PCBENCH: PHYSICAL COMMONSENSE BENCHMARK FOR AGENTIC VIDEO EVALUATION AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generators now render photorealistic clips, but do the objects in them move as physics requires? Released balls hover, pendulums stall and requested collisions never happen, yet common visual-quality metrics and vision-language judges do not directly measure conformity to explicit motion laws. We introduce PCBench, a physical-commonsense benchmark that pairs 3,093 real videos in ten phenomenon categories with 4,049 generated videos from six generator families and 945 simulations with perturbations of known severity. Four-axis annotations of generated clips separate prompt following from physical plausibility. Across all six generators, the requested phenomenon is realized less reliably than general physical plausibility. For mechanics, where motion laws are explicit and can be checked quantitatively, we propose agentic evaluation: a tool-using workflow grounds and tracks the target object, fits an executable motion program and reports an interpretable Motion-Law Conformity (MLC) score. Without fitting to human labels, the residual increases with severity across all 18 controlled perturbation pairs, and MLC has the highest point correlation with mean human ratings on 216 generated mechanics clips ($\rho = 0.38$ versus 0.21 for VideoScore2), although the paired interval for this difference includes zero. Because the program that scores a motion can also produce it, the same programs close the loop to generation: they render temporal controls for a video diffusion model, re-measure the output and repair measured failures. Program-controlled generation produces the requested motion where resampling rarely does, and three human raters on average favour it in all 16 matched pairs. Code is available through an anonymous link.[1]

1 INTRODUCTION

Video generators now render clips of striking visual realism, which has led to proposals to use them as world simulators and as data engines for physical AI (Meng et al., 2025; Zhou et al., 2025). These uses depend on a question that visual realism cannot answer: *do the objects in a generated video move as physics requires?* Often they do not. Asked for a ball launched horizontally, a generator can render a well-lit ball that simply hovers (Figure 1, left); asked for a collision, it can produce a clip in which nothing collides. Every frame of such a clip looks plausible. The failure appears only when the motion is checked against the law that the prompt implies, and measuring this reliably is a prerequisite both for tracking progress and for building better generators.

Why plausibility scores need motion measurements. Common perceptual and distributional metrics such as PSNR, SSIM, LPIPS and FVD do not directly test conformity to motion laws (Hore & Ziou, 2010; Wang et al., 2004; Zhang et al., 2018; Unterthiner et al., 2018); physics benchmarks collect one plausibility verdict per clip from people or vision-language models (Bansal et al., 2024; 2025; Meng et al., 2025); and learned judges return scores and textual rationales without an explicit fitted motion law (He et al., 2024; 2025). This has three consequences. A single verdict conflates whether the requested phenomenon happened with whether whatever happened is physically consistent: a correct free fall shown for a projectile prompt is sound physics but fails the request. Agreement is limited: on our generated mechanics clips, the three general-purpose evaluators reach at most $\rho = 0.21$ against mean human ratings of the requested phenomenon. A plausibility verdict

[1] https://anonymous.4open.science/r/pcbench-iclr27-review-code-2F15/

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.

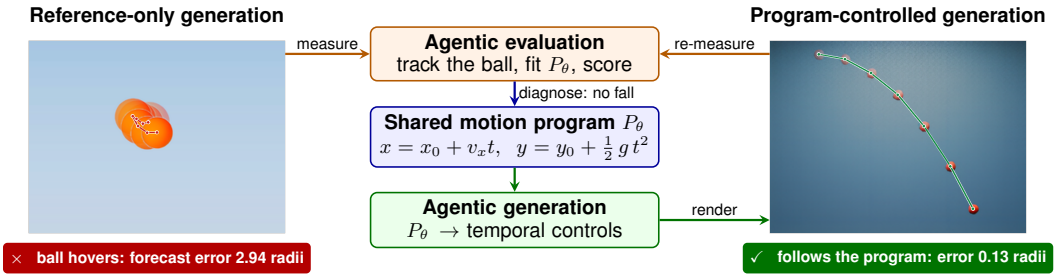

Figure 1: Measure, program, render, re-measure. For a horizontal-projectile request, the reference-only output (left; seven frames overlaid, earlier ones fainter, with the tracked path) hovers instead of falling, and agentic evaluation measures a large forecast error. The shared motion program that encodes the requested law is executed as temporal controls for the same video diffusion model, and the program-controlled output (right) follows the parabola, as re-measurement confirms. This matched pair illustrates the loop; diagnosis-directed repairs are evaluated on spring and pendulum failures (Section 5), and the evaluation itself is validated on PCBench against human labels and controlled perturbations.

also does not quantify deviations from a fitted motion law, limiting its use for verifying a regenerated clip.

Our thesis: measure the motion, with a program that can also produce it. We argue that physical commonsense in video should be *measured* rather than judged, and that the measurement should be *executable*. **A physics evaluator should output a program that explains the observed motion, can be checked against the video, and can be run to produce the motion that was requested.** For mechanics, where motion laws are explicit, this is achievable: a released ball, a sliding block or a swinging pendulum traces a trajectory, and the law it should obey is a short program. We cast the measurement as *agentic evaluation*, a tool-using workflow that decides which object to follow, whether it is observed well enough, which law the prompt calls for and how to fit it. It returns an auditable record instead of a verdict: the tracked trajectory, the fitted program and its normalized residual, summarized as the Motion-Law Conformity (MLC) score.

Closing the loop. Because the program that explains a motion can also produce it, measurement extends to generation (Figure 1). Executed forward, the program renders temporal controls for a controllable video diffusion model; the generated video is measured by the same evaluator; and an agent diagnoses the result and either acts on the controls or stops, citing measured evidence for every decision. Evaluation and generation thus share one representation and one measuring stick, which turns PCBench into a testbed for physics-aware generation in which every improvement is checked on the final video rather than asserted.

Key findings.

1. **Generators miss the requested phenomenon more than physics in general.** Every one of six generator families scores lower on the requested phenomenon than on general physical plausibility and object consistency, and interaction and constraint motions such as collisions and pendulums score 1.21 points lower than single-body kinematics (Section 3).
2. **An interpretable measurement tracks perturbations and human ratings.** The fitted residual increases with severity across all 18 controlled perturbation pairs. The corresponding MLC score reaches $\rho = 0.38$ against mean human ratings on 216 generated mechanics clips, versus 0.21 for VideoScore2, 0.18 for PhyGenEval and 0.10 for VideoPhy-2; its lead over VideoScore2 remains uncertain (Section 4).
3. **The program that scores a motion can also produce it.** With a fixed renderer, program-controlled outputs have lower motion error than reference-only outputs in every matched pair where both can be scored, whereas resampling reaches the requested motion in 1 of 25 attempts; measured failures are repaired, and three human raters on average favour program control in all 16 matched pairs (Section 5).

4. **Looking complete is not being consistent.** Scrambling the order of the controls leaves every visible event intact but raises the forecast error by two orders of magnitude. The evaluator detects this and human raters penalize it, whereas VideoScore2 does not even separate program-controlled from reference-only outputs (Section 5).

Contributions and scope. We contribute PCBench, with linked real, generated and controlled videos and a four-axis protocol; agentic physical evaluation with seven executable motion programs and the label-free MLC score; and shared motion programs that turn the evaluator into a verifiable control loop for video generation, with code for replaying our scores. We focus the measurement on mechanics because its laws are explicit and its trajectories observable, which permits quantitative checks, and because it contains motions that generators find especially hard; the human labels cover all ten categories. The closed loop is demonstrated in a controlled, matched study with one renderer and supplied programs, which tests whether each step can be verified on the final video rather than comparing agent policies. Section 2 reviews related work, and Sections 3–5 present the benchmark, the evaluator and closed-loop generation.

2 Related Work

Evaluating physics in generated video. Reconstruction, perceptual and distributional metrics such as PSNR, SSIM, LPIPS and FVD measure visual or distributional fidelity without directly testing conformity to physical laws (Hore & Ziou, 2010; Wang et al., 2004; Zhang et al., 2018; Unterthiner et al., 2018). Learned evaluators predict multi-aspect human judgments, including a physical-consistency aspect (He et al., 2024; 2025), and DEVIL scores the dynamics of generated video (Liao et al., 2024). Physics-focused benchmarks collect human and automatic judgments of physical commonsense: VideoPhy and VideoPhy-2 cover material interactions and everyday actions (Bansal et al., 2024; 2025), PhyGenBench spans several physical laws and proposes the PhyGenEval evaluator (Meng et al., 2025), Physics-IQ compares generated continuations with matched real experiments (Motamed et al., 2025), and PAI-Bench extends evaluation to broader physical-AI tasks (Zhou et al., 2025). Multi-axis, physics-focused evaluation is therefore not new. PCBench differs in linking every generated clip to a real clip of the same phenomenon, adding controlled perturbations of known severity, separating the requested phenomenon from general plausibility in its labels, and pairing the labels with a measurement that can also drive generation.

Trajectory-based physics evaluation. Open-vocabulary detection and promptable video segmentation make object trajectories extractable from a text prompt (Liu et al., 2024b; Ravi et al., 2024; IDEA Research, 2024). InterDyn measures trajectory-based motion fidelity (Akkerman et al., 2024), and Morpheus tracks objects in real and generated clips, filters unreliable observations and evaluates Newtonian dynamics with law-specific models (Tragoudaras et al., 2026), which overlaps our tracking and observability components. Our motion programs follow the executable equation format of LLM-SR (Shojaee et al., 2025), but use fixed, physically motivated program skeletons rather than open-ended equation discovery, and we validate the resulting score against both controlled perturbations and human labels of the requested phenomenon.

Physics-grounded and agentic generation. PhysGen combines image understanding, rigid-body simulation and generative rendering (Liu et al., 2024a). PhysAgent verifies and repairs simulation programs with reflective agents before synthesis (Li et al., 2026), and a separate PhysAgent refines force fields from tracked motion with multi-agent feedback for 4D synthesis (Lv et al., 2026). Simulation programs, controllable rendering with VACE-style conditioning (Jiang et al., 2025) and trajectory feedback thus all have precedents. Our contribution is the link between them: one shared motion record that is fitted to observed video for evaluation, executed to render temporal controls for generation, and used again to re-measure the final RGB output, so that generation is judged by the same measurement that evaluates real and generated clips.

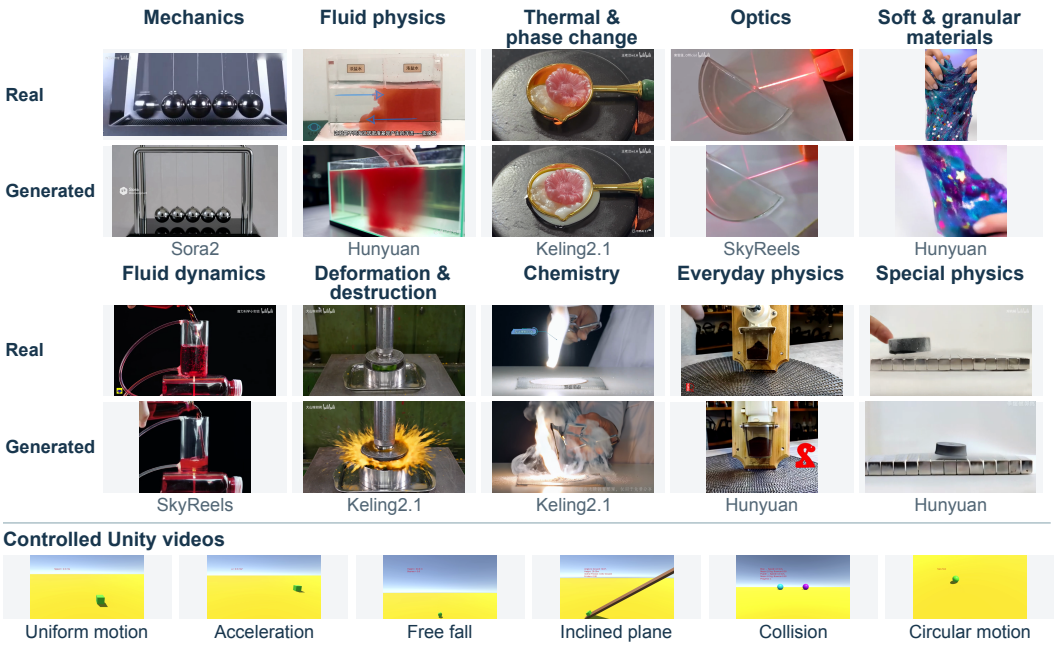

Figure 2: **A glimpse of PCBench.** Each category shows a collected real-video frame above a generated counterpart prompted by its caption; generator names appear below. The bottom strip shows the six nominal Unity motion scenes used for controlled perturbations. These selected, unretouched frames illustrate content diversity, not success rates; paired frames are not event-aligned. The collection workflow is in Appendix Figure 5.

3 THE PCBENCH BENCHMARK

PCBench links real, generated and controlled videos of the same phenomena and labels them on four axes, which reveals that generators get the requested phenomenon wrong more often than physics in general.

Data. PCBench covers ten physical-phenomenon categories, from mechanics, optics and fluids to thermal change, deformation and chemistry, each divided into fine-grained topics (Appendix A.1). It has three linked parts (Figure 2). **Real videos (3,093)** are 2–10 s clips from YouTube and Bilibili that show the target phenomenon, each with a caption describing the physical process. **Generated videos (4,049)** use these captions as prompts for six generator families (Seedance, SkyReels, Sora2, Hunyuan, Keling and Wan), so every generated clip has a real counterpart. **Controlled Unity videos (945)** simulate six elementary motions cleanly and with one of eight perturbations, such as acceleration jitter, an invisible wall or gravity drift, at controlled severity, giving ground truth for how wrong a motion is (Appendices A.2 and A.5).

Four-axis annotation. The initial annotation records one four-axis score set per generated clip on 1–5 scales (Appendix A.3). The *Prompt Consistency Score* (PCS) measures coverage of the prompt's objects, actions, attributes and order; the *specific Scientific Plausibility Score* (SCS-sp) whether the requested phenomenon is correctly realized; the *general Scientific Plausibility Score* (SCS-ge) whether whatever physical processes appear are self-consistent; and the *Object Consistency Score* (OCS) visual and temporal object integrity. Annotators score SCS-ge before seeing the prompt, so a correct free fall shown for a projectile prompt receives low PCS and SCS-sp but can receive high SCS-ge. SCS-ge and OCS carry information distinct from PCS and SCS-sp (Appendix A.4). For the 216 generated mechanics clips, four additional annotators supplied two further assessments per clip: 215 clips have three valid scores and one has two, after one unrateable return. We use their unrounded mean on each axis. Separately, when three raters apply the protocol to the 38 generation-study clips of Section 5, agreement on SCS-sp reaches Krippendorff's $\alpha = 0.82$.

Table 1: Human four-axis scores of six generator versions, using clips with all four valid scores and macro-averaging over each model's represented categories. The 216 re-rated clips use mean ratings; other clips retain the historical score. Overall averages the four axes. Every model scores lower on SCS-sp than on SCS-ge and OCS.

Model	Clips	Cat.	PCS	SCS-sp	SCS-ge	OCS	Overall
Sora2	904	10	**3.550**	**3.418**	**3.821**	**4.521**	**3.827**
Keling2.1	201	10	3.342	3.206	3.811	4.174	3.633
Wan2.2	413	8	3.347	2.730	3.338	3.840	3.314
Seedance1.0pro	69	8	3.023	2.717	3.318	4.072	3.282
SkyReels	657	8	2.468	2.652	3.403	3.647	3.042
Hunyuan	1467	10	2.322	2.216	3.256	3.436	2.808

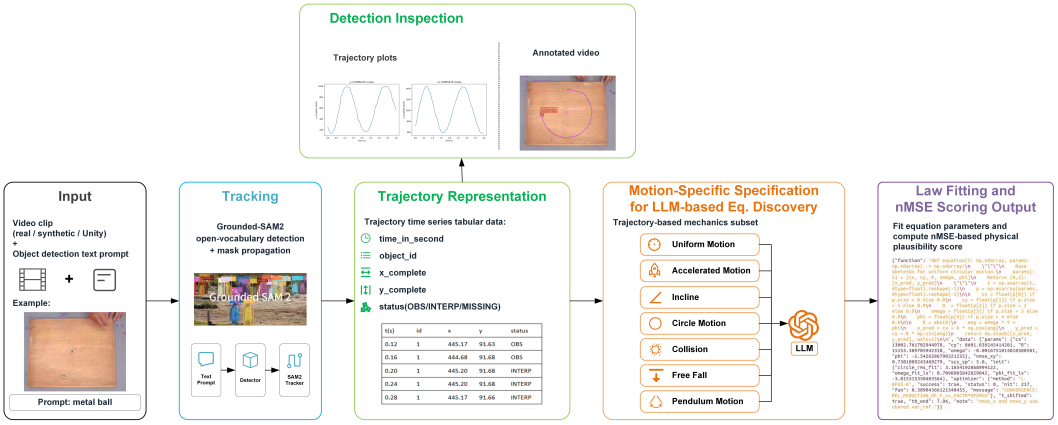

Figure 3: Agentic evaluation. Open-vocabulary grounding and tracking convert a clip and an object prompt into a trajectory table with observed, interpolated and missing frames. A motion program fits its equation to the trajectory and returns a normalized residual (nMSE), summarized as the MLC score; lower residuals indicate closer conformity to the motion law.

Generators miss the requested phenomenon. Table 1 benchmarks six generator versions. Sora2 has the highest point mean on every axis, with a small SCS-ge difference from Keling2.1; Hunyuan is lowest on every axis; an additive prompt-plus-generator model on the 1,010 prompts rendered by several models largely preserves this ordering (Appendix D). The more important pattern is shared by all models: *every generator scores lower on SCS-sp than on SCS-ge and OCS*. Generators keep objects coherent and dynamics broadly plausible more often than they realize the phenomenon they were asked for, which a single plausibility verdict would hide. Within mechanics, interaction and constraint motions (collision, pendulum and circular motion) receive SCS-sp scores 1.21 points lower than single-body kinematics ($[0.92, 1.50]$), and a gap of 0.56 ($[0.25, 0.87]$) persists with generator fixed effects. These failures concern the dynamics of specific objects, which motivates measuring them directly.

4 Agentic Physical Evaluation

Agentic evaluation turns a clip and its prompt into an auditable measurement, and without any labels it tracks both controlled perturbations and human judgments. It proceeds in three steps: ground and track the target, fit a motion program, and score the fit (Figure 3).

Ground and track. A Grounded-SAM2 pipeline (IDEA Research, 2024; Liu et al., 2024b; Ravi et al., 2024) detects the target on the first frame from a text prompt (for example "`metal ball.`") and propagates SAM2.1 masks through the clip. Mask centroids give image-coordinate positions; interior gaps are interpolated and labelled, and fitting uses the trajectory up to its first missing frame.

Table 2: Spearman agreement of automatic evaluators with human SCS-sp on the 421-video mechanics cohort (216 generated clips, with mean human ratings). No method uses PCB labels; results on the Test95 split, including VLM judges, are in Appendix D. Best per column in bold.

	Real + generated		Generated only	
Method	$\rho \uparrow$	p	$\rho \uparrow$	p
PhyGenEval (overall)	0.058	2.4×10^{-1}	0.181	7.7×10^{-3}
VideoScore2 (physical consistency)	0.115	1.8×10^{-2}	0.208	2.1×10^{-3}
VideoPhy-2 (physical commonsense)	0.203	2.6×10^{-5}	0.105	1.3×10^{-1}
MLC (ours)	**0.360**	2.7×10^{-14}	**0.382**	6.8×10^{-9}

Select and fit a motion program. Seven executable motion programs cover uniform linear, accelerated linear, uniform circular, collision, free-fall, inclined-sliding and single-cycle pendulum motion. Each declares its trajectory channels, an equation, preprocessing, valid-segment selection and a normalized residual; for example, accelerated motion fits $x(t) = b_x + v_x t + \frac{1}{2} a_x t^2$ and the analogous $y(t)$, and collision fits two linear regimes joined by a sigmoid changepoint. The programs use the executable format of LLM-SR (Shojaee et al., 2025) with fixed equations, fitted by closed-form initialization and deterministic L-BFGS-B refinement, so every score is reproducible and no language-model equation search enters it (Appendix B).

Score. For channel c with residuals $r_{c,i} = \hat{z}_{c,i} - z_{c,i}$ over N frames, the program returns

$$E_{\text{fit}} = \sum_c w_c \frac{\frac{1}{N} \sum_i r_{c,i}^2}{s_{\text{ref}}^2 + \varepsilon}, \tag{1}$$

with motion-specific weights w_c and reference scale s_{ref}. The *Motion-Law Conformity score* (MLC) $1 - \tilde{E}$ makes families comparable by min–max normalizing E_{fit} within each family, $\tilde{E} = (E_{\text{fit}} - \min_m)/(\max_m - \min_m)$, over the mechanics cohort; a higher score means closer conformity to the requested law. MLC uses no human labels, and computing the ranges from training and validation clips only changes held-out correlations by less than 0.002 (Appendix D). In the generation study, the shared record of Section 5 additionally returns *unknown* when the observations do not support a fit.

A controlled measurement setting. To isolate the measurement from upstream choices, the benchmark experiments take the motion family from the benchmark category and use per-clip object prompts, with the target track chosen manually when several objects are tracked (227 of 421 clips). Two programs are deliberately simple proxies: the pendulum fits a constant-angular-speed arc over one half cycle of a Newton's cradle's outer balls, and free fall fixes the horizontal position. These experiments therefore validate the measurement itself, and the agent's decisions are exercised in the generation loop of Section 5.

MLC tracks controlled perturbations. Within every one of the 18 Unity motion–perturbation pairs, the residual increases with perturbation severity. Averaged over base conditions, the within-condition Spearman correlation between severity and residual is at least 0.968 in 15 pairs; the weakest are free fall under a stop burst (0.742) and under acceleration jitter (0.817), because the free-fall program scores only the first descent (Appendix A.5).

Agreement with human ratings on 216 generated clips. Our primary comparison uses all 216 generated mechanics clips, with the mean of their available human ratings as the target (Section 3). MLC reaches $\rho = 0.382$ (95% source-group interval $[0.267, 0.484]$), versus 0.208 for VideoScore2, 0.181 for PhyGenEval and 0.105 for VideoPhy-2 (Table 2). Paired intervals exclude zero for the differences from VideoPhy-2 and PhyGenEval, but not VideoScore2 ($+0.173$, $[-0.020, 0.359]$). Thus MLC has the highest point estimate among these evaluators, with unresolved uncertainty relative to VideoScore2. Pooled agreement partly reflects differences between motion families: the within-family estimate is only 0.125 ($[-0.015, 0.260]$), so per-clip discrimination within a motion remains limited (Appendix D). The 421-clip view additionally assigns 205 real clips the conventional target

5. Appendix D retains the smaller held-out comparison with Qwen3-VL and the calibrated variant; it supplements the 216-clip result.

5 Agentic Generation with Shared Motion Programs

The program that scores a motion can also produce it. Executed as temporal controls, it makes a fixed video diffusion model generate the requested motion, and the same evaluator verifies the result.

Shared motion record. The evaluator's program becomes the core of a *shared motion record* that serves evaluation and generation alike. The record binds objects and their identities, native time and units, measurements with their visibility and uncertainty, declared physical and observation assumptions, an executable program P_θ with bounded parameters, observed and predicted events, and the evidence behind each decision (Appendix E.1). Keeping predicted events apart from observed ones means that a predicted contact is never mistaken for image evidence of contact, and when the observations do not identify the declared model, the record returns *unknown* rather than a physical score.

Generate and re-measure. Executing P_θ from its initial state produces temporal controls C, rendered in Blender as depth or RGB frames, and a fixed neural renderer produces $V = G_\phi(p, r, C; \xi)$ from a prompt p, an appearance reference r and a seed ξ. We compare *reference-only* generation (**Reference**: prompt and reference image) with *program-controlled* generation (**Controlled**: additionally receives C). Renderer, settings, seed and budget are matched. Each output is then measured again. Held-out forecast endpoints fit the first half of the observed trajectory and predict the second, reporting

$$D(\mathcal{R}, V) = \frac{1}{s_{\text{norm}}} \sqrt{\frac{1}{|I_{\text{valid}}|} \sum_{i \in I_{\text{valid}}} \|y_i - \hat{y}_i(P_{\hat{\theta}})\|^2} \tag{2}$$

in units of the observed object radius s_{norm}. Coverage gates (at least 18 valid centers per 24-frame half) and motion gates (for example, at least two radii of fitted motion for an oscillator) are fixed before measurement, so a small error on a static clip never counts as correct physics.

Agent loop. After measuring V_k, an agent selects an action from a frozen menu or stops, $a_k = \pi(\mathcal{R}_k, p)$, and the record is updated before the next render $V_{k+1} = G_\phi(p, r, C_{k+1}; \xi_{k+1})$. Actions apply or adjust program controls within a fixed range, repair a diagnosed visibility or timing issue, or stop; each must cite RGB evidence and state its expected effect before any new render. Decisions are made by an LLM agent operating the workflow, not by a trained policy.

Setup. All outputs use one renderer, Wan2.1-VACE-1.3B (Jiang et al., 2025), at 640×480 with 40 steps; 49 control frames yield a 48-frame, 4 s clip. Each output receives one observation pass (SAM2.1 from one RGB-grounded prompt) with no retry and no simulator coordinates, and unscorable outputs stay in every denominator. A *main* matched cohort has 20 outputs (Newton's cradle, two rebounds, a spring oscillator and a pendulum, two seeds each), and a *prospective* cohort has 12 outputs (horizontal and oblique projectiles and circular motion) whose programs, seeds and scorers were fixed before generation (Appendix E).

Program control produces the requested motion. Figure 4 shows the visible difference in two contact-driven examples. In all seven matched pairs where both arms can be scored, Controlled has the lower error: rebound residuals fall by 49.6–96.3%, and projectile forecast errors fall from 2.941 and 3.711 radii to 0.126 and 0.055 (Figures 1 and 11). All ten controlled oscillator, projectile and circular outputs pass their motion gates, whereas the matching reference-only outputs give three failed gates, five unscorable clips and two projectile forecasts with errors above 2.9 radii (Appendix Table 24). Resampling does not rescue the reference-only arm: over five seeds for each of the five RGB-controlled configurations, it reaches the requested motion in 1 of 25 attempts, and weak control (strength 0.25) fails in all ten cases, whereas strengths 0.5–1.0 pass in all ten (Appendix E.10). The program controls, not seed luck, produce the motion, and the gates reject outputs that ignore them. Program control shows no resolved paired difference in automatic consistency, flicker or imaging quality (aesthetic quality is lower in the prospective pairs), but controlled outputs inherit much of their appearance from the program render (Appendix E.9).

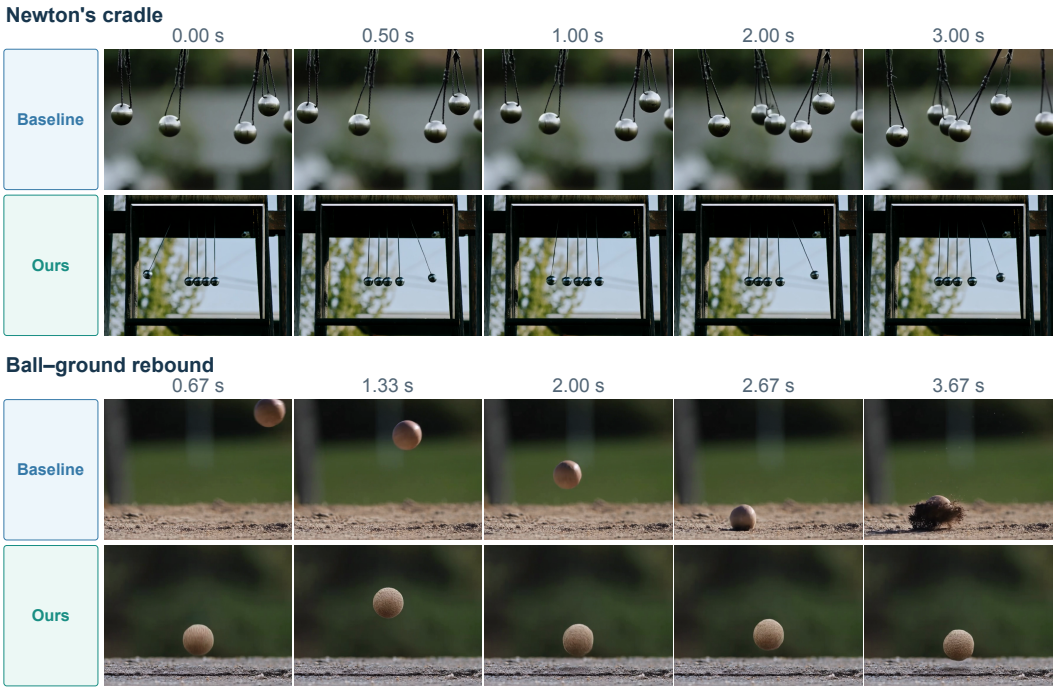

Figure 4: **Program controls recover structured motion that reference-only generation misses.** Selected, time-aligned frames show Newton's cradle (top) and ball–ground rebound (bottom). Baseline produces uncoordinated suspended-ball motion and a delayed impact with splash-like deformation; Ours preserves the cradle's ordered end-ball motion and repeated upward returns after contact. Baseline uses the prompt and reference image; Ours adds 49 program-rendered depth frames. Both use the same Wan2.1-VACE-1.3B renderer, settings and seed within each pair. These are qualitative examples; the cradle has no validated multi-object numerical score.

Diagnose, act and verify. We close the loop on two measured failures. For the reference-only spring failure, the agent diagnoses insufficient horizontal motion and applies the ordered program controls; an unchanged reference-only retry with the same new seed and budget is the comparator. The directed output completes three oscillation turns with 3.80 radii of amplitude and a 0.038-radius forecast error, whereas the retry stays static and fails the motion gate. For the reference-only pendulum failure, the directed output completes four turns (forecast error 0.052) while the retry moves only 0.38 radii. Two further one-step episodes from the same parents complete the requested motion and then stop. Every action the agent selected coincides with the fixed dense-control policy, so the gain comes from the executable program rather than from agent autonomy; what the agent adds is that each step is diagnosed, justified and verified on the final video.

Looking complete is not being consistent. To test whether the evaluator responds to timing rather than to image content, we permute the middle blocks of the 49 control frames, which preserves the first and last frames and the full set of images. Both permuted outputs still pass the visible turning-event rule, yet the forecast error rises from 0.038 to 4.369 radii for the spring and from 0.052 to 9.108 for the pendulum. The evaluator thus separates visible event completion from dynamically consistent timing, which a judgment of the visible events alone cannot do.

People agree; a learned judge does not. Three student raters scored all 38 clips with the four-axis protocol (Table 3). Agreement is high on the prompt and physics axes (ordinal α of 0.72–0.82) and low on OCS (0.24). The rater mean favours Controlled on SCS-sp in all 16 matched pairs, unanimously in 14, favours both directed regenerations over their retries, and penalizes the permuted-control clips (-2.2) even though they pass the visible event rule. Two raters used a portal version that showed an author's draft scores for five main-cohort clips as examples; excluding these clips leaves the conclusions unchanged (SCS-sp $\alpha = 0.86$ on 33 clips; all 12 remaining matched pairs

Table 3: Four-axis human ratings of the 38 generation-study clips by three raters (reference-score exposure is disclosed in the text): paired difference of the rater mean (Controlled minus Reference for the matched cohorts), with SCS-sp wins/ties/losses of the rater mean and the number of pairs favoured by all three raters. Last row: Krippendorff's ordinal α over the 38 clips.

Comparison	Pairs	ΔSCS-sp (W/T/L; all 3)	ΔPCS	ΔSCS-ge	ΔOCS
Main cohort	10	$+3.0$ (10/0/0; 8)	$+3.0$	$+2.0$	$+0.8$
Prospective cohort	6	$+3.2$ (6/0/0; 6)	$+3.4$	$+3.2$	$+0.2$
Directed − retry	2	$+3.7$ (2/0/0; 2)	$+4.0$	$+2.5$	$+1.2$
Permuted − ordered	2	-2.2 (0/0/2; 0)	-2.3	-3.0	-0.5
Inter-rater α (ordinal)	38 clips	0.82	0.80	0.72	0.24

unanimous), although this does not exclude cross-clip anchoring. The rater mean also separates clips that pass the evaluator's gates from those that fail (AUROC 0.93, 20 clips) and tracks the rebound residual ($\rho = -0.80$, 8 clips). VideoScore2, run with the benchmark settings, does not separate the arms (Controlled − Reference physical consistency -0.17, $[-0.54, 0.19]$, three sampling seeds) and is inversely related to the evaluator's labels (AUROC 0.20).

Toward real scenes. A separate controlled-only cohort conditioned on real source frames and captions retains 18 outputs from ten sources, 14 of which support their scoped object and action checks. It has no matched reference-only arm or human ratings, so it shows scoped feasibility rather than a matched effect (Appendix E.12).

6 Discussion and Conclusion

Scope and limitations. *Mechanics first.* The measurement covers seven trajectory-friendly mechanics families, where laws are explicit; fluids, deformation and severe occlusion need other representations, while the benchmark's human labels already span all ten categories. *Selective re-rating.* The primary 216-clip mechanics comparison uses mean ratings (215 clips with three valid scores, one with two); the wider corpus mostly retains single annotations. Within-family agreement remains weak, and the lead over VideoScore2 is uncertain. The separate generation study reaches $\alpha = 0.82$, but two raters saw draft scores for five examples. *A controlled closed loop.* The generation study uses one 1.3B renderer, supplied programs and eight configurations, and its outputs remain visibly synthetic; its matched and prospective design isolates the effect of program control, and scaling to larger renderers and real scenes is future work. *Verification before autonomy.* The agent's actions coincide with fixed dense control, so we establish a verifiable loop rather than a better policy (Appendix E).

Conclusion. Generated videos realize the requested physical phenomenon less reliably than they keep objects coherent and general physics plausible, and verdict-based evaluation neither exposes this clearly nor helps repair it. PCBench makes the failure visible with linked real, generated and controlled videos and a four-axis protocol that separates prompt following from physical plausibility. For mechanics, agentic evaluation turns a clip into an auditable measurement whose label-free MLC score tracks controlled perturbations and has higher point agreement with mean human ratings than the tested label-free baselines, with limited within-family discrimination. Because the same motion program can be executed as well as fitted, measurement extends to generation: program controls produce the requested motion in every scorable matched pair, failures can be diagnosed, repaired and re-verified, and the evaluator exposes motion that looks complete but is dynamically inconsistent. We hope that PCBench and shared motion programs help move the evaluation of generated video from judging whether physics looks right to measuring whether it is.

AI Use Statement

Generative AI tools assisted this work as follows. LLM-based coding agents helped design and implement the experiments, operated the computational workflow of the generation study, including

the diagnose, act or stop decisions of Section 5, which were taken from a frozen action menu, organized data and evidence records, ran statistical re-analyses of saved results, assisted the literature search and revised the manuscript. A video diffusion model (Wan2.1-VACE-1.3B) produced the outputs of the generation study, and the generator families listed in Section 3 produced the benchmark's generated videos. The human authors reviewed these contributions and take full responsibility for all claims, code, data and disclosures.

Ethics Statement

Real clips were collected from public video platforms for research evaluation. Media will be released only where per-item redistribution is permitted; otherwise source links and identifiers will be provided. Human ratings of benchmark and generation-study clips are stored under pseudonymous rater identifiers; participant names and private comments are excluded from the review artifacts. PCBench evaluates physical plausibility and is not intended to certify the safety or factual accuracy of generated videos.

Reproducibility Statement

Sections 4 and 5 and Appendices B, C and E specify the motion specifications, residual normalizers, calibration thresholds, programs, renderer settings, seeds, observation gates and scorers. The anonymous review repository at https://anonymous.4open.science/r/pcbench-iclr27-review-code-2F15/ contains the original single-rating numeric snapshot and code for MLC correlations, three program scene sources, twelve frozen generation conditions, 576 final-RGB observation rows, and a portable CPU scorer that replays all twelve frozen score records, including four unsupported cases. The updated aggregate targets and statistical replay scripts are archived with the manuscript source; the linked anonymous snapshot has not yet been refreshed to those targets. It does not include raw videos, model weights, or the full tracking and neural-rendering runtime.

References

Rick Akkerman, Haiwen Feng, Michael J. Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. Interdyn: Controllable interactive dynamics with video diffusion models. *arXiv preprint arXiv:2412.11785*, 2024.

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. *arXiv preprint arXiv:2406.03520*, 2024.

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*, 2025. URL https://arxiv.org/abs/2503.06800.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2333–2355, 2024.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, Qunshu Lin, Yuanxing Zhang, Ge Zhang, Wenhao Huang, and Wenhu Chen. VideoScore2: Think before you score in generative video evaluation. *arXiv preprint arXiv:2509.22799*, 2025. URL https://arxiv.org/abs/2509.22799.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *International Conference on Pattern Recognition*, pp. 2366–2369, 2010.

IDEA Research. Grounded SAM 2: Ground and track anything in videos. https://github.com/IDEA-Research/Grounded-SAM-2, 2024. GitHub repository.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Jiang_VACE_All-in-One_Video_Creation_and_Editing_ICCV_2025_paper.html.

Qirui Li, Jinkun Hao, Yibo Li, Ran Yi, Paul L. Rosin, and Yu-Kun Lai. PhysAgent: Reflective agentic physics control for physically plausible video generation. *arXiv preprint arXiv:2607.16355*, 2026. URL https://arxiv.org/abs/2607.16355.

Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective. *arXiv preprint arXiv:2407.01094*, 2024.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. *arXiv preprint arXiv:2409.18964*, 2024a. URL https://arxiv.org/abs/2409.18964.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024b.

Chunji Lv, Jiaxi Ye, Yuchen Jiang, Rexar Lin, and Changsheng Li. Physagent: Automating physics-based 4d synthesis via trajectory-grounded multi-agent feedback. *arXiv preprint arXiv:2606.08688*, 2026. URL https://arxiv.org/abs/2606.08688.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43781–43806. PMLR, 2025.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos? *arXiv preprint arXiv:2501.09038*, 2025. URL https://arxiv.org/abs/2501.09038.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. Llm-sr: Scientific equation discovery via programming with large language models. In *International Conference on Learning Representations*, 2025.

Antonios Tragoudaras, Chenyu Zhang, Daniil Cherniavskii, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Evaluating newtonian mechanics in video generative models with real physical systems. *arXiv preprint arXiv:2504.02918v3*, 2026. URL https://arxiv.org/abs/2504.02918v3. Version 3, revised June 29, 2026; first submitted April 3, 2025.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. Pai-bench: A comprehensive benchmark for physical ai. *arXiv preprint arXiv:2512.01989*, 2025.

A BENCHMARK DETAILS

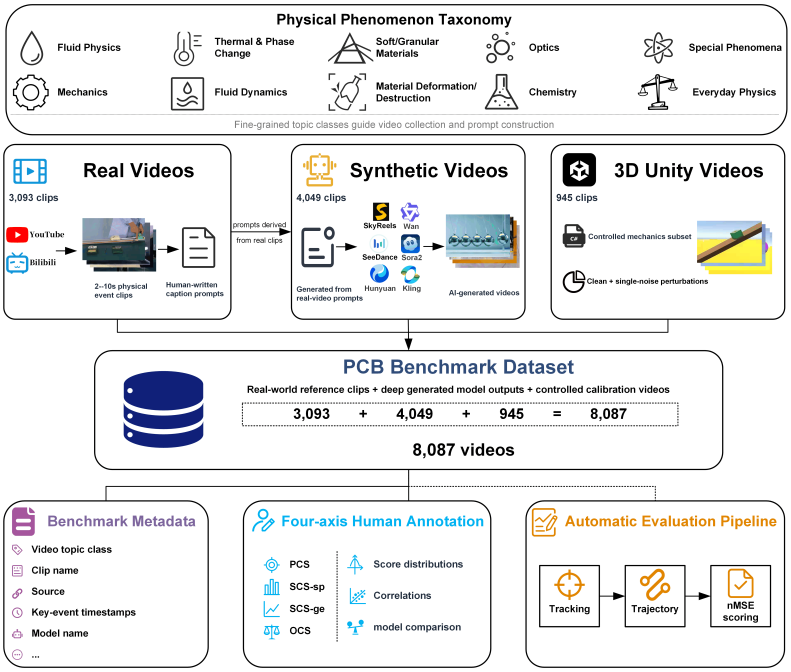

Figure 5: Construction of PCBench. A ten-category phenomenon taxonomy guides the collection of real clips, whose human-written captions become prompts for six generator families. Controlled Unity videos add clean and single-perturbation mechanics scenarios. Generated clips carry four-axis human annotations; the mechanics subset also supports the evaluator of Section 4.

A.1 PHENOMENON TAXONOMY

PCBench covers ten physical-phenomenon categories, each divided into fine-grained topic classes (Table 4). The topic classes serve as search keywords for collecting real videos and determine the phenomenon described by each caption.

Table 4: Taxonomy of Physics-Related Video Types

Physics Category	**Video Topic Classes**
Fluid Physics	Surface tension experiments, demonstrations of liquid stratification and density, buoyancy experiments with floating and sinking objects, centrifugal force demonstrations using a bucket of water, pouring liquid from a bottle, time-lapse diffusion of gas or pigment, slow-motion water splashes, close-up shots of salt dissolving in water, salt and sugar dissolution experiments, liquid splash experiments, coins floating on water due to surface tension, time-lapse water evaporation experiments.
Thermal & Phase Change	Ice melting demonstrations, boiling water experiments, melting of sealing wax stamps, heating metal rods, egg-frying/cooking experiments, demonstrations of metal melting and solidification, wax melting and solidification demonstrations.
Soft and Granular Materials	Slow-motion cloth falling, slime stretching and deformation.
Optics	Slow-motion colored smoke bomb videos, Tyndall effect demonstrations, light color decomposition, light refraction/reflection experiments, demonstrations of primary-color mixing, slow-motion smoke rings.
Special Physics Phenomena	Superconducting levitation demonstrations.

Mechanics	Slow-motion bouncing balls, elastic collisions of metal balls, inelastic collision experiments, tablecloth inertia experiments, friction demonstrations, projectile motion demonstrations, floating objects in space, centripetal force experiments, uniform circular motion demonstrations, uniform/accelerated linear motion experiments, slow-motion gravitational falling.
Fluid Dynamics	Siphon experiments, candle convection and smoke-flow demonstrations.
Material Deformation and Destruction	Videos of vases/cups/bowls breaking, high-speed balloon bursting, slow-motion watermelon cutting, object smashing and shattering, object deformation experiments, small-scale avalanche demonstrations, can-crushing experiments, pressure-based crushing demonstrations, hydraulic press crushing of fruits.
Chemistry	Foaming chemical reaction experiments, time-lapse metal rusting, magnesium ribbon combustion experiments, slow-motion combustion reactions, elephant toothpaste experiments, slow-motion small-scale explosions.
Everyday Physics	Towel wringing experiments, towel twisting processes, vegetable cutting experiments, close-up shots of coffee bean grinding.

A.2 Data statistics and release structure

Tables 5–7 give the number of videos per data source and category. Each real clip is 2–10 s long and is paired with its human-written caption. Generated clips are grouped by model family; the model versions compared in Table 1 are subsets of these families. The 945 Unity clips cover six motions, each with three perturbation types; each clip contains exactly one perturbation type at a controlled intensity.

Table 5: Collected real videos by physical category.

(a) Real videos

Physics Category	**Number of Videos**
Chemistry	214
Everyday Physics	114
Fluid Dynamics	96
Fluid Physics	760
Material Deformation and Destruction	528
Mechanics	695
Optics	230
Soft and Granular Materials	34
Special Physics Phenomena	44
Thermal and Phase Change	378
Total	**3093**

The release is organized into three components, subject to per-item redistribution permission: (i) video clips (real references, generated videos and Unity videos), or source links and identifiers for items that cannot be redistributed; (ii) per-clip metadata, including the category, topic class, clip name, source and page URL and key-event timestamps for real clips, the generator for generated clips, the caption or prompt, and the duration; and (iii) four-axis scores and per-axis justification comments for each generated clip.

A.3 Annotation protocol

The protocol scores generated clips; real clips define the phenomena and prompts and can serve as calibration references for annotators. Table 8 defines the four axes and Table 9 gives the score anchors.

Table 8: Four scoring axes in the PCB annotation protocol.

Axis	**Abbrev.**	**Definition**	**Primary Use**

Prompt Consistency	PCS	The degree to which the generated video faithfully presents the key objects, actions, attributes, and temporal order specified in the prompt.	Measuring prompt coverage and text-video alignment.
Scientific Plausibility-specific	SCS-sp	The degree to which the target physical phenomenon requested by the prompt is correctly and realistically realized in the generated video.	Measuring the quality of the requested physical phenomenon.
Scientific Plausibility-general	SCS-ge	The degree to which the physical or chemical processes that actually appear in the video are globally self-consistent, regardless of whether they are requested by the prompt.	Measuring the model's general physical commonsense.
Object Consistency	OCS	The degree to which objects remain visually and temporally coherent, without artifacts such as disappearance, flickering, clipping, texture collapse, unintended morphing, or geometry breakage.	Measuring visual and temporal object integrity.

Scoring flow. Annotators follow a fixed order: (1) SCS-ge, judging whether the physical or chemical processes that actually appear are self-consistent; (2) PCS, comparing the clip with the prompt; (3) SCS-sp, judging whether the requested phenomenon is correctly realized; (4) OCS, judged independently of the other axes; and (5) an optional short comment for each axis explaining the main reason for the score. Scoring what is shown before comparing it with the prompt is intended to reduce prompt-induced anchoring. For example, a projectile prompt answered by a physically consistent free fall should receive low PCS and SCS-sp but can receive high SCS-ge, and an irrelevant object that suddenly appears should mainly lower OCS.

Table 9: Score anchors for the four PCB annotation axes.

Score	PCS	SCS-sp	SCS-ge	OCS
5	All key nouns, verbs, attributes, and temporal order in the prompt are clearly presented.	The target physical phenomenon is fully realized with physically convincing details.	The entire video is consistent with real-world physical laws, including gravity, momentum, fluid behavior, phase change, optics, and other relevant processes.	No noticeable geometry break, texture popping, identity drift, clipping, or unintended morphing.
4	All major elements are present, with only a minor mismatch such as color, material, background, or camera viewpoint.	The main physical process is correct, with only slight quantitative deviations such as a small trajectory or timing error.	The dominant physical process is correct, with only one minor deviation.	A single minor local artifact appears, such as brief aliasing or a small flicker.
3	The core action is present, but one prominent element is missing or incorrect.	The core phenomenon is recognizable but incomplete or visibly distorted.	The physical principle is recognizable but contains an obvious imperfection, such as unrealistic speed, incorrect viscosity, or excessive energy gain/loss after collision.	A visible artifact appears but does not prevent understanding the scene, such as brief clipping, a one-time size jump, or several seconds of texture crawling.

2	The prompt is only partially matched, with several missing elements or scrambled timing.	The target phenomenon is severely distorted or only weakly suggested.	Key physical laws are largely wrong or mutually contradictory.	Multiple or persistent artifacts appear, such as object splitting, object appearance/disappearance, or long-lasting texture collapse.
1	The video is almost unrelated to the prompt.	The requested phenomenon is missing, impossible, or logically contradictory.	The physics completely collapses, or the video is too corrupted to judge.	Severe deformation or incoherent visual structure makes the action difficult or impossible to recognize.

Interpretation notes. The four axes distinguish different rating criteria. A low PCS does not automatically imply a low SCS-ge. If a video does not follow the prompt but shows a physically reasonable alternative event, its PCS and SCS-sp may be low while SCS-ge remains high. Conversely, a video may match the prompt well but still receive a low SCS score if the requested physical process is implausible.

SCS-sp and SCS-ge also capture different aspects of physical plausibility. SCS-sp focuses only on the target phenomenon specified by the prompt, while SCS-ge focuses on any physical or chemical process that actually appears in the video. For example, if a prompt asks for microgravity but the video shows a pen falling under normal gravity, SCS-sp should be low because the target phenomenon is absent. However, SCS-ge can remain high if the falling motion itself is physically consistent under normal gravity. If the same video also shows other objects floating in the same scene, SCS-ge should be reduced because the scene becomes physically contradictory.

OCS is scored separately from the scientific plausibility scores. Visual artifacts such as clipping, flickering, texture collapse, or object disappearance should primarily reduce OCS. They should only affect SCS-ge or SCS-sp when the artifact directly changes the interpretation of the physical process. For example, if a board physically crushes a can in a reasonable way but briefly clips through it, OCS should be reduced, while SCS-sp and SCS-ge may remain high. If the clipping makes the crushing process physically meaningless or impossible to interpret, then the scientific plausibility scores may also be reduced.

Annotator calibration. To improve scoring consistency, the protocol first familiarizes annotators with the four-axis definitions and the 1-5 score anchors. The protocol calls for calibration examples to illustrate common edge cases, such as a prompt mismatch with physically valid motion, a missing target phenomenon with otherwise normal physics, contradictory physical processes in the same scene, and visual artifacts that should mainly affect OCS. It instructs annotators to focus on the dominant reason for each score and to provide concise comments rather than exhaustive frame-level descriptions.

A.4 Annotation statistics

The expanded annotation export contains 4,217 generated-video records, including additional generator versions beyond the six-version benchmark summary. We replace the historical scores of the exact 216 re-rated mechanics IDs with the unrounded mean of available valid ratings, leaving other IDs unchanged. Four additional annotators supplied 432 returns, two per video: 431 are complete and one is explicitly unrateable. Thus 215 clips contribute three valid scores and one contributes two, including the historical score; missing ratings are not imputed. All 53 generated Test95 clips have three valid scores. Re-rating coverage is selective and does not make the whole corpus a three-rater dataset. The separate 38-video generation-study ratings are not pooled with this campaign.

Figures 6 and 7 now use the updated export's 4,181 complete, valid four-axis records; 36 incomplete or out-of-range records are excluded. PCS and SCS-sp have Pearson $r = 0.869$ (95% source/prompt-cluster bootstrap interval $[0.857, 0.880]$) and are identical on 67% of these clips. Their relation varies by category: $r = 0.95$ for fluid dynamics and 0.93 for fluid physics, versus 0.75 for thermal/phase change, 0.61 for soft/granular materials and 0.65 for special phenomena.

Table 6: Generated video inventory by category and model family. Model-family totals and version-level evaluation snapshots have distinct scopes.

(b) Synthetic videos generated by deep generative models

Category	Seedance	SkyReels	Sora2	Hunyuan	Keling	Wan	Total
Chemistry	2	31	87	143	29	31	323
Everyday Physics	5	12	43	68	15	33	176
Fluid Dynamics	2	70	34	27	2	2	137
Fluid Physics	13	261	199	312	57	180	1022
Material deformation / destruction	17	111	149	362	44	6	689
Mechanics	22	84	172	159	159	157	753
Optics	3	71	72	84	29	74	333
Soft / granular	0	17	8	17	2	0	44
Special physics	0	0	5	7	6	19	37
Thermal / phase change	6	0	135	293	43	58	535
Total	**70**	**657**	**904**	**1472**	**386**	**560**	**4049**

Table 7: Controlled Unity videos by motion and perturbation type.

(c) 3D Unity videos

Physics Category	Noise Type	Number of Videos
Uniform linear motion	AccelJitter	63
	AirWall	63
	StopBurst	63
Accelerated linear motion	AccelJitter	63
	AirWall	63
	StopBurst	63
Uniform circular motion	AccelJitter	42
	CameraJitter	42
	FrictionJitter	42
Elastic/inelastic collision	CollisionImpulseNoise	63
	CollisionLateralImpulse	63
	StopBurst	63
Free-fall motion	AccelJitter	63
	GravityDrift	63
	StopBurst	63
Inclined motion	AccelJitter	21
	GravityDrift	21
	StopBurst	21
Total		**945**

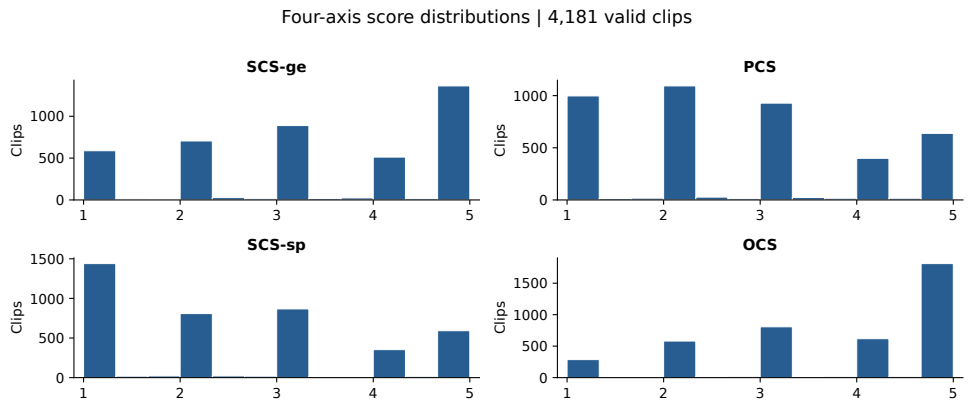

Figure 6: Updated four-axis score distributions on 4,181 complete records. Histograms use bins of width $1/3$ on the 1–5 scale. The 216 re-rated clips contribute mean ratings; the remainder retain their historical score.

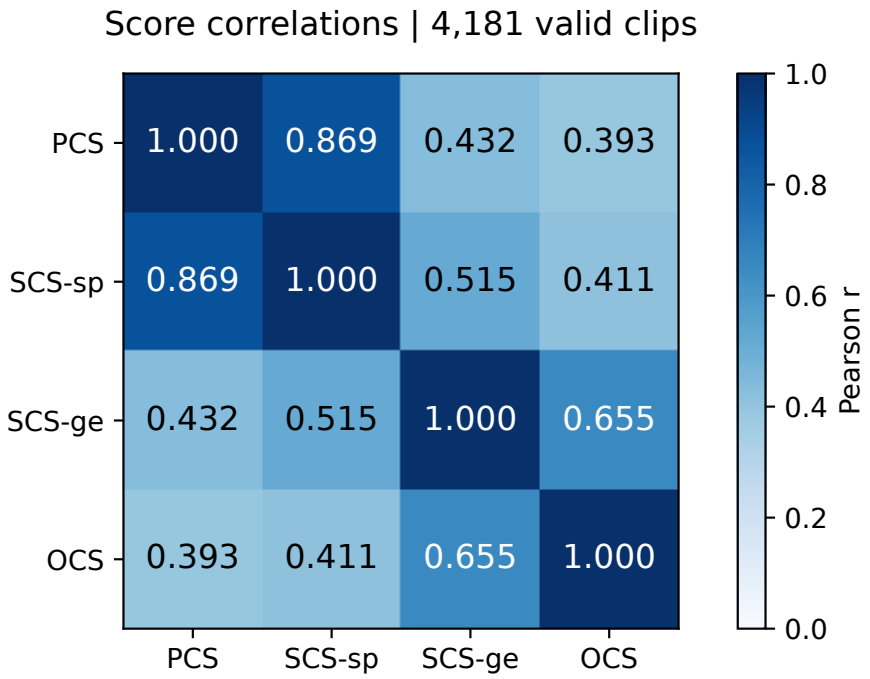

Figure 7: Pearson correlations among the four scores on the same 4,181-record updated export. Coefficients describe relationships between per-video scores, not inter-rater agreement.

These are associations between per-video scores, not measures of inter-rater reliability or proof that the four axes are independent. The histograms retain fractional mean scores; statistics never round the aggregated targets.

A.5 Controlled perturbations in Unity

Camera jitter is an observation perturbation rather than a change in physical dynamics; responsiveness to it is therefore reported separately from violations of motion laws. To make the 3D Unity videos useful for calibration and severity analysis, we implement a configurable noise system based on a unified noise profile. Each noise profile specifies a global severity factor and a set of motion-, collision-, friction-, gravity-, and camera-level perturbations. The final perturbation strength is

determined by multiplying the global severity with the corresponding noise-specific intensity and maximum magnitude. This design allows us to generate clean videos and systematically perturbed videos under comparable noise settings.

The injected perturbations are designed to simulate representative physical failure modes in generated videos. The specific controlled noise types used in the 3D Unity videos are shown in Table 11. These perturbations are applied to six controlled mechanical scenarios: uniform linear motion, accelerated linear motion, free fall, two-ball collision, inclined sliding, and uniform circular motion. Each motion–perturbation pair contains one to three base conditions that differ in a physical parameter of the target object, such as its restitution or friction coefficient (three conditions for most pairs, two for circular motion and one for inclined sliding), and the normalized perturbation level $\ell = (\alpha - \alpha_{\min})/(\alpha_{\max} - \alpha_{\min})$ of the swept noise parameter α varies from 0 to 1 within each condition. Table 10 computes the Spearman correlation between ℓ and nMSE and its p-value within each condition (21 perturbation levels) and averages both over the conditions of each pair, so that nMSE values from conditions with different physical parameters are never mixed. An averaged p-value is dominated by the weakest condition: for free fall under a stop burst, the mean p of 0.021 indicates that its conditions differ markedly in strength, whereas a single test of the averaged correlation over 21 levels would give $p \approx 10^{-4}$.

Table 10: Controlled Unity perturbations. Each condition sweeps the normalized perturbation level over 21 values; ρ is the Spearman correlation between perturbation level and nMSE within a condition, and p its two-sided p-value, both averaged over the conditions of the motion–perturbation pair. For single-condition pairs they are the ordinary within-condition values.

Motion	Perturbation	Conditions	Videos	Mean $\rho \uparrow$	Mean p
Uniform linear	AccelJitter	3	63	1.000	$< 10^{-16}$
	AirWall	3	63	0.994	$< 10^{-16}$
	StopBurst	3	63	1.000	$< 10^{-16}$
Accelerated linear	AccelJitter	3	63	1.000	$< 10^{-16}$
	AirWall	3	63	0.869	3.2×10^{-7}
	StopBurst	3	63	0.968	8.3×10^{-13}
Uniform circular	AccelJitter	2	42	1.000	$< 10^{-16}$
	CameraJitter	2	42	0.999	$< 10^{-16}$
	FrictionJitter	2	42	0.994	$< 10^{-16}$
Collision	ImpulseNoise	3	63	0.983	2.2×10^{-11}
	LateralImpulse	3	63	0.990	1.3×10^{-13}
	StopBurst	3	63	1.000	$< 10^{-16}$
Free fall	AccelJitter	3	63	0.817	1.4×10^{-4}
	GravityDrift	3	63	1.000	$< 10^{-16}$
	StopBurst	3	63	0.742	2.1×10^{-2}
Inclined sliding	AccelJitter	1	21	0.979	1.3×10^{-14}
	GravityDrift	1	21	0.971	2.5×10^{-13}
	StopBurst	1	21	0.968	2.6×10^{-12}

B Evaluator Details

Tracking and trajectory extraction. The tracker runs GroundingDINO-tiny on the first frame with the clip's text prompt (box and text thresholds 0.25 and 0.30), initializes SAM2.1 Hiera-L with the detected box and propagates masks through the clip. The centroid of each propagated mask gives the position $(x_o(t), y_o(t))$ of object o; masks below 30 pixels are recorded as missing. The raw table stores time, object ID, label, coordinates and a detection flag. A completed table fills interior gaps by linear interpolation and labels each frame as observed, interpolated or missing; leading and trailing gaps are not filled. The fitted trajectory is the selected object's completed trajectory up to its first missing frame. Per-object trajectory plots and an annotated overlay video are produced for inspection.

Table 11: Controlled noise types used in the 3D Unity videos.

Noise Type	**Intended Physical Effect**
Acceleration jitter	Adds random external acceleration, causing trajectory drift or motion jitter.
Stop burst	Applies sudden damping, producing abrupt slowdown or temporary stopping.
Air wall	Introduces an invisible boundary, causing unrealistic reflection.
Gravity drift	Changes the effective gravitational acceleration.
Collision impulse noise	Perturbs the normal collision impulse, changing rebound speed.
Collision lateral impulse	Adds tangential collision impulse, causing abnormal lateral deviation.
Friction jitter	Produces intermittent sticking or sudden motion freezing.
Camera jitter	Perturbs the camera pose, causing apparent visual shaking.

Fitting and residuals. Let $\mathcal{D} = \{(t_i, \mathbf{z}_i)\}_{i=1}^{N}$ be the trajectory, where $\mathbf{z}_i$ is one 2D track, two 1D collision tracks, or two 2D pendulum tracks. A specification with seed structure f predicts $\hat{\mathbf{z}}_i = f(t_i; \boldsymbol{\theta})$ and fits $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\text{fit}}(f, \boldsymbol{\theta}; \mathcal{D})$. Per-channel errors MSE_c are normalized by a shared motion scale, $\text{nMSE}_c = \text{MSE}_c/(s_{\text{ref}}^2 + \varepsilon)$, and combined as in Eq. (1). Pixel MSE alone is not comparable across videos because trajectory size depends on resolution, camera distance and motion amplitude. Some specifications sum and others average their channel errors, and the scale and aggregation of each family must be reproduced together. Table 12 summarizes the seven specifications, the paragraphs below give their equations, and Table 13 lists their preprocessing and auxiliary guards. The specifications used for the reported scores are the fixed seed equations; closed-form or least-squares initialization is followed by deterministic numerical refinement.

Table 12: Summary of the seven physics-guided motion specifications. Every specification returns a scalar normalized mean squared error (nMSE) as in Eq. (1), with a lower value indicating stronger conformity to the corresponding motion law.

Motion	**Trajectory Input**	**Seed Equation Structure**	**Reference Scale s_{ref}**
Uniform linear	(t, x, y)	$x(t) = x_0 + v_x t, \quad y(t) = y_0 + v_y t$	$\max\{d_N, \mathrm{RSpan}_{10,90}(\{d_i\}), \varepsilon\}$ †
Accelerated linear	(t, x, y)	$x(t) = b_x + v_{x,0} t + \frac{1}{2} a_x t^2$, $y(t) = b_y + v_{y,0} t + \frac{1}{2} a_y t^2$	$\sqrt{\mathrm{Var}(x) + \mathrm{Var}(y)}$
Uniform circular	(t, x, y)	$x(t) = c_x + R\cos(\omega t + \phi)$, $y(t) = c_y + R\sin(\omega t + \phi)$	$\sqrt{\mathrm{Var}(x) + \mathrm{Var}(y)}$
Collision	(t, x_L, x_R)	$s(t) = \sigma(g(t - t_c))$, ‡ $x(t) = [1 - s(t)]\,(v_1 t + b_1) + s(t)(v_2 t + b_2)$	$\sqrt{\mathrm{Var}(x_L^{\mathrm{pre}}) + \mathrm{Var}(x_R^{\mathrm{pre}})}$
Free fall	(t, x, y)	$x(t) = x_0, \quad y(t) = y_0 + v_{y,0} t + \frac{1}{2} g t^2$	$\max\{\mathrm{RSpan}_{2,98}(\{\hat{y}_i\}),\ 0.25 s_y,\ \varepsilon\}$
Inclined sliding	(t, x, y)	$x(t) = a_{x,2} t^2 + a_{x,1} t + a_{x,0}$, $y(t) = a_{y,2} t^2 + a_{y,1} t + a_{y,0}$	$\sqrt{\frac{1}{2}\,\mathrm{Var}(x) + \frac{1}{2}\,\mathrm{Var}(y)}$
Single-cycle pendulum	(t, x_L, y_L, x_R, y_R)	$x(t) = x_p + L\sin(\omega t + \phi)$, $y(t) = y_p - L\cos(\omega t + \phi), \quad \theta(t) = \omega t + \phi$	L

$d_i = \|\hat{\mathbf{p}}(t_i) - \hat{\mathbf{p}}(t_1)\|_2$; s_y is the robust span of the smoothed observed vertical track (defined below). † $\mathrm{RSpan}\left(\{x_i\}_{i=1}^N\right) = Q_{0.90}\left(\{x_i\}_{i=1}^N\right) - Q_{0.10}\left(\{x_i\}_{i=1}^N\right)$, where $Q_{0.90}$ and $Q_{0.10}$ denote the 90th and 10th percentiles, respectively. Unlike the full range, this percentile-based span is less sensitive to isolated endpoint or tracking errors. ‡ t_c is the collision time, g controls transition sharpness, and $\sigma(\cdot)$ is the sigmoid function.

Uniform linear motion. For constant-velocity motion, the seed equation structure is

$$x(t) = x_0 + v_x t, \qquad y(t) = y_0 + v_y t. \tag{3}$$

The four parameters are initialized by independent linear least-squares fits and refined jointly using a numerical optimizer. Since the last portion of a track can contain stopping behavior or unstable tracking, the specification optionally removes the final 0.5 seconds when sufficient earlier observations remain.

The shared scale is determined from the fitted motion itself. Write $d_i = \|\hat{\mathbf{p}}(t_i) - \hat{\mathbf{p}}(t_1)\|_2$. Then

$$s_{\text{ref}} = \max\left\{d_N, \text{RSpan}_{10,90}(\{d_i\}_{i=1}^N), \varepsilon\right\}. \tag{4}$$

Here $\text{RSpan}_{10,90}(\{d_i\}) = Q_{0.90}(\{d_i\}) - Q_{0.10}(\{d_i\})$, where Q_q denotes the q-quantile. This combines the fitted endpoint displacement with a robust span of fitted displacement magnitudes.

Therefore, the final residual is

$$E_{\text{uniform}} = \frac{1}{2}\left(\text{nMSE}_x + \text{nMSE}_y\right). \tag{5}$$

Accelerated linear motion. Constant-acceleration motion is represented in the position domain as

$$\begin{aligned} x(t) &= b_x + v_{x,0}t + \frac{1}{2}a_x t^2, \\ y(t) &= b_y + v_{y,0}t + \frac{1}{2}a_y t^2. \end{aligned} \tag{6}$$

A closed-form quadratic fit initializes the six interpretable parameters, and L-BFGS-B refinement minimizes a variance-weighted Huber loss. For scoring, both axes use the explicit shared variance reference

$$s_{\text{ref}}^2 = \text{Var}(x) + \text{Var}(y) + \varepsilon, \tag{7}$$

giving

$$E_{\text{acc,fit}} = \text{nMSE}_x + \text{nMSE}_y = \frac{\text{MSE}_x + \text{MSE}_y}{\text{Var}(x) + \text{Var}(y) + \varepsilon}. \tag{8}$$

Uniform circular motion. Uniform circular motion is modeled by

$$\begin{aligned} x(t) &= c_x + R\cos(\omega t + \phi), \\ y(t) &= c_y + R\sin(\omega t + \phi), \end{aligned} \tag{9}$$

with center (c_x, c_y), radius R, angular velocity ω, and phase ϕ. Initialization first applies a Kåsa circle fit to estimate (c_x, c_y, R), unwraps the observed polar angle around the fitted center, and then fits a line to angle versus time to initialize (ω, ϕ).

A dual-path strategy is used to avoid smoothing away the artifact being measured. The parameters are fitted on a moderately smoothed trajectory, but the principal residual is evaluated against the raw observations. The position residual uses

$$s_{\text{ref}}^2 = \text{Var}(x) + \text{Var}(y) + \varepsilon, \qquad E_{\text{circ,fit}} = \text{nMSE}_x + \text{nMSE}_y. \tag{10}$$

Collision motion. Collision is evaluated from either two horizontal trajectories $(x_L(t), x_R(t))$ or a single horizontal trajectory $x(t)$. The input format, rather than a heuristic, determines which mode is used. Each trajectory is represented by two linear regimes connected through a differentiable changepoint:

$$\begin{aligned} s(t) &= \sigma(g(t - t_c)), \\ x_j(t) &= [1 - s(t)]\,(v_{j,1}t + b_{j,1}) + s(t)(v_{j,2}t + b_{j,2}), \end{aligned} \tag{11}$$

where $j \in \{L, R\}$ for a two-ball track, t_c is the shared collision time, g controls transition sharpness, and $\sigma(\cdot)$ is the sigmoid function. For a single-ball track, only one branch is scored.

Candidate changepoints are initialized by searching for the split that minimizes the sum of the pre- and post-segment linear least-squares errors. The complete soft model is then refined using L-BFGS-B and a robust loss. To make the residual comparable across videos, the MSE is normalized using

the variance of the pre-collision signal, which captures the characteristic incoming-motion scale. In two-ball mode,

$$E_{\text{collision}} = \text{nMSE}_{L,\text{core}} + \text{nMSE}_{R,\text{core}}, \tag{12}$$

under a shared core evaluation window and pre-collision reference scale. In single-ball mode, the corresponding one-track nMSE is used.

Free-fall motion. Free fall is evaluated only over the first descent before the first ground contact. The specification detects the earliest plausible impact using the smoothed vertical trajectory and velocity reversal near the estimated ground level. If no reliable reversal is found, it searches for the beginning of a stable near-ground suffix. Post-impact rebound, rolling, and terminal rest are excluded from the main score because they are strongly affected by contact dynamics, perspective, and tracking noise.

The pre-impact segment is modeled as

$$x(t) = x_0, \qquad y(t) = y_0 + v_{y,0}t + \frac{1}{2}gt^2, \qquad g > 0, \tag{13}$$

where the image-coordinate convention takes downward motion as positive. Initial vertical parameters are obtained from a quadratic least-squares fit. The shared scale is $s_{\text{ref}} = \max\{\text{RSpan}_{2,98}(\{\hat{y}_i\}), 0.25s_y, \varepsilon\}$, where $s_y = Q_{0.98}(y^{\text{smooth}}) - Q_{0.02}(y^{\text{smooth}}) + \varepsilon$ is computed before truncation. Thus the predicted vertical span has a floor derived from the observed track. The vertical channel receives a larger weight:

$$E_{\text{freefall}} = 0.2\,\text{nMSE}_x + 0.8\,\text{nMSE}_y. \tag{14}$$

Impact, rebound, and restart-kink measurements are stored for analysis but do not contribute to the final fit-only score. Consequently, clips with different bounce settings are evaluated by the same pre-impact criterion.

Inclined sliding motion. For a cube sliding along an incline, both image-coordinate trajectories are approximated by quadratic functions:

$$\begin{aligned} x(t) &= a_{x,2}t^2 + a_{x,1}t + a_{x,0}, \\ y(t) &= a_{y,2}t^2 + a_{y,1}t + a_{y,0}. \end{aligned} \tag{15}$$

Before fitting, a trailing stationary segment is detected from the smoothed y trajectory. When a sufficiently long low-slope suffix is found, both x and y are synchronously truncated at its start, ensuring that the same physical interval is evaluated in both channels.

The coefficients are initialized by polynomial least squares and jointly refined with a numerical optimizer. The shared normalization is

$$s_{\text{ref}}^2 = \frac{1}{2}\,\text{Var}(x) + \frac{1}{2}\,\text{Var}(y) + \varepsilon, \tag{16}$$

and the balanced residual is

$$E_{\text{incline,fit}} = \frac{1}{2}\text{nMSE}_x + \frac{1}{2}\text{nMSE}_y. \tag{17}$$

Under the adopted image-coordinate convention, the expected clean trajectory has opposite quadratic signs in the horizontal and downward vertical coordinates. A lightweight sign-consistency term is therefore added only as a guard against a structurally incorrect solution:

$$\mathcal{L}_{\text{incline}} = E_{\text{incline,fit}} + \lambda_{\text{sign}}\Omega_{\text{sign}}. \tag{18}$$

Single-cycle pendulum motion. The linear-phase formula below is retained to describe the legacy geometric proxy used by this specification. It should not be interpreted as the physical pendulum law $\ddot{\theta} + (g/L)\sin\theta = 0$, nor does fitting it test conservation during Newton's-cradle contact. The new single-pendulum study uses a different declared dynamical family. The pendulum specification takes the 2D trajectories of the left and right balls of a Newton's cradle. To reduce contamination from hand release and subsequent interactions, extrema of the vertical trajectory are detected, and

only the second half-cycle is fitted for each side. The pivot and effective length are estimated from the trajectory range, after which the circular-arc model is

$$\begin{aligned} x_j(t) &= x_{p,j} + L_j \sin\theta_j(t), \\ y_j(t) &= y_{p,j} - L_j \cos\theta_j(t), \qquad j \in \{L, R\}. \\ \theta_j(t) &= \omega_j t + \phi_j, \end{aligned} \tag{19}$$

The angular parameters are optimized numerically for each side. Since the natural spatial scale of a pendulum trajectory is its radius, the per-side error is normalized by L_j^2:

$$E_j = \frac{1}{2}\left(\frac{\mathrm{MSE}_{x,j}}{L_j^2} + \frac{\mathrm{MSE}_{y,j}}{L_j^2}\right), \tag{20}$$

and the final two-side residual is

$$E_{\mathrm{pendulum}} = \frac{1}{2}E_L + \frac{1}{2}E_R. \tag{21}$$

Table 13: Motion-specific laws, preprocessing, and normalized residuals used by the physics-law evaluator. "Fit-only" indicates that auxiliary measurements are stored as diagnostics but are not added to the returned objective.

Motion	**Seed law**	**Preprocessing / initialization**	**Normalized residual**	**Final objective**
Uniform linear	2D linear position model	Optional final-$0.5\,\mathrm{s}$ truncation; linear least-squares initialization	Shared fitted displacement/path-span scale; equal x/y weighting	Fit-only
Accelerated linear	2D quadratic position model	Quadratic least-squares initialization; Huber-based refinement	Shared $\mathrm{Var}(x) + \mathrm{Var}(y)$ scale	nMSE plus light curvature, reversal, and collinearity guards
Uniform circular	Center–radius–angle model	Savitzky–Golay fit path; Kåsa circle and angle-line initialization; raw-data scoring	Shared $\mathrm{Var}(x) + \mathrm{Var}(y)$ scale	nMSE plus radius, angular, zero-speed, and jitter guards
Collision	Soft changepoint between pre/post linear regimes	One- or two-track mode; changepoint scan; robust piecewise-linear initialization	Pre-collision variance reference over a core window	Fit-only
Free fall	Constant x and quadratic y	First-impact detection; retain pre-impact descent only	Robust predicted-y span; weights $(0.2, 0.8)$	Fit-only
Inclined sliding	Joint quadratic $x(t), y(t)$	Detect y-plateau and synchronously truncate both axes; polynomial initialization	Shared mean of x/y variances; equal weighting	nMSE plus light sign guard
Single-cycle pendulum	Circular arc with linear phase	Detect extrema; fit second half-cycle; estimate pivot and length	Per-side normalization by L^2; equal side weighting	Fit-only

C CALIBRATED VARIANT

The calibrated variant is used only in the comparison with a judge fine-tuned on PCB labels (Appendix D). For a family-specific residual e and ordered cut points $t_1 < \cdots < t_4$, the calibrated score is $\widehat{s}(e) = 5 - \sum_{j=1}^{4} \mathbf{1}[e > t_j]$. The thresholds of each family are chosen to maximize the Spearman correlation between $\widehat{s}$ and SCS-sp on the calibration rows, subject to all five score levels being used. Calibration supervises only the mapping from residual to score; it does not train the tracker or change the equation family.

Test95 split. The historical split has 284 training, 42 validation and 95 test clips with disjoint source groups; exact normalized prompt text overlaps between seven training–validation, sixteen training–test and two validation–test groups. One threshold set per family is refitted on the original 326 training and validation IDs using the updated targets and applied unchanged to Test95 (Table 14). No test labels enter threshold selection. Test95 is unchanged in membership; it is the shared held-out set for this supervised threshold variant and the archived Qwen predictions.

Table 14: Updated Test95 thresholds fitted on the original 326 training/validation IDs with mean human targets where available (values rounded). Residual scales differ by family.

Family	Calib. N	Test N	t_1	t_2	t_3	t_4
Uniform linear	54	13	0.000114578	0.00330242	0.0403558	0.0747624
Accelerated linear	55	17	0.226211	0.292014	0.334311	0.646387
Uniform circular	42	11	0.81409	0.917692	0.938483	0.965602
Collision	29	11	0.00179754	0.0027428	0.00402794	0.0240784
Free fall	84	24	0.00133799	0.0658206	0.0748345	0.0840716
Inclined sliding	33	8	0.00422585	0.50147	0.518447	0.8832
Pendulum proxy	29	11	0.118557	0.153673	0.174321	0.302966

D ADDITIONAL BENCHMARK RESULTS

Per-family agreement. Table 15 reports per-family correlations using mean SCS-sp targets for generated clips. Within a family, the min–max score is a strictly monotone transform, so its correlation is the negative of the raw-residual correlation. The p-values are unadjusted per-video Spearman tests. Among generated clips, inclined sliding has the strongest agreement ($\rho = 0.537$); the other six families have weak or uncertain correlations, including slightly negative point estimates for circular motion and the pendulum proxy. With real clips included under the target-5 convention, the pendulum proxy is strongest.

Supplementary held-out comparison on Test95. The primary label-free comparison is the full 216-clip generated mechanics cohort (Table 2). Test95 is the historical test partition of the 284/42/95 train/validation/test split; its 53 generated clips are a subset of those 216, not independent replication, and its other 42 clips are real. We retain it because saved Qwen3-VL predictions and the calibrated variant permit a common held-out comparison only on these IDs. Every generated test clip has three valid ratings, whose mean is the target. All Test95 tables use min–max ranges determined only by the 326 training and validation clips; replacing them by full-cohort ranges changes ρ by less than 0.002. On the generated subset, MLC has the highest point correlation among the label-free methods in Table 16, but its differences from VideoScore2 and zero-shot Qwen3-VL are not resolved by paired source-group intervals.

Uncertainty on Test95. Table 17 reports 95% source-group percentile bootstrap intervals from 10,000 resamples of the 57 source groups in Test95 or the 34 groups in its generated subset; all methods use the same resampled groups. Intervals condition on saved predictions and aggregated targets: they do not resample raters or repeat model fitting. On generated clips, paired differences are MLC minus VideoScore2 $+0.289$ ($[-0.069, 0.617]$) and minus zero-shot Qwen3-VL $+0.150$ ($[-0.207, 0.491]$). On all 95 clips, the difference from VideoScore2 is $+0.320$ ($[0.071, 0.553]$), a mixed-cohort result using the target-5 convention for real clips. Replacing min–max normalization by within-family percentile rank against the same training/validation reference gives $\rho = 0.048$

Table 15: Per-family Spearman correlation with mean SCS-sp on the mechanics cohort. Real clips have the conventional target 5 in the upper block.

Family	N	MLC ρ	p
Real + generated (421)			
Uniform linear	67	0.231	6.0×10^{-2}
Accelerated linear	72	0.150	2.1×10^{-1}
Uniform circular	53	0.220	1.1×10^{-1}
Collision	40	0.319	4.5×10^{-2}
Free fall	108	0.259	6.9×10^{-3}
Inclined sliding	41	0.201	2.1×10^{-1}
Pendulum proxy	40	0.439	4.6×10^{-3}
Pooled	421	0.360	2.7×10^{-14}
Generated only (216)			
Uniform linear	28	0.115	5.6×10^{-1}
Accelerated linear	31	0.024	9.0×10^{-1}
Uniform circular	36	-0.073	6.7×10^{-1}
Collision	23	0.144	5.1×10^{-1}
Free fall	58	0.155	2.4×10^{-1}
Inclined sliding	25	0.537	5.7×10^{-3}
Pendulum proxy	15	-0.113	6.9×10^{-1}
Pooled	216	0.382	6.8×10^{-9}

Table 16: Supplementary Test95 agreement with updated SCS-sp targets. All methods use no PCB labels; best point estimates per column are bold. Qwen uses the archived zero-shot predictions, re-evaluated against the new targets.

	All ($N = 95$)		Generated ($N = 53$)	
Method	$\rho \uparrow$	p	$\rho \uparrow$	p
PhyGenEval (overall)	-0.059	5.7×10^{-1}	0.017	9.0×10^{-1}
VideoScore2 (physical consistency)	0.012	9.1×10^{-1}	0.092	5.1×10^{-1}
VideoPhy-2 (physical commonsense)	0.238	2.0×10^{-2}	0.136	3.3×10^{-1}
Qwen3-VL-8B zero-shot	0.297	3.4×10^{-3}	0.231	9.7×10^{-2}
MLC (ours)	**0.332**	9.9×10^{-4}	**0.381**	4.9×10^{-3}

on generated clips, indicating substantial sensitivity of pooled agreement to relative family scales. Per-video p-values and source-group intervals use different sampling assumptions; intervals are marginal, without multiplicity correction.

Table 17: Test95 agreement with updated SCS-sp and 95% source-group bootstrap intervals.

Method	All ($N = 95$) ρ [95% CI]	p	Generated ($N = 53$) ρ [95% CI]	p
MLC (ours)	0.332 [0.164, 0.471]	9.9×10^{-4}	0.381 [0.143, 0.571]	4.9×10^{-3}
Within-family percentile variant	0.171 [0.007, 0.327]	9.7×10^{-2}	0.048 [-0.237, 0.315]	7.3×10^{-1}
PhyGenEval (overall)	-0.059 [-0.259, 0.153]	5.7×10^{-1}	0.017 [-0.298, 0.310]	9.0×10^{-1}
VideoScore2 (physical consistency)	0.012 [-0.204, 0.218]	9.1×10^{-1}	0.092 [-0.173, 0.377]	5.1×10^{-1}
VideoPhy-2 (physical commonsense)	0.238 [0.044, 0.418]	2.0×10^{-2}	0.136 [-0.166, 0.449]	3.3×10^{-1}
Qwen3-VL-8B zero-shot	0.297 [0.076, 0.493]	3.4×10^{-3}	0.231 [-0.044, 0.482]	9.7×10^{-2}

Comparison with a judge fine-tuned on PCB labels. Qwen3-VL-8B with QLoRA is a supervised judge trained on historical PCB labels. We re-evaluate its saved epoch-3 predictions against the updated test targets; the model has not been retrained on consensus labels. Our calibrated variant instead refits family-specific score thresholds on the original 326 training/validation IDs with updated targets, then applies them unchanged to Test95 (Appendix C). These training-label versions differ, so this is a comparison of available predictors, not a controlled experiment training both on the same consensus labels. The frozen QLoRA judge is higher on the mixed 95-clip view, with the paired interval excluding zero. On the 53 generated clips, it is numerically higher than both MLC variants, but neither difference is resolved. Calibration does not improve the point estimate over label-free MLC (Table 18).

Table 18: Updated Test95 targets applied to the historical QLoRA checkpoint, label-free MLC and the refitted threshold variant. Intervals are conditional 95% source-group intervals; the difference is paired.

Method	PCB labels used	All ($N = 95$)	Generated ($N = 53$)
Qwen3-VL-8B + QLoRA	historical training labels	0.622 [0.464, 0.749]	0.408 [0.157, 0.601]
MLC (ours)	none	0.332 [0.164, 0.471]	0.381 [0.143, 0.571]
MLC, calibrated	updated train + val labels	0.210 [-0.043, 0.440]	0.274 [-0.091, 0.552]
MLC − QLoRA		-0.290 [-0.490, -0.098]	-0.028 [-0.310, 0.260]

Pooled versus within-family agreement. A pooled correlation mixes rankings within a motion family with differences between families. Table 19 reports a fixed-effect Fisher-z summary of per-family correlations and a null obtained by shuffling evaluator scores within each family. On the 216 generated clips the observed pooled correlation exceeds this null (one-sided permutation $p = 0.0009$), but the within-family summary is weak and its analytic interval includes zero. On the 53 generated test clips the observed pooled value does not clearly exceed the null ($p = 0.108$), and the within-family summary is near zero. Thus pooled agreement supports an aggregate signal, while strong, consistent discrimination within each motion family is not established. The permutation test and Fisher summary weight the data differently and should not be treated as interchangeable tests.

Real versus generated clips. The pooled views assign real clips the target 5. MLC separates real from generated clips only modestly, with AUROC 0.649 ($[0.599, 0.697]$) on the 421-video cohort and 0.628 ($[0.522, 0.728]$) on Test95. Pairs of one real and one generated clip nevertheless contribute 70% of the concordance behind the pooled MLC correlation on the 421 cohort, so pooled values largely describe real-versus-generated ordering. This is why we treat generated-only agreement as the primary view.

Which motions are hardest for generators. Table 20 uses the mean human scores of the 216 generated mechanics clips. SCS-sp differs across families (Kruskal–Wallis $H = 86.8$, $p = 1.4 \times 10^{-16}$). Interaction and constraint motions (collision, pendulum and circular motion) score 1.21

Table 19: Pooled and within-family agreement of MLC using updated targets. Within-family combines seven Spearman correlations with Fisher-z weights $n-3$ and an analytic 95% interval. The null is the mean and 95% range of 10,000 within-family permutations. Test95 uses training/validation normalization ranges.

Cohort	Pooled ρ	Within-family ρ [95% CI]	Permutation null
Mechanics, all (421)	0.360	0.249 [0.155, 0.339]	0.125 [0.045, 0.206]
Mechanics, generated (216)	0.382	0.125 [-0.015, 0.260]	0.220 [0.129, 0.315]
Test95, generated (53)	0.381	-0.013 [-0.344, 0.322]	0.274 [0.109, 0.446]

points lower on SCS-sp ($[0.92, 1.50]$; Cliff's $\delta = -0.58$) and 0.97 lower on SCS-ge than single-body kinematics. Generator coverage differs across families. With generator fixed effects, identified by the seven generators spanning both groups, the SCS-sp gap is 0.56 ($[0.25, 0.87]$) and the SCS-ge gap 0.42 ($[0.08, 0.75]$); the adjusted OCS difference is -0.15 ($[-0.49, 0.21]$). This supports a gap in physical scores after adjusting for generator mix, without resolving an adjusted difference in object consistency.

Table 20: Mean human scores of the 216 generated mechanics clips, with 95% source-group bootstrap intervals. Each clip contributes its mean available rating.

Family	n	SCS-sp	SCS-ge
Uniform linear	28	4.26 [3.85, 4.63]	4.25 [3.89, 4.55]
Accelerated linear	31	3.76 [3.34, 4.14]	4.12 [3.77, 4.44]
Free fall	58	3.01 [2.73, 3.28]	3.11 [2.81, 3.38]
Inclined sliding	25	2.72 [2.27, 3.16]	3.17 [2.70, 3.59]
Uniform circular	36	2.59 [2.32, 2.87]	2.93 [2.63, 3.23]
Pendulum proxy	15	1.67 [1.45, 1.87]	2.09 [1.81, 2.41]
Collision	23	1.81 [1.57, 2.12]	2.41 [2.19, 2.66]

Matched-prompt generator comparison. Table 1 macro-averages each model over its represented categories, so rows do not share prompts. Among the six versions, 1,010 prompts were rendered by at least two models, 125 by at least three and none by more than four. Table 21 fits an additive prompt-plus-generator model to the Overall score on shared prompts, using the same updated label export. Agreement between the macro and matched rankings is Kendall $\tau = 0.73$ for Overall, SCS-sp, PCS and OCS, but 0.33 for SCS-ge. Sora2 remains highest. Seedance1.0pro and Keling2.1 exchange order, with an unresolved Overall difference of $+0.007$ ($[-0.365, 0.387]$); SkyReels versus Hunyuan is also unresolved ($+0.096$, $[-0.038, 0.231]$). The Sora2–Hunyuan gap is 1.67 on matched prompts versus 1.02 in the macro table. These comparisons describe the covered prompts rather than a uniformly sampled test of generator capability.

Table 21: Matched-prompt Overall scores with the updated targets: additive prompt-plus-generator model and 95% prompt-bootstrap intervals.

Model	Matched prompts	Generator effect	Adjusted mean
Sora2	819	+1.029 [0.938, 1.120]	4.09 [3.99, 4.19]
Seedance1.0pro	42	+0.217 [-0.056, 0.506]	3.28 [3.00, 3.58]
Keling2.1	183	+0.210 [0.064, 0.358]	3.27 [3.12, 3.43]
Wan2.2	118	-0.272 [-0.459, -0.096]	2.79 [2.60, 2.97]
SkyReels	332	-0.543 [-0.661, -0.424]	2.52 [2.40, 2.64]
Hunyuan	656	-0.639 [-0.735, -0.543]	2.42 [2.33, 2.52]

Additional generator on linear-motion prompts. We also annotated 70 clips generated by Wan2.7 with the same four-axis protocol. They cover a single category, mechanics, and only uniform and accelerated linear-motion prompts, none of which was rendered by any of the six models in Table 1. Its scores (Table 22) are therefore not comparable with the main table: they describe one

generator on two single-body linear-motion families, which are among the highest-scoring families in the mechanics cohort (Table 20; most of that cohort's uniform and accelerated clips are Wan2.7 outputs). Five of the 70 clips have an invalid historical SCS-sp entry of 0 and are excluded from the four-axis summary. The remaining 65 use mean ratings where they overlap the 216-clip campaign; other scores remain historical.

Table 22: Wan2.7 on uniform and accelerated linear-motion prompts: 65 complete valid clips of 70, one category, with no prompts shared with the six main versions.

Model	PCS	SCS-sp	SCS-ge	OCS	Overall
Wan2.7 (65 valid)	4.318	4.318	4.333	4.692	4.415

Existing evaluators on the general benchmark. Table 23 evaluates the three general-purpose automatic evaluators on General224, a 224-clip sample of the broader benchmark with 112 generated and 112 real clips. Nine generated IDs overlap the re-rated mechanics set and use the updated means; the other generated targets remain historical. Their agreement with SCS-sp is weak: VideoScore2 reaches 0.240 on generated clips, while PhyGenEval and VideoPhy-2 are not significantly correlated with the labels. VideoPhy-2's physical-commonsense criterion is not conditioned on the prompt, whereas SCS-sp asks whether the requested phenomenon is realized, which may explain part of the gap. Morpheus (Tragoudaras et al., 2026), re-evaluated against the updated mean human ratings on a compatibility-restricted mechanics subset of 126 clips (59 generated), gives a separate result: its physical-invariance score has $\rho = 0.335$ ($p = 1.3 \times 10^{-4}$) on all clips and $\rho = -0.067$ ($p = 0.61$) on generated clips. The subset differs from Test95 and the full 421-clip cohort, so these values are not ranked against the common-cohort results in Table 2.

Table 23: Spearman agreement on General224 (112 real and 112 generated), updating the nine overlapping generated IDs with mean human ratings. Real targets remain 5.

	All (224)		Generated (112)	
Method	ρ	p	ρ	p
VideoScore2 (physical consistency)	0.194	3.5×10^{-3}	0.240	1.1×10^{-2}
PhyGenEval (overall)	-0.047	4.9×10^{-1}	0.113	2.4×10^{-1}
VideoPhy-2 (physical commonsense)	-0.006	9.3×10^{-1}	-0.102	2.9×10^{-1}

Baseline settings. VideoScore2 uses its physical-consistency output with 2 fps input sampling, at most 1,024 new tokens, temperature 0.7 and base seed 42 with a clip-dependent offset. Qwen3-VL uses the 8B Instruct model at 8 fps with at most 151,200 pixels per frame, four-bit NF4 quantization with double quantization and bfloat16 computation; the QLoRA model is the epoch-3 adapter fine-tuned on historical PCB training labels without consensus retraining, and both return a predicted SCS-sp. Scores on Test95 are extracted by exact clip ID from each method's saved outputs.

E Generation Study Details

E.1 Shared motion record

A motion record is a tuple $\mathcal{R} = (\mathcal{O}, \mathcal{T}, \mathcal{Y}, \mathcal{U}, \mathcal{H}, P_\theta, \mathcal{E}, \mathcal{Q})$. $\mathcal{O}$ binds objects and identities, $\mathcal{T}$ retains native time and units, $\mathcal{Y}$ stores measurements with their visibility, and $\mathcal{U}$ records uncertainty. $\mathcal{H}$ declares physical and observation assumptions, the executable program P_θ selects a supported motion family with bounded parameters, $\mathcal{E}$ stores observed and predicted events with their evidential status, and $\mathcal{Q}$ links each decision to the observations and assumptions that justify it. A predicted contact is therefore distinct from image evidence of contact. The program can express hybrid dynamics $\dot{x} = f_z(x; \theta)$ with guards $g_{z \to z'}(x) = 0$ and resets $x^+ = r_{z \to z'}(x^-; \theta)$ for events such as impacts, and an RGB measurement is modelled as $y_i = m(x(t_i); c) + \epsilon_i$ with explicit camera assumptions c. When the available observations do not identify a declared model, the record returns *unknown* rather than a physical score. Figure 8 shows how the record links evaluation, program-controlled generation and remeasurement.

Each endpoint declares its fitting and assessment intervals. Coverage and motion gates are fixed before measurement, and observation failure, visual artifact and motion deviation remain separate outcomes. The forecast errors D of Section 5 are in object radii and are not comparable with the nMSE of Eq. (1). In the feedback loop, $\mathcal{R}_{k+1} = \mathrm{Update}(\mathcal{R}_k, a_k)$ records each selected action together with the RGB evidence it cites and its expected effect before the next render.

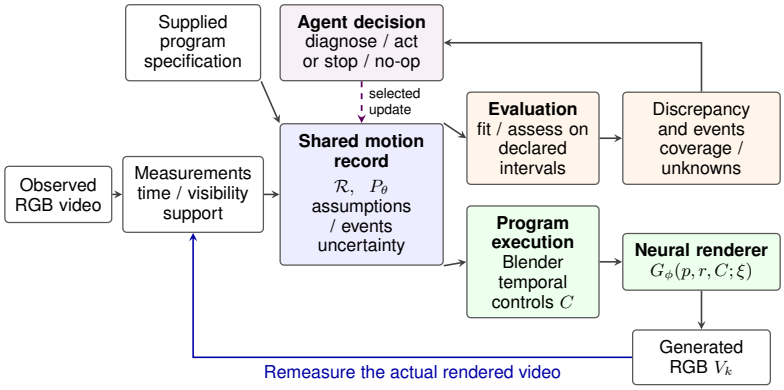

Figure 8: A shared motion record links evaluation, program-controlled generation and remeasurement of the final RGB output. Records can start from observations or from a supplied program; the generation experiments use supplied programs. Solid arrows show data flow; the dashed arrow is a selected update. Simulator targets and measured output trajectories are kept separate.

E.2 Protocol

All generation experiments use Wan2.1-VACE-1.3B (Jiang et al., 2025) at 640×480 with 40 sampling steps. Each call produces 49 frames from 49 control frames; the first 48 frames form the 12 fps, 4 s clip that is measured and rated. The first matched batch uses depth controls for the rebound and cradle scenes and RGB geometry controls for the spring and pendulum scenes; the prospective cohort uses RGB controls. The control modality is fixed for each configuration, so the experiments do not isolate its effect. The observer uses SAM2.1 Hiera-tiny (Ravi et al., 2024) with one RGB-grounded prompt point on the first frame and robust circle geometry for sphere centers. Each output receives one observation pass, and unknown frames and unscorable clips remain in all denominators. Fitting and assessment intervals, coverage gates, motion gates and scorer settings are fixed before measurement.

E.3 Summary of matched outcomes

Table 24 lists every matched outcome. In all seven pairs where both arms can be scored, Controlled has the lower error: rebound residuals fall by 49.6–96.3%, and projectile forecast errors fall from 2.941 and 3.711 radii to 0.126 and 0.055 (Figure 11). All four controlled oscillators complete their requested oscillation, while the reference-only arm has three failures and one unknown; a nearly static baseline spring has a small raw residual but cannot pass the motion gate. Resampling does not rescue the reference-only arm: over five seeds per RGB-controlled configuration it reaches the requested motion in 1 of 25 attempts (with a 3.7-radius forecast error), and at control strength 0.25 all ten outputs fail, whereas strengths 0.5–1.0 pass in all ten (Appendix E.10). In the prospective cohort, all six Controlled outputs have 48 accepted centers and held-out errors of 0.052–0.132 radii, and the two circular Controlled clips traverse 528.6 and 528.9 degrees of the requested 540. The other four baselines remain unscored: two projectile clips lack sufficient geometry in one half, and both circular baselines yield an orbit radius outside the admissible range. Scoring rebound pairs on shared timestamps preserves all four directions; a bounded refit that also changes time anchoring and bounds reverses one pair (Appendix E). On automatic appearance measures, program control shows no resolved paired difference in consistency, flicker or imaging quality (aesthetic quality is lower in the six prospective pairs); RGB-controlled outputs stay close to their control renders (LPIPS 0.20), so their motion is inherited from the program, and reference-only outputs drift from their own appearance reference (DINO similarity 0.43 versus 0.69; Appendix E.9).

Table 24: Matched Reference → Controlled outcomes for all 32 outputs, in object radii (lower is better). "fail": failed motion gate; "unk.": insufficient observations or inadmissible orbit fit; neither is converted into a number.

Cohort	Configuration	Endpoint	Seed 1	Seed 2
Main	Base rebound	in-sample D_y	0.054 → 0.027	1.056 → 0.039
Main	Higher rebound	in-sample D_y	0.103 → 0.043	0.741 → 0.045
Main	Spring	forecast	fail → 0.039	fail → 0.236
Main	Pendulum	forecast	unk. → 0.023	2.042 (fail) → 0.022
Main	Newton's cradle	none	no multi-object numerical scorer (4 clips)	
Prospective	Horizontal projectile	forecast	2.941 → 0.126	unk. → 0.052
Prospective	Oblique projectile	forecast	3.711 → 0.055	unk. → 0.092
Prospective	Uniform circular	forecast	unk. → 0.132	unk. → 0.116

Table 25: Held-out conditional NRMSE in two calibrated rebound states. The experimental units are two states; simulation truth belongs to a separate measurement audit.

State	Normal	Appearance-only	Temporal perturbation
State 1	0.3400%	0.3809%	6.5177%
State 2	2.0664%	2.3475%	18.2870%

E.4 Calibrated rebound sensitivity

The controlled evaluator uses six RGB clips from two predeclared rebound states with known camera and geometry. Within each state, a normal clip and an appearance-only variant share their motion, while a temporal perturbation changes the motion after the calibration interval. The evaluator fits an initial 0.20 s interval and an outgoing 0.25 s interval after the model-inferred first contact, then forecasts the remaining held-out observations. Table 25 passes the predeclared paired checks in both states: the error stays low under the appearance change and increases under the temporal perturbation. The experimental units are two physical states, so this establishes conditional sensitivity under known assumptions, not recognition of arbitrary scenes.

E.5 Main matched cohort

The first batch fixes three scenarios at seed 20260918: a five-ball Newton's cradle and two rebound configurations (base and higher rebound). The two conditions share the prompt, appearance reference, renderer, resolution, number of steps and output budget; Controlled additionally receives 49 dense controls. All six cells completed once without retries. A first-seed extension adds a horizontal spring oscillator and a nonlinear single pendulum with RGB geometry controls, and a second seed (20260921) repeats all five configurations without changing prompts, controls or backend. The cohort therefore contains 20 videos and ten pairs clustered in five configurations. The cradle pairs remain in the denominator but have no multi-object numerical scorer.

For the rebounds, a fixed observation and six-candidate fitting procedure measures horizontal variation D_x and an in-sample vertical residual D_y of a passive one-dimensional bounce family, both relative to the observed ball radius, under a vertical-release and static-camera interpretation (Table 26). For the oscillators, the observer fits frames 0–23 and forecasts frames 24–47, requiring at least 18 usable centers in each half and at least two radii of motion in the fitting window; simulator coordinates never initialize the observer (Table 27). Table 28 gives the second-seed outcomes. The second-seed rebound pairs also favor Controlled on horizontal variation over exactly shared times ($D_x = 0.180$ versus 0.010 for the base setting over 31 times, and 0.167 versus 0.017 for the higher setting over 26 times). Controlled spring forecast error varies from 0.039 to 0.236 radii across seeds. Across the cohort, 659 of 768 numerical observation slots have accepted geometry and 109 remain unknown; all 384 Controlled slots are supported.

Figure 9 complements the main-text comparison with the base-rebound setting. Both figures use unretouched, full-frame video samples at identical native timestamps within each Reference/Controlled pair; Figure 4 uses Newton's cradle seed 20260921 and higher rebound seed 20260918, while Fig-

Table 26: First-seed matched rebound measurements. D_x is horizontal variation and D_y is an in-sample image-coordinate model residual; both require the stated interpretation.

Scenario / arm	Measured frames	$D_x \downarrow$	$D_y \downarrow$
Base rebound Reference	37/48	0.980884	0.054087
Base rebound Controlled	48/48	0.014095	0.027249
Higher rebound Reference	35/48	1.630813	0.102642
Higher rebound Controlled	48/48	0.016372	0.043120

Table 27: First-seed spring and pendulum extension under the fixed observer. Unassessable is retained rather than converted to a numerical score.

Task / arm	Centers	Requested motion	Forecast error / radius
Spring Reference	43/48	Failed; insufficient horizontal motion	Not dynamically testable
Spring Controlled	48/48	Completed	0.0391
Pendulum Reference	7/48	Unassessable	Unassessable
Pendulum Controlled	48/48	Completed	0.0229

ure 9 uses seed 20260921 for both configurations. These selected examples illustrate the outcomes rather than estimate their frequency; Tables 24 and 3 retain the complete paired evidence.

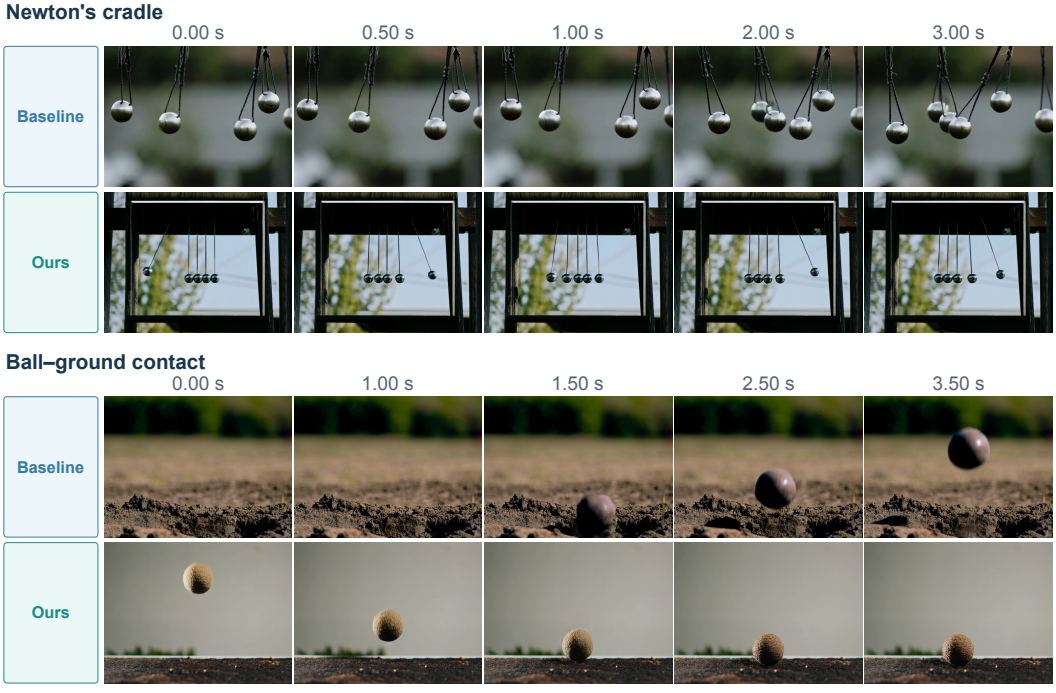

Figure 9: **Additional matched views of cradle motion and ball–ground contact.** Baseline receives the prompt and appearance reference; Ours adds 49 depth controls from the declared motion program, with the same renderer and seed within each pair. The cradle panel repeats the main-text example for context. In the base-rebound example, the controlled ball is visible from the start and descends toward the surface, whereas the reference-only ball appears later. The sampled base-rebound frames establish visibility and descent, not successful completion of the requested rebound.

Shared-time sensitivity of the rebound comparison. We restrict each rebound pair to identical accepted frame times, retaining 37, 35, 31 and 26 times and each arm's original radius normalizer.

Table 28: Fixed second-seed results. Errors are normalized by observed bob radius; lower is better conditional on the stated model. "Unknown" is retained rather than assigned a favorable score. Reference receives no temporal control.

Scenario	Component	Reference	Controlled	Event Reference / Controlled
Base rebound	In-sample D_y	1.05635	0.03875	Not separately scored
Higher rebound	In-sample D_y	0.74146	0.04450	Not separately scored
Spring	Held-out error	Unknown	0.23604	Failed / completed
Pendulum	Held-out error	2.04162	0.02195	Failed / completed
Cradle	Numeric scorer	N/A	N/A	No numerical scorer

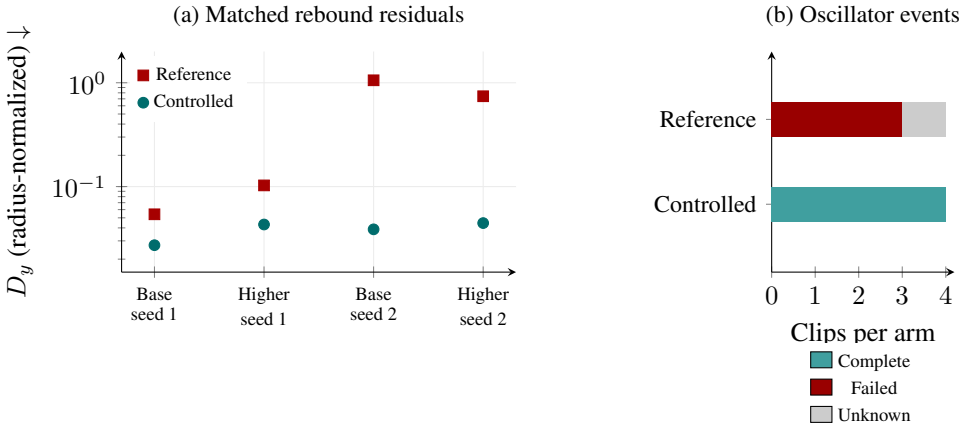

Figure 10: Main matched cohort across two seeds. (a) All four rebound pairs have a lower in-sample residual with program controls (relative reductions 49.6–96.3%). (b) All four controlled spring and pendulum clips complete the requested event; the reference-only arm has three failures and one unknown. Two rebound configurations and two oscillator types over two seeds are not eight independent tasks.

Holding the originally selected curves fixed and restricting only the scored times preserves all four directions (Table 29). A bounded refit with three deterministic starts per arm, which also resets the initial time to the first shared observation and recomputes interval-dependent bounds, preserves three directions and reverses the first-seed higher-rebound pair; all eight refits touch a bound. Restricting support alone therefore does not remove the advantage of the selected Controlled curves, but stability under refitting is not established.

Table 29: Shared-time rebound D_y (lower is better). Fixed-curve replay changes only score support. Refit also changes optimization budget, time origin and interval-derived bounds, so its difference is not attributable to support alone. All refits retain boundary-hit flags.

Seed	State	Replay Reference	Replay Controlled	Refit Reference	Refit Controlled
20260918	Base	.05409	.02273	.58318	.02257
20260918	Higher	.10264	.04738	.10264	.93163
20260921	Base	1.05635	.02299	2.16061	.01768
20260921	Higher	.74146	.04527	1.55677	.68905

E.6 Prospective projectile and circular cohort

The prospective cohort fixes three configurations before generation: horizontal projectile motion, oblique projectile motion and uniform circular motion on a horizontal table, each with seeds 20260923 and 20260924 and paired Reference/Controlled arms. The projectile programs use physical time $\tau = t/4$, where t is displayed time in seconds, so one simulated second spans the four

Table 30: All twelve fixed extension outcomes from actual RGB. Support lists valid fitting/prediction centers, out of 24 in each half. Motion flags describe fitting-half evidence, not whole-event certification. Error is the untouched second-half 2D RMSE divided by the median fitting-half sphere radius. Unknown denotes insufficient observations or an inadmissible circle fit; it is not an assigned low physical-plausibility score.

Seed	Motion	Arm	Support	Error / radius	Motion evidence
20260923	Horizontal	Reference	24/24	2.9414	Insufficient
20260923	Horizontal	Controlled	24/24	0.1256	Present
20260923	Circular	Reference	24/24	–	Unknown
20260923	Circular	Controlled	24/24	0.1319	Present
20260923	Oblique	Reference	24/24	3.7114	Present
20260923	Oblique	Controlled	24/24	0.0548	Present
20260924	Horizontal	Reference	5/24	–	Unknown
20260924	Horizontal	Controlled	24/24	0.0519	Present
20260924	Circular	Reference	24/24	–	Unknown
20260924	Circular	Controlled	24/24	0.1157	Present
20260924	Oblique	Reference	19/7	–	Unknown
20260924	Oblique	Controlled	24/24	0.0923	Present

displayed seconds; both set $x = -2.4 + 4.8\tau$ m, with $z = 6 - 4.905\tau^2$ for horizontal motion and $z = 1.2 + 4.9\tau - 4.905\tau^2$ for oblique motion, for a 0.16 m sphere seen by a static side camera. The circular program uses a 0.18 m sphere on a 2 m orbit at $\omega = 0.75\pi$ rad per displayed second, attached by a taut string to a fixed pivot on a frictionless table viewed from above, so the controls depict 1.5 revolutions. These are supplied programs, not recovered laws.

Frame slots 0–23 supply fitting observations and slots 24–47 prediction targets, with at least 18 valid centers per half and a 0.5 s fitting span. Projectile prediction fits $\hat{x}(u) = x_0 + v_x u$ and $\hat{y}(u) = y_0 + v_y u + \frac{1}{2} a u^2$ by linear least squares on image coordinates. Circular prediction fits a circle to the fitting centers and a line to the unwrapped phase, $(\hat{x}, \hat{y}) = (c_x + R\cos(\phi_0 + \omega u),\ c_y + R\sin(\phi_0 + \omega u))$, where u is elapsed time from the first valid fitting observation. Errors are normalized by the median observed sphere radius of the fitting half. Motion evidence is reported separately from predictability: projectile checks require at least one radius of fitted horizontal travel and downward curvature contributing at least half a radius, and circular checks require an admissible orbit radius and at least 30 degrees of fitted angular motion. All twelve renders and observation passes completed once without retries. Mean renderer wall time is 227 s for Reference and 232 s for Controlled, and the twelve observation passes take 214 s in total. Of 576 frame slots, 552 have observations and 535 pass the geometry check. Table 30 lists all twelve outcomes; the oblique first-seed baseline passes the coarse motion checks despite its large prediction error, which illustrates the difference between exhibiting a motion cue and following a consistent trajectory.

E.7 Endpoint coverage and trajectory sanity checks

Table 31 assigns each of the 32 matched outputs exactly one endpoint category; the development outputs are outside this denominator. A held-out forecast fits valid observations in the first 24 frame slots and evaluates valid observations in the last 24, whereas the rebound residual fits and assesses the same clip, and the two use different normalizers.

To check the scorers themselves, we applied eight fixed trajectory transforms to each of three existing projectile and circular clips, producing 24 derived trajectories and no new videos. The unmodified trajectories reproduce their saved forecast values within 10^{-10}. Coordinate noise of at most 0.25 and 0.5 pixels preserves support status, event flags and the ordering of the fully scored projectile pair, but raises the controlled circular forecast error by 10.3% and 21.0%. Static projectile traces yield almost zero forecast error while failing the motion flags, and a static circular trace is unidentifiable. A fixed temporal permutation sharply increases both controlled forecast errors, reversing a trace can preserve the fitted family, and removing half the observations makes all three endpoints unsupported. One residual is therefore neither a monotone corruption score nor a general plausibility label, which is why the endpoint combines coverage, motion and forecast checks. A

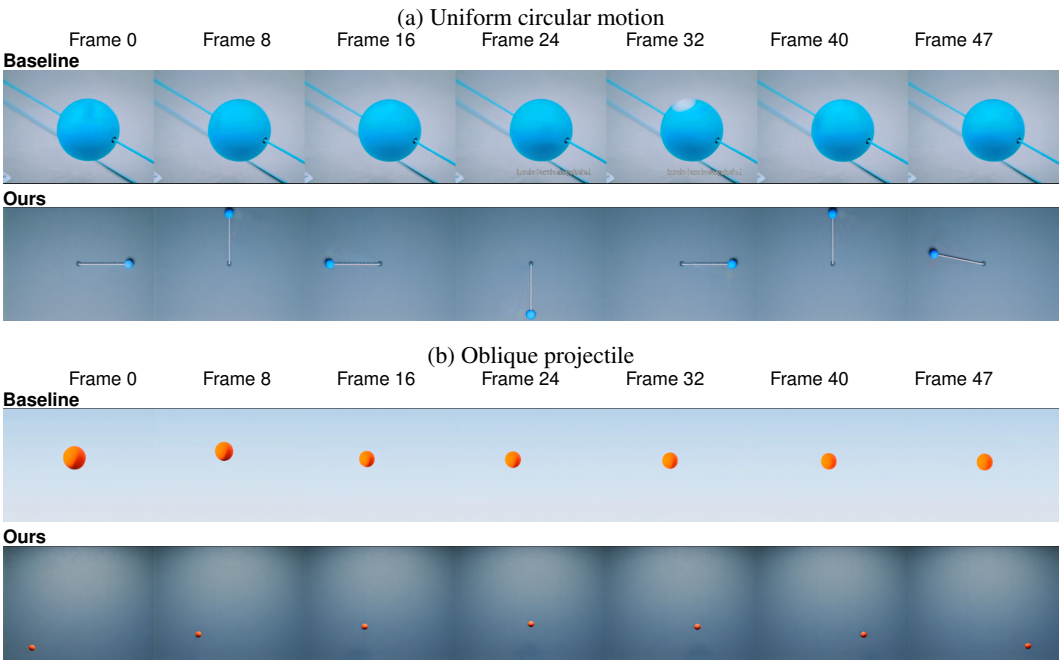

Figure 11: Matched previews of circular and oblique motion. Baseline (top row) receives the prompt and reference image; Ours (bottom row) adds program controls. Frames 0, 8, 16, 24, 32, 40 and 47 are shown. In the oblique pair, the baseline ball drifts without a ballistic arc; our controlled ball rises and falls. Held-out forecast errors are 3.711 and 0.055 object radii, whereas VideoScore2 rates the drifting ball higher (physical consistency 2.67 versus 2.00, mean of three sampling seeds).

Table 31: Saved endpoint availability across all 32 matched outputs. Categories are mutually exclusive; unknown and unassessed clips remain in the denominator. This table reports measurement coverage, not the number of physical successes.

Endpoint category	Outputs
Numeric held-out forecast	13
In-sample rebound diagnostic only	8
Dynamic event not testable	2
Unsupported or unknown endpoint	5
Cradle: no multi-object numeric endpoint	4
All matched outputs	32

Table 32: Completed spring diagnosis-to-action example, separate from the 20-video main cohort. Both new outputs have 48/48 supported centers. Motion amplitude and forecast error use each output's observed object radius. A static retry cannot pass the dynamics gate through a small numerical residual.

New arm	Amplitude	Turns	Event	Held-out error
Reference-only retry	0.0261	0	Failed	Not testable
Directed (49 controls)	3.7951	3	Completed	0.03809

Table 33: Fixed temporal-order intervention at seed 20260922. The same control images are presented in their physical order or with the prescribed middle-block exchange. All four outputs complete the coarse visible event. Errors are in observed object radii and use unchanged first-half fitting and second-half forecasting. There is one paired cell per motion; these development cells are separate from the main cohort.

	Fit discrepancy		Forecast discrepancy	
Motion	Ordered	Permuted	Ordered	Permuted
Spring	0.00930	0.35148	0.03809	4.36949
Pendulum	0.02881	2.50490	0.05207	9.10818

separate bounded, image-driven correction pilot on eight frames from eight existing clips recovered neither of two unsupported observations, and the original first-pass measurements are retained.

E.8 Diagnosis-directed regeneration and temporal-order interventions

Spring. The diagnostic parent is the first-seed reference-only spring failure, whose horizontal motion did not reach the requested oscillation. Before any new render, the episode record specifies the directed action (apply the ordered RGB program controls at strength 1) and its expected effect. The comparator is an unchanged reference-only retry. Both new arms use seed 20260922, the original prompt and reference, the same settings and one invocation each; the failed parent has a different seed and is not a third arm. The directed output completes the event under the unchanged observer, with a first-half frequency estimate of 0.498 Hz and a 0.038-radius forecast error, while the retry stays below the two-radius motion gate; its raw residual of only 0.014 radii illustrates why the motion gate is needed (Table 32).

Pendulum. The second parent is the second-seed reference-only pendulum output, which has 48/48 supported centers but no alternating turns. At seed 20260922, the directed output produces 19.97 radii of motion and four turns with fit and forecast errors of 0.029 and 0.052 radii, whereas the retry moves 0.38 radii with no turns and is not dynamically testable.

Temporal order. One spring and one pendulum output use a fixed permutation of their 49 RGB control frames, with zero-based source order $0{:}15, 32{:}47, 16{:}31, 48$, which preserves the first and last frames and the image multiset while exchanging the middle blocks. The ordered comparators are the directed outputs above; seed, prompt, reference, backend, control strength and budget are identical within each comparison, and no permutation was selected by outcome. All four outputs have 48/48 supported centers. The permuted spring fit selects the lower 0.1 Hz end of its frequency search, which is recorded as a fit warning (Table 33).

One-step feedback with an explicit stop. Two further episodes start from the same spring and pendulum parents and apply the same directed action at seed 20260923. Both outputs have 48/48 accepted observations. The spring output has 3.79 radii of motion, three turns and a 0.039-radius forecast error; the pendulum output has 20.11 radii of motion, four turns and a 0.021-radius forecast error. Because neither measurement reveals a remaining failure that a supported action could correct, the recorded next action is to stop, and no second render is made. These episodes show a measured stopping decision, not a multi-round improvement curve.

Additional pilots with retained failures. At one additional appearance seed, four predeclared Reference/Controlled pairs produced eight outputs. In the pendulum pair, Reference is nearly static and fails the requested event, while Controlled completes four turns with a 0.040-radius forecast error; in the projectile pair, Reference is scored as stationary, while Controlled supports horizontal and acceleration events with a 0.088-radius error. The spring and circular pairs are unsupported: the spring Reference output shows no ball in its first frame, the circular Reference output fails the admissible orbit-radius range, and both Controlled outputs have one accepted observation because their first-frame prompt points were placed away from the visible ball, an observer-initialization failure rather than a generation failure. A separate seed-20260926 pilot on uniform-velocity and uniform-acceleration motion gives a stationary uniform-velocity Reference and an unsupported Controlled (again a misplaced prompt point), and a uniform-acceleration Controlled that supports both events (0.054-radius error) while its reference-only counterpart shows no supported motion. A post hoc diagnostic that combines color proposals with SAM2.1 recovers 48/48 supported centers for the three misplaced-prompt clips, although the recovered spring Controlled trajectory is nearly stationary; these post hoc observations do not replace the frozen scores.

Isolated full-chain replay. In a fresh runtime, the saved horizontal-projectile seed-20260923 program was re-executed to 49 new control frames, both Reference and Controlled were regenerated, and both outputs were re-observed and rescored. All 48 playback frames of both arms match the original outputs byte for byte, and both scores are reproduced exactly (Reference: 2.941 radii without supported events; Controlled: 0.126 radii with horizontal and acceleration events), using 495 GPU-seconds. This verifies one complete execution path; identical outputs are not an independent stochastic replicate.

E.9 Appearance and control fidelity

Table 34 reports automatic appearance measures for all 32 matched outputs, using reimplementations of the VBench quality dimensions (DINO subject consistency, CLIP background consistency, temporal flickering, LAION aesthetic quality and MUSIQ imaging quality) and similarity to the appearance reference and to the rendered control frames. The paired intervals do not resolve differences in consistency, flicker or imaging quality; aesthetic quality is lower in all six prospective pairs. Program-controlled outputs stay close to the rendered controls and have no-reference quality close to the raw Blender render, so their visual realism is bounded by the program scene. The reference-only outputs depart from their own appearance reference, so their deficit concerns scene fidelity as well as motion. VBench's dynamic degree scores the reference-only outputs as more dynamic (mean top-5% flow 15.1 versus 11.5 pixels) because several contain unrequested drift that the trajectory evaluator rejects; the aesthetic and imaging predictors are trained on natural images. These are automatic image statistics, not human realism judgments.

Table 34: Paired appearance measures on the 32 matched outputs (16 seed-specific pairs in 8 configurations; Δ = Controlled − Reference with 95% intervals that resample configurations). Control-fidelity rows use the 10 RGB-control pairs.

Measure	Reference	Controlled	Δ [95% CI]	Controlled higher
Subject consistency (DINO)	0.931	0.931	0.000 [−0.050, 0.043]	8/16
Background consistency (CLIP)	0.966	0.966	0.001 [−0.014, 0.014]	10/16
Temporal flickering	0.992	0.991	−0.001 [−0.004, 0.003]	6/16
Aesthetic quality (LAION)	0.452	0.429	−0.023 [−0.072, 0.029]	5/16
prospective pairs	0.453	0.395	−0.058 [−0.131, −0.016]	0/6
Imaging quality (MUSIQ)	0.522	0.562	0.040 [−0.019, 0.103]	10/16
Similarity to appearance reference (DINO)	0.430	0.687	0.257 [0.167, 0.347]	15/16
LPIPS to control frames ↓	0.589	0.204	−0.386 [−0.529, −0.226]	0/10
SSIM to control frames	0.698	0.902	0.204 [0.134, 0.306]	10/10
PSNR to control frames (dB)	9.9	22.4	12.5 [9.4, 16.0]	10/10

E.10 CONTROL STRENGTH AND REFERENCE-ONLY RESAMPLING

To test whether the renderer must follow the program exactly, a predeclared sweep renders the five RGB-control configurations (spring, pendulum, horizontal and oblique projectile, circular) at control strengths 0.25, 0.5 and 0.75 for both original seeds, and three additional reference-only seeds per configuration provide a best-of-five resampling baseline at matched renderer budget. The same observer and scorers are applied unchanged. An evaluation-gated policy that tries strengths in increasing order and stops at the first output that passes its event and coverage gates can be replayed exactly on this grid. Table 35 summarizes the sweep.

Table 35: Control-strength sweep and reference-only resampling on five RGB-control configurations (two seeds each for Controlled; five seeds for Reference). Entries: outputs passing the event and coverage gates, and median held-out forecast error in object radii among passing outputs.

	Reference (best of 5)	0.25	0.5	0.75	1.0 (full)	Gated policy
Passing outputs	1/25	0/10	10/10	10/10	10/10	10/10
Median forecast error	3.711	—	0.121	0.071	0.074	0.121
LPIPS to program render	0.589	0.633	0.595	0.202	0.204	0.595
MUSIQ imaging quality	0.529	0.432	0.400	0.489	0.515	0.400
CLIP P(photograph)	0.211	0.562	0.442	0.286	0.197	0.442

The appearance rows use the ten original-seed units per strength (Reference: its ten original outputs). Resampling does not rescue the reference-only arm: across 25 samples, one oblique-projectile output passes the event checks, with a 3.711-radius forecast error. The projectile samples are mostly static balls, whose trivially predictable tracks give small forecast errors but no requested motion. Circular samples are close-ups with inadmissible orbit geometry, and oscillator samples fail the motion gate. At strength 0.25 the renderer composes its own reference-only-like scenes and every unit fails. Strengths 0.5, 0.75 and 1.0 pass in every unit. Strength 0.75 matches full control (median forecast ratio 0.97). Strength 0.5 has a median forecast ratio of 2.4 and departs from the program render; the projectile scenes show background artifacts. The gated policy tries 0.25, then 0.5, and stops at 0.5 in all ten units, after two renders each. It rejects every strength-0.25 output and 24 of 25 reference-only samples. The admitted reference-only sample shows that the event gate must be read together with the numerical forecast. The automatic realism measures disagree (CLIP rates strength-0.5 outputs as more photographic, MUSIQ rates their technical quality lower in all ten units). On inspection, all passing outputs remain clearly synthetic, so we make no perceptual-realism claim.

E.11 AN INDEPENDENT LEARNED JUDGE ON THE GENERATION STUDY

We scored all 32 matched outputs, the 45 sweep outputs and the five program renders with VideoScore2. We used the script, checkpoint and settings of the benchmark comparison: 2 fps, at most 1,024 new tokens, sampled decoding at temperature 0.7, and base seed 42 plus a per-clip offset. The only change places the vision and language modules on two 22 GB GPUs; the benchmark run used four 24 GB GPUs. On three saved Test95 clips the run reproduced two of the three saved scores exactly. Because decoding is sampled, we also ran base seeds 43 and 44 and report means over the three seeds. A clip's physical-consistency score varies by about one point across seeds on average.

- **Matched 32:** Controlled − Reference physical consistency is -0.17 ($[-0.54, 0.19]$; Controlled higher in 7 of 16 pairs, lower in 9), and visual quality -0.18 ($[-0.45, 0.08]$).
- **Evaluator pass/fail:** across the 65 scored clips of the five RGB-controlled configurations, VideoScore2 physical consistency is inversely related to the evaluator's pass/fail outcome (AUROC 0.22).
- **Program renders:** the raw renders receive 2.8.

VideoScore2 therefore does not register the requested-motion differences that the trajectory evaluator and the three-rater means identify, subject to the disclosed reference-score exposure.

E.12 REAL-REFERENCE COHORT

A separate controlled-only cohort retains 18 outputs from ten real-labelled source cases across seven nominal motion classes, using exact source captions, verified real frames and source-background composites. Fourteen outputs support their scoped object/action checks; two block orbits leave disk co-rotation unobserved and two earlier cradles retain scene defects. Added cases preserve road traffic, human slider handling, five reflective cradle bodies and the person activating a rotating disk. The final disk pair supports local body/material co-motion with two occluded frames left unknown, but retains visible lighting bands. Only ten of 89 audited sources have generated outputs; 78 contracts remain incomplete and one caption is held. These selected cases, including two illustration sources, have no matched reference-only arm or independent human ratings (Appendix Table 36). They establish scoped feasibility, not physical truth or perceptual realism.

Each output uses the exact metadata caption and a verified frame from its original video (frame 0, except frame 7 for the rebound). Candidate sets, source-specific controls, observers and seeds were fixed before their corresponding outputs. The final source roster comprises all 42 Test95 real sources plus 47 extension candidates; 88 of 89 captions match the master metadata exactly, and the unmatched case remains held. Only ten source cases have generated outputs. The 18 outputs belong to this separate prospective cohort and do not augment the historical 44 or the frozen Test95 score set.

Controls are 49-frame source-background composites of original object appearance under prescribed motion, using explicit donor pixels or background filling. The renderer uses Wan2.1-VACE-1.3B, 640×480, 40 steps, CFG 5 and conditioning strength 1; each call emits 49 raw frames, and frames 0–47 form the evaluated 48-frame playback at 12 fps (raw frame 48 is omitted to retain the prespecified four-second playback). All 18 first passes remain retained, with no scientific retry or reference-only arm. One early observer dependency failure was recovered on unchanged saved RGB with no regeneration. For the deadline extension, two cradle and two disk seeds were frozen together within each source before seeing either result; all four ended normally, each with 49 verified raw hashes and 48 decoded playback frames. Their earlier preparation failures remain logged. Traffic coverage uses the 56 vehicle-frame slots where the prescribed motion places a vehicle fully in view; program visibility defines only that denominator, not proposal selection or identity linking.

The automatic observers use source-frozen appearance or geometry eligibility without selecting identities by fit to the intended trajectory. The rotating-disk observer measures the white body and material optical flow separately. The earlier block-orbit observer has no material channel, so its co-rotation outcome is unobserved rather than a measured failure; source-derived camera and plane geometry are conditioning information, not recovered metric physical truth. For the rotating-disk pair, median one-step material-advection errors are 0.098 and 0.097 radii, with 90th percentiles 0.176 and 0.204; each output has 46/48 body observations, 47/47 material-flow steps and 29 eligible post-activation co-motion steps. Body phase spans 6.098/6.098 radians while integrated material flow spans 5.560/5.552 radians, so this supports the frozen local co-motion criterion, not exactly zero accumulated slip. Separately frozen static-disk and static-body control fixtures both fail that criterion. The cradle observer recovers five bodies in all 48 frames and tests one-in/one-out alternation separately from its held-out forecast; a static cradle control fails the event gate despite near-zero forecast error.

Operational visual review was performed by an LLM multimodal-agent workflow on saved RGB; it is distinct from both automatic motion tests and independent human ratings. Earlier cases have saved sampled RGB reviews; the six traffic, slider, new-cradle and new-disk outputs additionally have complete 48-frame scene reviews. The original two-in/two-out cradle scene failures are retained: string remnants and background patches were already present in the accepted control input and missed at preflight, so this is not attributable solely to the generator. The new cradle has fixed reflective cutout appearance and approximate off-frame supports; the disk retains visible low-frequency lighting bands and interpolated hidden texture. Source composites prescribe motion and retime appearance; they are not source-video reconstruction, calibrated force/material inference or perceptual-realism evidence. Rights to redistribute source appearance and independent human ratings for this cohort remain unavailable.

The separate appearance probe covers the 16 outputs predating the final disk pair, with mean CLIP photograph-versus-render probability 0.573. It uses different scenes and prompts from the synthetic-

reference cohort and supplies no matched comparison or human realism evidence; it does not cover all 18 outputs.

Summed elapsed time across all 18 model calls is 2836.0 seconds (47.27 minutes), with CPU offload and parallel calls on recorded shared hosts. This excludes downloads, adapter preparation and interactive work; it is neither GPU-kernel time nor a complete compute-cost account.

Table 36: All 18 real-reference Controlled outputs from ten sources; nominal class and count in parentheses. Two seeds per source except one for traffic and slider. Paired numbers follow seed order 20260925/20260926 unless labelled outward/return. Coverage is per output and unknowns are retained. r is each observer's radius unit (area-equivalent, nominal or fitted); w is the tracked source object's half-width (per vehicle or slider). Rebound branch-fit residuals, cradle cycle-repeat discrepancies, held-out trajectory forecasts and disk one-step material-advection errors are distinct measurements; values are compared only within each row.

Source case (class; outputs)	Observation support	Scoped result and limitation
Equal-mass collision (illustration) (collision; 2)	48/48, both bodies	Velocity exchange; struck body outgoing 73.4/73.4 versus incident body incoming 73.6/73.5 image px/s.
Unequal-mass collision (illustration) (collision; 2)	48/48, both bodies	Struck light body leaves at 130.2/130.3, incident heavy body at 43.4/43.3, versus incoming 86.9/86.9 image px/s; programmed mass ratio 3.
Vertical rebound (fall; 2)	48/48	Fall, reversal and return; branch-fit residual 0.010–0.016 r. Background seam/shadow remnant retained.
Block orbit (circular; 2)	38/48; 10 unknown	Held-out orbit forecast 0.049/0.048 r; body–disk co-rotation unobserved.
Incline cart (incline; 2)	48/48	Slowing and stop; held-out forecast 0.115/0.126 r.
Two-in/two-out cradle (pendulum; 2)	48/48, five bodies	Phase gate passes, cycle-repeat discrepancy 0.008/0.009 r; string/background defects present in composite input fail scene acceptance.
Road traffic (uniform speed; 1)	50/56 program-visible vehicle-frame slots; 6 unknown	Six directed passages, entries/exits and empty intervals retained; six track forecasts 0.007–0.017 w.
Accelerating slider (acceleration; 1)	35/48; 13 unknown	Human placement/release, acceleration, end impact and rebound; outward/return forecasts 0.194/0.121 w.
One-in/one-out cradle (pendulum; 2)	48/48, five bodies	One-in/one-out, middle three at rest and connected strings; held-out forecast 0.011/0.014 r; fixed reflections and approximate off-frame supports.
Ball and rotating disk (circular; 2)	46/48 body; 47/47 material steps	Human activation and fixture retained; median material-advection error 0.098/0.097 r; conditional forecast 0.077/0.092 r. Lighting bands remain.

E.13 Real-source program reconstruction

We also attempted to recover a program from an observed real video before generation. For one real rebound, one observation pass over 48 native-time samples yields 45 supported centers (22 of 24 fitting and 23 of 24 forecast samples); two samples near contact fail the geometry check, so contact is unverified and restitution is fixed at an assumed $e = 0.7$. Three fixed starts under the inherited bounds converge to a high-residual solution with no impact in the fitting interval (fit and forecast errors 11.00 and 21.13 radii). A post hoc change of only the initial-velocity bounds and starts places an impact inside the fitting interval, but the errors remain 3.20 and 15.92 radii and the velocity reaches its new bound. The reconstruction is therefore unsupported under this numerical policy and no descendants were generated; this is a limitation of the reconstruction procedure, not evidence that the source violates physics. Two later source-conditioned edits of locally generated reference-only videos show that a moving ball can be composited into a simple scene, while in a richer circular-motion scene the original large sphere remained in all 48 frames, failing the prespecified single-ball criterion.

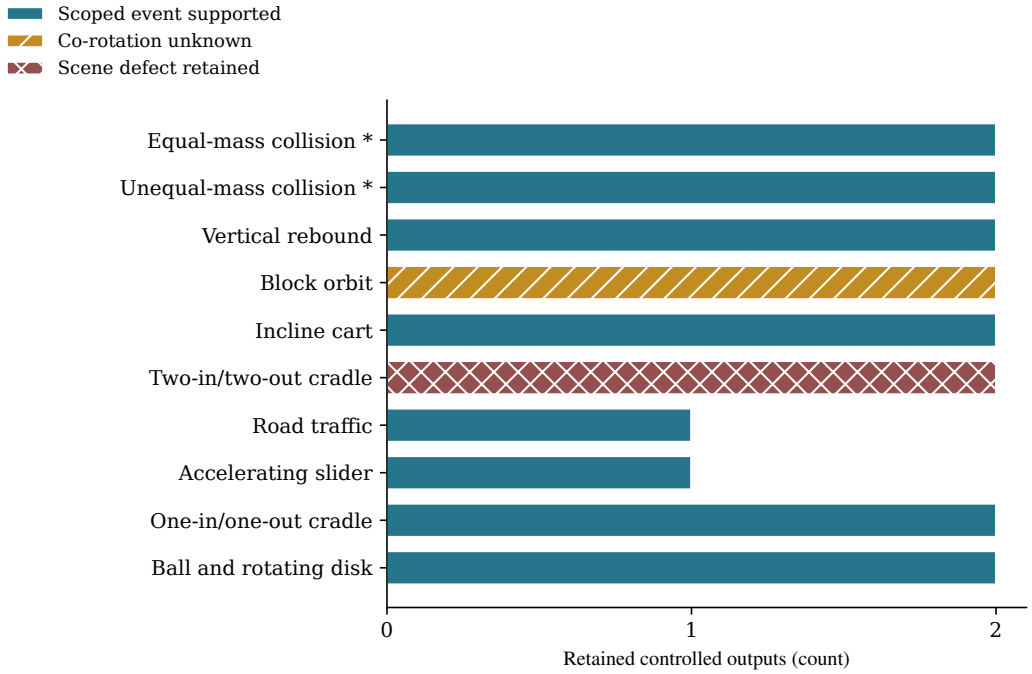

Figure 12: Disposition of every real-reference output under its source-specific operational gates. Fourteen have scoped object/action evidence; two leave disk co-rotation unknown and two retain scene defects. Asterisks mark real-labelled illustration sources. Counts describe selected source cases, not an estimate of general success probability or calibrated physical correctness.

Table 37: RGB observation coverage in the rebound development cohort, which is separate from the matched cohorts.

	State 1			State 2	
Condition	Measured	Unknown	Condition	Measured	Unknown
Reference	35/48	13/48	Reference	35/48	13/48
8 controls	48/48	0/48	8 controls	48/48	0/48
49 controls	48/48	0/48	49 controls	48/48	0/48

E.14 Prior development cohorts

Rebound development cohort. A separate cohort contains two rebound states with reference-only, eight-control and dense 49-control clips. One observation pass preserves all 288 timeline rows (Table 37). Reference-only clips have horizontal variation $D_x = 1.171$ (State 1) and 1.204 (State 2), versus 0.010 and 0.012 for 49 controls; the ordering between 8 controls and 49 controls varies by state. The vertical in-sample residual is 0.031 and 0.053 for 49 controls, 0.298 and 0.515 for 8 controls, and 0.476 and 0.484 for reference-only. These are local in-sample fits to a declared family, not held-out forecasts.

Native-time pilot on existing benchmark videos. Before measurement, a roster of 28 fall and rebound benchmark videos was fixed; the 16 compatible or uncertain candidates were retained. Fourteen SAM calls produced 669 supported measurements and 14 finite passive-bounce fits; two candidates without a supported initialization were not measured. On identical generated IDs, the negative rebound residual correlates with the historical single-rating SCS-sp labels at $\rho = 0.918$ ($n = 7$, four metadata groups), against 0.154 for stored VideoScore2 outputs; on five clips in two groups the residual reaches 0.975, against 0.000 for zero-shot Qwen3-VL and 0.803 for QLoRA. Twelve of the fourteen finite fits carry a boundary, rank or unobserved-impact warning, and the tiny, clustered samples support no population-level ranking.

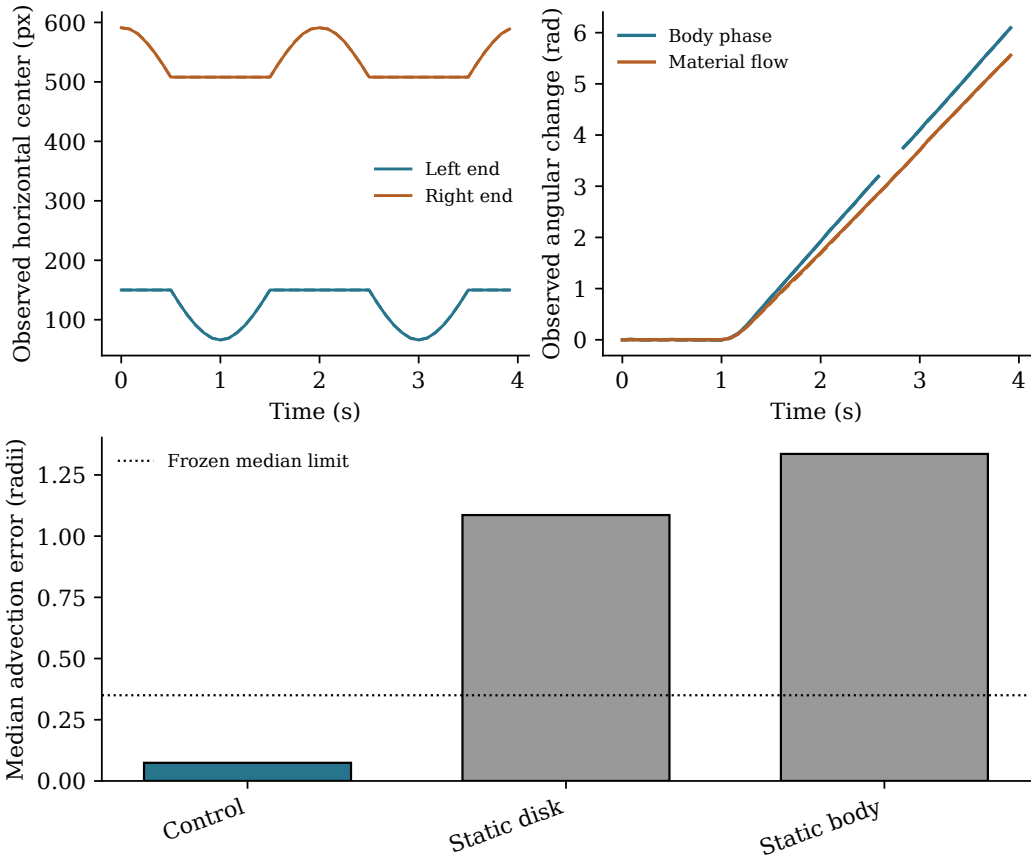

Figure 13: Final-RGB checks for the deadline extension. Top left: measured one-in/one-out cradle end-body centers for both frozen seeds (solid/dashed), with the three middle bodies stationary. Top right: both disk seeds (solid/dashed) show body phase and independently measured material flow; gaps preserve two occluded body frames, and the accumulated discrepancy precludes a zero-slip claim. Both panels use observed RGB, with source-derived, non-metric camera/plane geometry for the disk. Bottom: the original source-control composite (Control bar, 0.074 radii) passes; the static-disk and static-body fixtures fail the frozen local co-motion criterion when either the disk or the body remains stationary; the dotted line is only the median-error component of the multi-condition gate. These controls are not additional model outputs.

Table 38: Completed generated outputs organized as process data. Counts describe completed videos, not independent source groups or newly collected human labels.

Cohort	Videos	Role
Main matched cohort	20	Five configurations, two seeds, paired conditions
Rebound development	6	Two rebound states, three control variants
Diagnostic and order interventions	6	Spring/pendulum: retry, ordered and permuted controls
Prospective motion extension	12	Projectile/circular motion, two seeds, paired conditions
Total	44	Seven proposed conservative grouping blocks

Applicability of the generation observer to Test95. Of the 19 generated Test95 clips in the fall and pendulum classes, applicability was determined from RGB evidence before measurement. Five support the observation proxy, two are uncertain, and twelve do not match the single-object adapter (five Newton's cradles, two deforming coils, three phones and two articulated people). This bounds the scope of the generation observer rather than the physical quality of these clips.

E.15 Process records

Each process record separates the source group, generated clip, matched pair and actual parent–child feedback relation, and stores the caption and its provenance, native timestamps and observation support, the executable family and parameter status (estimated, assumed, fixed or unknown), simulator states and camera mapping, control modality, renderer settings and seed, measured output components, the diagnosis and action or no-op, rating identifiers, and stage costs and failures. Measured RGB, simulator targets and human judgments are separate fields. The catalog contains 44 completed outputs (Table 38); counts are videos, not independent physical systems. For future training and test partitions, related seeds, control variants and feedback descendants are grouped into seven conservative program-state blocks. The later one-step feedback, appearance, linear-motion and real-reference outputs are indexed in separate overlays.

E.16 Human rating protocol

The 38 generation-study clips are packaged in four neutral batches: the first and second seeds of the main cohort (ten clips each), the six development clips (the two directed outputs, the two retries and the two permuted outputs) and the twelve prospective clips. Clip files have neutral names, presentation order is fixed per batch, and method, condition, seed and automatic scores are hidden. Raters apply the benchmark's four-axis protocol, scoring SCS-ge before the prompt is revealed, then PCS, SCS-sp and OCS, and confirm that they watched the full clip. The primary endpoint is the within-rater, within-pair Controlled − Reference difference in SCS-sp; PCS, SCS-ge and OCS are secondary. Human–metric associations are computed only on matching clips within one compatible endpoint family and are undefined for fewer than three clips. The analysis is descriptive; pairs are clustered within configurations and seeds, and no significance test is applied.

Two raters used the reference-assisted portal: all their exported rows carry that flag, and the intake record identifies displayed author examples for five first-seed main-cohort clips. The all-38 analysis is accompanied by a 33-clip sensitivity analysis excluding those examples and any pairs touching them. This exclusion does not establish absence of cross-clip anchoring. The third rater returned an independent assessment; its exports have no reference-assistance field. These ratings concern the separate 38-clip study, not the real-reference cohort.