VERIFIABLE ≠ VALID: WHAT PROGRAMMATIC PHYSICS REWARDS FOR VIDEO GENERATION ACTUALLY OPTIMIZE

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training video generators with verifiable physics rewards, meaning programmatic quantities read deterministically off frozen perception models, has become a leading alternative to VLM judges. Such rewards are validated by showing that they separate real video from rule-corrupted video. We find that this validation systematically overstates their validity, that what the two rewards we could measure actually optimize is motion magnitude, and that removing the motion confound is necessary but far from sufficient. Across two human-rated datasets, 13 generators and two architectures, a reimplemented kinematic-residual reward correlates with peak optical flow at $\rho = 0.85$–0.92. Once motion is partialled out it retains no association with human ratings in 0 of 12 generators under a single motion control, and a pure motion statistic containing no physics predicts human ratings better than it does. Optimization reproduces this causally: in two GRPO runs identical but for the reward, the motion-correlated reward drives peak flow down 46% (95% CI $[-58, -28]$) while a motion-orthogonal reward drives it up 67%, the arms separating monotonically in dose at $p = 10^{-8}$, with three motion-orthogonal signals pushed in no consistent direction; a lower-fidelity second prompt set replicates it, with peak flow down 88% at the highest dose and the arms separating at all five doses. The identical test applied to a VLM physics judge answers oppositely: the judge retains 88.0% of its agreement with human ratings once motion is partialled out, against -1.7% for the programmatic reward. **The dividing line is therefore not verifiability but whether the quantity is a motion statistic.** Escaping the confound is not sufficient either: the most structured physics reward published to date, which grounds recovered 3D bodies in a physics simulator, does escape it ($\rho = 0.18$ with motion, against 0.92) yet, tested outside its design domain, carries no measurable validity beyond motion on this rating (partial $\rho = -0.033$, $p = 0.26$). We stop short of the last step. We do not show that the invalid reward makes videos worse for people: we have no human ratings on our own generations, the two runs disagree about the automatic judge, and treating a rising automatic physics score as evidence about human judgement is the very inference this paper is written against.

1 INTRODUCTION

Post-training video generators against *verifiable* physics rewards has become a leading alternative to using a vision–language model as a judge. A reward computed programmatically from frozen perception models is deterministic, inspectable and cheap, and does not require one black box to grade another. Such rewards read physical quantities off optical flow, tracked keypoints, masks or a recovered 3D body, and score a clip by how far they depart from a law of motion.

Validation in this literature rests on a separation test: show that the reward distinguishes real video from video a rule has corrupted by reversing, shuffling or interpolating its frames. That test poses an easier question than the one that matters. Our own trajectory residual separates real clips from rule-corrupted ones at AUC 0.7422, and then ranks the clips a generator actually produces at AUC 0.5381, which is the distribution a reward would be optimized against. The gap is 0.2041 $[0.028, 0.366]$, reported as an interval because its corrupted-versus-real side rests on only 15 real clips (Appendix L).

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.

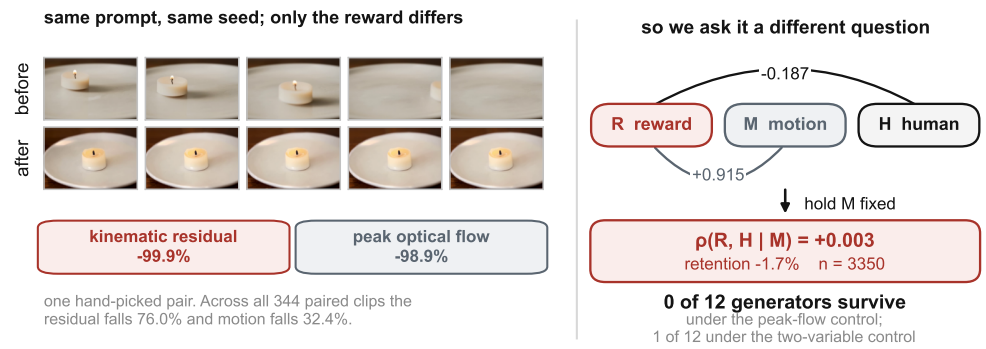


Figure 1: **A verifiable physics reward can be driven almost to zero by a clip in which nothing happens. Left:** one prompt under CogVideoX-5B (Yang et al., 2025) and under VideoREPA (Zhang et al., 2025), same seed, five frames taken at the same indices from each clip. This pair was chosen for its size and is an illustration, not evidence; the population contrast printed beneath it is what carries the claim. **Right:** the question we ask instead of the usual separation test, and its answer on VideoPhy-2. Section 3 defines the protocol and Section 4 reports it across two datasets and twelve generators.

What the dominant family measures instead is motion magnitude. A reimplementation of a published kinematic-residual reward correlates with peak optical-flow magnitude at $\rho = 0.85$–0.92 across five independent sources spanning 13 generators and two architectures (Appendix K). Partialling motion out removes its association with human judgement: under a single motion control no generator retains any association on either dataset, and under a two-variable control one of twelve does. The relationship is asymmetric, since motion survives controlling for the reward while the reward does not survive the converse, which is what licenses reading it directionally. A statistic with no physics content in it at all, peak optical flow, is the *better* predictor of human physical-plausibility ratings. That motion might confound such rewards has been noted once before, in an appendix; what we add is not the observation but a protocol any reward can be run through and the causal evidence that follows. The benchmark we work on reports the opposite conclusion, and Appendix C shows the answer turns on *which statistic of the flow field* is taken.

Correlation is not the interesting part; what happens when the reward is optimized is. We post-train with GRPO under two rewards that are identical in every respect except the reward itself, one correlated with motion at $\rho \approx 0.9$ and one at 0.0035, and measure five checkpoints against a shared, per-clip-paired baseline (Figure 2). The motion-correlated reward drives peak flow down 46.3% while the motion-orthogonal reward drives it *up* 66.6%; the arms separate at $p = 10^{-8}$, monotonically in dose. Three motion-orthogonal physical signals are not pushed in any consistent direction, stated as a bound rather than as a null: any systematic change in them sits within about 18%, an interval that does not overlap the motion interval. A second run on a harder prompt family replicates and strengthens the dissociation. We are careful about what this does not show. Human ratings do rise as motion falls over the range anyone has measured, and optimization drives the policy past the edge of that range, where the relationship has never been observed. Our claim is that the extrapolation is unwarranted, not that its direction is known.

Removing the confound turns out to be necessary and far from sufficient. Run on the most structured physics reward published to date, with its authors' own code, the protocol finds a reward that genuinely escapes the confound ($\rho = 0.18$ with motion, against 0.92), which already locates the problem in motion-proxy rewards rather than in verifiability as such, but that on this rating carries no information beyond motion either. It is built for human-motion realism and we test it on general physical commonsense, so we do not claim it fails at its own task. Every motion-orthogonal signal we could construct is similarly weak, and a reader is entitled to suspect the protocol is simply unpassable.

It is not. No benchmark in this area turns the question back on its own evaluator; we do, with the identical protocol on the same clips. The judge retains 88.0% of its agreement with human ratings after motion is partialled out, significant in 6 of 7 generators, against -1.7% for the reward, and it

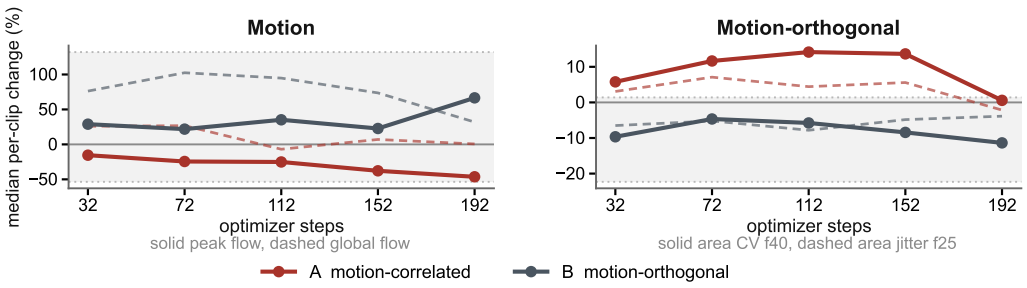


Figure 2: **Optimizing a verifiable physics reward moves the policy along a motion axis; a motion-orthogonal reward does not.** Two GRPO arms identical except for the reward, five doses, 80 per-clip-paired clips per point. The dotted band is what no intervention looks like, the same prompt at a different seed: a magnitude reference, not a significance test. Appendix J adds the manipulation check and the judge scores.

replicates on a second dataset of 9569 clips from five different generators. On the same rows, pure motion significantly beats the reward and significantly loses to the judge. **The dividing line is not verifiability. It is whether the quantity is a motion statistic.**

Contributions. A motion-orthogonal **validity protocol** any reward can be run through; a **diagnosis** with it across two human-rated datasets and 13 generators; **causal evidence from the optimization side**, a double dissociation over five doses replicated on a second prompt family; the same protocol applied to **two published rewards** and then **turned on a VLM judge**, which answers oppositely; and an account of **why this was not already noticed**.

2 Related Work

Verifiable physics rewards for video post-training. The family this paper audits computes a scalar from frozen perception models and uses it as a post-training reward. The two we can measure are NewtonRewards (Le et al., 2025), which reads velocity off optical flow and mass off appearance features and penalises departures from Newtonian structure, positioning itself explicitly against human and VLM feedback, and PhyMotion (Huang et al., 2026), which recovers 3D human bodies, retargets them into a physics simulator, and scores kinematic plausibility, contact and dynamic feasibility. Several of the rest score against a reference, and a reference-defined reward cannot be computed at all on unpaired open-domain generations, the setting in which a reward has to rank a policy's own samples. That is what bounds our coverage to two. Appendix I places the family in full, and Section 5 treats the pattern as a finding rather than as an excuse.

Physics benchmarks and their automatic evaluators. VideoPhy (Bansal et al., 2025a) and VideoPhy-2 (Bansal et al., 2025b) collect human physical-commonsense ratings and train automatic evaluators to extend them, and Physion-Eval (Zhang et al., 2026a) collects reasoning-grounded severity annotations over newer generators. A dozen further benchmarks build finer-grained physics evaluations, and Appendix I lists them. Two things are common to all of them: the automatic evaluator is a VLM judge or human annotation, and it is validated by its agreement with human ratings. Benchmarks that score against measured ground truth instead (Wang et al., 2026a) escape that, but they need a paired reference and so cannot be computed on the open-domain generations a reward has to rank. Agreement is a correlation, and a motion confound is exactly what a correlation does not separate out. **None of this work asks the question in the other direction, namely what is left of that agreement once motion is held fixed.** We ask it of an evaluator from this family and of a programmatic reward, with the same test on the same clips; the two answers differ, and that difference is this paper's constructive result.

Motion as a confound, and the question before reward hacking. We are not the first to raise the confound. Proprio (Hassan et al., 2026) observes in an appendix that motion magnitude alone is not a proxy for physical correctness and stratifies its own reported gains by motion; that observation is theirs. What we add is a measurement rather than a caveat: a quantity for how much of a

reward's apparent validity is motion, a bidirectional test for which way the dependence runs, and an optimization experiment. A separate line of work asks whether a trained policy exploits a reward. **We ask the question that comes before it:** whether the reward measures the property it is named for, prior to any policy being trained, and we make no reward-hacking claim anywhere in this paper. Appendix I places both lines, and the shortcut-learning literature, against ours.

3 METHOD

3.1 WHAT WE REQUIRE OF A PHYSICS REWARD

Let `R` be a programmatic reward that maps a video clip to a scalar, `H` a human physical-plausibility rating, and `M` a statistic of motion magnitude. The standard validation asks whether `R` separates real clips from artificially corrupted ones. We instead ask, on clips a generator actually produces, a question we call *motion-orthogonalized on-policy validity*: does `R` predict `H` once `M` is held fixed? We report the rank partial correlation $\rho(\mathrm{R}, \mathrm{H} \mid \mathrm{M})$, the retention $|\rho(\mathrm{R}, \mathrm{H} \mid \mathrm{M})| \,/\, |\rho(\mathrm{R}, \mathrm{H})|$, and the reverse direction $\rho(\mathrm{M}, \mathrm{H} \mid \mathrm{R})$.

The reverse direction is what licenses a directional reading. If `R` loses its association with `H` under motion control while `M` keeps its own under control for `R`, the shared variance is attributable to motion; if both survive, the correct description is a common cause, not a proxy.

The control set matters and must be reported. A single motion statistic does not span "how much moved": controlling for peak optical flow alone leaves 16% of the Newton residual's apparent validity ($p = 0.081$), while controlling for peak flow together with mean velocity magnitude leaves 2% ($p = 0.857$). Every number we report names its control set.

The protocol is reward-agnostic: it takes per-clip values of `R`, `M` and `H` and never computes a reward itself. A self-test validates it in both directions, on a synthetic pure-motion reward where retention must collapse and on a genuinely valid motion-independent one where it must be preserved. Only the second case can falsify the instrument, since a test that always answers "it is just motion" passes the first.

3.2 DATA

Two human-rated sets, no overlap in generators. VideoPhy-2 (Bansal et al., 2025b) supplies 3350–3372 clips (the count differs by motion pipeline; we state which throughout) from seven 2024–25 generators, each with 1–5 human ratings for semantic adherence and physical commonsense. Physion-Eval (Zhang et al., 2026a) supplies 9569 clips from five 2025–26 generators (Sora 2, Veo 3.1 Fast, Kling 2.5, Wan 2.2, Hailuo 2.3), labelled instead with a 0–5 glitch severity. The two label constructions differ on purpose: a result that holds under both is not an artifact of one rating scheme. Because severity is higher-is-worse we correlate against its negation.

Two data facts are easy to get wrong and we state them rather than assume them. We use VideoPhy-2's *benchmark* split, verified before running anything on it, and we keep the two out-of-range severity values in Physion-Eval rather than winsorising them silently; Appendix A gives both in full.

3.3 SIGNALS

Motion. Peak optical-flow magnitude (`flow_peak`) as the primary statistic, with mean velocity magnitude as a second control. Optical flow from SEA-RAFT (Wang et al., 2024).

The audited rewards. (i) A reimplementation of the NewtonRewards (Le et al., 2025) kinematic residual from its published equations, the code not being released. (ii) PhyMotion (Huang et al., 2026), run with the authors' released code, on the 1168 VideoPhy-2 clips containing a full human body.

The automatic judge. VideoCon-Physics, the auto-evaluator released with VideoPhy (Bansal et al., 2025a), scoring each clip against its caption on two axes; we use the physical-commonsense head. Scores are `P(yes) / (P(yes)+P(no))` from the repository's own unmodified inference code, verified byte-identical to the official release. It is VideoPhy's evaluator rather than VideoPhy-2's successor, and was not trained on either set we audit it on. It appears in two roles, an outcome

measure for the causal experiment (§4.3) and an audit *target* under the same protocol as the rewards (§4.5), and we keep them separate.

The motion-orthogonal comparison signal. `area_cv_f40`, the coefficient of variation of the moving-region area, whose rank correlation with `flow_peak` is 0.0035. Its role is to be a reward that is verifiable in exactly the same sense but demonstrably not a motion statistic.

3.4 The causal experiment

If a reward is largely a motion statistic, optimizing it should move the policy along a motion axis and leave motion-orthogonal signals where they were. That is falsifiable: if motion barely moved while orthogonal physics improved, the claim would be wrong.

Two arms, identical except for the reward. We post-train Wan2.1-1.3B (Team Wan et al., 2025) with GRPO (Shao et al., 2024) on 99 restricted rigid-body prompts (five physical categories, 79/20 train/test split, static camera, single dominant object). Arm A uses the Newton residual, arm B `area_cv_f40`; the two run configurations were diffed field by field and differ only in the reward function, the run name and the output directory.

Dose–response, and paired measurement. Checkpoints are evaluated at five doses (32, 72, 112, 152, 192 parameter updates), so the claim is about a trend; the highest available is 192 of 232, because the schedule never writes the final epoch. Each measurement generates 80 clips, 20 held-out prompts by 4 repeats, with the seed determined solely by (prompt, repeat), so every post-training clip pairs with a pre-training clip at the same 50 denoising steps and the difference reflects the checkpoint only.

Reward scaling. The reward is `clip(-log10(x + ϵ), lo, hi)`. We use `ϵ = 1e-12` and bounds (-4, 12), under which no sample saturates; the implementation's own defaults saturate 31% of the training distribution and would have produced a clean-looking null result agreeing with our hypothesis (Appendix A).

3.5 Statistics

Per-clip paired, against a null we measured first. All before/after quantities are reported as per-clip paired changes with up/down counts and a Wilcoxon test on matched clips, never as a ratio of medians, which on this prompt set is unusable for heavy-tailed signals: with no intervention at all the Newton residual's median-ratio spans -86.7% to +1050%. Before reading any result we measured what "no intervention" looks like on the same repeats. Changing only the seed moves the Newton residual by -86.3% ($p = 0.024$) while the motion-orthogonal signals stay far tighter, which is what makes a null prediction about them decidable at all (Appendix A).

What counts as surviving motion control. A generator is counted as retaining validity only if its raw correlation is significant, its partial correlation is significant, and the two share a sign. The third condition is not decoration, and Appendix B gives the artifact it excludes.

The remaining choices. "Not significant" is not "unchanged", so for each orthogonal signal we bootstrap a confidence interval on the per-clip median change and report what can be excluded. Dependent correlations are compared with the Hotelling–Williams test with polarity aligned to magnitude, reported beside a bootstrap of the difference in $|\rho|$ that carries no distributional assumption. Multiple comparisons use Benjamini–Hochberg. Running the two public codebases required patches, one of them a bug in the upstream training code rather than an incompatibility of our own; Appendix A lists every departure with the reason it was forced.

4 Results

4.1 The standard validation overstates validity

Our own trajectory residual (Appendix L) separates real clips from rule-corrupted ones at AUC 0.7422 (95% CI [0.576, 0.884], permutation $p = 0.0020$), which is a result that would ordinarily be reported as validation. On 3350 human-rated clips from seven generators, the same reward

ranks clips at AUC 0.5381. The gap is 0.2041 ([0.028, 0.366]), reported as an interval because the corrupted-versus-real side rests on only 15 real clips.

The sharper version is categorical rather than aggregate. Its parabola term, built for uniformly accelerated free fall, correlates with human judgement at +0.002 on the 186 clips whose human annotation names gravity or free-fall as the violated law. The audited kinematic residual, whose primitives are free fall, projectile and inclined-plane motion, does the same on its own categories: ρ = -0.075 (p = 0.32) on gravity and free-fall and -0.001 (p = 0.99) on momentum/velocity. **Neither residual has a measurable relationship with human judgement precisely where its own primitives apply.**

4.2 The reward is largely a motion statistic

A gap of that size says the validation is optimistic; it does not say what the reward is tracking instead. It is tracking motion.

A reimplementation of the NewtonRewards kinematic residual correlates with peak optical-flow magnitude at ρ = 0.85–0.92 across five independent sources spanning 13 generators and two architectures: VideoPhy-2 0.9032 (n = 3372), Physion-Eval 0.8747 (n = 9569), our CogVideoX (Yang et al., 2025) generations at two seeds 0.8811 and 0.9101, and Wan2.1-1.3B on our own constrained prompts 0.8511. A second, independently implemented pipeline over the same flow field gives 0.9149 (n = 3350), so the coupling does not turn on the downstream statistic.

Controlling for motion removes the reward's association with human judgement. On VideoPhy-2 the raw correlation is ρ = -0.1867 (p = 1.16e-27); controlling for peak flow and mean velocity magnitude leaves ρ = +0.0031 (p = 0.857).

A generator counts as *surviving* only under the three conditions of §3.5, and the control set changes the answer:

control set	VideoPhy-2	Physion-Eval	total
peak flow only	0 of 7	0 of 5	**0 of 12**
peak flow + mean velocity	0 of 7	1 of 5	**1 of 12**

Under the single control no generator survives on either dataset; under the two-variable control one does, and the pooled Physion-Eval association that survives with it is carried by that same generator (Appendix B). Across two datasets and twelve generators, then, at most one generator retains any association with human judgement after motion is controlled, and none does under the stricter single-control reading.

The relationship is asymmetric, which is what licenses a directional reading. Motion retains its association with human judgement after controlling for the reward (ρ = -0.1166, p = 1.31e-11), while the reward does not survive the converse. Motion explains the reward's apparent validity; the reward does not explain motion's.

Finally, a statistic with no physics content is the better predictor. On VideoPhy-2, peak optical flow correlates with human physical-plausibility ratings at ρ = 0.2171 against the reward's 0.1867 (Hotelling–Williams p = 1.331e-05); on Physion-Eval, 0.0714 against 0.0546 (p = 9.8e-4). Against our own parabola term the gap is an order of magnitude larger (p = 2.867e-14). The difference is individually significant in 5 of 12 generator strata with the direction consistent in 11 of 12, and we state it as pooled significance with a minority of strata individually significant.

This confound has been reported absent once, in the benchmark's own paper, which obtains 0.002 correlating human physical commonsense against motion. We reproduce all fourteen of its published per-generator scores, so we work from the same ratings, and the answer turns on which flow statistic is taken: an aggregate gives -0.008 (p = 0.74), the peak gives -0.217 (p = 5.2e-37). Appendix C gives the comparison and states what we do and do not claim about their number.

4.3 Optimizing the reward moves the policy along a motion axis

Everything so far is correlational, and a correlation does not say what happens when the reward is optimized. That takes an intervention: two GRPO arms identical except for the reward, five doses, a shared paired baseline.

	dose 32	72	112	152	**192**
A (motion-correlated reward), peak flow	-15.5%	-24.5%	-25.1%	-37.8%	**-46.3%**
B (motion-orthogonal reward), peak flow	+28.9%	+21.8%	+35.2%	+22.7%	**+66.6%**
A - B, same clip, Wilcoxon	p=1e-3	2e-4	3e-6	6e-7	**1e-8**

Both effect size and significance are monotone in dose. Each arm also drives its own reward and pushes the other arm's in the opposite direction: the Newton residual falls 81.1% under A and rises 382% under B.

The three motion-orthogonal physical signals do not move in any consistent direction. At the highest dose `area_cv_f40` sits at +0.6%, `area_jitter_f25` at -2.2%, `area_cv_f25` at -1.7%. We state this as a bounded, directional claim rather than as "unchanged": motion falls 46.3% with a bootstrap CI of [-58.2%, -28.1%] excluding zero, while any consistent change in the orthogonal signals can be bounded within about 18%, and the two intervals do not overlap. Per clip, the orthogonal signals do move by roughly 40%, but they move that much with no intervention at all (40–68% from changing the seed alone), so per-clip magnitude carries no information here and we do not use it.

The training pushes the policy outside the range where the motion–rating relationship was ever measured. Human ratings do fall monotonically as motion rises, so suppressing motion raises human scores over the observed range. But the *lowest* motion decile in the human-rated data still has a median peak flow of 0.0068, not zero, and shows no interior optimum. GRPO drives toward zero: at the highest dose we can measure, motion is down 46.3%, still monotone in dose with no sign of levelling off. What the relationship does below the observed floor no one has measured, and this reward's optimum lies there, which is the sense in which its gain cannot be extrapolated.

What we cannot claim. Neither judge score separates the arms at any dose (physical commonsense minimum p = 0.2, semantic adherence minimum p = 0.09). We therefore do not claim the reward bought judge score, and we do not claim reward hacking. Nor do we claim arm A damaged physics: the orthogonal signals return to baseline, they do not degrade.

Every number above is read against the null of §3.5, measured before we read any result: changing only the seed moves the Newton residual by -86.3% (p = 0.024). That is why a reader shown "the residual fell 80%" should not be persuaded, and why every number here is a per-clip paired change with up/down counts.

It replicates on a second prompt family, and there the judge behaves differently. The dissociation on motion is if anything stronger: peak flow falls to -88.3% at the highest dose and the arms separate at 5 of 5 doses. The judge, however, reverses, separating the arms at every dose where in the first run it separated at none. We report that rather than resolve it (Appendix D); the direction of the judge result is not stable across our two runs, and nothing here rests on the judge moving a particular way.

4.4 The same test applied to a third-party reward

Both runs so far audit one reward, ours, so the obvious objection is that we built the thing we found wanting. We apply the identical protocol, with its authors' own code, to the most structured physics reward published to date, on the 1168 VideoPhy-2 clips containing a full body.

PhyMotion is **not** a motion proxy, and that is to the method's credit: it correlates with peak flow at +0.181 where the kinematic residual reaches +0.920 on the same clips (Appendix F). **But on this rating it carries no information beyond motion either.** Its raw correlation with human ratings is -0.068 (p = 0.020) and vanishes under the two motion controls (-0.033, p = 0.255), while the bidirectional test again favours motion (ρ(flow, rating | PhyMotion) = -0.189, p = 7.8e-11). Two outcomes were written down before this was run, that it would also be motion-confounded or that it would survive; neither happened, and we report the third rather than fitting it to one of the two. The test is outside PhyMotion's design domain, and Section 5 states what that means for the claim.

4.5 THE SAME PROTOCOL APPLIED TO A VLM JUDGE

That leaves an uncomfortable possibility: nothing measures physical plausibility on these clips and the protocol is simply too strict to pass. The next result rules it out. We apply the identical protocol, on the same clips and rows and under the same two motion controls, to VideoCon-Physics, the auto-evaluator released with VideoPhy.

	ρ, human PC	ρ, motion	partial ρ	retention	per gen.
VideoCon-Physics (PC head)	**+0.314**	-0.227	**+0.276**	**88.0%** [85.2, 90.7]	**6 of 7**
kinematic-residual reward	-0.187	**+0.915**	+0.003	**-1.7%**	**0 of 7**
peak flow (pure motion)	-0.217	n/a	n/a	n/a	n/a

n = 3350; controls are peak flow and mean velocity magnitude; per-generator counts are after BH correction; the retention interval is a 2000-sample bootstrap.

Pure motion is a significantly *better* predictor of human ratings than the reward is (Hotelling–Williams p = 1.33e-05), and a significantly *worse* one than the judge (p = 2.12e-06). **The same test, on the same rows, answers oppositely for the two.**

This is not one dataset. On Physion-Eval's 9569 clips, from five generators absent from VideoPhy-2 and under a differently constructed label, the judge retains 82.2% after the same two controls, significant in 4 of 5 generators, and pure motion again loses to it (p = 7.09e-05). Its coupling to motion is stronger there (ρ = -0.360 against -0.227) and it retains anyway. Unlike the reward (§4.2), its retention does not move with the control set: 82.2% and 83.9%.

The judge does not read the arrow of time. Playing a clip backwards makes its physics illegal while leaving objects, scene and caption untouched: liquid climbs back into the bottle and splinters reassemble. Against a control that went through the identical ffmpeg pass without the reversal, the judge's physical-commonsense score moves by +0.00075 on 434 clips after the same two motion controls (95% CI [-0.00185, +0.00350]); a placebo re-judging the unchanged originals gives +0.00005. The design, and the equivalence bound this interval falls well inside, were fixed before any outcome was read (Appendix G).

This is the answer we can give to "what should I use instead". It is not a method of ours, and it is not an argument that automatic evaluation is sound in general. **Verifiability is not the line that separates these two; whether the quantity is a motion statistic is.** What the judge retains after motion is controlled is therefore neither motion nor time direction, and we do not claim to have identified it.

Four limits travel with this result, each with its number in Appendix E: the semantic-adherence head is uninformative here, so its retention is the retention of a null; this is VideoPhy's evaluator, not VideoPhy-2's successor; it is not a motion proxy but is still motion-sensitive, so the second run's rising score cannot be decomposed without human ratings; and one judge is not the family.

5 LIMITATIONS

Each limitation below is stated here; the arguments behind several are in Appendix H, which adds no further limitation.

Coverage: two rewards, not a family. We set out to audit four and could compute only two: PhysRVG (Zhang et al., 2026b) and PISA (Li et al., 2025) score against ground truth, and VGGRPO's (An et al., 2026) rewards are camera smoothness and geometry reprojection rather than the plausibility of the depicted events. Our claim about this family is therefore an argument from shared construction with two measured instances, not an exhaustive sweep.

PhyMotion is evaluated outside its design domain. It is built for *human motion realism* and we evaluate it against *general* physical commonsense on the 1168 clips containing a full human body, whose violations often concern objects rather than bodies. We therefore do **not** claim that PhyMotion does not work: on this rating it carries no information beyond motion magnitude (partial ρ = -0.033, p = 0.255).

One base model, one scale, and a confounded second prompt family. The causal experiment uses Wan2.1-1.3B, the only model our GRPO implementation supports. On the second prompt family the base model's prompt adherence is substantially lower (0.13 vs 0.49), so scene complexity and generation fidelity are confounded there, and we do not describe it as "replicating on more complex scenes".

We measure up to 83% of training, and we have no human ratings on our own generations. No checkpoint beyond step 192 of 232 exists; effects are monotone up to there, so no claim here is about a fully trained model.

The two runs disagree about the judge, and the pre-registered prediction failed in both. We predicted that semantic adherence would degrade before physical commonsense, and it did not, in either run (first-run minima $p = 0.2$ and $p = 0.09$). Any account needing the judge to move a particular way is unsupported, and for the same reason we do not claim reward hacking.

Our null result is about direction, not magnitude. Per-clip magnitude carries no information here, and we withdrew an earlier test built on it. The defensible statement is bounded rather than null: the motion interval in Section 4.3 excludes zero, any consistent change in the orthogonal signals is bounded within about **18%**, and the two intervals do not overlap. We do not claim the orthogonal signals were "unaffected."

Reducing motion does improve human ratings, which weakens the practical force of our critique. In both human-rated datasets, ratings fall monotonically as motion rises, so a reward that suppresses motion will raise human scores. Our claim is about the *pathway*, not the effect. This was pre-registered as a branch that would weaken our position, and it triggered.

Three methodological dependencies we report rather than bury. The motion-controlled result depends on which motion statistics are controlled, and on Physion-Eval that choice moves the count of surviving generators, not only the retained fraction, so every survival count here carries its control set. Second, the Hotelling–Williams test was derived for Pearson correlations and is liberal when fed ranks. Third, the kinematic residual and the motion statistics we control for come from the same optical-flow network, so their correlation shares an estimator. We measured that with a second, non-learned estimator rather than conceding it: the coupling survives across networks but weakens (Appendix H).

The reversal result is about one evaluator, and its first design failed. The probe was built as a degradation that preserves motion and failed that pre-registered gate, because forward and backward optical flow are not mirror images to a real estimator; the question survives without that assumption and is reported under motion control instead. The judge it answers for is a video-language model, a family whose weak temporal modelling is documented, so the null may be a property of this evaluator (Appendix H).

We do not offer a working replacement of our own. The constructive half of this work was pre-registered with a kill criterion and it triggered: our object-identity signal detects human-annotated object-permanence violations at AUC 0.53, no better than chance, and every motion-orthogonal signal we could construct is weak ($\rho \leq 0.06$ after motion control). That failure is the most informative thing we can say about the design space: escaping the motion confound and carrying measurable validity are separate problems, and the second is not solved by solving the first.

6 Conclusion

Verifiability is not the property that matters. A reward can be deterministic, inspectable, and able to separate real video from rule-corrupted video, and still carry nothing about human judgements of physical plausibility once motion is held fixed; an opaque judge carries it anyway. What separates them is not how the quantity is computed but whether it is a statistic of motion.

That distinction can be measured before any policy is trained, and measuring it is cheap. The protocol needs only per-clip values of the reward, a motion statistic and a human rating, and it answers in both directions, so it reports a common cause rather than a proxy when that is what the data show. Before optimizing a generator against a physics reward, run it on the reward: report what survives motion control, name the control set, and say whether pure motion predicts the human ratings better on the same rows.

AI USE STATEMENT

We used generative AI throughout this work, in the form of an AI coding assistant used interactively by the authors, and we disclose that use here against the categories the ICLR policy names. What this paper claims, which results it reports, what would have counted as falsification, and how the argument is organised were decided by the authors, and everything the assistant produced was checked before it was used.

Of the uses whose disclosure is required, the assistant helped develop the conceptual framework, proposed and refined hypotheses, gave feedback on the methodology and the experimental design, implemented the methods, cleaned and reformatted the datasets into the per-clip tables the protocol consumes, translated the manuscript out of the authors' working language into English, and drafted interpretations of the results. In each of these the assistant drafted and the authors accepted, altered or rejected. The clips analysed here that are not distributed with the benchmarks were generated by video models, which in this paper are the object of study rather than an aid to producing it; the runs that produced them are specified in Section 4.3 and, for the clips sampled from a third party's released checkpoint, in Appendix I. We did not use generative AI to produce or alter any measured value, to select which results this paper reports, or to create or edit any human rating: every rating analysed here was collected by the authors of the benchmarks and is used as published. The remaining required categories do not arise, since this work states no mathematical claims and contains no proofs, and its analysis is quantitative rather than qualitative or thematic.

Of the uses whose disclosure is recommended, the assistant wrote and edited the software, namely the motion-control protocol and its self-test, the analysis scripts, the self-checks that guard this manuscript and the pipeline that assembles this document from the section drafts. It ran the analyses, drew every figure, drafted and edited prose, formatted the references, searched for candidate related work and summarised it for internal triage, and proposed candidate titles and section structures, among which the authors chose.

The checking was not a formality. Every number in this paper is checked by a script against the experiment record, and the check fails if the manuscript states a number the record does not contain; a further check fails if the same group of numbers is printed in two places, because a quantity with two homes is one that can drift out of agreement with itself. Our hold on the dataset was gated on reproducing its authors' own published table before anything was run on it (Appendix C), and every deviation from the reference implementation we reimplement is listed individually rather than summarised (Appendix A). The constructive half of this work was pre-registered with a kill criterion, the criterion triggered, and we report that outcome. Every paper cited here was read by the authors from its PDF before being described, and no description of prior work rests on a machine-generated summary.

Separately, and not as a form of AI assistance: the vision–language judge examined in Section 4.5 is an object of study here rather than a tool used to produce this work. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work runs no study on human subjects. The human physical-plausibility ratings we analyse were collected by the authors of VideoPhy-2 and Physion-Eval and are used as published; we add no annotation of our own and identify no annotator. The clips we score are model-generated video from those benchmarks together with our own generations, and we introduce no new recording of any real person.

This paper is critical of published work, and we have tried to hold that criticism to what we measured. We do not claim that any method examined here fails at the task its authors designed it for. Where we measure a method outside its stated domain we say so, and where a reported gain is consistent with our finding rather than contradicted by it we say that as well. Our own constructive attempt failed against a criterion we set in advance, and we report it rather than the alternative of quietly redirecting the paper around it.

REPRODUCIBILITY STATEMENT

The protocol is specified in Section 3 and is reward-agnostic: it consumes per-clip values of the reward, the motion statistic and the human rating and never computes a reward itself, so it can be applied to a reward we have not seen. Its self-test validates it in both directions, on a synthetic pure-motion reward whose retention must collapse and on a motion-independent one whose retention must be preserved, and only the second case can falsify the instrument. Section 3.5 fixes the survival definition, and every survival count here is reported with the control set it was computed under, because the count moves with that set.

Appendix A lists each deviation from the reference implementation we reimplement, with its reason. Appendix C reproduces the published per-generator table of the dataset we use, which is how we establish that we hold the same human ratings its authors did. The ratings and the benchmark clips are those distributed with VideoPhy-2 and Physion-Eval; the remaining clips are our own generations, and the two-arm intervention is specified in Section 4.3 with the second run reported in full in Appendix D. The pre-registered criteria were written before the data they concern existed, and the one that triggered is reported as a negative result rather than dropped.

REFERENCES

Zhaochong An, Orest Kupyn, Théo Uscidda, Andrea Colaco, Karan Ahuja, Serge Belongie, Mar Gonzalez-Franco, and Marta Tintore Gazulla. VGGRPO: Towards world-consistent video generation with 4D latent reward. In *European Conference on Computer Vision (ECCV)*, 2026.

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)*, 2025a.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025b.

León Begiristain, Olaf Dünkel, and Adam Kortylewski. CRONOS: Benchmarking counterfactual physical consistency in video models. *arXiv preprint arXiv:2605.23699*, 2026.

Harold Haodong Chen, Disen Lan, Wen-Jie Shu, Qingyang Liu, Zihan Wang, Sirui Chen, Wenkai Cheng, Kanghao Chen, Hongfei Zhang, Zixin Zhang, Rongjin Guo, Yu Cheng, and Ying-Cong Chen. TiViBench: Benchmarking think-in-video reasoning for video generative models. *arXiv preprint arXiv:2511.13704*, 2025.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Eric Wang. PhyWorldBench: A comprehensive evaluation of physical realism in text-to-video models. In *International Conference on Learning Representations (ICLR)*, 2026.

Mariam Hassan, Kaouther Messaoud, Wuyang Li, and Alexandre Alahi. Proprio: Latent self-scoring and inference-time refinement for physically plausible video generation. *arXiv preprint arXiv:2605.28230*, 2026.

Yidong Huang, Zun Wang, Han Lin, Dong-Ki Kim, Shayegan Omidshafiei, Jaehong Yoon, Jaemin Cho, Yue Zhang, and Mohit Bansal. PhyMotion: Structured 3D motion reward for physics-grounded human video generation. *arXiv preprint arXiv:2605.14269*, 2026.

Haozhe Jia, Pengyu Yin, Wenshuo Chen, Shaofeng Liang, Lei Wang, Bowen Tian, Xiucheng Wang, Nanqian Jia, and Yutao Yue. Learning to think in physics: Breaking shortcut learning in scientific diffusion via representation alignment. In *International Conference on Machine Learning (ICML)*, 2026.

Bo Jiang, Depu Meng, Yihan Hu, Yichen Xie, Tianshuo Xu, and Wei Zhan. LaMo: Self-supervised latent motion priors for physical realism in video generation. *arXiv preprint arXiv:2605.23878*, 2026.

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)*, 2025.

Minh-Quan Le, Yuanzhi Zhu, Vicky Kalogeiton, and Dimitris Samaras. What about gravity in video generation? Post-training Newton's laws with verifiable rewards. *arXiv preprint arXiv:2512.00425*, 2025.

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. PISA experiments: Exploring physics post-training for video diffusion models by watching stuff drop. In *International Conference on Machine Learning (ICML)*, 2025.

Jiesong Lian, Ruizhe Zhong, Zixiang Zhou, Xiaoyue Mi, Long Hu, Yuan Zhou, Qinglin Lu, Yixue Hao, and Junchi Yan. SoliReward: Mitigating susceptibility to reward hacking and annotation noise in video generation reward models. *arXiv preprint arXiv:2512.22170*, 2025.

Juyi Lin, Arash Akbari, Yumei He, Lin Zhao, Haichao Zhang, Arman Akbari, Xingchen Xu, Zoe Y. Lu, Enfu Nan, Hokin Deng, Edmund Yeh, Sarah Ostadabbas, Yun Fu, Jennifer Dy, Pu Zhao, and Yanzhi Wang. PhyGround: Benchmarking physical reasoning in generative world models. *arXiv preprint arXiv:2605.10806*, 2026.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *International Conference on Machine Learning (ICML)*, 2025.

Atin Pothiraj, Jaemin Cho, Yue Zhang, Elias Stengel-Eskin, and Mohit Bansal. Physics question scene graph: Fine-grained evaluation of physical plausibility in text-to-video generation. In *European Conference on Computer Vision (ECCV)*, 2026.

Akshat Rakheja, Aarsh Ashdhir, Aryan Bhattacharjee, and Vanshika Sharma. World consistency score: A unified metric for video generation quality. *arXiv preprint arXiv:2508.00144*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

Peiyao Wang, Weining Wang, and Qi Li. PhysCorr: Dual-reward DPO for physics-constrained text-to-video generation with automated preference selection. *arXiv preprint arXiv:2511.03997*, 2025.

Shuai Wang, Yaxin Feng, Xuekun Jiang, Shihan Tian, Ningyu Yan, Xing Shen, Chaoyang Lyu, Hui Wang, Yunsong Zhou, Hanqing Wang, Jiangmiao Pang, Yang Xiang, Xing Gao, Chunhua Shen, and Weinan Zhang. GAUGE: A measurement-grounded benchmark for physical fidelity in simulation engines and video world models. *arXiv preprint arXiv:2608.05948*, 2026a.

Xinpeng Wang, Nitish Joshi, Barbara Plank, Rico Angell, and He He. Is it thinking or cheating? Detecting implicit reward hacking by measuring reasoning effort. In *International Conference on Learning Representations (ICLR)*, 2026b.

Yifan Wang, Yanyu Li, Gordon Guocheng Qian, Sergey Tulyakov, Yun Fu, and Anil Kag. Diffusion-DRF: Free, rich, and differentiable reward for video diffusion fine-tuning. *arXiv preprint arXiv:2601.04153*, 2026c.

Yihan Wang, Lahav Lipson, and Jia Deng. SEA-RAFT: Simple, efficient, accurate RAFT for optical flow. In *European Conference on Computer Vision (ECCV)*, 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.

Qin Zhang, Peiyu Jing, Hong-Xing Yu, Fangqiang Ding, Fan Nie, Weimin Wang, Yilun Du, James Zou, Jiajun Wu, and Bing Shuai. Physion-Eval: Evaluating physical realism in generated video via human reasoning. *arXiv preprint arXiv:2603.19607*, 2026a.

Qiyuan Zhang, Biao Gong, Shuai Tan, Zheng Zhang, Yujun Shen, Xing Zhu, Yuyuan Li, Kelu Yao, Chunhua Shen, and Changqing Zou. PhysRVG: Physics-aware unified reinforcement learning for video generative models. *arXiv preprint arXiv:2601.11087*, 2026b.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. VideoREPA: Learning physics for video generation through relational alignment with foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Tinghui Zhu, Sheng Zhang, James Y. Huang, Selena Song, Xiaofei Wen, Yuankai Li, Hoifung Poon, and Muhao Chen. Video models can reason with verifiable rewards. *arXiv preprint arXiv:2605.15458*, 2026.

A Deviations from the reference implementation, and the seed-pair null bands

Which VideoPhy-2 split, and the out-of-range severities. We use the *benchmark* split. This matters more than it sounds, because the authors' auto-evaluator was trained on a different set that also contains 3350 clips, from three generators rather than seven. We verified that we hold the benchmark split by reproducing their published per-generator joint scores (Appendix C) before running anything on it. Physion-Eval's severity column contains two out-of-range values, an 8 and a 44 on a 0–5 scale. We keep them, since the published analysis of that dataset kept them and every test we run is rank-based, but we report their presence rather than winsorising silently.

Reward scaling. With the implementation's original ϵ = `1e-6` and cap 5, 31% of the training distribution saturates at the cap, which drives within-group variance to zero and makes GRPO advantages identically zero. That failure would have produced a clean-looking null result consistent with our own hypothesis, which is the shape of error this paper is least entitled to make.

The seed-pair null bands. Motion is no tighter than the residual under a change of seed alone: `flow_peak` spans -53.5% to +132.1%, which is why every claim about motion in this paper rests on paired tests and up/down counts rather than on magnitude. Every band is the per-clip paired change, the same statistic as the curves, taken over six same-prompt different-seed pairs; a ratio of medians is not used anywhere. The two motion-orthogonal signals drawn beside it are far tighter, -22.3% to +1.4% for `area_cv_f40` and -38.8% to -11.8% for `area_jitter_f25`, which is what makes a null prediction about them decidable at all. We verified that the repeats are exchangeable (per-signal Friedman tests; none survives BH correction).

The GRPO implementation we build on has only ever been run in its authors' 8-GPU default shape, and every departure from that shape met an untested coupling. We list each deviation and why it was forced:

parameter	default	ours	why
`num_processes`	8	**3** per arm	six GPUs available, two arms concurrent
`train_batch_size`	2	**1**	at 2 the backward pass OOMs (42.53 of 44.40 GiB); at 1 it peaks at 42.67 GiB, against 26.9 GiB during sampling
`gradient_accumulation_steps`	1	**1** (unchanged)	`>1` trips a DeepSpeed ZeRO-2 `no_sync` assertion, so the divisibility constraint below had to be met another way
`num_image_per_prompt` (`k`)	4	**3**	the sampler asserts `(num_processes × 10 × batch) % k == 0`; at 3 processes only `k = 3` satisfies it. **We note that `k = 4` is legal at 2 processes and would have been the better shape, with fewer deviations and a larger GRPO group. We kept `k = 3` because the smoke test that validated the pipeline ran at that shape, and changing it would have meant validating one configuration and running another.**
`num_steps`	20	**20** (unchanged)	a hard-coded 10-entry timestep table pins this; `timestep_fraction = 0.51` is likewise unchanged
`num_epochs`	n/a	**60**	compute budget
reward bounds	(-1, 5), ϵ=1e-6	**(-4, 12), ϵ=1e-12**	31% of samples saturated the cap

A GRPO group is therefore 6 samples of one prompt: `k = 3` sample slots, all holding the same prompt because the prompt is identical on every process, times the two videos per slot that the training script generates unconditionally.

The public code does not train as released, and we name what we patched rather than count it: the `assets` directory and the prompt-embedding module are both absent from the repository, `config/dgx.py` imports `imp`, which Python 3.12 removed, and the log-probability computation indexes a timestep tensor on one axis where the tensor beside it is indexed on two. The last is a bug in the released repository and is not of our making; it crashes training on a clean checkout.

B THE ONE SURVIVING GENERATOR

We count a generator as *surviving* only if its raw correlation is significant, its partial correlation is significant, and the two share a sign. A signal whose raw association is absent and whose conditional association appears with the opposite sign has not retained validity; it has acquired an artifact of conditioning, and counting it would inflate the count.

The one generator that survives the two-variable control is Kling 2.5 on Physion-Eval, retaining 66% of a raw ρ of -0.10. Pooled over Physion-Eval that control leaves a small surviving association (ρ =

-0.047, p = 5.4e-06 at n = 9569), but it is carried by that single generator, which is the pattern we distrust elsewhere in this paper and distrust here too.

C WHY THIS HAS NOT ALREADY BEEN NOTICED

The benchmark we use reports the opposite. VideoPhy-2's own paper correlates its human ratings against "motion quality (measured using optical flow from RAFT)", obtains 0.002, and concludes that "physical commonsense is not well-correlated with any of these video metrics" (their Table 3). That single value is, as far as we can tell, the field's basis for not treating motion as a confound on this benchmark.

We hold the same ratings. Their Table 2 publishes joint performance (SA $\geq 4 \wedge$ PC ≥ 4) for each of seven generators on both the full and hard splits, and we reproduce all fourteen, published against ours: 32.6/32.7, 25.0/25.0, 24.1/24.1, 17.2/17.2, 10.5/10.5, 20.3/20.3 and 23.3/23.3 on the full split, and 21.9/21.9, 0.0/0.0, 10.9/10.9, 6.2/6.2, 2.9/3.0, 8.3/8.3 and 5.3/5.3 on hard. The benchmark's own joint column agrees with (SA $\geq 4 \wedge$ PC ≥ 4) on every clip. So the disagreement is not about the data.

On those ratings, the answer depends almost entirely on *which statistic of the flow field* is used:

motion statistic	Pearson	Spearman	p
tracked motion energy	-0.008	+0.006	0.74
tracked centroid travel	-0.026	-0.041	0.017
global (median) flow	-0.047	-0.109	2.9e-10
99th-percentile flow	-0.126	-0.190	1.2e-28
peak flow	**-0.148**	**-0.217**	**5.2e-37**

The peak-flow result is not fragile: it is individually significant in 6 of the 7 generators, it reproduces under a second motion pipeline over the same flow field (-0.211, n = 3372), and its magnitude exceeds all 200 draws of a label-permutation null (largest null $|\rho| = 0.046$). What none of these vary is the flow estimator itself, which Section 5 states as a dependency.

We are careful about what this does and does not establish. We cannot reproduce their 0.002 with any statistic we compute, and we cannot recover what their metric was: the paper names the sensor but not the statistic, and the analysis is not in their released code. It is entirely possible that "motion quality" denotes motion *smoothness* rather than motion *magnitude*, in which case the two numbers measure different quantities and neither is wrong. Our claim is therefore not that their value is incorrect. It is narrower and sufficient. **On these ratings, whether physical commonsense appears confounded with motion is decided by the choice of flow statistic, and the statistic that a single published value rests on is not specified.**

D THE SECOND RUN IN FULL

We repeated the whole two-arm experiment on a second prompt set, chosen to be harder to generate and more cluttered. The design is identical: same base model, same two rewards, same five checkpoints, arms differing only in the reward. Ten measurements, 80 paired clips each.

In this run scene complexity and generation fidelity are confounded, because the harder prompts are also the ones the base model renders worse. So this is evidence that the effect survives *a lower-fidelity, more cluttered regime*, not that it survives *greater scene complexity*. We do not make the second claim.

What replicates. The dissociation on motion is if anything stronger. Peak flow falls monotonically in the motion-coupled arm (-54.7%, -56.1%, -72.8%, -83.3%, -88.3% across the five doses) and barely moves in the control arm; the arm difference is BH-significant at 5 of 5 doses. Each arm drives its own reward down and not the other's. The pre-registered motion-orthogonal signal, `area_cv_f40`, separates the arms at 0 of 5 doses after correction.

What does not. Two sibling motion-orthogonal statistics, `area_jitter_f25` and `area_cv_f25`, *do* fall in the treated arm, at 4 of 5 doses each, where in the first run neither moved. We report

all three, because "the signal we pre-registered did not move" and "no motion-orthogonal signal moved" are different sentences and only the first is true here.

And the judge reverses. In the first run neither judge score separated the arms at any dose. Here physical commonsense separates them at all five, monotonically in dose (q from 1.1e-3 to 3.3e-9), while semantic adherence separates at none, so the pre-registered ordering fails here too and in the opposite direction. Within the treated arm, the clips that lost the most motion gained the most judge score, with median per-clip differences of +0.012, +0.050, +0.066, +0.124 and +0.122 (absolute, on a 0–1 score with baseline mean 0.188) and a coupling that deepens monotonically with dose (Spearman -0.24, -0.28, -0.33, -0.37, -0.43); in the control arm the same correlation is -0.20 and -0.14, neither significant. That contrast does not generalise, and we say so rather than leave it to be read as a treatment effect: in the first run the same coupling is present in every arm, including the motion-orthogonal control, at -0.33 to -0.49 with every dose significant. It is a property of this judge on these clips.

We also tried to explain the difference between the two runs with our own protocol and could not. Holding the per-clip motion change fixed under the same two-variable control used throughout, the between-arm judge difference in the second run is essentially unchanged and remains significant at all five doses after correction, while in the first run it stays absent at all five. The between-arm difference in motion does not account for the between-arm difference in judge score.

We draw the narrow conclusion only, and §4.5 constrains it in both directions. It would be convenient to read the rising judge score as the reward buying credit with stillness. We cannot: §4.5 shows this judge retains 88% of its agreement with human ratings once motion is partialled out, so its score is not a motion proxy and a rise in it is not nothing. It would be equally convenient to read it as the arm genuinely improving physics. We cannot do that either, because the same judge is still motion-sensitive (ρ = -0.227 with peak flow), so a collapse in motion reaching 88% at the highest dose predicts *some* of the rise on its own. What the second run establishes is the size and monotonicity of the effect, not its composition, and decomposing the gain requires human ratings on these outputs, which we do not have.

A note on units. The per-clip *percent* change in judge score is not monotone in dose (+29%, +61%, +96%, +177%, +142%) while the *absolute* change is. The judge score is bounded and has a floor, so percent change is the wrong unit for it, and reading the percent column would have produced a different qualitative claim.

E THE FOUR LIMITS ON THE JUDGE RESULT

(i) The judge's semantic-adherence head correlates with human semantic adherence at ρ = +0.018 (p = 0.32) on these clips, so its 97.9% "retention" is the retention of a null and carries no content; we report it only so that the number is not quoted without its raw correlation. **(ii)** This judge is VideoPhy's evaluator rather than VideoPhy-2's, and VideoPhy-2's own authors state that existing video-language models "struggle to achieve strong agreement with human annotators", which is why they trained a successor. We are therefore auditing a judge that its own successors call insufficient, which bears on its accuracy rather than on the question we asked, namely what kind of quantity it is. **(iii)** A judge that is not a motion proxy is still motion-sensitive (ρ = -0.227), so the second-run finding in Appendix D, judge score rising monotonically as motion collapses, cannot be decomposed into a motion-driven part and a remainder without human ratings on those outputs. We do not claim that gain was bought with stillness, and we do not claim it was earned. **(iv)** One judge is not the judge family, and we do not generalise beyond the one we measured.

F THE THIRD-PARTY REWARD ON THE SAME ROWS

For the comparison in Section 4.4, the kinematic residual evaluated on the same body-containing clips correlates with human ratings at ρ = -0.180, and at +0.006 (p = 0.845) once the two motion controls are applied. The two rewards therefore reach the same place from different starting points: one is a near-perfect motion proxy and the other is not, and neither retains an association with the rating once motion is held fixed.

G THE TIME-REVERSAL TEST

Time reversal changes the physics of a clip while leaving its content, its caption and its objects alone, so it separates the two in a way a correlation cannot. The comparison is against a re-encoded control rather than the original, because the ffmpeg pass by itself moves peak flow by -2.2% (p = 1.1e-07) and that effect would otherwise be attributed to the reversal. Both sides went through the same pass; the paired difference is regressed on the paired changes in peak flow and mean velocity magnitude, and the intercept is the effect at zero motion change. The design and the four possible outcomes were fixed before any reversed or re-encoded score was read. The equivalence bound of ± 0.02 was set later, after measuring what the pipeline does to unchanged clips and before reading any outcome, and it is that measurement rather than any outcome that fixes its value.

Two measurement facts shape the test. The judge is deterministic given a clip, a caption and a batch: re-running an identical job on a different GPU reproduces every score to 0.000e+00. It is not deterministic across batch layouts, where one clip moved by 2.5e-02. Both arms were therefore judged under the same shard layout, and a placebo arm re-judges the unchanged originals under that layout to measure what the pipeline produces when the video does not change at all. One clip is excluded from the analysis because ffmpeg's reverse filter dropped 14 frames on it; it is kept in the judging job, since removing it would have reshuffled every batch. The same sensitivity applies to any comparison drawn across separate judging jobs, including the dose-by-dose judge scores of Appendix D: there it adds noise rather than structure, since batch composition is unrelated to dose, but it is one reason those judge scores resolve less finely than the per-clip paired signals reported beside them.

Two limits remain. Reversal is a strong manipulation on these clips, and classifying all 26 action categories by dissipation found almost none that reverse into legal physics, which is why the stratified result is reported alongside; but the stratification was fixed after the main result was known and is secondary. And this judge is a video-language model, a family whose weak temporal modelling is documented, so the null may be a property of this evaluator rather than of physical validity.

H LIMITATIONS IN DETAIL

This appendix carries the arguments behind several of the limitations stated in Section 5. Nothing here is a further limitation; each entry expands one that the main text already claims.

Why the first version of the reversal probe failed. The reason is a fact about the instrument rather than about our implementation: **flow statistics are invariant neither to time reversal nor to re-encoding**, which any such probe needs to establish before it can be trusted. What that ruled out was reversal as a motion-preserving degradation, not the question of whether these quantities respond to time direction, which needs no such invariance and is answered in Section 4.5 by controlling the measured motion change instead.

The second optical-flow estimator, in numbers. On all 3372 VideoPhy-2 clips, with motion recomputed by Farneback, a classical estimator that learns nothing and shares no training data with the RAFT family: the reward read off one network correlates with motion read off the other at 0.736, the two motion statistics correlate with each other at 0.708, both quantities read off the second estimator correlate at 0.854, and the published figure, with both read off the same network, is 0.9032. So a shared estimator is not what produces the coupling, but part of the 0.9032 does depend on one. Two cautions travel with this. The agreement between the estimators, 0.708, is the ceiling this test can speak to, and it is low enough that the comparison is only just interpretable. And the second estimator is the noisier of the two, which attenuates both 0.736 and 0.708, so the cross-estimator coupling is a lower bound rather than a point estimate; we did not measure the reliability that would let us correct it.

Dose-by-dose behaviour of the judge in the second run. The five doses are reported in full in Appendix D; two things about them belong here rather than there. Reporting that judge's approval as evidence that the policy improved would be the same error we document in Section 4.1, committed by us. And the coupling reported there, between motion lost and judge score gained, **is a statement about which channel the reward acts through, not about whether the videos got better.**

What PhyMotion's own reported gain measures. PhyMotion reports a VideoPhy-PC gain of +7.1%, obtained by using it as a *training* reward. That is a different quantity from the one we measure, which is how well its score ranks clips a generator already produced, and the two are not in conflict.

Why this is not evidence of a motion-biased judge. We tested that separately and it does not hold: on 3350 human-annotated clips, human ratings move in the same direction and by a similar amount as the judge, at a `flow_peak` AUC of 0.398 against the judge's 0.349 to 0.399 across the four arms. Humans and the VLM agree that less local motion reads as better physics. The problem is not a biased judge; it is that both of them, and the reward, are reading the same channel.

The time-reversal ablation, in full. The design assumed that reversing a clip negates every displacement and therefore leaves motion magnitude untouched while changing the physics, since a falling object rises and spilled liquid returns. Reversing the decoded frames, with no re-encoding anywhere in the loop, shifts peak flow by a median -2.6% (p = 1.6e-05, n = 435), against a pre-registered gate of $|\cdot| < 1\%$ and $p > 0.05$. Occlusion and the aperture problem break the symmetry the design assumed. A null `ffmpeg` pass alone moves peak flow by -2.2% (p = 1.1e-07), which is the second measurement worth having. Any degradation-based probe of a motion-derived reward needs both of these before trusting its own contrast.

Calibration of the Hotelling–Williams test on ranks. The test was derived for Pearson correlations. Fed rank correlations it is liberal when both predictors track the outcome strongly, with a measured false-positive rate of up to 0.079 at a nominal 0.05. Our comparisons sit at $\rho \approx 0.2$, where we measure it to be exact (0.052), and every such comparison is reported beside a bootstrap of the difference in $|\rho|$ that carries no distributional assumption.

I Related work in detail

This appendix carries the per-work discussion behind Section 2. Every claim the main text makes about this literature is stated there; what is here is which paper does what.

The rest of the reward family. PISA (Li et al., 2025) compares a generation to a ground-truth simulation through frozen segmentation, flow and depth models, and PhysRVG (Zhang et al., 2026b) optimizes the deviation of tracked centroids from ground-truth positions with GRPO (Shao et al., 2024). Both therefore need a paired reference and cannot rank a policy's own open-domain samples. Nearby, PhysCorr (Wang et al., 2025) learns a reward model over object stability and interaction, VGGRPO (An et al., 2026) runs GRPO in latent space against geometric rewards, and VideoRLVR (Zhu et al., 2026) carries verifiable rewards into video reasoning with objective success criteria.

The benchmarks. PhyGenBench (Meng et al., 2025), PhyWorldBench (Gu et al., 2026), PhyGround (Lin et al., 2026), PQSG (Pothiraj et al., 2026) and TiViBench (Chen et al., 2025) each build a finer-grained physics evaluation, CRONOS (Begiristain et al., 2026) tests stability under interventions on viewpoint, scene and appearance, and WCS (Rakheja et al., 2025) proposes object permanence, relation stability and causal compliance as sub-metrics, deferring validation to future work.

Motion as a confound, in more detail. A definitional example already sits inside the reward literature, since one of VGGRPO's two rewards is a camera-motion smoothness penalty, a motion statistic by construction, whereas our claim about the kinematic residual is a statistical one about a reward that does not present itself that way. Two methods also move motion in opposite directions while both reporting improved physical plausibility on the same benchmarks: we sample VideoREPA's (Zhang et al., 2025) released checkpoint against its own base model and find motion down (Figure 3), while LaMo (Jiang et al., 2026) describes the failure it repairs as collapsed temporal structure with no observable motion. We have not run the two side by side and do not claim they contradict each other.

Reward hacking. TRACE (Wang et al., 2026b) detects implicit reward hacking by measuring how little of a chain of thought is needed to collect the reward, SoliReward (Lian et al., 2025) targets the

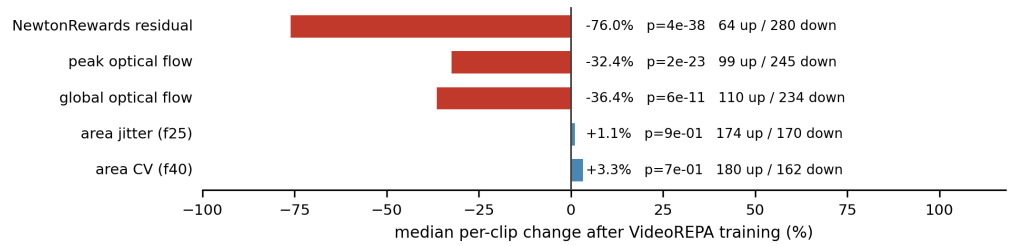


Figure 3: **An independent check on a model we did not train** (seed 1234, $n = 344$). We sample VideoREPA's released checkpoint and its own base model on the same prompts and seed. The kinematic residual moves by a large margin while the motion-orthogonal signals do not, which is the signature the controlled experiment produces, here on a training run that is not ours.

susceptibility of learned video reward models to hacking, and Diffusion-DRF (Wang et al., 2026c) motivates dense feedback partly by reward exploitation. The mechanism is analogous to ours, which is why the boundary is worth stating: TRACE's premise is that a loophole takes less effort than the task, ours that motion is easier to measure than physics.

Shortcut learning in physics-aware generative models. REPA-P (Jia et al., 2026) documents shortcut learning in PDE-constrained diffusion, where constraints imposed only on the output leave intermediate representations free to fit spurious cues, and Kang et al. (2025) find a video model on a deterministic 2D simulator generalising by colour before size before velocity before shape. Both are generators preferring a cheaper attribute, one level below the failure we report, and the second supports a caution we state about our own, since optimization drives the policy outside the range in which the motion–rating relation has been observed. Their confound is not ours: rich appearance distracting a model from physics is not motion inflating a measurement, and we draw that line rather than claim the precedent.

J THE MANIPULATION CHECK AND THE JUDGE, DOSE BY DOSE

These are the two panels that Figure 2 leaves out, on the same two arms and the same five doses. Neither carries the claim: the first is a manipulation check and the second is a null.

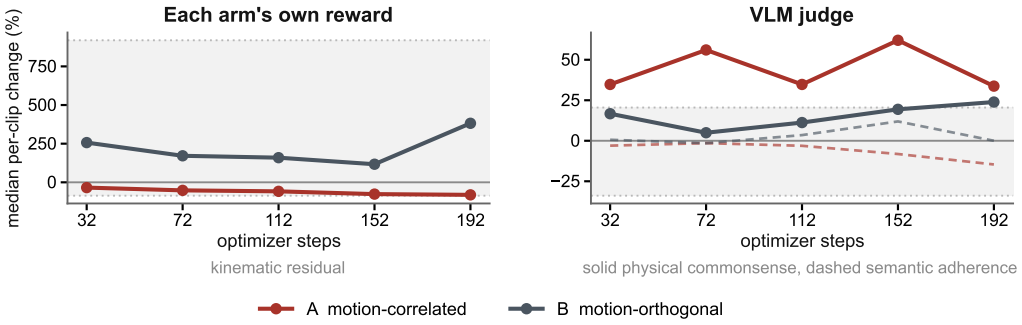


Figure 4: **Each arm optimizes its own reward, and the judge separates neither. Left:** the manipulation check. Arm A drives the kinematic residual down and arm B drives it up, which is what makes Figure 2 a comparison between two working optimizations rather than between an optimization and a failure to train. **Right:** neither judge head separates the arms at any dose in this run, and the second run reverses the direction, which is why no claim in this paper rests on the judge moving a particular way.

K THE COUPLING, SOURCE BY SOURCE

Figure 5 shows the correlation behind Section 4.2 one source and one generator at a time, so that the pooled range can be read against the spread it summarises. No source and no individual generator sits outside it, which is what rules out the reading that this is a property of one weak model or one benchmark.

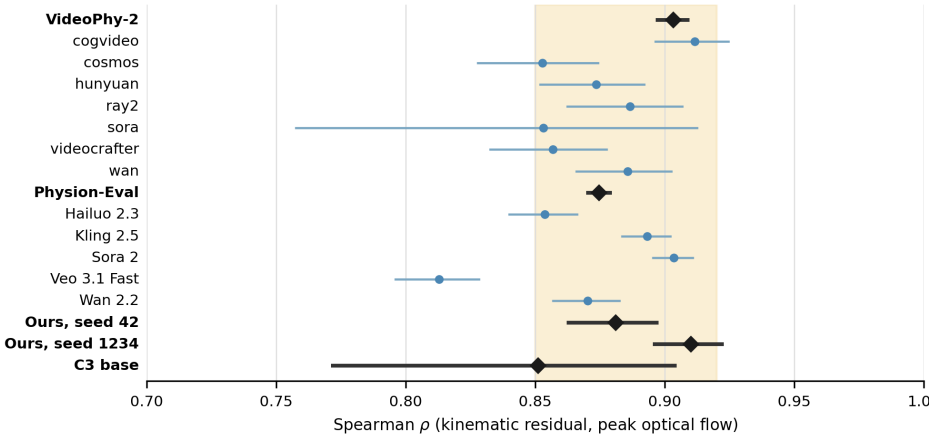


Figure 5: **A verifiable physics reward tracks motion magnitude, across every source we can measure it on.** Spearman ρ between a reimplemented kinematic-residual reward and peak optical flow; diamonds are sources, circles the generators within them, bars 95% intervals. The VideoPhy-2 value shown is 0.9032 ($n = 3372$); Section 4 reports 0.9149 under a second, independently implemented pipeline over **the same flow field**, so the coupling does not turn on the downstream statistic.

L THE VERIFICATION GAP, AND WHY ITS TWO HALVES ARE NOT COMPARABLE

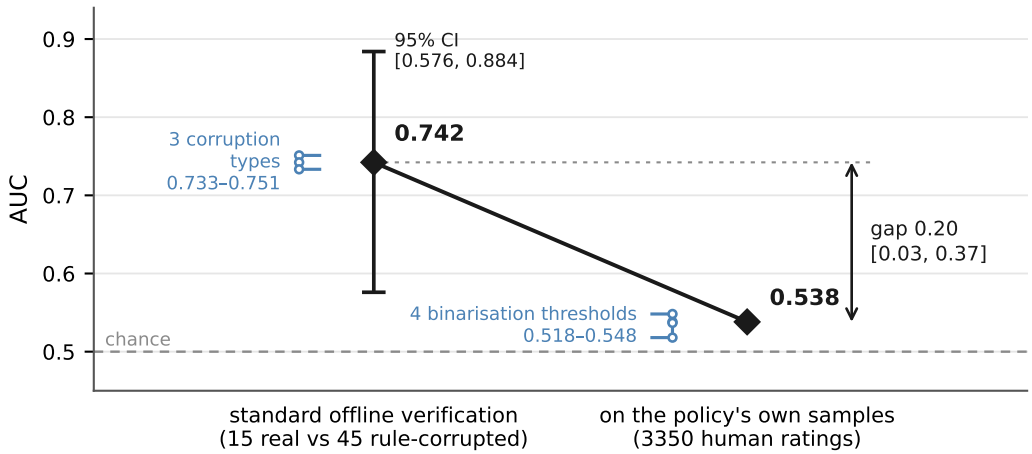


Figure 6: **What the standard protocol reports, and what the same reward scores on the samples a policy actually produces.** Values are given in Section 4.1.

The reward the gap is measured on. Our own trajectory residual is computed on the tracked centroid of the dominant moving object and is the sum of three terms: a least-squares parabola fit to the centroid path, a consistency term between the optical-flow displacement and the tracked displacement, and a motion-smoothness term entered at weight 0.05. It is mapped to a reward by

$-\log_{10}$ of that sum. It is ours and is not one of the audited published rewards; we report it because it is the reward we could run the standard separation protocol on end to end, from the videos. Where a single term is meant rather than the sum, we say so: the parabola term is what Section 4.1 reports on gravity and free-fall clips, and what Section 4.2 compares against pure motion.

Figure 6 draws the comparison that opens Section 4.1. The two vertical marks carry different meanings and are drawn differently for that reason. The capped bar on the left is sampling error, and it is wide because the corrupted-versus-real comparison rests on fifteen real clips: the standard protocol we are criticising is usually run at about that size, so it cannot measure its own optimism precisely either. The bracket on the right is not sampling error. On thousands of ratings the estimate is tight, and the range shown is over the four ways of binarising the human score, which is a robustness check rather than an uncertainty. The open circles behind each mark are the individual values, three corruption types on the left and four thresholds on the right, so that neither summary rests on a single choice.