludes simple RGB, color histogram, and frame difference summaries; stronger static, optical flow, image-only, and random-feature shortcuts remain possible. The sections below separately examine how appearance affects

200 representation geometry and how native prediction error responds to visual and physical interven-
 201 tions.

202 4.3 Geometry separates magnitude, direction, and subspace overlap

203 Linear readouts show what information is accessible. We next ask how appearance and physical
 204 differences are organized in latent space: whether appearance changes become smaller and more
 205 consistently directed across depth, and how much physical differences overlap with appearance
 206 directions.

207 Using V-JEPA2-G and VideoMAE2-G on the same 57-scene sealed IntPhys2 representation package,
 208 we measure three controlled properties at five relative depths. Magnitude divides each displacement
 209 by the median distance between unrelated clean videos. Direction measures mean resultant length
 210 beyond shuffled correspondence. Subspace overlap measures the fraction of a held-out physical
 211 displacement captured by an appearance basis fitted on training scenes. We also report transformed-
 212 to-clean retrieval as a descriptive alignment check. Definitions and statistical procedures were fixed
 213 on corrected development before the completed sealed evaluation. Appendix C.2.4 gives the formu-
 214 las, null constructions, complete layerwise values, and sensitivity analyses.

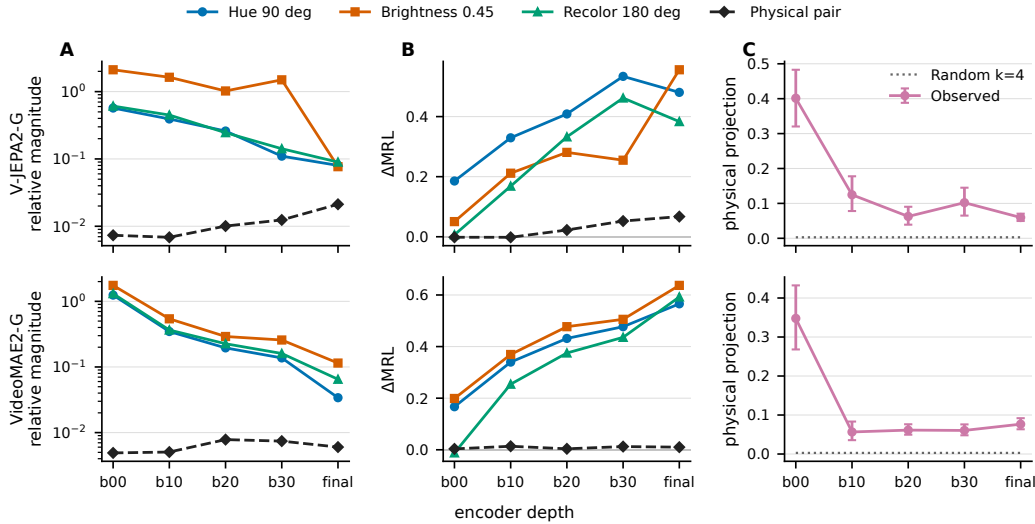


Figure 2: Development-frozen geometry on 57 sealed scenes. Columns show (A) displacement relative to unrelated clean videos (log axis), (B) MRL above shuffled correspondence, and (C) cross-fitted $k = 4$ physical projection into appearance space with scene-bootstrap intervals and a matched random reference. These quantities do not establish physical understanding or residual specificity.

215 Figure 2 shows that appearance displacement ends substantially smaller than at block 00 in both
 216 models after normalization by unrelated clean videos. Appearance directions are more consistent
 217 across scenes in deeper layers, while possible versus impossible directions remain much closer to
 218 shuffled correspondence. In V-JEPA2-G, physical projection into the appearance subspace falls from
 219 0.401 [0.320, 0.483] at block 00 to 0.060 [0.050, 0.071] at the final representation, while remaining
 220 above the matched random reference. VideoMAE2-G shows the same broad pattern. Transformed
 221 to clean top one retrieval reaches 0.656 for V-JEPA2-G and 0.446 for VideoMAE2-G at the final
 222 representation.

223 Together, direction and retrieval show that the reduced appearance displacement remains organized,
 224 preserving information about both the source video and the transformation. Physical differences
 225 occupy less of the appearance subspace at greater depth, although the final overlap remains larger
 226 than expected for a random subspace. The shared broad pattern in VideoMAE2-G precludes a JEPA-
 227 specific causal interpretation. Smaller magnitude supports suppression relative to background scale
 228 and leaves invariance unresolved. Magnitude, direction, retrieval, and subspace overlap characterize

distinct properties of the representation geometry; semantic factorization and calibrated physical surprise remain separate questions.

4.4 Native errors are action sensitive while appearance ordering differs by model

We test whether native prediction error changes when the model receives an incorrect action but is evaluated against the exact same target frame, and whether an appearance change with approximately matched visible magnitude can produce a comparable effect.

Each task or scorer uses 256 episodes. With released LeWorldModel (LeWM) checkpoints on TwoRoom and PushT, we reuse identical target bytes under lawful, zero, and opposite actions; hue changes alter only color and preserve state and trajectory. DINO-WM and the post-hoc JEPA-WM analysis reuse the PushT targets, rendered effect matches, and bootstrap. DINO-WM scores final step visual latent MSE, while JEPA-WM uses its own preprocessing, context, conditioning, and predictor to score final one step visual embedding MSE.

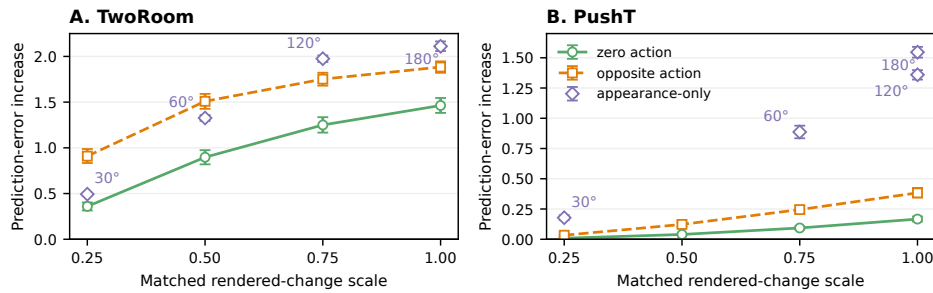


Figure 3: Controlled LeWorldModel prediction error audits. Incorrect actions raise native MSE. Hue changes that leave simulator state fixed can match or exceed action mismatch effects. The lawful action contrast is zero by definition and omitted from the legend. Hue points are relative to the unchanged static target.

Table 3: Selected high-dose native prediction-error contrasts (256 episodes per task or scorer; 95% episode-bootstrap intervals). Native magnitudes are model specific; Appendix Table 20 gives low and high doses, zero-action controls, and score definitions.

Model/task	Opposite — lawful at 1.0	Hue — opposite at higher match
LeWM TwoRoom	1.884 [1.821, 1.946]	0.220 [0.147, 0.295] at 180°/1.0
LeWM PushT	0.3834 [0.3456, 0.4242]	0.642 [0.585, 0.696] at 60°/0.75
DINO-WM PushT	0.00364 [0.00304, 0.00428]	−0.10987 [−0.12416, −0.09614] at 60°/0.75
JEPA-WM PushT	0.00351 [0.00293, 0.00411]	−0.12481 [−0.13905, −0.11114] at 60°/0.75

Figure 3 shows that zero action and opposite action counterfactuals increase LeWorldModel prediction error, with opposite actions producing the larger effect. Appearance only interventions can match or exceed these action mismatch effects in LeWorldModel. Table 3 shows positive opposite action minus lawful action contrasts for DINO-WM and JEPA-WM, while their fixed state hue changes remain below the matched opposite action residuals. Across the three native scores, action mismatch increases error, while appearance versus action ordering differs by model and score.

The common action response supports action transition compatibility. The differing within score ordering shows that native error semantics depend on the model and score. The predeclared pairs are closely matched in rendered RGB change, while hue and displacement remain non-equivalent interventions. We interpret each native score on its own scale. The matched color and exact rerender experiments below examine how visual and physical interventions contribute to prediction error.

4.5 Color direction affects error with simulator state and rendered magnitude fixed

We test whether color direction affects LeWorldModel’s native error after matching simulator state and rendered magnitude, using target only and persistent recolorings across 256 fresh PushT

episodes; Appendix B gives the full design, exact contrasts, and latent decomposition. Task associated colors consistently produce larger errors than matched novel colors, and in target only conditions the predictor is unchanged while shifts in the target representation account for almost all of the difference. Native prediction error therefore depends on the direction of an appearance change in the target encoder, even when simulator state and rendered magnitude are fixed. Because PushT colors may carry task information, the experiment leaves open whether the larger errors reflect task role encoding.

4.6 Appearance and physical interventions can produce similar native error changes

We test whether native error separates appearance change from physical displacement under exact simulator control. Using V-JEPA2-G, we evaluate eight Kubric MOVi-A scenes that pass predefined rendering and visibility checks; Appendix C.4.3 gives the exact contrasts. The first 12 context frames are byte-identical. Appearance changes preserve simulator state and trajectory; physical interventions introduce persistent displacement matched approximately in rendered magnitude. Both predeclared physical-minus-appearance confidence intervals include zero, so neither comparison resolves which intervention produces a larger mean error change. With identical context, the changed futures are scored against the same prediction. Together, the controls show that native error responds to appearance as well as physical displacement, even with the observed history and approximate magnitude of rendered change held fixed.

5 Discussion

The experiments separate information accessibility, representation geometry, and native error semantics. The post-hoc IntPhys2 readout exposes modest information about possible versus impossible labels alongside highly accessible appearance. Across both evaluated encoders, deeper representations show smaller appearance displacement and less overlap with possible versus impossible differences while preserving structured appearance information.

Across LeWorldModel, DINO-WM, and JEPA-WM, native error responds to inconsistent actions, while appearance versus action ordering depends on the model and score. Matched PushT recolorings localize one abrupt effect to target encoder color direction. Physion++ static controls and exact Kubric rerenders show that static content and appearance can produce large or overlapping residuals. Residual meaning therefore depends on the control, model, score, and population.

The audit treats readout, geometry, action sensitivity, specificity, and calibration as distinct evidential stages, preserving the difference between physical information and physically specific error.

6 Limitations

The post hoc IntPhys2 readout supports only modest accessibility, and its fixed prediction intervals omit probe refitting and fold assignment variability. A confirmatory study should preregister scene folds on a new sealed population and quantify uncertainty from both sources.

The geometry analysis was selected on development and covers five depths and three perturbations; its MRL and projection metrics characterize directional coherence and subspace overlap. The multi-dose extension is exploratory. Preregistered evaluations across additional transformations and models, together with controlled interventions on fitted subspaces, could test whether these patterns have semantic or functional consequences.

Broader generalization will depend on shared task evaluations across additional self-supervised video and action-conditioned world models. The current comparisons vary jointly in objective, architecture, pretraining, capacity, resolution, receptive field, and native score, and stronger static and motion baselines remain unevaluated. A common protocol can address these differences while preserving each model’s native score. For physical specificity, a larger Kubric study can define equivalence margins and compare rendered effect matched lawful and impossible trajectories.

Broader impact. World model residuals may serve as anomaly signals, rewards, or safety signals when sensitivity, specificity, and calibration are validated for the intended task, environment, and nuisance variation. Their reliability and transfer depend on measurements in each application.

References

- Ali J. Alrasheed, Aryan Yazdan Parast, Basim Azam, James Bailey, and Naveed Akhtar. Latent video prediction learns better world models. *arXiv preprint arXiv:2605.15618*, 2026. URL <https://arxiv.org/abs/2605.15618>.
- Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL <https://arxiv.org/abs/2506.09985>.
- Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. Back to the features: DINO as a foundation for video world models. *arXiv preprint arXiv:2507.19468*, 2025. URL <https://arxiv.org/abs/2507.19468>.
- 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. *arXiv preprint arXiv:2404.08471*, 2024. URL <https://arxiv.org/abs/2404.08471>.
- Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021. URL <https://arxiv.org/abs/2106.08261>.
- León Begiristain, Olaf Dünkel, and Adam Kortylewski. CRONOS: Benchmarking counterfactual physical consistency in video models. *arXiv preprint arXiv:2605.23699*, 2026. URL <https://arxiv.org/abs/2605.23699>.
- Åke Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973. doi: 10.1090/S0025-5718-1973-0348991-3.
- Florian Bordes, Quentin Garrido, Justine T. Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025. URL <https://arxiv.org/abs/2506.09849>.
- Bo Dai, Linge Wang, Baoxiong Jia, Zeyu Zhang, Song-Chun Zhu, Chi Zhang, and Yixin Zhu. X-VoE: Measuring explanatory violation of expectation in physical events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL <https://arxiv.org/abs/2308.10441>.
- 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. URL <https://arxiv.org/abs/2502.11831>.
- Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL <https://arxiv.org/abs/2203.03570>.
- Ian T. Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065):20150202, 2016. doi: 10.1098/rsta.2015.0202.
- Sonia Joseph, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake Richards, and Mike Rabbat. Interpreting physics in video world models. *arXiv preprint arXiv:2602.07050*, 2026. URL <https://arxiv.org/abs/2602.07050>.
- Yann LeCun. A path towards autonomous machine intelligence. OpenReview, June 2022. URL <https://openreview.net/forum?id=BZ5a1r-kVsf>. Version 0.9.2.
- 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*, pages 991–999, 2015. doi: 10.1109/CVPR.2015.7298701.

353 Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorld-
354 Model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint*
355 *arXiv:2603.19312*, 2026. URL <https://arxiv.org/abs/2603.19312>.

356 Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley, 1999. doi: 10.1002/
357 9780470316979.

358 Nilaksh, Saurav Jha, Artem Zholus, and Sarath Chandar. Reconstruction or semantics? what makes
359 a latent space useful for robotic world models. *arXiv preprint arXiv:2605.06388*, 2026. URL
360 <https://arxiv.org/abs/2605.06388>.

361 Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khali-
362 dov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud As-
363 sran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Is-
364 han Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien
365 Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust vi-
366 sual features without supervision. *Transactions on Machine Learning Research*, 2024. URL
367 <https://openreview.net/forum?id=a68SUt6zFt>.

368 Luis S. Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning
369 in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*,
370 6:1257–1267, 2022. doi: 10.1038/s41562-022-01394-8. URL <https://doi.org/10.1038/s41562-022-01394-8>.

372 Samuele Punzo, Niccolò Caselli, Ippokratis Pantelidis, Francesco Massafra, Salvatore Lo Sardo, and
373 Mohammadreza Salehi. Do video foundation models understand intuitive physics? a layerwise
374 probing analysis. *arXiv preprint arXiv:2606.09646*, 2026. URL <https://arxiv.org/abs/2606.09646>.

376 Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique
377 Izard, and Emmanuel Dupoux. IntPhys: A framework and benchmark for visual intuitive physics
378 reasoning. *arXiv preprint arXiv:1803.07616*, 2018. URL <https://arxiv.org/abs/1803.07616>.

380 Kevin A. Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth S. Spelke, Joshua B.
381 Tenenbaum, and Tomer D. Ullman. Modeling expectation violation in intuitive physics
382 with coarse probabilistic object representations. In *Advances in Neural Informa-
383 tion Processing Systems*, 2019. URL <https://papers.nips.cc/paper/2019/hash/e88f243bf341ded9b4ced444795c3f17-Abstract.html>.

385 Basile Terver, Tsung-Yen Yang, Jean Ponce, Adrien Bardes, and Yann LeCun. What drives success
386 in physical planning with joint-embedding predictive world models? *Transactions on Machine
387 Learning Research*, 2026. URL <https://arxiv.org/abs/2512.24497>.

388 Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-
389 efficient learners for self-supervised video pre-training. In *Advances in Neural Information Pro-
390 cessing Systems*, 2022. URL <https://arxiv.org/abs/2203.12602>.

391 Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, et al. Physion++: Evaluating physical
392 scene understanding that requires online inference of different physical properties. *arXiv preprint*
393 *arXiv:2306.15668*, 2023. URL <https://arxiv.org/abs/2306.15668>.

394 Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao.
395 VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the
396 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023.
397 URL <https://arxiv.org/abs/2303.16727>.

398 Zaishuo Xia, Yukuan Lu, Xinyi Li, Yifan Xu, and Yubei Chen. Cloning deterministic worlds: The
399 critical role of latent geometry in long-horizon world models. In *Proceedings of the IEEE/CVF
400 Conference on Computer Vision and Pattern Recognition*, pages 32602–32612, 2026. URL
401 <https://arxiv.org/abs/2510.26782>.

402 Jewon Yeom, Hanseul Kim, Jeongjae Park, Sungmok Jung, Jaejin Lee, and Taesup Kim. What
403 makes video world model latents action-relevant: Prediction over reconstruction. *arXiv preprint*
404 *arXiv:2606.07687*, 2026. URL <https://arxiv.org/abs/2606.07687>.

405 Jianhao Yuan, Fabio Pizzati, Francesco Pinto, Lars Kunze, Ivan Laptev, Paul Newman, Philip Torr,
406 and Daniele De Martini. LikePhys: Evaluating intuitive physics understanding in video diffusion
407 models via likelihood preference. *arXiv preprint arXiv:2510.11512*, 2025. URL <https://arxiv.org/abs/2510.11512>.

409 Jianhao Yuan, Xiaofeng Zhang, Felix Friedrich, Nicolas Beltran-Velez, Melissa Hall, Reyhane
410 Askari-Hemmat, Xiaochuang Han, Nicolas Ballas, Michal Drozdal, and Adriana Romero-
411 Soriano. Inference-time physics alignment of video generative models with latent world models.
412 *arXiv preprint arXiv:2601.10553*, 2026. URL <https://arxiv.org/abs/2601.10553>.

413 Xi Zeng, Haojie Ren, and Ziyang Song. PhyLatent: Learning dynamics-relevant representations for
414 JEPA world models. *arXiv preprint arXiv:2608.05720*, 2026. URL <https://arxiv.org/abs/2608.05720>.

416 Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-
417 trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024. URL
418 <https://arxiv.org/abs/2411.04983>.

A Evidence map and Physion++ static-content audit

A.1 Evidence interpretation

The experiments deliberately answer different questions. Table 4 records the independent unit, evidential status, supported inference, and stopping point for each block. The main conclusion depends on their separation: representation accessibility, representation geometry, action-conditioned sensitivity, and error specificity are not interchangeable.

Table 4: Evidence map after the scene-overlap audit. Status labels describe the executed protocol and bound the valid inference.

Evidence block	Unit/status	Supports	Does not support
IntPhys2 readout	57 scenes / 114 pairs; post-hoc five-fold scene cross-fit	Modest out-of-fold physics-label accessibility advantage under the evaluated probe budget	Preregistered confirmation, native-error specificity, calibrated surprise, or general physical reasoning
IntPhys2 frozen geometry	57 scenes / 114 pairs; development-frozen, completed sealed geometry evaluation	Controlled relative magnitude, null-adjusted direction, and cross-fitted subspace overlap for the evaluated perturbations/models	Semantic disentanglement, causal factorization, residual specificity, calibration, or general physical understanding
V-JEPA2-G multi-dose geometry	38 development scenes / 76 pairs; post-sealed exploratory	Factor-specific trajectories and matched- k tangent versus endpoint projection	Confirmation, hue circularity, semantic axes, or generality beyond tested factors
LeWorldModel/ DINO-WM action controls	256 episodes per task or scorer; confirmatory	Action-conditioned prediction sensitivity in LeWM TwoRoom, LeWM PushT, and DINO-WM PushT	Specificity of large error to dynamics or physical violation
JEPA-WM PushT action control	256 episodes; post-hoc exploratory	Action sensitivity and qualitative robustness of the DINO-WM ordering under a revised predictor recipe over the same encoder family	Independent encoder-family replication, causal attribution to one model choice, broad cross-model generality, or calibrated physical surprise
LeWorldModel appearance controls	256 fresh PushT episodes; confirmatory plus exact decomposition	Appearance sensitivity, color-direction anisotropy, and partial persistent transport	Pure nuisance invariance or role-aware dynamics
Kubric rerendering	8 scenes; exploratory	Exact rerenders with fixed state for appearance controls and displaced state for physical controls	A calibrated effect size or population-level V-JEPA2 conclusion
Physion++ window audit	10 videos, 273 windows; diagnostic	Need for pixel-motion and frozen-frame controls	False-positive rate, benchmark accuracy, or event localization

A.2 Selection and scoring

The local static-content audit uses the released V-JEPA2-L checkpoint facebook/vjepa2-vitl-fpc64-256 (checkpoint SHA-256 25466aef85727d16546c6cf8c99f12fcfad9cbca8225d45f23685e2e025b786b). Ten additional Physion++

Mass test videos were selected deterministically as evenly spaced files after excluding the earlier single-video debug example. Every complete 16-frame segment was scored, yielding 273 segments. For each segment we record the native latent residual, mean consecutive absolute RGB difference on the 0–255 scale, and the fraction of pixels whose consecutive-frame change exceeds 5.

For static calibration, the first, middle, or last decoded frame is repeated across all 16 input positions and rescored. The first-frame event residual is the original score minus the corresponding frozen-frame score. This control does not estimate a universal corrected surprise score; it asks only whether a local raw peak requires temporal change.

A.3 Motion association and static candidates

Across all segments, raw score and mean consecutive RGB change have Pearson correlation $r = 0.596$. Each video’s highest raw-score segment contains substantial motion; their mean pixel-motion value is 3.712 and their first-frame residuals are positive, from 0.0165 to 0.0492. However, 6 of 10 videos also contain a high-scoring near-static segment with mean RGB change below 0.01. Table 5 shows that all six have nonpositive frozen-frame residuals.

Table 5: Highest-scoring near-static candidate in each affected Physion++ video. RGB change is on the 0–255 scale; residual is original minus first-frame-frozen score.

Video	Segment	Raw score	Mean RGB change	Static residual
0	12	0.5898	0.001867	−0.0488
1	9	0.6021	0.001294	−0.0400
2	11	0.6083	0.009740	−0.0313
3	21	0.6380	0.005892	−0.0267
4	16	0.6215	0.001611	−0.0367
5	20	0.6455	0.001180	−0.0266

The mean first-frame residual over these six candidates is −0.0350. The threshold and highest-candidate rule were diagnostic choices, so 6/10 is not a population false-positive rate. Raw V-JEPA2-L error partially tracks visible motion but also admits high-error static-content candidates. This motivates the main paper’s pixel-motion and static baselines; it is not evidence about V-JEPA2-G benchmark accuracy or a general event detector.

B PushT role-color decomposition

B.1 Question and matched design

PushT uses blue for the agent, gray for the block, and green for the goal. Color is a potentially task-relevant cue in PushT. We test whether native LeWorldModel prediction MSE depends only on the amount of rendered change or also on color direction. The frozen confirmatory run uses 256 independent episodes excluded from all earlier PushT pilots, the released checkpoint, official four-frame and frame-skip-five preprocessing, and the installed Pymunk/Pygame renderer. Because the installed renderer is close but not byte-identical to stored observations, every contrast uses the paired current-renderer output as its baseline.

The block is recolored to agent blue and compared with a novel orange selected to match rendered magnitude; the agent is recolored to block gray and compared with a matched novel purple. Each condition is applied either only to the target or persistently to all four frames. Across colors, the maximum realized simulator-state difference is zero. In target-only conditions, context pixels, actions, and predictor outputs are exactly identical; only the target-encoder representation changes. Table 6 summarizes the matched rendered changes.

B.2 Matched contrasts

For each object and scope, the predeclared contrast is role-collision color minus matched novel color. Episode-level 95% intervals use 10,000 paired-bootstrap draws. All four directional contrasts are positive in the mean (Table 7), including substantially matched target RGB change and affected-pixel fraction.

Table 6: Matched PushT recolorings. Values are episode means relative to the paired current-renderer baseline.

Object	Role color	Matched novel	Role/novel RGB L1	Role/novel changed pixels
Block	agent blue	orange	0.006157/0.006188	0.02924/0.02924
Agent	block gray	purple	0.000563/0.000563	0.002844/0.002841

Table 7: Role-color minus matched-novel prediction-MSE contrasts over 256 fresh episodes.

Contrast	Scope	Mean [95% CI]	Positive episodes
Block blue — orange	target only	0.493 [0.450, 0.535]	233/256
Block blue — orange	persistent	0.111 [0.092, 0.131]	191/256
Agent gray — purple	target only	0.397 [0.355, 0.440]	245/256
Agent gray — purple	persistent	0.088 [0.054, 0.123]	164/256

468 B.3 Exact latent decomposition

469 Let $r_0 = p_0 - z_0$ be the baseline residual and $q = \Delta p - \Delta z$ the intervention-induced residual shift.
 470 Because the native score is mean squared latent error,

$$\Delta S = 2 \text{ mean}(r_0 \odot q) + \text{mean}(q^2).$$

471 Reconstructing every intervention score from this identity has maximum absolute error 2.38×10^{-7} .
 472 In target-only conditions $\Delta p = 0$ exactly, and the role-minus-novel score gaps nearly equal the
 473 corresponding target-encoder shift gaps (Table 8). This locates the abrupt effect primarily in target-
 474 encoder color-direction sensitivity.

Table 8: Matched-color decomposition. Cosines compare persistent predictor and target shifts for the role-color and matched-novel conditions.

Pair	Target score gap	Target-shift gap	Persistent score gap	Shift cosines
Block blue/orange	0.493	0.495	0.111	0.916/0.928
Agent gray/purple	0.397	0.400	0.088	0.815/0.705

475 Persistent predictor and target shifts are directionally aligned, especially for block recoloring, but
 476 transport is incomplete. The experiment therefore supports color-direction sensitivity and partial
 477 persistent-appearance transport. It does not establish that LeWorldModel encodes task roles or that
 478 these colors are irrelevant nuisances. The released weights have SHA-256 48938400ae3464c9
 479 680731287f583a9cb516f55a8ec64ea13a91be47fb15b607; the evaluated dataset has SHA-256
 480 b6ebd9ac94bbe9e383f6e7a9cd92d74e9aa665ea57b758ed3717b0ee7df8d4fb.

481 C Experimental details, audit trail, and asset scope

482 C.1 Resources, provenance, and asset scope

483 C.1.1 Checkpoint scope

484 Section 4.1 uses V-JEPA2-L only as a local diagnostic checkpoint for dense Physion++ window in-
 485 spection. The main IntPhys2 representation result and the exact-rerender Kubric pilot use V-JEPA2-
 486 G. The LeWorldModel studies use separate released task checkpoints and native prediction MSE.
 487 We therefore use each experiment for its own evidential role: V-JEPA2-G for the main represen-
 488 tation and Kubric rerender claims, LeWorldModel for action-conditioned prediction-error controls,
 489 and V-JEPA2-L only to motivate pixel-motion and static-baseline diagnostics. The V-JEPA2-L audit
 490 is not pooled with, or treated as a substitute for, the V-JEPA2-G results. DINO-WM and the post-hoc
 491 JEPA-WM extension provide separate frozen-DINO PushT visual embedding error checks; neither
 492 score is pooled with LeWorldModel native prediction MSE.

C.1.2 Compute and software environments

The experiments were executed in several recorded environments rather than a single undifferentiated compute stack. Table 9 maps each result to the hardware, runtime, and numerical precision actually used, while Table 10 reports the recorded stage times. CUDA versions are the versions reported by `nvidia-smi` or the PyTorch build suffix, as applicable.

Table 9: Compute and software environments for each experiment block.

Experiment block	Compute	Runtime, libraries, and precision
Corrected IntPhys2 representation study	NVIDIA GeForce RTX 3080 workstation; CUDA 12.8	Python 3.11.7; PyTorch 2.8.0+cu128; Transformers 5.13.1 for VideoMAE2-G; bfloat16 autocast and float16 stored features; PixelSummary uses NumPy 1.26.4.
IntPhys2 corrective probes/statistics	ARM64 Mac mini CPU; 16 GiB memory	Python 3.14.3; NumPy 2.4.3, pandas 3.0.2, and scikit-learn 1.9.0.
V-JEPA2-G multi-dose grid	NVIDIA GeForce RTX 3050 Laptop GPU; 4 GiB VRAM	Python 3.10.11; PyTorch 2.5.1+cu121; exact float32 weights with bfloat16 autocast, SDPA, and sequential transformer-block CPU/GPU offload; no quantization; unchanged 64-frame, 256-pixel preprocessing; peak reserved VRAM 704 MiB.
Physion++ static diagnostic	NVIDIA GeForce RTX 3080; CUDA 12.8	PyTorch 2.8.0+cu128 in float16; OpenCV 4.13.0, NumPy 1.26.4, pandas 2.2.1, and Matplotlib 3.10.8.
Kubric exact-rerender pilot	Cloud NVIDIA A40 (46,068 MiB); driver 570.211.01; CUDA 12.8	Rendering: Python 3.11.13, bpy 4.3.0, NumPy 1.26.4, Cycles OptiX, 64 samples/pixel, no denoiser. V-JEPA2-G scoring: Python 3.13.14, PyTorch 2.11.0+cu128, xFormers 0.0.35, timm 1.0.28, and einops 0.8.2; bfloat16 autocast with float32 error reduction.
LeWorldModel controls	ARM64 Mac mini CPU; macOS 26.5.2	Python 3.12.13; PyTorch 2.13.0, torchvision 0.28.0, NumPy 2.2.6, and stable-worldmodel 0.1.1; batch size 16.
DINO-WM PushT control	NVIDIA GeForce RTX 3080 workstation; CUDA 12.8	Python 3.11.7; PyTorch 2.8.0+cu128, DINO-WM repo revision <code>0a9492fa</code> , stable-worldmodel 0.1.1, NumPy 1.26.4, pandas 2.2.1; batch size 8.
JEPA-WM PushT control	Cloud NVIDIA A40 (46,068 MiB); driver 580.159.04	Python 3.10.12; PyTorch 2.7.0+cu126, torchvision 0.22.0+cu126, NumPy 2.2.6, pandas 2.3.3, Pymunk 7.3.0, JEPA-WM revision <code>13cf1d9c</code> , and stable-worldmodel revision <code>0ef38568</code> ; float32, batch size 8.

The Windows workstation and cloud worker host CPU models and system memory were not captured; their accelerator models and memory were recorded. The supplementary compute record marks these fields unavailable rather than estimating them. The exact low-memory multi-dose extraction reproduced overlapping clean features with minimum cosine 0.99999988. This engineering detail supports reproducibility and carries no scientific inference.

Table 10: Recorded wall times for paper-facing stages; unrecorded project-wide compute is excluded.

Stage	Compute	Recorded wall time
IntPhys2 development extraction	RTX 3080, CUDA 12.8	V-JEPA2-G 967.25 s; VideoMAE2-G 1296.84 s; PixelSummary 1227.36 s
IntPhys2 sealed extraction	RTX 3080, CUDA 12.8	V-JEPA2-G 1207.04 s; VideoMAE2-G 1263.22 s; PixelSummary 1769.81 s
IntPhys2 five-fold cross-fit	ARM64 Mac mini CPU	45.07 s total: development probes/statistics 17.85 s; sealed probes/statistics 27.22 s
V-JEPA2-G multi-dose grid	RTX 3050 Laptop GPU plus CPU analysis	912 clips extracted and 608 reused; extraction inference 3580.4 s; cached-feature geometry/statistics 5038.1 s
Physion++ diagnostic	RTX 3080, CUDA 12.8	48.20 s
LeWorldModel controlled tasks	ARM64 Mac mini CPU	TwoRoom 115.23 s; PushT 143.38 s
DINO-WM PushT control	RTX 3080, CUDA 12.8	143.02 s for the 256-episode controlled-displacement run
JEPA-WM PushT control	A40, PyTorch CUDA 12.6	200.90 s for 256 episodes; 13.76 s for the successful eight-episode gate
PushT role color/decomposition	ARM64 Mac mini CPU	139.12 s / 140.41 s
Kubric renderer gate	Cloud A40, CUDA 12.8	Selected/repeat 39.65/57.97 s; two denoiser trials 118.98/37.84 s
Kubric pilot render/validation/scoring	Cloud A40, CUDA 12.8	Render process 3348.14 s (72 jobs sum 3318.23 s); validation 21.16 s; scoring process 167.01 s (load 50.26 s; inference sum 35.90 s)

C.1.3 Executable provenance

The released experiment records preserve exact commands, code or file hashes, versioned configuration files, random seeds, selection manifests, output hashes, and checkpoint/data SHA-256 values. The anonymous supplementary archive at <https://anonymous.4open.science/r/Latent-Surprise-Audit-8730/README.md> provides source, manifests, processed metrics, integrity records, and a smoke script; the SHA-256 of `archive_manifest.json` is 56f9a8a7756806d917c075e0cb336181b8b84663689f57521773402407900b75; third-party datasets and checkpoints are not redistributed and must be obtained under their licenses or access terms. In particular, the V-JEPA2-G and VideoMAE2-G checkpoints used for IntPhys2 hash to 67129f011434e605d894e69f2c8e13d9db118deabe59d54bf6e0fa62c2c5cb8e and a231169cd13307aeb29bc56259a494ef273bfd9b7232f881ac4e8d7e39713add, respectively. The TwoRoom and PushT LeWorldModel weights hash to 566f223624ea4bfb39dbfe6ae731198dd6ea73b7b8919fed6b1ecafca810f7dd and 48938400ae3464c9680731287f583a9cb516f55a8ec64ea13a91be47fb15b607; the released DINO-WM PushT checkpoint hashes to e909f0cec958fc0b49a79f2e85730ae6b5f83dee61a224705f91883eb3bb74c5. Full package freezes are retained for the Kubric renderer and each confirmatory LeWorldModel run, and the A40 record includes the complete `nvidia-smi -q` capture. The corrected IntPhys2 Windows record captured the core runtime versions rather than a complete package freeze; it additionally preserves strict-load status, checkpoint hashes, feature hashes, deterministic repeat checks, and cross-model shard parity. The corrective CPU run records Python, NumPy, pandas, and scikit-learn versions plus input, code, and output SHA-256 values, but no contemporaneous complete transitive package freeze.

The post-hoc JEPA-WM record pins source commit 13cf1d9c7e476f53c17714d2e0f1dc239a883ce0, Hugging Face revision 9b9c41ef249466630dbf1a20e78391865d07b3b9, checkpoint SHA-256 9beca3eafe0739c3b3adb5d734fa435ccbd40fea8a65d53d4cccec176aaaa0eb, DINOv2 source commit 7764ea0f912e53c92e82eb78a2a1631e92725fc8, DINOv2 weight SHA-256 b938bf1bc15cd2ec0feacfe3a1bb553fe8ea9ca46a7e1d8d00217f29aef60cd9, and stable-worldmodel commit 0ef3856875e70a1283e637fcd2ab936eae6c4e6f. The 256-episode

record preserves the exact command, full package freeze, source-data and target hashes, and successful strict-load and numerical gates.

C.1.4 Asset licenses and terms of use

We use publicly released research assets, cite their original sources, and do not redistribute third-party datasets or checkpoints in this paper. Table 11 summarizes the licenses or stated access terms verified for the assets used in the reported experiments.

Table 11: Existing assets used in the experiments and their verified licenses or stated terms.

Asset	Use in this paper	License or stated terms
V-JEPA2-G checkpoint	Main IntPhys2 representation result and Kubric rerender pilot	Released Meta checkpoint; the corresponding Giant Hugging Face model card lists Apache-2.0.
V-JEPA2-L checkpoint	Local Physion++ static/surprise diagnostic	Hugging Face Large model card lists MIT.
V-JEPA2 code	Feature extraction and V-JEPA prediction utilities	Official repository is primarily MIT, with specified utility files under Apache-2.0.
VideoMAE V2 code and VideoMAE2-G checkpoint	IntPhys2 comparator	Official VideoMAE V2 code is MIT; the exact OpenGVLab/VideoMAEv2-giant Hugging Face checkpoint/model card is CC BY-NC 4.0 and is used only for noncommercial research evaluation.
IntPhys2	Evaluation videos and pair labels	CC BY-NC 4.0; stated terms restrict use to evaluation and prohibit generative uses. Our use is benchmark evaluation/probing.
LeWorldModel code, checkpoints, and task data	TwoRoom and PushT native-error controls	LeWorldModel source is MIT; the quentinll LeWM TwoRoom and PushT model/data cards list MIT.
DINO-WM code and PushT checkpoint	Matched PushT native-error control	The official DINO-WM source repository includes an MIT license; the released OSF PushT checkpoint is used only for research evaluation and is not redistributed.
JEPA-WM code and PushT checkpoint	Post-hoc matched PushT native-error control	The official JEPA-WM source and model release list CC BY-NC 4.0; the frozen DINOv2 source is Apache-2.0. Checkpoints are used only for noncommercial research evaluation and are not redistributed.
stable-worldmodel and PushT environment stack	Environment loading and PushT rendering for LeWM controls	stable-worldmodel is MIT; gym-pusht, which the PushT environment is based on, is Apache-2.0.
Kubric/MOVi-A generator	Exact rerendering pilot	Kubric code is Apache-2.0; generated MOVi-A-style pilot videos are internal evaluation artifacts.
Physion++	Local diagnostic videos in Section 4.1	The project page publicly provides download links but states a license only for the website source; no separate dataset license file was found in the local archive. We use it only for local research evaluation and cite the original work.

C.2 IntPhys2 representation reproducibility

C.2.1 Representation readouts and feature pipelines

The corrected IntPhys2 sealed test contains 114 matched possible/impossible pairs from 57 held-out scenes, 228 source videos, and 912 appearance-conditioned samples. Each sample uses exactly 64 decoded source frames sampled by rounded linear spacing. The four appearance conditions are clean, 90-degree hue rotation, brightness multiplier 0.45, and 180-degree selective recoloring. V-JEPA2-G and VideoMAE2-G are extracted at five frozen depths (block_00, block_10, block_20, block_30, final); PixelSummary is a 100-dimensional low-level baseline of RGB summary, color-histogram, and frame-difference statistics. The probe is StandardScaler plus balanced logistic regression with $C = 1.0$ and at most 2000 iterations. The paper-facing post-hoc correction hash-orders scenes with seed 0 into five folds; fold index 0–4 is also the classifier random state. Fold k is test, fold $(k + 1) \bmod 5$ is validation, and the remaining folds are training. Sealed test folds contain 12, 12, 11, 11, and 11 scenes; development test folds contain 8, 8, 8, 7, and 7, with all pair IDs and appearance variants from one scene held together. Primary layers preserve the original development choices: V-JEPA2-G block_20 for appearance and block_30 for physics, VideoMAE2-

552 G block_30 for appearance, block_00 for all-appearance physics, and block_30 for clean-only
 553 physics, with PixelSummary using its only layer. The cross-fit development check retains V-JEPA2-
 554 G block_20 for appearance and block_30 for physics_all; its all-appearance comparator deltas
 555 include zero. Table 12 summarizes the model-specific input and temporal pipelines.

Table 12: Checkpoint-family feature pipelines. The comparison does not isolate training objective or architecture.

Family	Input	Native window	Patch/tubelet	Layers	Temporal scope
V-JEPA2-G	64×256^2	64 frames	16/2	0,10,20,30,40	one joint 64-frame clip
VideoMAE2-G	$4 \times 16 \times 224^2$	16 frames	14/2	0,10,20,30,40	four independent chunks
PixelSummary	64 frames	n/a	n/a	one	100 RGB/histogram/difference statistics

556 Both neural pipelines receive the same decoded frames, use ImageNet RGB normalization, and
 557 expose 1408-dimensional pooled features at matched relative depths. V-JEPA2-G resizes without
 558 cropping; VideoMAE2-G uses its pinned short-side resize and center crop. The shared coarse token
 559 grid is $4 \times 4 \times 4$, obtained by mean pooling native spatiotemporal tokens. Neural extraction uses
 560 bfloat16 autocast and stores float16 features.

561 C.2.2 Superseded original split audit

562 The corrected development and sealed populations are scene-disjoint. The historical pair-held-out
 563 probe splitter, however, hashed pair IDs rather than scene IDs. All appearances and both pair mem-
 564 bers remained together, but other pairs from the same scene could cross folds. Table 13 reports a
 565 mechanical join of those superseded split assignments to the sample manifest. "Any overlap" counts
 566 scenes occurring in more than one fold.

Table 13: Scene overlap in the five superseded original pair-held-out probe splits.

Seed	Train	Val	Test	Tr-val	Tr-test	Val-test	Any overlap
0	46	22	18	16	8	5	29
1	51	18	23	12	22	1	35
2	50	21	21	16	16	3	35
3	45	20	21	10	12	7	29
4	47	21	20	14	12	5	31

567 A subsequent repeated scene-held-out correction kept scenes disjoint within each split but yielded
 568 60 test scene-fold slots from only 42 unique sealed scenes: 15 scenes were never tested, while 15
 569 recurred two or three times. Its point estimates therefore weighted scenes by test frequency, and
 570 its within-fold bootstrap did not preserve dependence across recurring scenes. Those outputs are
 571 retained only as a superseded audit trail.

572 C.2.3 Sealed-test integrity

573 Before creating the corrected split, we excluded the union of every inspected selection: 193 pairs
 574 across 158 scenes. The remaining Main partition was divided by scene into development and sealed
 575 test; the sealed set was exposed only after freezing the original extraction, layer, and probe specifica-
 576 tion. The five-fold cross-fit was designed after sealed-population inspection and is a post-hoc correc-
 577 tion outside the frozen specification. The original specification SHA-256 is f7e0ca1651c9586c
 578 180ba89c8be347d02d99de9045a8b1b9b75ed6d9fac7b465; the sealed sample-manifest SHA-
 579 256 is 1ddc3677781e605f6d298992584768e414801dc1479b03184150dd2af11fd32a. Both
 580 model extractors produced 912 finite shards, repeated the first sample with maximum absolute dif-
 581 ference 0, triggered no out-of-memory fallback, and passed cross-model parity with zero missing,
 582 extra, or metadata-mismatched samples.

583 C.2.4 Estimand, uncertainty, and geometry

584 For physics, correctness is averaged over samples within each out-of-fold pair and then over both
 585 pairs in each unique scene. Each of 10,000 bootstrap draws samples all 57 sealed scene values glob-
 586 ally with replacement. Probe predictions remain fixed: neither scaler nor classifier is refit. Model
 587 deltas preserve fold, scene, and pair alignment before scene aggregation and paired resampling.
 588 These intervals represent global scene-clustered fixed-prediction uncertainty, but do not include
 589 probe-refitting or fold-assignment variability. Near- or below-chance comparator estimates are not
 590 interpreted as meaningful anti-physics signals.

591 **Geometry estimands and nulls.** Let $\hat{z}_i^{(\ell)} = z_i^{(\ell)} / \|z_i^{(\ell)}\|_2$ and $d(x, y) = 1 - \cos(x, y)$. The
 592 unrelated clean scale and relative displacement magnitude are

$$G_\ell = \text{median}_{i,j: s_i \neq s_j} d(\hat{z}_i^{\text{clean}}, \hat{z}_j^{\text{clean}}), \quad R_\ell(T; i) = \frac{d(\hat{z}_i^{\text{clean}}, \hat{z}_i^T)}{G_\ell},$$

593 with the same G_ℓ used for matched physical pairs. For $\Delta_i^T = \hat{z}_i^T - \hat{z}_i^{\text{clean}}$ and non-negligible
 594 $u_i^T = \Delta_i^T / \|\Delta_i^T\|_2$, directional coherence is the standard mean resultant length [Mardia and Jupp,
 595 1999],

$$\text{MRL}_\ell(T) = \left\| N^{-1} \sum_i u_i^T \right\|_2, \quad \Delta \text{MRL} = \text{MRL}_{\text{obs}} - \mathbb{E}[\text{MRL}_{\text{shuffle}}].$$

596 Our operational control shuffles source–transform correspondence only across scenes, retaining the
 597 factor, model, and layer. An orthonormal appearance basis $U_{\text{app},k}^{(\ell)}$ is obtained by standard SVD/PCA
 598 [Jolliffe and Cadima, 2016] on training-scene displacements. For held-out $\Delta_i^{\text{phys}} = \hat{z}_i^{\text{impossible}} -$
 599 $\hat{z}_i^{\text{possible}}$,

$$P_{\text{app},k}^{(\ell)}(i) = \frac{\|(U_{\text{app},k}^{(\ell)})^\top \Delta_i^{\text{phys}}\|_2^2}{\|\Delta_i^{\text{phys}}\|_2^2}.$$

600 Five-fold scene cross-fitting excludes evaluation scenes when fitting $U_{\text{app},k}$; dimension-matched
 601 Haar-random subspaces provide the reference. The primary $k = 4$ was retained from develop-
 602 ment before sealed effects were observed. Every unrelated pair uses different scenes, possible-
 603 to-impossible orientation is fixed, and paper-facing intervals/nulls use 10,000 scene-level draws.
 604 Principal-angle diagnostics follow standard subspace comparison [Björck and Golub, 1973]; tiny-
 605 displacement, raw-feature, local-normalization, and anisotropy checks remain supplementary.

606 An initial sealed-package integrity run aborted before any representation-derived geometry metric
 607 was computed because an order-sensitive validator compared equivalent condition lists. Before any
 608 sealed geometry outcome was observed, the validator was changed to strict order-insensitive multiset
 609 equality and regression-tested; no metric or statistical code changed. Development reproduction then
 610 passed; Tables 14, 15, and 16 report the completed frozen geometry run.

Table 14: Complete sealed magnitude and cross-fitted projection curves. Magnitude cells are hue/brightness/recolor/clean-physical, each divided by G_ℓ . Projection is $P_{\text{app},4}$ with a dimension-matched random mean near 0.00284 at every layer; brackets are 95% scene-bootstrap intervals.

Model	Layer	G_ℓ	Relative magnitude H/B/R/P	$P_{\text{app},4}$ [95% CI]
V-JEPA2-G	b00	.607285	.571/2.111/.610/.007	.401 [.320, .483]
	b10	.007055	.392/1.634/.449/.007	.125 [.078, .178]
	b20	.006974	.260/1.022/.245/.010	.063 [.039, .090]
	b30	.015104	.110/1.497/.142/.012	.102 [.065, .145]
	final	.235825	.080/.077/.090/.021	.060 [.050, .071]
VideoMAE2-G	b00	.010641	1.245/1.750/1.310/.005	.348 [.268, .432]
	b10	.017612	.347/.542/.366/.005	.057 [.036, .083]
	b20	.031948	.195/.291/.226/.008	.061 [.050, .076]
	b30	.028516	.137/.258/.161/.007	.060 [.048, .076]
	final	.057018	.034/.115/.065/.006	.077 [.063, .092]

Table 15: Development-to-sealed trajectory classification. Metrics and labels were fixed before the completed sealed geometry run.

Property	V-JEPA2-G	VideoMAE2-G
Relative appearance magnitude	strong replication	strong replication
Null-adjusted direction	strong replication	strong replication
Appearance–physics overlap	qualitative replication	strong replication

Table 16: Final-layer sealed projection sensitivity. Cells report observed $P_{\text{app},k}$ [95% scene-bootstrap interval] / random mean. The paper-facing $k = 4$ remained fixed.

Model	$k = 1$	$k = 2$	$k = 4$	$k = 8$
V-JEPA2-G	.005 [.004,.006]/.0007	.012 [.010,.014]/.0014	.060 [.050,.071]/.0028	.153 [.132,.174]/.0057
VideoMAE2-G	.024 [.016,.034]/.0007	.058 [.046,.073]/.0014	.077 [.063,.092]/.0028	.130 [.113,.148]/.0057

611 **Directional robustness.** Raw MRL is insufficient: the shuffled V-JEPA2-G block-00 brightness
612 population has first-PC energy 0.886, showing strong background anisotropy. Figure 2 therefore
613 reports ΔMRL ; Table 17 gives full curves. Major observed MRL effects are not driven by one
614 scene: leave-one-scene-out ranges are [0.781, 0.797] for V-JEPA2-G hue at block 30 (full 0.784),
615 [0.953, 0.961] for brightness at block 30 (0.954), and [0.708, 0.720] for recoloring at block 30
616 (0.712). Final projection is likewise stable at 0.060 with leave-one-scene-out range [0.057, 0.060].

Table 17: Complete sealed directional curves. Each cell is observed MRL / shuffled-correspondence mean / difference. H, B, R, and P denote hue, brightness, selective recoloring, and clean possible-to-impossible displacement. MRL intervals and empirical permutation values are in the supplementary tables.

Model	Layer	Hue	Brightness	Recolor	Physics
V-JEPA2-G	b00	.651/.466/.186	.835/.785/.050	.427/.420/.007	.065/.066/-.002
	b10	.714/.384/.329	.953/.741/.212	.499/.330/.168	.064/.066/-.002
	b20	.758/.349/.409	.941/.660/.281	.615/.282/.333	.088/.066/.022
	b30	.784/.250/.534	.954/.699/.255	.712/.250/.462	.119/.066/.052
	final	.693/.212/.481	.792/.236/.556	.564/.180/.383	.134/.067/.067
VideoMAE2-G	b00	.666/.499/.167	.962/.763/.199	.449/.461/-.012	.070/.067/.003
	b10	.720/.380/.340	.947/.578/.369	.578/.323/.254	.079/.065/.014
	b20	.742/.310/.432	.942/.465/.477	.658/.283/.375	.069/.065/.004
	b30	.747/.270/.477	.955/.450/.506	.693/.257/.436	.078/.065/.012
	final	.704/.139/.566	.965/.327/.637	.783/.191/.592	.075/.065/.010

617 C.3 Post-sealed exploratory transformation trajectories

618 **Status: post-sealed exploratory.** This analysis does not alter or replace the development-frozen
619 sealed geometry result. It applies the exact paper V-JEPA2-G checkpoint to corrected development
620 (38 scenes, 76 pairs, 152 source videos) at the five reported depths. Each source has clean video
621 plus hue rotations of 15° , 45° , and 90° ; brightness multipliers 0.85, 0.65, and 0.45; and selective
622 recoloring rotations of 45° , 90° , and 180° . Brightness strength is $1 - b$. Scene is the independent
623 unit and intervals use 10,000 bootstrap draws. We use these trajectories to probe transformation
624 geometry and local equivariance descriptively [Lenc and Vedaldi, 2015].

625 For an ordered trajectory $z(a_0), \dots, z(a_m)$, straightness is endpoint distance divided by path length,
626 $\|z(a_m) - z(a_0)\|_2 / \sum_j \|z(a_{j+1}) - z(a_j)\|_2$. Displacement dimensionality uses entropy effective
627 rank, $\exp[-\sum_j p_j \log p_j]$, where p_j is normalized squared singular-value energy. These quantities
628 describe ray-likeness and dimensionality; neither identifies a semantic axis.

629 Figure 4 shows that at final, brightness is the most ray-like evaluated factor: straightness is 0.907
630 [0.899, 0.914], versus 0.692 [0.673, 0.710] for hue and 0.553 [0.520, 0.587] for selective recoloring.
631 Final effective ranks are 5.43, 9.83, and 13.72, respectively. Thus selective recoloring is higher-

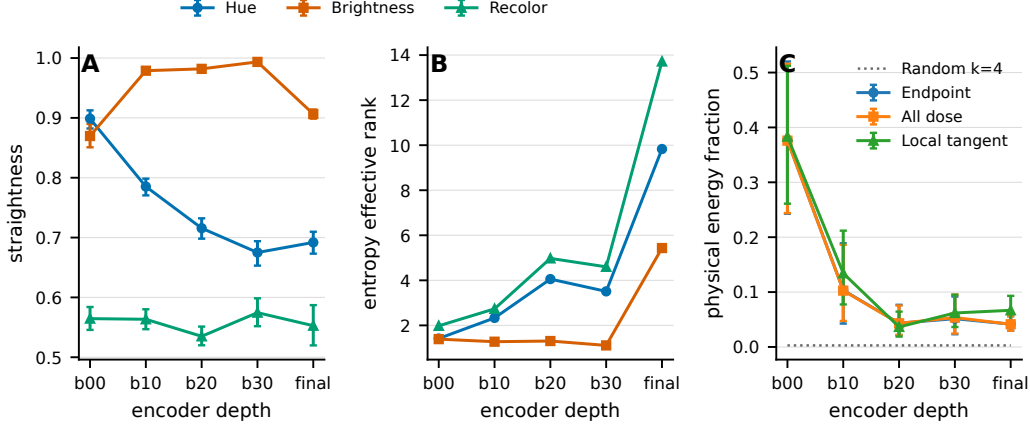


Figure 4: **Post-sealed exploratory** V-JEPA2-G multi-dose geometry on corrected development. (A) Scene-mean trajectory straightness with 95% scene-bootstrap intervals. (B) entropy effective rank of the pooled scene-dose displacement matrix. (C) five-fold held-out physical projection into strongest-endpoint, all-dose, and local-tangent appearance spaces, all matched at $k = 4$, with scene-bootstrap intervals and a dimension-matched random reference. The tangent advantage is layer dependent and does not replace the frozen sealed projection result.

632 dimensional than the other evaluated factors, while brightness remains only *relatively* ray-like and
633 has more than one effective dimension.

634 For hue, held-out relative error is 0.798 for a shared-ray model and 0.747 for a harmonic curved
635 model. This is modest evidence that the observed 0° – 90° arc is better represented by a curved model
636 than one common ray, but the final effective rank of 9.83 makes the low-dimensional part of the
637 hypothesis only partial. Three nonzero doses spanning at most 90° cannot test closure, periodicity,
638 circularity, or a latent color wheel.

639 The strongest new exploratory result is the matched-dimensional local comparison. At final and
640 $k = 4$, held-out physical projection is 0.0416 [0.0297, 0.0559] into the strongest-endpoint appear-
641 ance space, 0.0414 [0.0295, 0.0564] into the all-dose space, and 0.0668 [0.0448, 0.0932] into the
642 local appearance-tangent space; the random reference is approximately 0.00284. The paired tangent-
643 minus-endpoint difference is 0.0252 [0.0072, 0.0450]. The advantage is not universal across depth:
644 at block 20, tangent projection (0.037) is below endpoint projection (0.043). The final result there-
645 fore says only that a matched- k local appearance-tangent basis captures more held-out physical-
646 displacement energy than the strongest-endpoint basis at that layer; it does not establish entangle-
647 ment, causal factorization, or physical understanding.

Table 18: Qualitative V-JEPA2-L pilot to exact V-JEPA2-G replication check at the final representa-
tion. The models differ in scale and the G run uses the corrected 38-scene development population;
this is not a causal model-size comparison.

Factor	L straightness	G straightness	L effective rank	G effective rank
Hue	.739	.692	3.61	9.83
Brightness	.887	.907	2.98	5.43
Selective recolor	.324	.553	5.94	13.72

648 Table 18 gives the qualitative classifications: factor-specific geometry replicated; brightness as the
649 most ray-like factor replicated; recoloring as higher-dimensional replicated; hue as curved and low-
650 dimensional replicated only partially; and final tangent projection exceeding endpoint projection
651 replicated. Table 19 reports direction by dose. These labels provide a post-sealed cross-check and
652 do not identify causal effects of model scale.

Table 19: Post-sealed exploratory final-layer direction by dose. Each row is observed MRL / shuffled mean / Δ MRL.

Factor	Mild dose	Medium dose	Strong dose
Hue ($15^\circ/45^\circ/90^\circ$)	.640/.064/.576	.685/.160/.525	.734/.234/.500
Brightness (.85/.65/.45)	.810/.081/.729	.830/.164/.667	.816/.257/.559
Recolor ($45^\circ/90^\circ/180^\circ$)	.598/.138/.459	.636/.195/.442	.625/.213/.412

653 C.4 Additional experiment protocols

654 C.4.1 Physion++ static audit

655 The V-JEPA2-L static audit used 10 Physion++ videos and 273 complete 16-frame windows. Raw
656 V-JEPA2-L surprise was moderately associated with pixel motion ($r = 0.596$), but 6/10 videos
657 also contained a high-surprise near-static window with mean consecutive RGB difference below
658 0.01. All six near-static candidates had nonpositive original-minus-frozen-frame residuals, with
659 mean residual -0.0350 ; Appendix Sections A.2–A.3 give the protocol and candidate-level values.

660 C.4.2 LeWorldModel, DINO-WM, and JEPA-WM controls

661 The TwoRoom and PushT studies strict-load the released LeWorldModel checkpoints, preserve the
662 official four-frame, frame-skip-five preprocessing and action normalization, and use 256 fresh in-
663 dependent episodes per task. For each episode we simulate lawful targets from the third context
664 state under scaled five-action blocks ca for $c \in \{0, 0.25, 0.5, 0.75, 1\}$, render the lawful target once,
665 and reuse the exact same target bytes under lawful ca , zero, and opposite $-ca$ action blocks. The
666 LeWM native score is the model’s prediction MSE. The DINO-WM PushT extension uses the same
667 256 PushT episodes, target bytes, action blocks, and rendered-effect matches; it normalizes actions
668 and proprioception with the released DINO-WM PushT statistics and scores final-step visual-latent
669 MSE between the predictor output and encoded target. The JEPA-WM PushT extension uses the
670 same controlled construction and three-frame teacher-forcing context with the released 224-pixel
671 preprocessing, action and proprioception statistics, AdaLN plus RoPE predictor, and final one-step
672 visual embedding MSE over DINOv2 patch tokens. The visual score excludes proprioceptive pre-
673 diction loss, and the analysis is post-hoc exploratory. Action contrasts are same-target differences
674 relative to the lawful-action baseline; appearance contrasts rotate the static target hue and com-
675 pare against predeclared rendered-effect matches. Episode-level 95% intervals use 10,000 bootstrap
676 draws over episodes. Appendix Sections B.1–B.3 separately test matched color directions and de-
677 compose their latent source.

Table 20: Complete native prediction-error contrasts (256 episodes per task or scorer; 95% episode-bootstrap intervals). LeWM rows use native prediction MSE; DINO-WM PushT rows use final-step visual-latent MSE. JEPA-WM PushT rows use final one-step visual embedding MSE and are post-hoc exploratory. Native magnitudes are not comparable across models. Action rows are relative to lawful action; hue rows to the matched opposite action.

Model/task	Contrast	Low or matched scale	High or matched scale
LeWM TwoRoom	zero vs. lawful	0.359 [0.315, 0.403] at 0.25	1.463 [1.382, 1.545] at 1.0
LeWM TwoRoom	opposite vs. lawful	0.909 [0.834, 0.987] at 0.25	1.884 [1.821, 1.946] at 1.0
LeWM TwoRoom	hue vs. opposite	0.222 [0.150, 0.296] at $120^\circ/0.75$	0.220 [0.147, 0.295] at $180^\circ/1.0$
LeWM PushT	zero vs. lawful	0.0085 [0.0070, 0.0102] at 0.25	0.1668 [0.1469, 0.1880] at 1.0
LeWM PushT	opposite vs. lawful	0.0332 [0.0279, 0.0391] at 0.25	0.3834 [0.3456, 0.4242] at 1.0
LeWM PushT	hue vs. opposite	0.146 [0.125, 0.167] at $30^\circ/0.25$	0.642 [0.585, 0.696] at $60^\circ/0.75$
DINO-WM PushT	zero vs. lawful	0.00052 [0.00046, 0.00059] at 0.25	0.00183 [0.00152, 0.00215] at 1.0
DINO-WM PushT	opposite vs. lawful	0.00104 [0.00091, 0.00118] at 0.25	0.00364 [0.00304, 0.00428] at 1.0
DINO-WM PushT	hue vs. opposite	$-0.03082 [-0.04028, -0.02273]$ at $30^\circ/0.25$	$-0.10987 [-0.12416, -0.09614]$ at $60^\circ/0.75$
JEPA-WM PushT	zero vs. lawful	0.00049 [0.00042, 0.00056] at 0.25	0.00179 [0.00151, 0.00209] at 1.0
JEPA-WM PushT	opposite vs. lawful	0.00098 [0.00084, 0.00112] at 0.25	0.00351 [0.00293, 0.00411] at 1.0
JEPA-WM PushT	hue vs. opposite	$-0.03670 [-0.04434, -0.03008]$ at $30^\circ/0.25$	$-0.12481 [-0.13905, -0.11114]$ at $60^\circ/0.75$

678 C.4.3 Kubric exact rerendering

679 The Kubric pilot uses V-JEPA2-G with eight accepted MOVi-A scenes rendered from saved simula-
680 tor trajectories. The first 12 frames are identical context bytes across all conditions, and the target
681 is the remaining 36 frames represented as 18 two-frame tubelets. Appearance conditions rotate only
682 the target object’s post-event base color; physical conditions apply a persistent target-object $+z$ dis-
683 placement of 0.25, 0.50, 1.00, or 1.50 object diameters while preserving non-target states. The scalar
684 score is mean absolute error between predictor output and layer-normalized EMA-target-encoder to-
685 kens. We report scene-level paired differences from clean with 10,000 bootstrap draws; with eight
686 scenes, the matched hue-versus-displacement contrasts are exploratory and used only to test scalar
687 specificity under exact state and rendering controls. The two predeclared physical minus appearance
688 contrasts are 0.000259 $[-0.000632, 0.001145]$ for a 120° hue change versus a displacement of 0.25
689 object diameters, and 0.000201 $[-0.000791, 0.001035]$ for a 180° hue change versus a displacement
690 of 0.50 object diameters. Both intervals include zero.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The title, abstract, results, and limitations distinguish the post-hoc IntPhys2 readout, development-frozen sealed geometry, and post-sealed exploratory trajectory analysis; no checkpoint establishes every construct confirmatorily. The JEPA-WM extension is also identified as post-hoc exploratory.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Section 6 states the uncertainty boundary of the post hoc IntPhys2 readout, the development-selected and exploratory scope of the geometry analyses, the factors confounded across model comparisons, the missing static and motion baselines, and the need for a larger Kubric study. It specifies preregistered confirmation, shared task evaluation, and a rendered effect matched lawful versus impossible comparison as next tests.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper introduces a measurement framework and empirical controls, but no theorem or formal proof.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 3 and Appendix C.2 specify the cross-fit ordering and rotation, probe, exactly-once held-out scene coverage, pair/scene aggregation, global fixed-prediction scene bootstrap, feature construction, unrelated-scale normalization, shuffled directional null, cross-fitted subspace projection, and score definitions.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The anonymous supplementary archive at <https://anonymous.4open.science/r/Latent-Surprise-Audit-8730/README.md> provides source, manifests, processed metrics, integrity records, and a smoke script. The SHA-256 of `archive_manifest.json` is 56f9a8a7756806d917c075e0cb336181b8b84663689f57521773402407900b75. Third-party datasets/checkpoints are not redistributed and must be obtained under their licenses or access terms.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Appendix Sections C.2, C.3, and C.4 give the probe classifier, scene cross-fitting, pre-served layers, transformations, geometry nulls and subspaces, rendered-effect matching, bootstrap units, LeWM and DINO-WM scores, the post-hoc JEPA-WM score and model-native preprocessing, and Kubric score construction.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

744 Justification: Primary inferential results report 95% confidence intervals. The post-hoc IntPhys2
745 correction holds all 57 sealed scenes out exactly once, aggregates each scene’s pair predictions, and
746 bootstraps those unique scene values globally; paired deltas preserve scene and pair support. These
747 intervals condition on fitted probes and omit refitting and fold-assignment variability. Development-
748 frozen geometry uses scene-level intervals plus 10,000 shuffled and random-subspace draws. LeWM,
749 DINO-WM, and JEPA-WM intervals bootstrap episodes, and Kubric intervals bootstrap scenes.

750 **8. Experiments compute resources**

751 Question: For each experiment, does the paper provide sufficient information on the computer re-
752 sources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

753 Answer: [Yes]

754 Justification: Tables 9 and 10 report hardware, software, precision, and measured wall times for the
755 paper-facing experiment stages, including the exact multi-dose V-JEPA2-G extraction. Supplemen-
756 tary provenance records worker memory, workloads, peak accelerator memory where captured, tim-
757 ing semantics, and evidence paths; remaining untimed deterministic CPU postprocessing is explicitly
758 delimited.

759 **9. Code of ethics**

760 Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS
761 Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

762 Answer: [Yes]

763 Justification: This work evaluates public datasets/checkpoints and generated simulator variants, con-
764 tains no human-subject data collection, and is framed as measurement rather than deployment.

765 **10. Broader impacts**

766 Question: Does the paper discuss both potential positive societal impacts and negative societal im-
767 pacts of the work performed?

768 Answer: [Yes]

769 Justification: The broader impact paragraph at the end of Section 6 identifies potential benefits from
770 using validated residuals as anomaly, reward, and safety signals, and the corresponding risk that such
771 signals will not be reliable or transferable outside the task, environment, and nuisance conditions in
772 which they were measured.

773 **11. Safeguards**

774 Question: Does the paper describe safeguards that have been put in place for responsible release of
775 data or models that have a high risk for misuse (e.g., pre-trained language models, image generators,
776 or scraped datasets)?

777 Answer: [N/A]

778 Justification: The paper does not release a high-risk model, scraped dataset, or deployed system
779 requiring safeguards.

780 **12. Licenses for existing assets**

781 Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper,
782 properly credited and are the license and terms of use explicitly mentioned and properly respected?

783 Answer: [Yes]

784 Justification: Appendix C.1.4 lists verified licenses and stated terms, including the unresolved Phys-
785 ion++ dataset-license boundary; assets are cited and used for research evaluation.

786 **13. New assets**

787 Question: Are new assets introduced in the paper well documented and is the documentation provided
788 alongside the assets?

789 Answer: [N/A]

790 Justification: The current paper does not release a new public dataset, model, or benchmark asset;
791 generated artifacts are used as experiment records.

792 **14. Crowdsourcing and research with human subjects**

793 Question: For crowdsourcing experiments and research with human subjects, does the paper include
794 the full text of instructions given to participants and screenshots, if applicable, as well as details about
795 compensation (if any)?

796 Answer: [N/A]

797 Justification: The work does not involve crowdsourcing or human subjects.

798 **15. Institutional review board (IRB) approvals or equivalent for research with human subjects**
799 Question: Does the paper describe potential risks incurred by study participants, whether such risks
800 were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equiv-
801 alent approval/review based on the requirements of your country or institution) were obtained?
802 Answer: [N/A]
803 Justification: The work does not involve crowdsourcing or human subjects, so IRB approval is not
804 applicable.
805 **16. Declaration of LLM usage**
806 Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard
807 component of the core methods in this research? Note that if the LLM is used only for writing, editing,
808 or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of
809 the research, declaration is not required.
810 Answer: [N/A]
811 Justification: LLMs were not used as a core methodological component; any language-model assis-
812 tance was limited to writing or editing and does not affect the scientific method.