

How Does Geometry Enter Generated Motion?

Weihan Li¹

Junhao Wu²

Yuhan Song¹

Xiaofeng Lin¹

Xinlei Chen³

¹The University of Tokyo

²RWTH Aachen University

³Harbin Institute of Technology, Shenzhen

liweihan1107@g.ecc.u-tokyo.ac.jp

Abstract

Under a fixed physical law, the visible geometry of a scene determines how motion must change. We ask how video generators realize this relationship. We fix the law and the initial state and change only the geometry drawn in the first frame, within matched families of tracks and deflectors, and compare each generated trajectory with the simulator prediction for that geometry. Paired interventions change one thing at a time: a local bump, the height of a barrier, the words of the prompt, the length of the clip. Across nine image-to-video models, geometry is preserved and shapes the motion: the speed of the ball follows the drawn undulation of a track. A physical state would carry this response forward, and here the generated motion parts from the law. The mean slope barely accelerates the ball, successive contacts fail to compose through a consistent state, an edit ahead of the ball alters its motion before it arrives, and the ball climbs over barriers higher than its release point. Two global conditions organize the global trajectory: text strongly controls the destination, while clip length strongly controls timing in the open-weight models tested. The pattern persists with photographed first frames. Current video generation thus behaves as geometry-conditioned motion synthesis whose evolution of state differs systematically from that of a fixed physical law.

The project page is available at

<https://liweihan1107.github.io/samelaw/>.

1. Introduction

A ball on a track speeds up where the track falls and slows where it rises. Tilt a plate in its path and it leaves in a different direction. Under a fixed physical law and a fixed initial state, the visible geometry of a scene determines how the motion must change, and it does so through the state of the ball: each piece of geometry acts when the ball reaches it, and what it does depends on the speed and direction the ball brings. This gives a direct test for a video generator. When the geometry in the first frame changes, does the generated motion change in the way the same law requires?

The evaluation of physics in generated video has become

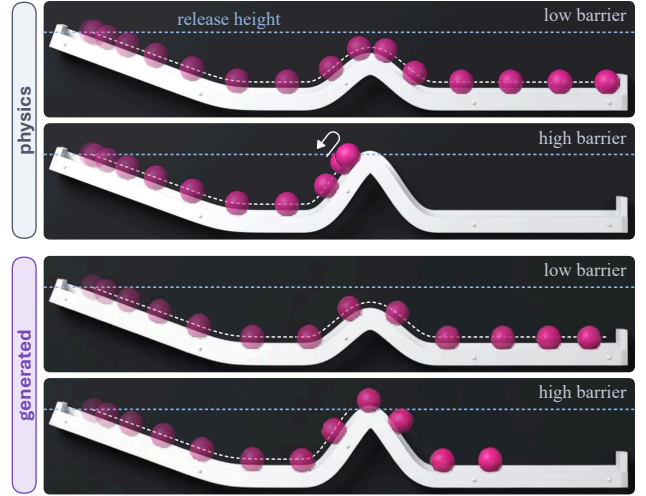


Figure 1. **Same law, different geometry.** One ball, one law, one starting point; only the height of a barrier changes. In the simulated motion (top) the ball crosses the lower barrier and turns back at the one that rises above its release height (dashed line). In the generated motion (bottom; MiniMax-H3, one seed) it crosses both, as it does with every seed.

steadily more quantitative. Benchmarks first asked whether generated clips are physically plausible [3, 19, 21]. Later work tests equations of motion and conserved quantities on trajectories extracted from the videos [31], compares them with simulator rollouts from the same first frame [9], and reads relational quantities that need no calibration of scale or time [29, 30]. Controlled studies now vary one factor at a time: a physical parameter conveyed by the prompt or the conditioning frames [10], the length of a chain of dependent collisions [5], or a counterfactual branch under shared noise [16]. Together, these studies make physical correctness measurable. They leave open a complementary question: how visible geometry enters the generated dynamics when the law is held fixed.

We study this question through *same law, different geometry*. A world is a triple $W = (\Omega, L, x_0)$ of visible geometry, local law and initial state. We fix L and x_0 and vary Ω within families whose global statistics are matched, so that two first frames differ only in shape, and a simulator gives

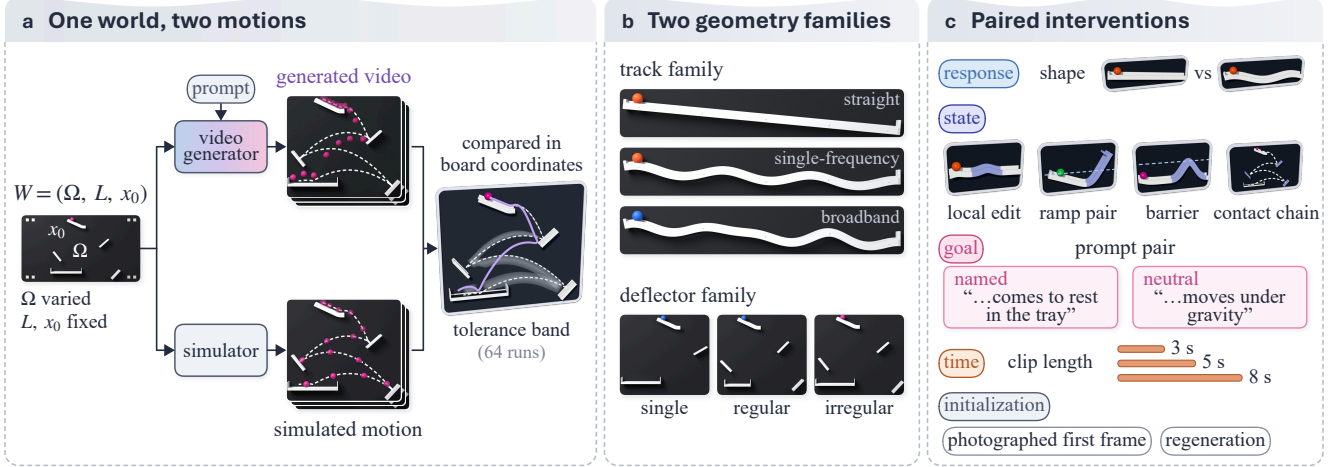


Figure 2. **Study design.** *Left:* a world $W = (\Omega, L, x_0)$ is rendered as a first frame; the video generator produces a video and the simulator the physical motion with its tolerance band; both are compared in the coordinates of the board. *Middle:* two families of geometry with matched global statistics. *Right:* paired interventions, each changing one thing: shape (Sec. 4.1); local edit, ramp pair, barrier and contact chain (Sec. 4.2); prompt pair and clip length (Sec. 4.3); photographed first frame and regeneration (Sec. 4.4).

the physical motion for each. Two families cover continuous and discrete geometry: a ball rolling on a track with a drawn undulation, and a ball bouncing through flat deflectors into a tray (Fig. 2). Around them, paired interventions change one thing at a time: a local bump on the track, the height of a barrier, the words of the prompt, the length of the clip. We generate more than 5,000 videos with nine image-to-video models, six open-weight and three closed, and track the ball in the coordinates of the board. The design separates three questions that a plausibility score merges: whether the geometry is preserved, whether it changes the motion, and whether the change follows one local dynamics.

Geometry enters the motion. Across all nine models, the drawn track is largely preserved and the ball remains constrained to it. The speed of the ball rises and falls with the drawn undulation, at 0.5 to 1.2 of the physical gain and, in eight models, ahead of the physical phase by 15 to 41 degrees. The mean slope, which supplies most of the speed in the physical motion, contributes little: the ball settles to a preferred speed, and the undulation modulates that speed.

A physical state would carry this response forward, and three tests ask for it. Composition: with three deflectors, the physical order of contacts appears in at most a quarter of a model’s videos, and each contact deviates on its own. Locality: a bump changes the speed at the bump in one model, at a fifth of the physical size, while in three models an edit ahead of the ball alters its path before it arrives. Energy: the ball rises above its release height in 63 of 72 videos and crosses 43 of 49 barriers that are physically impassable (Fig. 1). The response on the track is therefore a modulation of motion by shape, and the role of state is taken by

something else.

Two global conditions instead organize the global trajectory. The outcome named in the prompt controls the destination. When the prompt says that the ball comes to rest in the tray, it does so in nearly every video, more often than its own motion, continued by the simulator, would bring it there. With the same first frame and a prompt that leaves the outcome unnamed, arrival falls by 33 to 96 percentage points in all eight models tested, and the drawn tray by itself barely attracts the ball. Clip length controls the timing in the open-weight models tested. From 3 to 8 seconds the same event stretches almost in proportion, and one model reaches the end of the track at nine tenths of the clip at every length.

Taken together, current video generators use the visible geometry to shape motion locally, the text to choose its outcome, and the clip length to pace it. A fixed law prescribes a different organization: a state that is updated where the ball is, composed across events and bounded by energy. The pattern persists when the rendered first frames are replaced by photographs of the built scenes, and a regeneration probe returns each model to its own version of the motion. These results motivate evaluating physics through how controlled changes propagate through generated motion, beyond whether the resulting clip appears physically plausible.

2. Related Work

Physics in generated video. Benchmarks first scored generated clips for physical plausibility, with human or vision-language judges [3, 19] or against real footage of the same event [21, 25]. Trajectory-based evaluation tests the motion of tracked objects against equations of motion and

conserved quantities [31], and relational readings remove the dependence on unknown scale and frame rate [29, 30]. Simulator rollouts from the same first frame separate errors of trajectory, contact and appearance [9]. These studies make the gap between generated and physical motion measurable for a single scene. We study how that gap depends on the visible geometry when the law and the initial state are held fixed.

Controlled response and composition. One-factor interventions reveal what a generator responds to. Physical parameters conveyed by the prompt or by the conditioning frames barely move the generated dynamics [10, 34]. Video diffusion models trained on hard-sphere dynamics degrade with the length of a chain of dependent collisions [5]. Noise-coupled factual and counterfactual branches define locality for interactive world models [16], and single-factor pairs show that video-to-audio models take physics from the caption more than from the video [12]. This makes the response to a controlled change an object of measurement in its own right. Our controlled variable is the shape of the tracks and deflectors, which appear in earlier suites as fixed stimuli [21, 26], and we ask of the response whether it composes across contacts, acts where the ball is, and respects energy.

What organizes generated motion. Generated motion carries an organization of its own. Falls are slower than physical ones and the implied time scale is unstable [6, 30], trajectory spectra are weak at nonzero temporal frequencies [22], motion is planned in the early denoising steps [13] and its phase erodes during later refinement, so that two steps can be more physically consistent than fifty [8], and events start late relative to their cause [2]. We assign parts of the trajectory to conditions: the drawn geometry shapes the local speed, the outcome named in the prompt controls the destination, and the clip length controls the timing. The contrast between simulating a state and applying a regularity has a long history in the study of human intuitive physics [4, 11, 17], and the preferred-speed motion we observe resembles the impetus-like regularities described there.

3. Same Law, Different Geometry

Object. A world is $W = (\Omega, L, x_0)$: visible geometry, a local law and an initial state. Physics defines a flow $\Phi_\Omega(x_0)$. An image-to-video model defines a distribution over trajectories given a first frame that shows Ω and x_0 , and a prompt. We hold L and x_0 fixed and vary Ω . Within a family the variants share their global statistics, such as the length of a track and its physical transit time or the drop and the end points of a deflector course, and the prompt is identical, so

the variant is conveyed by the first frame alone. All scenes share one visual world: a dark board, white geometry, one coloured ball, a fixed frontal camera and four corner markers that give the scale.

Two families. In the *track family* a ball rolls on a twin-rail track whose height follows a mean slope plus an undulation. The levels are a straight track, a single frequency of 3 or 4 cycles per metre at 15 mm amplitude, and a broadband profile with three components of 10 mm, matched in length and physical transit time (40 scenes). In the *deflector family* the ball leaves a short launch ramp, strikes flat deflectors and lands in a tray. The levels are one deflector at four angles, three deflectors in a regular zigzag, and three with irregular angles and spacing, matched in drop, start, tray position and mean angle (24 scenes). Smooth tracks and flat contacts keep the physical trajectory stable: the exit direction depends on the visible normal and the irregularity lies in the layout of the geometry. The reference is therefore one trajectory with a narrow band.

Paired interventions. Each intervention changes one thing between generations that share everything else, including the seed where the model allows it. *Shape*: the levels of a family. *Local edit*: a 15 mm bump or dip on one section of a track, with an edit of equal size placed off the track as a null. *Ramp pair*: a descent followed by an ascent of 0 to 60 degrees. *Barrier*: a hump whose height is 0.70 to 1.20 of the release drop. *Prompt pair*: the same first frame with a prompt that names the outcome (the ball comes to rest in the tray) or a neutral one. *Clip length*: 3, 5 and 8 s. *Photographed first frame*: a photograph of the built scene in place of the render. *Regeneration*: a source video is partially noised and regenerated [18].

Reference and readings. A simulator integrates L on the drawn geometry, and a 64-member ensemble over material parameters and the starting point gives a tolerance band. We track the ball into board coordinates. On a track, the ball’s speed along the arc is regressed on the height profile, split into mean slope and undulation. This yields the *trend ratio*, the acceleration along the mean slope relative to the acceleration implied by the ball’s own transit time, and a complex *gain* H for the undulation. Both are normalized by the tempo α , the ratio of generated to physical transit time, so that any time-rescaled Newtonian motion has trend ratio 1, $|H| = 1$ and phase 0. At a deflector we compare the exit direction with the simulator’s prediction for the generated incoming direction, and fit the apparent normal *restitution* e_n . We also read the order of contacts, whether the ball comes to rest in the tray, and, on ramps, the *height ratio* of the highest point reached to the release height. Exact renders passed through the same chain give each reading

Table 1. **Nine models under the same law.** Medians over scenes; the first row is the physical oracle, the simulator read through the same measurement chain. Intact: share of intact videos in the track and the deflector family. Gain and phase are read on tracks with 3 cycles per metre, the trend ratio on straight tracks, e_n at the first contact, the order of contacts on three deflectors, and the height ratio on ramp pairs. Tray arrival is given under the named and the neutral prompt, and γ is the exponent of tempo against clip length ($\dagger$: twice as fast at 3 s, unchanged from 5 to 8 s). A dash marks a reading that was not taken for that model.

Model	Intact (%)		Response			State			Goal: tray (%)	Time
	track	deflector	gain $ H $	phase ($^\circ$)	trend	e_n	order (%)	height	named $\rightarrow$ neutral	γ
★ Physical oracle	–	–	1.00	0	1.00	0.85	100	0.96	91 $\rightarrow$ 91	0
<i>Open-weight</i>										
▲ LTX-2.3	87	1	1.24	+38	0.15	0.35	0	1.23	100 $\rightarrow$ 25	0.85
▼ LTX-2.5 (distilled)	100	75	0.97	+35	−0.10	0.21	0	2.36	100 $\rightarrow$ 56	0.91
● LTX-2.5	83	6	0.98	+36	0.02	–	–	1.60	–	–
◆ Cosmos 3 Nano	88	67	0.57	+41	−0.02	0.10	0	1.15	100 $\rightarrow$ 7	–
◆ MiniMax-H3	100	88	0.73	−3	0.22	0.45	0	1.56	98 $\rightarrow$ 2	1.04
■ Wan 2.2 (A14B)	96	81	0.54	+15	0.22	0.15	0	1.76	47 $\rightarrow$ 3	0.89 [†]
<i>Closed</i>										
○ Wan 3.0	100	88	1.11	+27	0.25	0.67	12	1.94	100 $\rightarrow$ 67	–
◀ Seedance 2.0	100	72	0.60	+19	0.62	0.74	24	1.66	100 $\rightarrow$ 33	–
▶ Seedance 2.0 mini	100	86	0.59	+36	0.65	0.67	14	1.82	100 $\rightarrow$ 33	–

a measurement floor, and differences are read against the floor and the band.

Models. We study nine image-to-video models: six open-weight models (LTX-2.3, LTX-2.5 and its distilled version, Wan 2.2, MiniMax-H3, Cosmos 3 Nano) and three closed models accessed through their APIs (Seedance 2.0, Seedance 2.0 mini, Wan 3.0) [1, 7, 14, 15, 20, 23, 24, 27, 28, 32, 33]. Unless otherwise stated, each model generates 5 s clips at its native resolution and frame rate, with three seeds per scene. Interventions that pair videos by seed use the open-weight models. A video is *intact* when the scene stays what the first frame shows: one ball, static apparatus, steady camera. It stays *readable* when a part changes only after the ball has left it. Dynamics are read on readable videos, and the intact share is reported with them.

4. Results

Each subsection takes up the question left by the one before. Table 1 collects the readings per model, and the supplement reports every reading with its interval, floor and denominator.

4.1. How does visible geometry modulate motion?

The drawn geometry is preserved. In the track family, 83–100% of each model’s videos are intact. In them the drawn undulation keeps its amplitude to within 1% and the ball stays on the track. In the deflector family, seven models keep 67–88% of their videos intact, with the deflectors within 0.2° of their drawn angle. LTX-2.3 and LTX-2.5 move parts of the deflector scenes or the camera, and enter

the analysis through the track family. Whatever separates generated from physical motion below arises with the geometry in place.

The speed follows the drawn shape. On undulating tracks the speed of the generated ball rises and falls with the drawn profile (Fig. 3a). Relative to each model’s own tempo, the gain at 3 cycles per metre is 0.54–1.24 of the physical one. Eight models lead the physical phase by 15° – 41° , at least twice the floor of 6° – 7.5° , and MiniMax-H3 responds in phase. The response repeats across seeds. When tracks share their first 0.4 m and differ afterwards, the response in the later section follows the phase drawn there, with a slope of 0.80–1.07 in six models. The models read the undulation from the first frame, and the motion carries it where it is drawn.

The mean slope contributes little. In the physical motion the mean slope supplies most of the ball’s speed. In the generated motion the acceleration along the straight track is at most 0.65 of what the ball’s own transit time implies, and below 0.3 in seven models (Tab. 1); on the undulating tracks it is at most 0.42 (Fig. 3b). On a straight ramp the distance covered grows with time with an exponent of 0.5–1.6, where uniform acceleration gives 2 and a constant speed gives 1 (Fig. 3c). A ball that relaxes toward a preferred speed reproduces this signature as a whole: a flat trend, a reduced gain and a phase lead. Among ten candidate mechanisms passed through the same camera and estimator, four models are best described by a preferred speed that varies with the local slope and five by a constant preferred speed

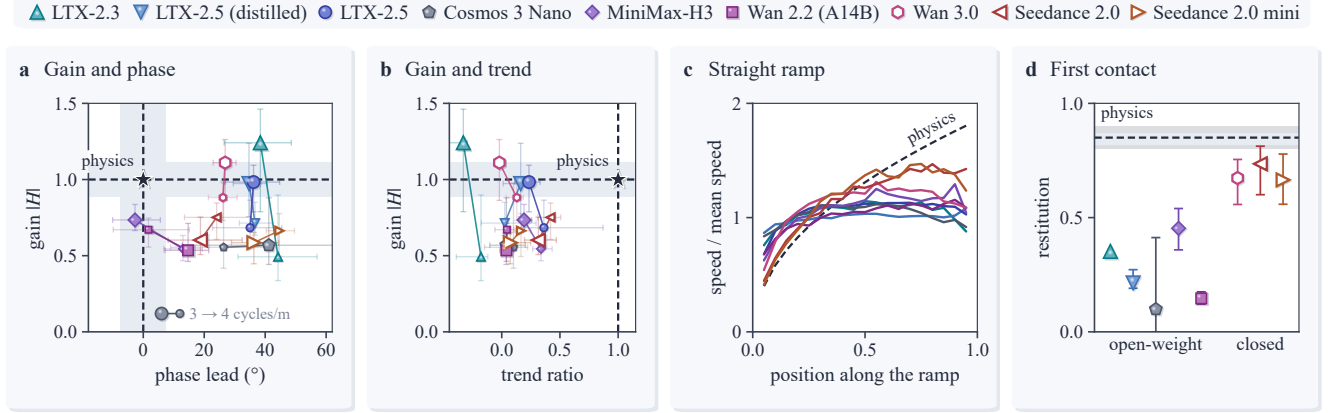


Figure 3. **Geometry modulates the motion.** (a) Gain and phase of the speed’s response to the drawn undulation; physics sits at gain 1 and phase 0. (b) The same gain against the trend ratio: the undulation is followed while the mean slope adds little. (c) Speed along a straight ramp, normalized: physics accelerates, the models hold a preferred speed. (d) Restitution at the first contact; the prescribed value is 0.85. Filled markers are open-weight models and hollow markers closed ones; grey bands are measurement floors.

with temporal smoothing. Time-rescaled Newtonian motion fits none of the nine (supplement).

Closed models bounce, open-weight models slide. The deflector family gives the discrete counterpart. At the first contact the three closed models bounce, with a restitution of 0.67–0.74 against the prescribed 0.85 and an exit direction 16°–35° from the physical one. MiniMax-H3 bounces in part, with a restitution of 0.45. Cosmos 3 Nano, the distilled LTX-2.5 and Wan 2.2 slide along the deflector, with a restitution of 0.10–0.21 and exits 57°–78° away (Fig. 3d). In both groups the ball meets the drawn surface where it is drawn.

4.2. Does the modulation form a local dynamics?

A state-mediated physical dynamics makes three testable predictions here: successive events compose through the state, only geometry the ball has reached can affect it, and its evolution obeys the energy bound.

Successive contacts fail to compose. With three deflectors the simulator’s ensemble keeps a single order of contacts. The closed models reproduce it in 12–24% of their videos and the open-weight models in none (Fig. 5a). In MiniMax-H3, Wan 2.2 and Cosmos 3 Nano the ball seldom reaches a second contact: it slides off the first deflector and drops. In the two Seedance models the exit direction is off by 19°–31° at the first, the second and the third contact alike, so the chain is lost through independent local deviations, and in Wan 3.0 the deviation grows from 10° to 28°. A contact law calibrated on each model’s single-deflector videos improves the prediction of the three-deflector scenes for one model only, so the deviations do not reduce to a

consistent wrong material.

Edits act ahead of the ball. We add a 15 mm bump or dip to one section of a track and generate with the same seed as for the unedited track. Physics changes the speed at the edit by –11% and +8% and leaves the motion before it unchanged. MiniMax-H3 slows at the bump by 2.4%, a fifth of the physical change, and reaches the edit at the unchanged time. In LTX-2.5 the ball reaches either edit 3–6% sooner and slows at the dip, where physics speeds up. In three open-weight models the path before the ball reaches the edit changes 1.7–4.0 times as much as under an edit of equal size placed off the track, beyond the seed-matched null (Fig. 5b). The shape ahead of the ball takes part in the motion from the start.

The ball climbs past its release height. On a ramp pair the physical ball turns back at 0.93–0.98 of its release height. In 63 of 72 readable videos the generated ball rises above the release height, and in 37 it runs to the top of the ramp (Figs. 4 and 6a). The mean height ratio per model is 1.15–2.36. Barriers show the same result as a threshold. Physics passes the barriers below the release height and none above it. Six models cross every barrier they were given, and overall 43 of 49 physically impassable crossings are made (Figs. 1 and 6b). In physics the speed kept over the top falls from 0.53 of the speed at the foot to zero as the barrier rises. In the models it stays at the same level at every height, for example 0.42–0.46 in Seedance 2.0 mini. The ball slows on a rise, as the shape suggests, by an amount unrelated to the height gained. The preferred-speed description of Sec. 4.1 reaches its limit here: no single rule of that kind reproduces one model’s behaviour on the undulating

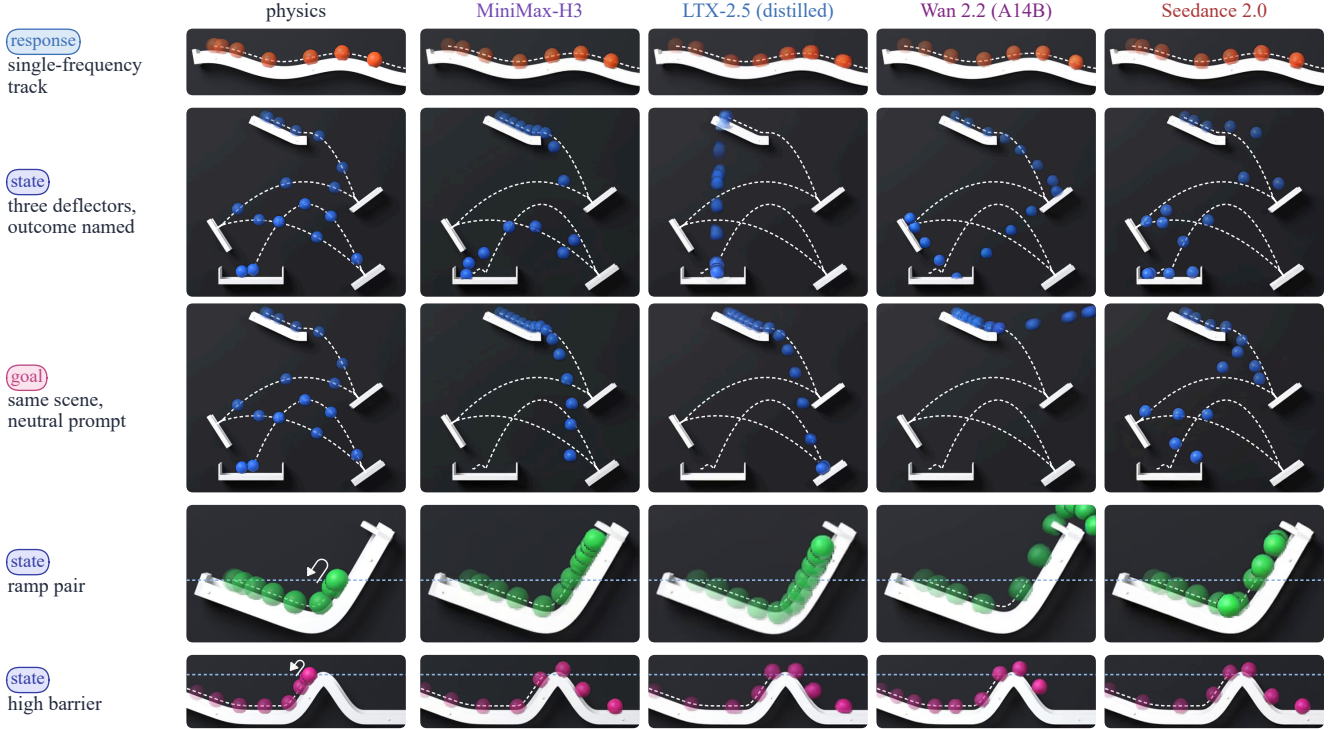


Figure 4. **Cases.** Five scenes (rows) in the simulated motion and in four models (columns). The ball is drawn at equal time steps, from light to full, so that its spacing shows the speed; the white dashed curve is the physical path and the blue dashed line the release height. The third row repeats the scene of the second with a prompt that leaves the outcome unnamed. Each model contributes the video closest to its median reading on that scene.

track, at the local edit and on the ramp pair together (supplement).

4.3. What organizes the global trajectory?

The local response does not propagate as a consistent physical state. What, then, organizes where the ball ends and when? Two paired interventions identify strong global controls.

The named outcome controls the destination. The prompt of the deflector family says that the ball bounces on the deflectors until it comes to rest in the tray. Under this prompt the ball reaches the tray in 98–100% of the trackable videos of seven models, whether or not the contacts were right. The generated motion itself leads there less often: continued by the simulator from the state in which the ball leaves its last correct contact, it reaches the tray in 11–84% of the videos of six of these models, which exceed the simulator by 16–87 percentage points. We then generate the same scenes, first frames and seeds with a prompt that describes the scene and leaves the outcome unnamed. Tray arrival falls in all eight models, by 33 to 96 points (Fig. 7a). The excess over the simulator is then at most 27 points in every model, and in MiniMax-H3 and the two Seedance mod-

els it is gone. Wan 3.0 retains the most, with 67% arrival under the neutral prompt. The drawn tray by itself barely attracts the ball: when the tray is shifted, the landing point follows it by at most 0.42 of the shift, and with the tray removed no model brings the ball to rest where it was.

Clip length controls the timing. For four open-weight models we change only the length of the clip, from 5 s to 3 or 8 s, with the scene, prompt and seed fixed. A physical event keeps its duration, so its tempo is independent of the clip length. The generated tempo grows with the clip length with an exponent γ of 0.85–1.04 in three models, where 1 means that the event is scaled to the clip (Fig. 7b). MiniMax-H3 reaches the end of the track at 90–93% of the clip at every length, LTX-2.3 at 55–59% and the distilled LTX-2.5 at about half. Plotted against normalized time, the three motions of one seed fall on one curve (Fig. 7c). The response to the undulation keeps its gain and phase through this stretching, and a longer clip brings neither more arrivals at the tray nor a better order of contacts. The time left after the event is filled by a ball at rest or by a second run of the event. Wan 2.2 doubles its speed at 3 s and keeps its 5 s pace at 8 s. The named outcome and the clip length act together: under the neutral prompt the same scenes run 1.2–2.4 times

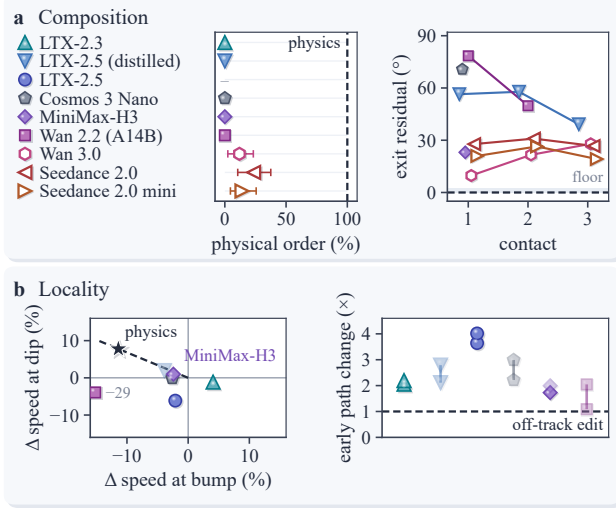


Figure 5. **Composition and locality.** (a) Share of three-deflector videos with the physical order of contacts, and exit residual by contact index; the rows name the models, and a dash marks a reading that was not taken. (b) Change of speed at a bump or dip against physics, and change of the path before the ball reaches the edit, relative to an edit of equal size off the track; faded marks do not separate from the off-track edit.

slower in seven of eight models, and many are still under way when the clip ends.

4.4. Photographed first frames and regeneration

The pattern holds on photographs. We built eight of the scenes, photographed them, measured their geometry from the photographs and ran the simulator on the measured geometry. With the photographs as first frames, the readings of six of the nine models stay within their seed-to-seed spread of the readings on renders of the same geometry (Tab. 2a). The three changes are specific: LTX-2.5 moves the camera on photographs, MiniMax-H3 shows larger exit residuals, and Seedance 2.0 a higher gain. The pattern of Secs. 4.1 to 4.3 is reproduced on photographs. The speed follows the drawn undulation, with a phase lead in eight models. The trend ratio stays below 0.5. The order of three contacts appears in at most one of six videos per model. The ball reaches the named tray in every trackable video of every model except Wan 2.2, which misses it on renders as well.

Regeneration returns the model’s own motion. Finally we give two models a source video whose structure is intact and whose dynamics are wrong, add noise and let the model regenerate it. We read λ , the displacement of the regenerated motion from the source toward physics: 0 keeps the source, 1 reaches physics. A motion slowed twofold keeps its tempo. A constant-speed motion keeps its tim-

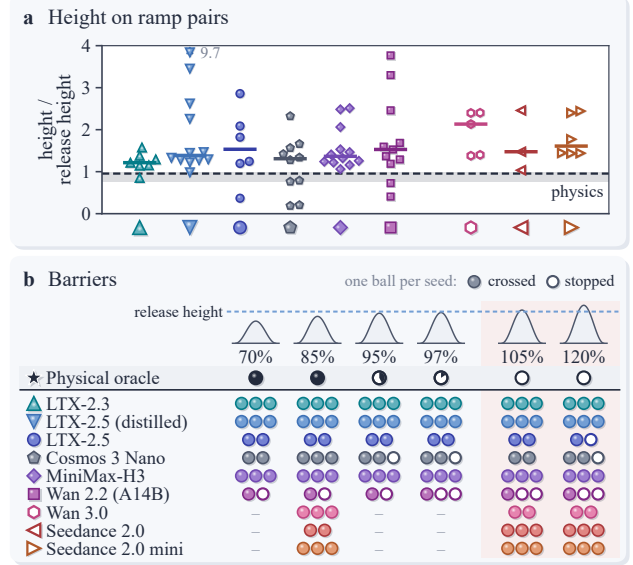


Figure 6. **Energy.** (a) Height reached over release height on ramp pairs, one mark per video with bars at the median; physics lies in the band below 1, and the glyphs under the swarms name the models in the order of the rows of (b). (b) Barriers, one ball per seed, solid where the ball crossed and hollow where it stopped; columns give the barrier height in percent of the release height, tinted above it. The physical oracle crosses below the release height and none above; its ball is filled by the share of the 64-member ensemble that crosses.

ing, and the response to the undulation is added back. A phase shifted by 90° returns to the model’s own lead. For these sources λ stays below 0.1, except at the highest noise level of the distilled LTX-2.5, where the output drifts toward the model’s own video (Tab. 2b). A violated contact is changed: a ball that passes through a drawn deflector is made to hit it, with $\lambda \approx 1$. What replaces the source is the model’s own trajectory for that scene, with the order of contacts still wrong and a restitution of 0.16–0.45 in place of the prescribed 0.85. For a source with a wrong restitution, the regenerated motion of Wan 2.2 moves 0.66 of the way toward the model’s own image-to-video output of the same scene and seed, and 0.10 toward physics.

5. Discussion

Three signals, three roles. The measurements assign a role to each input of an image-to-video model. The geometry drawn in the first frame is kept, and it modulates the motion where it is drawn. The text names an outcome, and the motion is brought to it, whereas text that names a physical parameter leaves the dynamics in place [10]. The clip length fixes a span, and the event is fitted into it. A similar division appears in video-to-audio generation, where the caption decides what is generated and the video decides when [12]. A

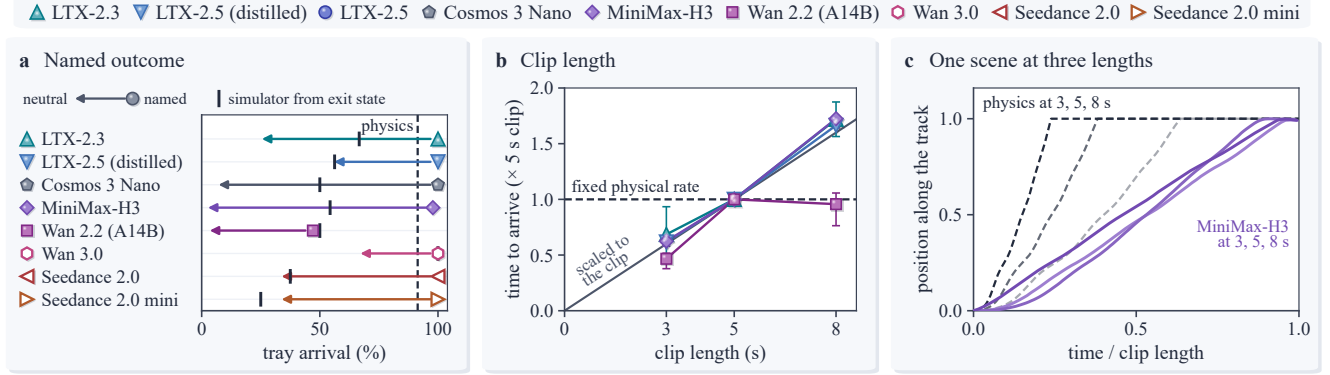


Figure 7. **Text controls the destination, clip length the timing.** (a) Tray arrival under the prompt that names the outcome (marker) and under the neutral prompt (arrow head), for the same first frames and seeds; ticks mark the arrival the simulator reaches from the model’s own exit state, and the dashed line the simulator from the initial state. (b) Arrival time against clip length; the dashed line shows a fixed physical rate, the solid line scaling with the clip. (c) One scene and seed at 3, 5 and 8 s: position against normalized time.

Table 2. **Photographed first frames and regeneration.** (a) Change of each reading from rendered to photographed first frames of the same geometry: a check where the reading stays within the seed-to-seed spread and the change where it leaves it, bold when it is attributed to the photograph and in parentheses when it lies within the range of the weaker control; a dash where there is no reading. (b) Displacement λ of the regenerated motion from the source (0) toward physics (1), by source and noise level.

(a) Photographed first frames

Model	gain $ H $	Track tempo α	Deflector exit residual	tray
▲ LTX-2.3	✓	✓	✓	—
▼ LTX-2.5 (distilled)	✓	✓	✓	✓
● LTX-2.5	✓	✓	−42 %	—
⬢ Cosmos 3 Nano	✓	✓	✓	✓
◆ MiniMax-H3	✓	✓	+18°	✓
■ Wan 2.2 (A14B)	✓	✓	✓	✓
○ Wan 3.0	(+0.35)	(+0.26)	✓	✓
◁ Seedance 2.0	+0.66	✓	✓	✓
▷ Seedance 2.0 mini	✓	✓	✓	✓

(b) Regeneration λ : 0 source $\square\square\square\square\square$ 1 physics

Source	noise level	▼ LTX-2.5 (distilled)	■ Wan 2.2 (A14B)
		0.70 0.80 0.90	0.90 0.94 0.95
constant speed		0.01 0.00 0.00	0.01 0.02 0.02
phase +90°		0.00 0.02 0.10	0.01 0.07 0.08
slowed 2×		0.00 0.06 0.24	0.00 0.00 0.01
restitution 0.3		0.01 0.03 0.32	0.01 0.99 1.40
transparent deflector		0.07 1.01 1.01	0.28 1.09 1.12

fixed law organizes motion differently: through a state that is updated where the ball is, composed across events and bounded by energy. The generated motion reproduces the local imprint of geometry on speed and direction, and the destination and the pace come from the conditions on the whole clip. This is consistent with generation that shapes the clip as a whole: a destination and a duration are available as global conditions, while a state has to be computed

event by event [5, 13].

Evaluating physics in generated video. Two common readings are affected. A score based on the outcome of an event can be satisfied through the prompt: with the outcome named, the ball arrives more often than its own motion allows. A score normalized for tempo absorbs the clip length: the same event is slower in a longer clip. The paired interventions used here address both, because the quantity read is the change of the motion under a controlled change of one input, with the law as the reference for what that change should be. This matters when video generators are used to predict the consequence of an action: a description that names the intended outcome can bias an outcome-based evaluation even when the intervening dynamics are wrong.

Scope. The study covers one rigid ball in two planar families of geometry, nine image-to-video models and clips of 3 to 8 s. The reference is a simulator that is exact for the drawn geometry and the stated law, and photographed first frames test the visual domain. The design carries over to other laws and scenes: it needs a family of geometries with matched statistics, a simulator for the reference, and pairs of generations that differ in one thing.

References

- [1] Alibaba Cloud. Wan3.0 video generation API reference. <https://www.alibabacloud.com/help/en/model-studio/wan3-video-generation-api-reference>, 2026. Alibaba Cloud Model Studio documentation for the closed Wan 3.0 video model (wan3.0-video). Accessed 2026-09-30. 4, 21
- [2] Estela Monserrat Arriaga Santana, Julian Rosas Scull, Ehécatl Sacamch’en Núñez Rico, and Hugo Jair Escalante. Foresight at the Event Boundary: Evaluating

- physical prediction in video world models. *arXiv preprint arXiv:2609.36531*, 2026. 3
- [3] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations (ICLR)*, 2025. 1, 2
 - [4] Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013. 3
 - [5] Jorge Diaz Chao, Konpat Preechakul, Yuxi Liu, and Yutong Bai. The seriality gap in video diffusion models. *arXiv preprint arXiv:2607.13031*, 2026. 1, 3, 8
 - [6] Xiangbo Gao, Mingyang Wu, Siyuan Yang, Jiongze Yu, Pardis Taghavi, Fangzhou Lin, and Zhengzhong Tu. The Pulse of Motion: Measuring physical frame rate from visual dynamics. *arXiv preprint arXiv:2603.14375*, 2026. 3
 - [7] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, et al. LTX-2: Efficient joint Audio-Visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026. Technical report of the LTX-2 family (base of LTX-2.3 and LTX-2.5). 4, 21
 - [8] Woojung Han, Seil Kang, Youngjun Jun, Min-Hung Chen, Fu-En Yang, and Seong Jae Hwang. Physics in 2-Steps: Locking motion priors before visual refinement erases them. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. 3
 - [9] Swarnim Jain and Shangzhe Wu. RigidBench: Evaluating Rigid-Body physics in video generation models. *arXiv preprint arXiv:2608.15555*, 2026. 1, 3
 - [10] Rasul Khanbayov and Hasan Kurban. PhysWeep: Does a video generator realize the physics you ask for? *arXiv preprint arXiv:2609.06207*, 2026. 1, 3, 7
 - [11] Maria Kozhevnikov and Mary Hegarty. Impetus beliefs as default heuristics: Dissociation between explicit and implicit knowledge about motion. *Psychonomic Bulletin & Review*, 8(3):439–453, 2001. 3
 - [12] Tingle Li, Siddharth Gururani, Kevin J. Shih, Gantavya Bhatt, Sang-gil Lee, Zhifeng Kong, Arushi Goel, Gopala Anumanchipalli, and Ming-Yu Liu. Benchmarking Single-Factor physical Video-to-Audio generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1939–1949, 2026. 3, 7
 - [13] Yueyan Li, Haibo Wang, Caixia Yuan, and Xiaojie Wang. Why do video diffusion models violate physics? unveiling the flaws in attention mechanisms. *arXiv preprint arXiv:2609.23658*, 2026. 3, 8
 - [14] Lightricks. LTX-2.3 model card. <https://huggingface.co/Lightricks/LTX-2.3>, 2026. Open-weight 22B audio-video model (dev and distilled checkpoints). Accessed 2026-09-30. 4, 21
 - [15] Lightricks. LTX-2.5 model card. <https://huggingface.co/Lightricks/LTX-2.5>, 2026. Open-weight 22B model with full (dev) and distilled checkpoints; see also <https://ltx.io/model/ltx-2-5>. Accessed 2026-09-30. 4, 21
 - [16] Yu Ma, Hongli Shi, and Xinran Xu. Twin Rollouts: Noise-Coupled counterfactual branching in interactive video world models. *arXiv preprint arXiv:2608.08982*, 2026. 1, 3
 - [17] Michael McCloskey, Alfonso Caramazza, and Bert Green. Curvilinear motion in the absence of external forces: Naïve beliefs about the motion of objects. *Science*, 210(4474):1139–1141, 1980. 3
 - [18] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022. 3, 22
 - [19] 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 (ICML)*, pages 43781–43806, 2025. 1, 2
 - [20] MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities. <https://www.minimax.io/blog/minimax-h3>, 2026. Release note, 2026-07-31. Open weights and model card: <https://huggingface.co/MiniMaxAI/MiniMax-H3>. Accessed 2026-09-30. 4, 21
 - [21] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 948–958, 2026. 1, 2, 3
 - [22] Ziqi Ni, Rui Li, Shiqi Jiang, and Wei Zhou. MotionSpec: Spectral trajectory supervision for Motion-Consistent video generation. *arXiv preprint arXiv:2609.28095*, 2026. 3
 - [23] NVIDIA. Cosmos3-Nano model card. <https://huggingface.co/nvidia/Cosmos3-Nano>, 2026. 16B omnimodal world model of the Cosmos 3 family. Accessed 2026-09-30. 4, 21
 - [24] NVIDIA et al. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026. 4, 21
 - [25] Tim Rädtsch, Yuki M Asano, Hilde Kuehne, Stefan Bauer, Priyank Jaini, Robert Geirhos, and Carsten T. Luth. Physics-IQ verified. *arXiv preprint arXiv:2606.18943*, 2026. 2
 - [26] Jonathan Sadeghi, Jenny Seidenschwarz, Jesse Allardice, Sirish Srinivasan, Benjamin Graham, and Jeffrey Hawke. CaliBench: Are the stochastic dynamics of video world models physically calibrated? *Transactions on Machine Learning Research*, 2026. 3
 - [27] Team Seedance et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026. Seedance 2.0 model card; covers Seedance 2.0 and 2.0 Fast. 4, 21
 - [28] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, et al. Wan: Open and advanced Large-Scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 4, 21

- [29] Varun Varma Thozhiyoor, Shivam Tripathi, Venkatesh Babu Radhakrishnan, and Anand Bhattad. Principia: Relational physics tests for video models. *arXiv preprint arXiv:2609.04200*, 2026. 1, 3
- [30] Varun Varma Thozhiyoor, Shivam Tripathi, Venkatesh Babu Radhakrishnan, and Anand Bhattad. Objects in generated videos are slower than they appear: Models suffer Sub-Earth gravity and don’t know Galileo’s principle...for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 3830–3839, 2026. 1, 3
- [31] Antonios Tragoudaras, Chenyu Zhang, Daniil Cherniavskii, Antonis Vozikis, Thijmen Nijdam, Derck Prinzhorn, Márk Bodrácska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Evaluating Newtonian mechanics in video generative models with real physical systems. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. 1, 3
- [32] Volcano Engine. Model list (ark): doubao-seedance-2-0-mini-260615. <https://www.volcengine.com/docs/82379/1330310>, 2026. Official model list of the Ark platform, which lists Seedance 2.0 mini (model id doubao-seedance-2-0-mini-260615). In Chinese. Accessed 2026-09-30. 4, 21
- [33] Wan-AI. Wan2.2-I2V-A14B model card. <https://huggingface.co/Wan-AI/Wan2.2-I2V-A14B>, 2025. Open-weight image-to-video mixture-of-experts model (27B total, 14B active). Accessed 2026-09-30. 4, 21
- [34] Meng Zhu and Airui Zhang. MultiEcho: An experimental science of learned worlds. *arXiv preprint arXiv:2609.33347*, 2026. 3

How Does Geometry Enter Generated Motion?

Supplementary Material

Key

open-weight models (filled marks)	closed models (hollow marks)
▲ LTX-2.3	○ Wan 3.0
▼ LTX-2.5 (distilled)	◁ Seedance 2.0
● LTX-2.5	▷ Seedance 2.0 mini
★ Cosmos 3 Nano	integrity only:
◆ MiniMax-H3	× Wan 2.2 (T12V-5B)
■ Wan 2.2 (A14B)	
★ Physical oracle	the simulator's reading through the same measurement chain
1.24 [0.79, 1.46] (24)	median, 95% interval over scenes, denominator a value and its interval on a common scale, with a tick at the physical value
	intact (dark) and readable (light) share of the videos
	tray arrival under the named (filled) and the neutral prompt (hollow)
1.6	a value singled out in the caption
	the model did not run this arm (–: no reading)
Header bands follow the four questions: response state goal time integrity, setup	

A. Scene families

Every scene is a world $W = (\Omega, L, x_0)$ drawn on a vertical board: a ball released from rest at x_0 , a fixed law L , and white parts whose shape is the geometry Ω . The two main families vary Ω over matched levels while the prompt stays fixed within a family, so a level reaches the model through the first frame alone. Each intervention family changes one element of a base scene and keeps the rest. For each we state what is held fixed, what changes, what the stated law predicts and what the family isolates.

A.1. Common visual world

Common visual world setup



board	matte black, 1.20×0.675 m (16:9), on a wooden table before a wall
markers	four 30 mm squares per corner
parts	white acrylic
ball	40 mm, orange, green, blue or magenta, named in the prompt
camera	locked, 50 mm lens, 1.85 m from the board centre

A checker strip of four 30 mm squares in each corner gives the scale and a per-frame reference for camera motion. Frames are path-traced at 1280×720 and downsampled to 832×480 . Exact renders of the simulated motion exist at 24 and 30 fps and, for the model that generates at 16 fps, at exactly 16 fps. The first frame given to every model is the frame at $t = 0$ of the exact render. Board coordinates have their origin at the lower-left corner, x to the right and y up.

A.2. Track family

Track family response

The track is two laser-cut white acrylic plates 5 mm thick, and the ball rides on their sharp inner top edges, 24 mm apart, with the far plate 8 mm higher. Over $x \in [0, 1.0]$ m the upper edge follows

$$h(x) = H_0 - \bar{a}x + \sum_k A_k \sin(2\pi f_k x + \varphi_k) \quad (\text{A.1})$$

and ends at a stop block (levels in Table A.1, with random phases on the broadband level). The track starts at a local crest of h . With rolling resistance c_{rr} the energy balance gives

$$v^2(s) = v_0^2 - 2g_{\text{eff}} [h(x(s)) - h(x_0) + c_{rr} (x(s) - x_0)] \quad (\text{A.2})$$

(g_{eff} in Sec. B).

Table A.1. Track family levels, ten scenes each. f , A : frequency and amplitude of each component. $N_{\min}$: lowest normal force on the nominal trajectories, in units of mg . Last column: scenes in which all 64 ensemble members reach the stop (62–63 in the others).

level (first frame)	f (m^{-1})	A (mm)	$\bar{a}$	$N_{\min}$	64/64
straight	–	–	0.076–0.132	0.99	10
single-frequency	3	15	0.033–0.115	0.52	7
single-frequency	4	15	0.031–0.085	0.40	10
broadband	2, 3, 4	10 each	0.030–0.100	0.30	7

held fixed	The length, the release rule (the ball is released where the downhill slope first reaches 0.025, at least 10 mm past the crest), the prompt and the physical transit time (1.62–2.15 s). Every nominal trajectory keeps a normal force of at least 0.3 mg , which
------------	---

	bounds Af^2 and sets the physical bandwidth of the family, and a speed of at least 0.25 m/s beyond $x = 0.16$ m.
changed	The drawn profile. The mean slope $\bar{a}$ is drawn per scene and enters as a covariate, and the straight level needs a steeper $\bar{a}$ because it lacks the first downhill of a crest start.
law predicts	By Eq. (A.2) the speed follows the drawn height with unit gain and zero phase at every frequency of the family, and the mean slope accelerates the ball ($s \propto t^2$ on the straight level).
isolates	The transfer from drawn shape to speed, resolved by spatial frequency, at matched transit time.

A.3. Deflector family

Deflector family response, state

A launch ramp (a 25° descent of 0.15 m ending in a horizontal lip at $x = 0.647$ m) releases the ball into free flight toward K flat deflectors (120×10 mm, with 5 mm lips on the front and back that keep the ball in the board plane) and a white tray with 40 mm walls and a felt floor. Each deflector faces the launch ramp, so every contact throws the ball upward and flights between contacts last 0.2–0.4 s (levels in Table A.2).

Table A.2. Deflector family levels, eight scenes each. Order and tray: shares of the 64 ensemble members that keep the nominal sequence and that rest in the tray. Last: time of the last contact.

level	K	angles	order	tray	last (s)
single					
	1	20, 30, 45, 60°	1.00	0.86–1.00	0.54–0.57
regular					
	3	$+45/-45/+45^\circ$	0.94–1.00	0.81–0.98	1.12–1.16
irregular					
	3	$\pm(20-60)^\circ$	0.91–1.00	0.83–0.98	1.05–1.15

held fixed Each irregular layout shares launch ramp, tray position and total drop with the regular layout of the same index, with mean $|\theta|$ of 43.1 – 49.2° . Every nominal path hits each deflector at least 1.5 ball radii from its ends, flies at least 0.19 s between contacts (0.22 s between irregular deflectors, at least four frames on each side of a contact at 24 fps), makes its last contact before 4.5 s and rests in the tray before 5 s. The tray widens from 200 to 300 mm when fewer than 80% of the ensemble rest in it.

changed The deflectors. The two single-deflector scenes of an angle slide the deflector by up to 20 mm along its axis about a fixed contact point (190 mm below the lip, 250 mm at 20°), so they are physically identical. The regular layout is $+45/-45/+45^\circ$ at equal spacing (140 mm $\pm 15\%$). The irregular layout alternates in sign with magnitudes in $[20^\circ, 60^\circ]$ at random spacing.

law predicts The exit of each contact is set by the drawn normal and the prescribed restitution. On the single

deflector the physical exit angles at 20, 30, 45 and 60° are 111.2, 130.0, 162.1 and 193.1° , a slope of 2.06 (a mirror gives 2). On three deflectors the ball visits $p_1 \rightarrow p_2 \rightarrow p_3$ and rests in the tray.

isolates A discrete geometric event and the composition of three, and the paired layouts separate the physical path from a zigzag template with the same drop, start, tray and mean angle.

A.4. Local edits and shared-front tracks

Local edits state

Twelve tracks share their first 0.25 m and release point and then blend into a straight ramp of slope $\bar{a} \in \{0.03, 0.04, 0.05, 0.06\}$. An edit $\pm\epsilon(1 - u^2)^3$ ($\epsilon = 15$ mm, width 0.15 m) is centred at 0.42, 0.50 or 0.58 m.



Table A.3. Local edits: physical arrival shift at the stop, bump / dip (ms). The ball enters the edit at 0.99, 1.11–1.12 and 1.23–1.25 s for the three centres and reaches the base stop at 1.94, 1.89, 1.85 and 1.82 s for the four slopes. The speed at the edit centre is 0.62–0.74 (base), 0.45–0.61 (bump) and 0.75–0.85 m/s (dip), and 62–64 members reach the stop. *: also run on the closed models.

$\bar{a}$	centre 0.42 m	centre 0.50 m	centre 0.58 m
0.03	+45 / -14*	+41 / -13	+37 / -11
0.04	+41 / -13	+36 / -11	+32 / -10*
0.05	+38 / -12	+33 / -10	+29 / -8
0.06	+35 / -11	+30 / -9*	+25 / -7

held fixed The first 0.25 m and the release point of each track (an $f = 3 \text{ m}^{-1}$ crest start, $A = 15$ mm, $\bar{a} = 0.06$), the seed, and one ensemble per track whose members coincide until the edit.

changed One element per arm. Each track has five arms: base, bump ($+\epsilon$), dip ($-\epsilon$), a null edit (the same shape added to the lower edges of the plates, area 10.3 cm^2 , leaving the rolling edge and the simulated motion unchanged) and a rerun of the base with the same seed.

law predicts Identical motion up to the edit, $\Delta v^2 = \mp 2g_{\text{eff}}\epsilon = \mp 0.177 \text{ m}^2/\text{s}^2$ across it, and recovered speed after it with a persisting arrival shift (Table A.3).

isolates Whether a model changes speed at the edit in the physical direction, whether an edit ahead of the ball changes the motion before arrival, and, through the null edit and the rerun, how much an unrelated change of equal size and the sampler itself move the generation.

Shared-front tracks response, state



held fixed	The first 0.40 m of twelve $f = 3 \text{ m}^{-1}$ tracks ($A = 15 \text{ mm}$, $\bar{a} = 0.06$), all within the rules of the family.
changed	The back section, blended in over 0.40–0.50 m, with a random phase (3.61, 0.58, 0.32, 2.84, 5.29, 4.24, 5.47, 0.35, 1.39, 6.24, 3.09 and 1.51 rad for tracks 0–11).
law predicts	The motion is identical until the front section ends at 1.12 s, and the back-section undulation follows the drawn phase.
isolates	Whether the back-section response follows the drawn phase or continues the front, and whether the motion over the shared front changes with geometry the ball has not reached.

A.5. Ramp pairs, barriers and chutes

These families test energy and state, and each prediction holds at any tempo (Table A.4).

Ramp pair state



A 20° descent into a rounded valley and an ascent of 60, 30, 15, 5 or 0° , with release 45 mm above the valley and a stop block on each ascent at 2.52, 1.61, 1.46 and 1.32 times the release height.

held fixed	The descent, the valley and the release height.
changed	The ascent angle.
law predicts	The ball turns around just below its release height, lower on longer ascents through rolling resistance.
isolates	Whether the climb is bounded by the release height.

Barrier state



The same descent into a flat valley and a hump $(1 - u^2)^3$ 220 mm wide whose height is 0.70–1.20 of the 80 mm release drop.

held fixed	The descent, the valley, the hump width and the 80 mm release drop.
changed	The height of the hump, 0.70–1.20 of the release drop.
law predicts	The ball passes below the rolling-resistance threshold 0.971 ($1 - c_{rr}L/h_{rel}$ for the 0.46 m path to the crest) with a crest speed that falls to zero at the threshold, and the ensemble splits at 0.95 and 0.971 by design.
isolates	Whether the outcome has a threshold at the release height.

Chute state



A chute set holds the exit point (0.55, 0.45) m, the exit tangent (-20°) and the exit speed (0.900 m/s) fixed and changes only the curvature κ of the last 60 mm.

held fixed	The exit point, the exit tangent and the exit speed.
changed	The curvature κ of the last 60 mm.
law predicts	The four flights are one parabola, and their nominal landing points agree to 10^{-10} m.
isolates	Whether a drawn shape that the stated law ignores moves the outcome.

A.6. Goal interventions

Goal interventions: trays and prompt pairs goal

held fixed	The neutral prompt of the tray interventions (Sec. E.2), which names neither deflectors nor tray. Scenes built on a main-set scene keep its colour and ensemble members, so the two first frames differ only in the changed part. The wooden table enters the simulator because the ball can miss the tray (Table A.5).
changed	The tray, a second tray or the prompt. <i>Two trays</i> place a second, identical tray under the launch lip. The three-deflector layouts leave no room for it, so all four scenes use single deflectors. <i>No tray</i> removes the tray. <i>Prompt pairs</i> regenerate the 16 regular and irregular main-set scenes with the neutral prompt, with first frame, ensemble and seed unchanged.
law predicts	In the <i>shifted-tray</i> scenes the stated law gives a capture gain of zero: the landing point does not move with the tray, and in the beyond group the ball flies over the moved tray to the table. The stated law with its nominal parameters ($\mu = 0$) lands at the left edge of a one-sided ensemble spread (+9 to +145 mm at 20° with median +107, -11 to +128 at 30° , -23 to +70 at 45° and -42 to +16 at 60°), so a model that lands at the ensemble median sits up to 107 mm off the nominal point, and the primary reading compares two tray positions of the same base and seed (Sec. C.2).
isolates	Whether a drawn tray attracts the ball, which of two trays receives it, whether a model keeps a stopping place that is no longer drawn, and whether the destination follows the text.

A.7. State controls

State controls: near misses and permuted layouts state

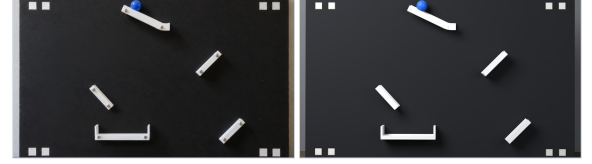
held fixed	The neutral prompt (Sec. E.2), the free-flight path in the near misses, and the three deflector magnitudes in a permuted pair.
------------	--

Table A.4. Ramp pairs, barriers and chutes. Times are physical, counts are ensemble members out of 64, and * marks scenes also run on the closed models.

ramp pairs				barriers				chutes				
ascent	turn (s)	ratio	travel (mm)	ratio	outcome	time (s)	speed (m/s)	pass	κ (m ⁻¹)	exit, land (s)	$N_{\min}$	landing (m)
60°*	0.54	0.980	35 (33–37)	0.70	pass	0.78	0.51 (0.53)	64	0*	0.31, 0.53	0.82	0.714–0.744
30°	0.64	0.975	92 (84–96)	0.85*	pass	0.83	0.34 (0.35)	64	+4	0.29, 0.51	0.82	0.714–0.742
15°*	0.87	0.965	183 (162–193)	0.95	pass	0.90	0.14 (0.15)	27	−4*	0.34, 0.56	0.61	0.714–0.741
5°	1.74	0.929	509 (411–572)	0.971	pass	1.03	0.03 (0.03)	10	+8*	0.26, 0.48	0.49	0.715–0.741
0°*	stop 1.36	–	–	1.05*	turn back	0.86	–	0				
				1.20*	turn back	0.83	–	0				

Ramps: turn-around time, height ratio and horizontal travel up the ascent (ensemble range in brackets). Barriers: time at the crest (or turn-around), crest speed with its ratio to the speed at the foot (reached at 0.61 s). Chutes: $N_{\min}$ in units of mg , nominal landing 0.7377 m for all four.

changed	A deflector near the path or the order of the deflectors. <i>Near misses</i> place a deflector end 10 mm or 40 mm from the free-flight path, with a hit control. <i>Permuted layouts</i> use the same three magnitudes in the orders $(+a, -b, +c)$ and $(+a, -c, +b)$, each with its tray at its own landing point.
law predicts	No contact at the near misses and a contact at the hit control, and on each permuted layout the order in which its deflectors are drawn (Table A.5).
isolates	The radius within which a deflector attracts a contact (zero under the stated law), and whether the contact order follows the drawn order.



photograph rendered twin

held fixed	The geometry. It is measured from the photograph (the laser-cut end blocks set the scale, the corner strips the board homography, and the white parts are fitted in board coordinates, Table A.6), and the simulator runs its ensemble on it. The prompt is the main-set prompt. The control of an open-weight model is a rendered twin (measured geometry, the photograph's camera, our appearance) wherever the as-built motion leaves the nominal band, and the main-set render where it stays inside. Closed models run only the photograph rows and are compared with their own nominal-scene readings, a weaker control.
changed	The visual domain of the first frame. Each photograph is cropped to 16:9 and scaled to 1280×720 as the first frame. The build departs from the drawing: the board is 1166×723 mm, the checker squares are 27.6–29.2 mm and the balls 46–50 mm.
law predicts	The as-built ensemble of Table A.6.
isolates	The visual domain of the first frame with the geometry held fixed.

A.8. Clip length

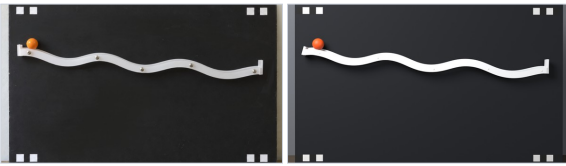
Clip length time

held fixed	The first frame, the prompt and the seed of eight scenes (one per track level, the 30° and 60° single deflectors, and a regular layout with its paired irregular layout). The 5 s clips are the main-set videos.
changed	The length of the clip: 3, 5 and 8 s.
law predicts	The stated law keeps its events at fixed times, so its tempo is independent of the clip length.
isolates	Whether timing is set by the motion or by the clip.

A.9. Photographed first frames

Photographed first frames integrity, visual domain

The same eight scenes were built from laser-cut parts and photographed from a tripod with the camera parallel to the board within 0.7°.



A.10. Regeneration sources

Regeneration sources probe

Six scenes of a twelve-scene pilot set serve as sources: three tracks with $A = 20$ mm ($f = 1$ and twice $f = 3$ m⁻¹) and three deflector layouts (one regular, two irregular), each with its exact render.



Table A.5. Tray interventions, near misses, permuted layouts and prompt pairs, all with the neutral prompt. Shifts are relative to the nominal landing point, positive toward the incoming ball. Counts are ensemble members out of 64. *: also run on the closed models.

first frame	scene	base	change	physical outcome (ensemble)
	shifted tray, in: 20, 30°, 45, 60°	single deflector	240 mm tray at +60 mm	lands in the tray off-centre (64 each)
	shifted tray, beyond: 20, 30°, 45, 60°	single deflector	100 mm tray at +240, +220, +160, +120 mm	flies over it to the table (64), touches it (0), leaves the view
	two trays: 45° (2 scenes, 1*)	single deflector	second 200 mm tray under the lip	physical tray (63; 64), second tray (0)
	two trays: 60° (2 scenes, 1*)	single deflector	second tray, both 140 mm	physical tray (64; 64), second tray (0)
	no tray: 30° and irregular	main-set scenes	tray removed	lands on the table and leaves the view (64; 63 within 5 s)
	near miss: 10 mm above*, 40 mm above, 10 mm below	free flight	deflector end at the stated clearance	no contact (0); smallest clearance 4.3, 35.7, 6.9 mm
	near miss: hit control*	free flight	30° deflector on the path	contact, then tray (64)
	permuted pairs (3)	new irregular	magnitudes 38/46/54, 40/47/55, 37/47/52°, two orders	order kept 58–63, tray 53–63
	prompt pairs (16; 4*)	regular, irregular	neutral prompt	as in the main set

Table A.7. Regeneration sources: how each version is built and the reading that identifies it, on the source and for the physical oracle.

version	construction	source / oracle
constant speed	tracks: uniform speed over the physical transit time	$ H $ 0.02 / 1.0
phase +90°	tracks: undulation of $v^2(s)$ advanced by a quarter period	phase +86 / 0°
slowed 2×	both: the simulated motion at half speed	α 2.0 / 1.0
restitution 0.3	deflectors: every contact with $e = 0.3$	e_n 0.30 / 0.85
transparent	deflectors: the ball passes through deflector 2	order 0 / 1

held fixed	The drawn structure. Sources are rendered at 24 fps and at exactly 16 fps, with a first frame equal to the exact render's within one grey level.
changed	The dynamics. Each scene has three versions that keep the drawn structure and break the dynamics (Table A.7). A deflector ball that misses the tray lands on the table and rolls to a stop, so no version passes through a drawn surface other than the transparent deflector.
law predicts	The physical readings of Table A.7.
isolates	Which parts of a motion a model keeps when it re-renders it and whether a replacement moves toward the stated law or toward the model's own version of the scene.

B. The reference

The physical reference is a simulator that integrates the stated law on the geometry drawn in the first frame, and a 64-member ensemble that perturbs what the drawing leaves open gives each scene a tolerance band. The reference is exact for the drawn geometry and the stated law. All physical comparisons in the paper are defined relative to this oracle.

The stated law. On tracks, ramps and chutes the ball (radius $r = 20$ mm) rolls without slipping on the two inner plate edges, whose chord lies $r_{\text{eff}} = 15.5$ mm from the ball centre. A solid ball ($k = 2/5$) then carries the inertia factor $1 + \chi = 1 + kr^2/r_{\text{eff}}^2 = 5/3$ and accelerates with $g_{\text{eff}} = g/(1 + \chi)$. Rolling resistance is a force $c_{\text{rr}} mg \cos \theta$ with $c_{\text{rr}} = 0.005$, the ball stops at the end block, and every nominal trajectory keeps a positive normal force. Between contacts the ball flies freely, and a deflector contact reverses and scales the normal velocity by $e = 0.85$ and keeps the tangential one ($\mu = 0$). The tray has a felt floor ($e = 0.25$) and 40 mm walls, and in the tray interventions the table below the board has restitution 0.60, friction 0.30 and rolling resistance 0.010.

The ensemble. Each scene draws 64 members from a key of its own (Table B.1), perturbing what a first frame cannot show: the material of the ball, restitution and friction (with

Table A.6. Photographed first frames: as-built geometry against the design (track offset in x/y and rotation, launch exit speed with design 0.760 m/s), the as-built ensemble (on tracks all 64 members reach the stop, on deflectors the shares keeping the design order and resting in the tray), the nominal band, and the rendered control.

photograph	scene	as-built against design	ensemble	nominal band	control
	straight	$-16.3/+52.3$ mm, $+0.07^\circ$; $\bar{a}$ 0.093 $\rightarrow$ 0.092	stable	outside, 4% faster	twin
	$f = 3$	$-14.9/+42.5$ mm, -1.36° ; $\bar{a}$ 0.055 $\rightarrow$ 0.076; A 15 $\rightarrow$ 17.5 mm	stable	outside, 19% faster	twin
	$f = 4$	$-10.8/+28.6$ mm, -1.03° ; $\bar{a}$ 0.048 $\rightarrow$ 0.077; A 15 $\rightarrow$ 14.7 mm	stable	outside 20% of the time	twin
	broadband	$-13.8/+47.7$ mm, -0.62° ; $\bar{a}$ 0.069 $\rightarrow$ 0.078; A 10.6/11.5/10.2 mm	stable	outside, 15% faster	twin
	single, 30°	deflector 31.7° , 149×22 mm; exit 0.815 m/s	1.00 / 1.00	inside	main-set render
	single, 60°	deflector 61.1° , 149×23 mm; exit 0.872 m/s	1.00 / 1.00	contact 17 ms late	twin
	regular	deflectors $+44.8/-47.1/+44.8^\circ$; exit 0.839 m/s	0.95 / 0.95	contacts 1, 3 early	twin
	irregular	deflectors $+39.1/-54.7/+36.4^\circ$ (design $+38.8/-55.5/+36.2$); exit 0.764 m/s	0.95 / 1.00	contact 1 5 ms late	twin

Table B.1. The stated law (nominal) and the ensemble (64 members). $\mathcal{U}$: uniform, $\mathcal{N}$: normal. The inertia factor is that of a solid (2/5) or a hollow (2/3) ball. The bars show each uniform range (band) with the nominal value (tick) on a scale from 0 to the range's upper end.

family	quantity	nominal	ensemble
rolling	inertia factor k	2/5	{2/5, 2/3}
	rolling resistance c_{rr}	0.005	$\mathcal{U}[0, 0.02]$
	release point	release rule	$+\mathcal{N}(0, 3 \text{ mm})$
deflectors	restitution e	0.85	$\mathcal{U}[0.80, 0.90]$
	friction μ	0	$\mathcal{U}[0, 0.3]$
	release point	launch ramp	$+\mathcal{N}(0, 3 \text{ mm})$
table	restitution	0.60	$\mathcal{U}[0.5, 0.7]$
	friction	0.30	$\mathcal{U}[0.2, 0.4]$
	rolling resistance	0.010	$\mathcal{U}[0.005, 0.02]$

spin), rolling resistance and a few millimetres of the release point. The arms of an intervention share their key, so the arms of a local edit coincide until the edit and a tray intervention reuses the members of its base: differences between arms come from the changed geometry alone.

How the band is used. Smooth tracks and flat contacts make each exit depend on the drawn normal and the incoming state, so the families are irregular in layout and regular in dynamics, and a scene enters the study only when its ensemble stays together: on tracks every member reaches

the stop (up to two stop at the release point), on deflectors at least 90% keep the nominal sequence and 80% rest in the tray. The physical value of every reading is its reading on the exact render of the nominal trajectory, measured through the same chain as the videos, and the tables give the ensemble spread where it matters (ramp heights, landing ranges, barrier passes). The barrier heights 0.95 and 0.971 split the ensemble (27 and 10 of 64 pass), so there a model is read against a share. The simulator also runs from states measured in a video: the destination counterfactual rolls the stated law forward from a model's own exit state, and geometric forgiveness runs the ensemble with one deflector removed. For photographed first frames the ensemble runs on the measured geometry and decides whether the as-built scene is stable and whether it stays in the nominal band (Table A.6).

C. Measurement

Every video is reduced to the trajectory of the ball in board coordinates, and every reading is a function of that trajectory and the drawn geometry. This part defines the readings, shows that the chain returns the physical values on exact renders and names each error injected into motions built to be wrong, and gives the floors against which model readings are compared.

C.1. From pixels to the board

The eight squares of the corner strips give a per-frame homography from the board to the image, which absorbs small camera motion. Letterboxing is undone by the frame map of each model’s canvas. The ball is segmented inside the board with a hue window for its colour (orange 355–40°, green 105–157°, blue 200–244°, magenta 300–345°) and located by its chroma-weighted centroid. White parts are bright pixels with saturation below 0.25, so the ball or its blur never counts as a part. The ball centre lies 27.9 mm in front of the board, and its plane is mapped by scaling the board homography about the principal point. On a track the centre is projected onto the offset centreline to give the arc length $s(t)$, differentiated by a Savitzky–Golay filter (seven frames at 24 fps), and on deflectors the planar path (x, y) is used. The drawn geometry is read from the white mask of every frame (the upper track edge refitted at the input frequencies, the axis and centroid of each deflector), minus the same extraction on the exact render.

C.2. Readings

Table C.1 summarizes the readings of the main paper.

Track readings. The tempo α is the video time from first motion to a common arc length over the physical time. On deflectors the launch tempo times the ball from first motion to leaving the launch ramp. For the gain, v^2 is regressed jointly on $[1, s - \bar{s}, (s - \bar{s})^2]$ and, per drawn component, on $A_k \sin \psi_k$ and $A_k \cos \psi_k$ with $\psi_k = 2\pi f_k x(s) + \varphi_k$, and the coefficient pair $z_k = c_k + i d_k$ gives

$$H_k = z_k / z_k^{\text{ref}}, \quad (\text{C.1})$$

where the reference is the simulated motion rescaled to the video’s tempo, aligned at first motion and read through the same filter. $|H|$ is the gain and $\arg H$ the phase, positive when the speed undulation leads the drawn height. The stated law and any uniformly slowed physical motion give $|H| = 1$ and phase 0. A component is read when the video traverses at least 1.5 of its cycles with 60% of the frames in motion, and a $2f$ term enters when its frequency is not another component (intervals in Sec. H). The trend ratio compares two gravities, $g_{\text{trend}} = g_{\text{eff}} T / T_{\text{ref}}$ from the mean slope T of the fitted v^2 trend and $g_{\text{time}} = g_{\text{eff}} / \alpha^2$ from the tempo. It is 1 for a time-rescaled physical motion and near 0 for a ball held at a preferred speed. On the straight level $s = s_0 + c(t - t_0)^p$ gives the exponent p (2 under the stated law, 1 at constant speed). The relaxation rate comes from an output-error fit of $s(t)$ itself, with $s = s_0$ before t_0 and

$$\ddot{s} = a_c - g_w W(s) - b \dot{s} \quad (\text{C.2})$$

afterwards (W the undulating part of $\sin \theta$, fourth-order Runge–Kutta at 1/240 s). The stated law has $b = 0$ and

$g_w = g_{\text{eff}}$. The balanced accuracy is the share of moving frames on which the tangential acceleration has the sign of the local downhill direction, averaged over downhill and uphill frames ($|a| < 0.06 \text{ m/s}^2$ counts one half). The kinematic gain regresses the measured centre’s vertical offset about its trend on the exact one, and it is 1 when the ball stays on the drawn rails.

Deflector readings. A contact is a minimum of the ball-to-deflector distance within $r + 8 \text{ mm}$ where a bounce model (two parabolas sharing contact point and vertical acceleration) fits at least twice as well as one parabola, with windows bounded by the neighbouring events. The contact order merges adjacent repeats and is compared with the physical sequence. The residual is the angle between the measured exit and the stated law’s exit for the measured incoming direction, and $e_n \in [0, 1.2]$ and $e_t \in [-1, 1]$ are fitted along the deflector normal and tangent (0.85 and 1 under the stated law). The calibrated law fits (e_n, μ) on a model’s single-deflector contacts, freezes it and predicts its three-deflector contacts, which helps only a model with a wrong but consistent contact law. A ball has *arrived* when it is visible in half of the last 0.3 s and its median centre lies in the tray (between the inner walls, from 10 mm below the floor to 20 mm above the walls). For a wrong route, the destination counterfactual rolls the stated law forward for 5 s from the exit state after the last correct contact (or the launch exit). The difference between the model’s arrival and that of the rollouts is the steering beyond the ball’s own state, and it is tested with the state two frames later and with the velocity rescaled by the apparent gravity or by the launch tempo.

Intervention readings. *Local edits:* with the anchor t_a at 0.10 m of travel, $T = t_{\text{end}} - t_a$, $T_{\text{pre}} = t_{\text{in}} - t_a$, v_{pulse} the mean speed in the edit and v_{pre} that from 50 mm after release to 30 mm before the edit, each arm is paired with the base of the same track and seed,

$$dv = \frac{(v_{\text{pulse}}/v_{\text{pre}})_{\text{arm}}}{(v_{\text{pulse}}/v_{\text{pre}})_{\text{base}}} - 1, \quad dt = \frac{(T/T_{\text{pre}})_{\text{arm}}}{(T/T_{\text{pre}})_{\text{base}}} - 1, \quad (\text{C.3})$$

so that the tempo of the two videos cancels. Locality is read from the change of T_{pre} and the RMS path difference before the edit. *Shared fronts:* the gain is read on the back section ($x \geq 0.52 \text{ m}$) with its own phase φ_b , and the circular-linear slope of $\Phi = \delta + \varphi_b$ on φ_b is 1 when the response follows the drawing and 0 when it continues the front. An exploratory fit $|H|e^{i\Phi} = G e^{i\varphi_b} + C$ across tracks splits a locked part $|G|$ with lead $\arg G$ from a part $|C|$ with the same phase on every track. *Ramps:* the turn-around is the highest centre after the valley (quadratic fit within $\pm 0.12 \text{ s}$), and the height ratio divides its height above the valley by

that of the release. Climbs are classed as turning back, reaching the top stop, still climbing or stopped. *Barriers*: a pass moves the centre 20 mm beyond the crest, and the crest-speed ratio divides the crest speed by the speed at the foot. *Chutes and trays*: the landing is the x where the centre first descends through 50 mm above the tray rest height (720 mm on chutes), and the chute exit direction comes from a parabola fitted to at most 12 flight frames. The capture contrast pairs the in-tray and beyond scenes of one base and seed, $(x_{\text{beyond}} - x_{\text{in}})/(\text{shift}_{\text{beyond}} - \text{shift}_{\text{in}})$, 0 when the landing ignores the tray and 1 when it follows it, so that the model's own flight bias cancels. *Clip length*: γ is the slope of $\log \alpha$ on $\log T_{\text{clip}}$ within each scene and seed (0 for a fixed physical rate, 1 when the event is stretched to the clip). *Regeneration*: with r the regenerated motion, s the source and p the physical motion, summed over visible frames,

$$\lambda = \frac{\sum \langle r - s, p - s \rangle}{\sum \|p - s\|^2}, \quad (\text{C.4})$$

0 when the output stays on the source and 1 when it is pulled onto physics. The copy ratio is $\text{RMS}\|r - s\|/\|s - p\|$, κ is the same projection onto the model's own image-to-video video q , and $r - s = \lambda_2(p - s) + \kappa_2(q - s) + \epsilon$ separates the two directions.

C.3. Validation

Exact renders. Exact renders go through the full chain as if they were generated videos. On the 120 track renders the fiducials agree with the scene camera to 0.04–0.07 px, every ball stays on the track and reaches the stop, the tempo reads 1.000–1.004 and the gain 1.02–1.10 (a pixel-level bias of at most +9%). On the 72 deflector renders every sequence is correct with residuals below 0.32° , and pilot renders read $e_n = 0.848\text{--}0.852$ on all 42 contacts. The drawn-geometry extraction recovers the track amplitude to 1.2×10^{-4} , its phase to 0.017° and deflector angles to 0.004° . Table C.7 gives the intervention readings on exact renders.

Synthetic mechanisms. Seven candidate dynamics, simulated on tracks and read through the same estimator, leave distinct fingerprints (Table C.2). Spatial smoothing lowers the gain at zero phase, temporal smoothing lowers it with a phase lead, and a response lag produces a phase lag that grows with frequency. A constant-speed cruise flattens the trend and its b is recovered exactly, a speed that follows the local slope reads $+90^\circ$ at chance accuracy, and a curvature heuristic produces a large $2f$ term. A single real gain would misread three of them (lag as bandwidth, first order as no response, cruise as an unbounded ratio), which is why gain, phase, trend ratio and b are read together.

Motions with a known error. Motions that keep the drawn structure and break one piece of dynamics show that

each reading names the error it was built with (Table C.3, where the reading that names each error is tinted). Constant speed reads as no response, a phase advanced by 90° as a phase error of $+85^\circ$ at full amplitude, and a motion slowed $2\times$ keeps unit gain and zero phase with $\alpha = 1.98$. Rendered and measured, they read gains of 0.024, 1.52 and 1.03 and phases of $+4$, $+86$ and $+1^\circ$. On deflectors the calibrated law recovers a fixed restitution of 0.5 ($e_n = 0.49$) and then fits the three-deflector contacts to $1\text{--}2^\circ$ against $15\text{--}16^\circ$ for the stated law, and a transparent second deflector breaks every sequence. Rendered and measured, a restitution of 0.3 reads 0.300–0.304, and a motion slowed $2\times$ keeps the order with $e_n = 0.848\text{--}0.853$ and a contact gravity of $2.45\text{--}2.48 \text{ m/s}^2$, a quarter of g .

Integrity rule. On 208 exact renders mapped onto every model's canvas and frame rate the integrity rule (Sec. D) raises no flag, with part displacements up to 0.15 mm and rotations up to 0.20° . Injected perturbations are read as intended (Table C.4).

Photographed first frames. The rendered twins measured with the photograph's camera set the floor of the photograph rows: with the correction for the ball's height above the board, tracking errors are 1.3–1.4 mm on tracks (2.4 mm on the broadband scene, whose ball is the largest) and 2.1–2.2 mm on deflectors, the gain is biased by $+3$ to $+5\%$ ($+10\%$ at most on the broadband scene), the phase by at most 2.5° , and every order is correct, as on the nominal scenes. Pose-level floors match the nominal ones except on two track scenes at 16 fps (gain floor up to 0.31, phase floor up to 25°).

C.4. Floors

A floor is the spread of a reading on the simulated motion with 1 px Gaussian detection noise (three noise seeds) at each native frame rate and output size, and a model reading is interpreted only when its interval clears the floor at the model's rate and size (Tables C.5 and C.6). For the interventions the simulated motion is first slowed $2.5\times$, a typical model tempo (Table C.7). For clip-length pairs the floor of the 3 s minus 5 s difference includes the truncation of the short clip (0.08 for α , 0.15 for $|H|$, 5° for phase, 0.07 s^{-1} for b , 0.08 for p and for accuracy), and the 8 s minus 5 s differences on the simulated motion are zero.

D. Integrity

A generated video is read only while the scene stays the one that the first frame shows. This part states the rule, reports the intact and readable shares of every model in every experiment, and describes the failure forms that recur, which

Table C.1. Readings used in the main paper, as definition cards: symbol and name, definition, the physical oracle and the floor at the three native frame rates (16 / 24 / 30 fps).

$ H $ gain undulation of $v^2(s)$ against the simulated motion at the same tempo oracle 1 floor 0.09 / 0.11 / 0.06	$\arg H$ phase lead of the speed undulation over the drawn height oracle 0° floor 6.8 / 7.5 / 6.0°	$g_{\text{trend}}/g_{\text{time}}$ trend ratio mean-slope acceleration over that implied by the tempo oracle 1 floor –	b relaxation rate rate of return to a preferred speed, Eq. (C.2) oracle 0 floor 0.09 / 0.06 / 0.05 s ⁻¹
α tempo generated over physical time of the same motion oracle 1 floor 0.08 (clip-length pairs)	p straight-ramp exponent $s \propto (t - t_0)^p$ on the straight level oracle 2 floor 1.96–2.02	e_n restitution apparent normal restitution of a contact oracle 0.85 (band 0.80–0.90) floor 0.81–0.89 / 0.68–0.91 / 0.79–0.90	contact order merged contact sequence equals the physical one oracle 100% floor 100%
residual exit against the stated law's exit for the measured incoming state oracle 0° floor p90 4.2 / 3.2 / 2.3°	h/h_0 height ratio highest point after the valley over the release height oracle 0.93–0.98 floor 0.006, single video 0.055–0.091 (tempo 2.5)	tray arrival ball at rest inside the tray over the last 0.3 s oracle 100% floor exact on renders	γ duration exponent slope of log α on log clip length oracle 0 floor from the α floor
λ regeneration pull projection of $r - s$ on $p - s$, Eq. (C.4) oracle 1 (0 on the source) floor distance floor 1.6–2.3 mm	The floor of a reading is its spread on the simulated motion with 1 px detection noise at 832×480 for the three native frame rates (95th percentile of the absolute error, the 90th percentile for the residual, or the stated range). Track floors are those of the single-frequency level with $f = 3 \text{ m}^{-1}$ (all levels in Table C.5). The prescribed restitution 0.85 has the ensemble band 0.80–0.90. The height-ratio floor is that of the mean over videos, followed by the single-video floor at a model tempo of 2.5.		

Table C.2. Fingerprints of synthetic mechanisms on tracks with $A = 20 \text{ mm}$ and $\bar{a} = 0.06$ at $f = 2, 3, 4, 5 \text{ m}^{-1}$ (values listed in that order and drawn as a line from $f = 2$ to 5, with the physical value dashed: gain on 0 to 2.1, phase on -80 to $+100^\circ$), 24 fps, 1280×720 , 1 px noise, every motion slowed $2\times$ and started 1 s late. Accuracy at $f = 3$, medians of three seeds.

mechanism	gain $ H $	phase ($^\circ$)	$2f/1f$	trend ratio	$b (\text{s}^{-1}) / \text{accuracy}$
★ Physical oracle	0.98, 0.99, 0.98, 0.95	-1, 0, +2, -1	0.03, 0.06, 0.11, 0.09	0.99, 0.99, 0.99, 0.98	0.0 / 0.92
spatial smoothing, $\sigma = 40 \text{ mm}$	0.90, 0.83, 0.72, 0.59	-1, +1, -2, +4	0.03, 0.05, 0.08, 0.10	1.01, 1.14, 1.23, 1.32	0.0 / 0.93
temporal smoothing, $\sigma = 0.2 \text{ s}$	0.91, 0.70, 0.47, 0.28	+7, +19, +30, +34	0.18, 0.18, 0.20, 0.25	0.68, 0.69, 0.73, 0.74	0.3–0.5 / 0.84
response lag 0.1 s	0.98, 0.87, 0.74, 0.59	-26, -40, -56, -72	0.09, 0.11, 0.17, 0.24	1.11, 1.16, 1.15, 1.08	-0.1–0.6 / 0.80
cruise, $b = 5 \text{ s}^{-1}$	0.71, 0.88, 0.97, 0.97	+53, +39, +30, +26	0.05, 0.06, 0.04, 0.10	0.10, 0.13, 0.15, 0.15	5.0–5.1 / 0.78
first order (speed from slope)	0.90, 1.25, 1.65, 2.00	+90, +90, +91, +92	0.05, 0.06, 0.10, 0.10	0.01, 0.01, 0.00, 0.00	≥ 80 / 0.52
curvature slowdown, $v/(1 + c \kappa)$	0.87, 0.72, 0.66, 0.56	-2, -5, +4, +9	0.31, 0.88, 1.34, 2.16	0.99, 1.09, 1.01, 1.20	0.0 / 0.66

are findings in their own right about how each model treats the drawn apparatus.

D.1. The rule

Intact and readable integrity

A video is *intact* when five flags stay clear.

count a second blob of the ball's colour with a quarter of its area for more than three frames, or a resting blob of 1.6 times its area and aspect ratio 1.4 (two touching balls)

Table C.3. Track readings on motions with a known error ($f = 3 \text{ m}^{-1}$ and straight tracks, 24 fps, 832×480 , 1 px noise, median of three seeds). Tinted: the reading that names the error.

motion	$ H $	phase	α	acc.	b	p
* Physical oracle	1.01	0°	1.00	0.99	0.03	2.00
constant speed	0.02	-27°	1.08	0.50	23.2	1.00
phase +90°	1.51	+85°	1.26	0.54	-0.62	—
slowed 2×	0.96	+1°	1.98	0.90	0.00	1.99

Table C.4. Injected perturbations and how the integrity rule reads them: ✓ passes, ✗ flagged.

passes	flagged
✓ a deflector shifted by 3 mm	✗ a deflector shifted by 8 mm
✓ a deflector turned by 1.5°	✗ a deflector turned by 3°
✓ a board drifting by 2–6 px	✗ a deflector area grown by 60%
✓ a darkened and blurred deflector	✗ a removed launch ramp
✓ a ball enlarged by 30% in area	✗ a rail plate moved by 8 mm
	✗ half an outline moved by 6 mm
	✗ a white blob of 0.6 ball areas
	✗ a second ball

Table C.5. Track floors at 832×480 : 95th percentiles of $||H| - 1|$, $|phase|$, $2f/1f$, $|b|$ and the kinematic-gain error, and the lowest balanced accuracy. On the straight level p reads 1.96–2.02 and $|b| \leq 0.06 \text{ s}^{-1}$.

level	fps	$ H $	phase	$2f/1f$	$ b $	kin.	acc.
$f = 3$	16	0.09	6.8°	0.21	0.09	0.049	0.81
	24	0.11	7.5°	0.15	0.06	0.027	0.86
	30	0.06	6.0°	0.13	0.05	0.059	0.86
$f = 4$	16	0.08	4.5°	0.16	0.09	0.038	0.93
	24	0.07	5.6°	0.15	0.08	0.026	0.90
	30	0.04	4.1°	0.10	0.05	0.032	0.94
broadband	16	0.23	14.1°	0.23	0.18	0.028	0.84
	24	0.17	8.8°	0.15	0.13	0.040	0.88
	30	0.15	9.8°	0.19	0.07	0.036	0.79

Table C.6. Deflector floors on the 24 main-set layouts (20 noise seeds per scene): exact order, largest residual, largest per-scene 90th percentile of the residual, and the range of e_n (stated law 0.85).

fps	size	order	residual max	p90 max	e_n
16	1280×720	100%	4.8°	2.9°	0.83–0.88
	832×480	100%	8.6°	4.2°	0.81–0.89
24	1280×720	100%	2.8°	2.1°	0.82–0.89
	832×480	100%	4.3°	3.2°	0.68–0.91
30	1280×720	100%	3.0°	1.5°	0.81–0.88
	832×480	100%	4.4°	2.3°	0.79–0.90

missing	the ball not found for more than 12 frames, a ball leaving the view being an outcome
apparatus	a part (each deflector, the tray, the launch ramp, each rail plate), tracked as a rigid body in board coordinates, that for six or more frames shifts by more than a quarter of the ball radius (5 mm), turns by more than 2°, changes its white area outside $[0.6, 1.4]$, moves its outline by more than 5 mm or disappears
extra	a new white region of half the ball’s area for more than six frames
fiducial drift	a board moving by more than 8 px at 832×480 , or found in fewer than 90% of the frames

The apparatus thresholds sit at the edge of what a reading can see. A 2° rotation turns the exit by 4°, inside the 4–9° residual floor, a 5 mm shift keeps the normal and the 8 mm contact window, and 40% of area is the light on lips and bevels.

A video is *readable* when every flag but apparatus is clear and every apparatus change starts after the ball’s last use of that part (the launch ramp after the ball leaves it, a struck deflector after its last contact, the tray and unstruck deflectors after the landing, the rails after the last measured point). The intact share is the integrity statistic, and dynamics are read on readable videos, with denominators in the order all, intact, readable, trackable (ball visible in half of the frames), reached (the ball gets to where the reading is taken) and estimable.

D.2. Shares per model and experiment

Six models keep 96–100% of their track videos intact (Table D.1), and the deflector scenes, whose parts are struck and whose tray receives the ball, cost every model more. LTX-2.3 keeps the tracks and rarely the deflector scenes, and LTX-2.5 moves its camera in 64 of 72 deflector videos, which are therefore reported here only. A tenth model, Wan 2.2 (TI2V-5B), generated on the two main families, keeps too few videos intact for any reading and appears only in this table.

D.3. Recurring failure forms

Table D.2 lists the forms that recur, and Fig. G.5 shows examples. Struck parts separate the models: the Seedance models let a struck deflector turn, LTX-2.3 and MiniMax-H3 let the tray tilt after the landing, LTX-2.5 (distilled) rotates the launch ramp, mostly after the ball has left it, and LTX-2.3 treats deflectors as loose objects in every visual domain. Other forms belong to the camera or the object count (LTX-2.5 dollies in, Cosmos 3 Nano duplicates the ball), and some grow with the clip (LTX-2.3 changes more parts in longer clips, and Wan 2.2 (A14B) lets the ball vanish after it has arrived).

E. Models and generation settings

Nine image-to-video models generate the videos: six open-weight models run locally and three closed models are called through APIs. This part gives what a reader needs

Table C.7. Intervention readings on exact renders (four frame-rate and size combinations) and their floors at a model-like tempo (simulated motion slowed 2.5 $\times$, 1 px noise, 16 / 24 / 30 fps): median over 36 pairs, and the 95th percentile (p95) of the error of a single pair or video.

reading	simulator	exact renders	at tempo 2.5
local edit, bump: dv / dt	-11.4 / +2.8%	-10.9 to -11.8 / +2.7 to +3.0%	-11.1 to -11.4 / +2.6 to +3.0%; p95 2.4-3.9 / 1.1-1.5%
local edit, dip: dv / dt	+7.8 / -0.8%	+7.3 to +8.1 / -0.6 to -1.3%	+7.5 to +8.2 / -0.5 to -1.1%; p95 2.8-5.6 / 1.3-1.6%
local edit: dt of the bump over its simulated value	1	1.07-1.10	0.94-1.06
null edit minus base: dv / dt	0	$\leq 0.08 / \leq 0.03\%$	p95 3.0-4.6 / 1.0-1.6%; flagged in 0 of 120 runs
path RMS before the edit	0	p95 0.07-0.09 mm	p95 4.2-4.8 mm
shared front: phase slope; front dispersion	1; 0	1.00; 0.03-0.05 mm	0.92-1.02 (single-video phase p95 19-32 $^\circ$); 1.6-1.7 mm
ramp pair: mean height ratio	0.962	0.959-0.961	single-video p95 0.055-0.091
barrier: correct outcomes; 50% crossing	6 of 6	6 of 6; 1.010	18 of 18
chute: landing offset; slope per unit κ	0; 0	-0.4 to -0.5 mm; -0.004 to +0.008 mm	landing p95 1.9-2.9 mm; exit angle p95 3.6-4.0 $^\circ$
regeneration: λ of the physical motion; of the source	1; 0	0.998-1.004; -0.002 to +0.004	distance floor 1.6-2.3 mm
shifted tray: capture contrast	0	-0.001 to +0.001	-
two trays: physical; second tray	100%; 0	100%; 0	-
near miss: contacts; deflection	0; 0 $^\circ$	0 of 3; 0.00-0.15 $^\circ$	-
prompt pairs, permuted layouts	0	paired differences 0	-

Table D.1. Intact / readable videos per model and experiment, with the number of videos in brackets. The bars draw the intact (dark) and readable (light) shares of that number. Clip length: the new 3 s and 8 s clips (the 5 s clips are main-set videos). Photographed: photograph and twin rows together. Regeneration: the three noise levels of each model. -: not read (LTX-2.5 on deflectors).

model	track main set	deflector main set	local edit, shared front	ramps, barriers, chutes	trays, prompt pairs	clip length		photographed	regeneration
						3 s	8 s		
▲ LTX-2.3	104/107 (120)	1/14 (72)	155/167 (216)	28/32 (45)	20/33 (120)	17/19 (24)	8/9 (24)	18/19 (45)	not run
▼ LTX-2.5 (distilled)	120/120 (120)	54/57 (72)	216/216 (216)	36/39 (45)	56/67 (120)	22/22 (24)	18/20 (24)	35/39 (45)	63/68, 68/68, 58/60 (72)
● LTX-2.5	99/100 (120)	4/- (72)	209/209 (216)	20/20 (33)	not run	not run	not run	13/13 (24)	not run
◆ Cosmos 3 Nano	106/106 (120)	48/48 (72)	174/174 (216)	33/33 (45)	49/57 (120)	not run	not run	40/40 (45)	not run
◆ MiniMax-H3	120/120 (120)	63/68 (72)	216/216 (216)	45/45 (45)	119/119 (120)	21/21 (24)	23/23 (24)	42/42 (45)	not run
■ Wan 2.2 (A14B)	115/115 (120)	58/58 (72)	199/199 (216)	43/43 (45)	101/105 (120)	22/22 (24)	19/19 (24)	41/41 (45)	62/66, 60/60, 62/63 (72)
○ Wan 3.0	120/120 (120)	63/63 (72)	45/45 (45)	23/23 (27)	35/36 (36)	not run	not run	24/24 (24)	not run
◀ Seedance 2.0	120/120 (120)	52/52 (72)	45/45 (45)	22/22 (27)	35/35 (36)	not run	not run	24/24 (24)	not run
▶ Seedance 2.0 mini	120/120 (120)	62/62 (72)	45/45 (45)	27/27 (27)	31/35 (36)	not run	not run	23/23 (24)	not run
× Wan 2.2 (TI2V-5B)	36/36 (120)	1/1 (72)	not run	not run	not run	not run	not run	not run	not run

to reproduce a generation: weights and precision, output size, frame rate and length, sampler, steps and guidance, the prompts, the seeds, and how the closed models were called.

E.1. Models

Table E.1 lists the settings of LTX-2.3 and LTX-2.5 [7, 14, 15], Wan 2.2 [28, 33], MiniMax-H3 [20], Cosmos 3 Nano [23, 24], the Seedance models [27, 32] and Wan 3.0 [1]. The open-weight models run through their packaged ComfyUI workflows at the 480p class of each model (Cosmos 3 Nano through its Diffusers pipeline), prompt rewriting is off for every model, and the sound tracks that some models produce are not used. The first

frame is the 1280 $\times$ 720 render, letterboxed without cropping to each open-weight model’s canvas (bars of at most 18 px) and sent unchanged to the closed models. Models that sample with classifier-free guidance above 1 receive a structural negative prompt. LTX-2.5 (distilled) samples without guidance and MiniMax-H3 has no negative input.

Clip length. The 3 s and 8 s clips use 73 and 193 frames for LTX-2.3 and LTX-2.5 (distilled), 73 and 192 for MiniMax-H3, and 49 and 129 (16 fps) for Wan 2.2 (A14B). All other settings equal those of the 5 s clips.

Table D.2. Recurring failure forms, with the later frame of the example in Fig. G.5 where one is shown (a moved part’s first-frame position dashed). Counts are flagged videos over all videos of the experiment.

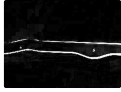
example	form	model	where and how often
	struck deflector rotates at or after impact	◀ Seedance 2.0, ▶ Seedance mini	deflector main set, apparatus 14/72 and 7/72; one deflector turned 26° and moved 17 mm for 92 frames
	tray tilts or shifts after the landing	▲ LTX-2.3, ◆ MiniMax-H3	deflector main set (LTX-2.3: 13–24 mm and 6.6–8.5°); MiniMax-H3 apparatus 9/72, mostly after the landing
	launch ramp rotates to another angle	▼ LTX-2.5 dist.	deflector main set, 3–9°, 13 of its 18 flagged videos, 9 of them after the ball has left the ramp; trays and prompt pairs 62/120
	deflectors move as loose objects	▲ LTX-2.3	apparatus on 71/72 deflector main-set and 100/120 tray and prompt videos; on photographs a deflector is knocked 125 mm and turned 124°, on twins it ends in the tray; chute trays move 7–27 mm in 12/12
	rails shift against the board	▲ LTX-2.3	tracks: 16/120 in the main set, 58/180 local-edit videos (4.5–5.2 mm, about 1°), in every arm including the null edit
	camera dollies in	● LTX-2.5	fiducial drift on 64/72 deflector videos, 12/120 track videos, 9/12 photographed track videos against 0/12 rendered twins
	second ball	◆ Cosmos 3 Nano	count flags 23/72 deflector main set, 51/120 trays and prompt pairs, 41/216 local edits and shared fronts (a second ball at the left edge of the frame), 10/45 ramps, barriers and chutes
	new objects appear	▲ LTX-2.3	photographed deflector scenes, extra 10/12 against 3/9 on twins (a large white tray enters from below)
	ball vanishes after the event	■ Wan 2.2	8 s clips: 4/7 readable track videos lose the ball after it reaches the stop; missing 13/72 on the deflector main set
	apparatus changes grow with clip length	▲ LTX-2.3	deflector clips: apparatus 6, 12 and 12 of 12 at 3, 5 and 8 s
	scene breaks down	× Wan 2.2 5B	count 33/120 and fiducial drift 48/120 on tracks; apparatus 70/72 on deflectors

Table E.1. Generation settings. Frames: frame rate × number of frames (duration). Neg.: structural negative prompt (Sec. E.2). Wan 2.2 (TI2V-5B) enters only the integrity table.

model	weights, precision	output	frames	steps	guidance	sampler	neg.
▲ LTX-2.3	22B dev, fp8; distilled LoRA 384 v1.1 at 0.2	832 × 480	24 × 121 (5.04 s)	20	video CFG 3, STG 1, rescale 0.9	res_2s, LTX scheduler	☑
▼ LTX-2.5 (distilled)	22B distilled, int8	832 × 448	24 × 121 (5.04 s)	8 + 3, fixed	none (CFG 1)	Euler ancestral, distilled sigmas	☐
● LTX-2.5	22B dev, bf16; first frame at strength 0.7	832 × 480	24 × 121 (5.04 s)	20	video CFG 3, STG 1, rescale 0.9	res_2s, LTX scheduler	☑
◆ Cosmos 3 Nano	16B, bf16	832 × 480	24 × 121 (5.04 s)	35	6.0	UniPC, flow shift 10	☑
◆ MiniMax-H3	FL2VA, pruned int8	864 × 480	24 × 124 (5.17 s)	20	none (guidance-distilled)	res_multistep, shift 12	no input
■ Wan 2.2 (A14B)	I2V-A14B, two experts, fp8	832 × 480	16 × 81 (5.06 s)	8 + 8	CFG 3.5	Euler, shift 5	☑
○ Wan 3.0	API wan3.0-video	854 × 480	30 × 150 (5.00 s)	–	–	480P, 16:9, 5 s, no prompt extension	n/a
◀ Seedance 2.0	API doubao-seedance-2-0-260128	864 × 496	24 × 121 (5.04 s)	–	–	480p, 16:9, 5 s, no audio	n/a
▶ Seedance 2.0 mini	API doubao-seedance-2-0-mini-260615	864 × 496	24 × 121 (5.04 s)	–	–	480p, 16:9, 5 s, no audio	n/a
× Wan 2.2 (TI2V-5B)	TI2V-5B, fp16	832 × 480	24 × 121 (5.04 s)	16	CFG 5	UniPC, shift 8	☑

Regeneration. The source video is encoded by the model’s autoencoder, noised to the level σ of flow matching and denoised from there [18]. LTX-2.5 (distilled) uses the 24 fps source on its 832×448 canvas with $\sigma \in$

$\{0.7, 0.8, 0.9\}$, and Wan 2.2 (A14B) the exact 16 fps source at 832×480 with $\sigma \in \{0.90, 0.94, 0.95\}$. The levels are set per model, as the share of the model’s own image-to-video deviation that each level reaches on a calibration scene: 16,

50 and 67% for LTX-2.5 (distilled) and 14, 40 and 98% for Wan 2.2 (A14B), whose deviation stays at 36–40% from $\sigma = 0.92$ to 0.94 and jumps to 98% at 0.95. The prompts are those of the pilot set, identical to the main-set prompts, and no negative prompt is used.

Closed models. The closed models were called through their providers’ APIs one task at a time, with the request parameters in Table E.1 and the watermark off. Wan 3.0 receives the first frame inline and the Seedance models receive a link to the same file. They run the two main sets, a subset of each intervention that needs no seed control, and the photograph rows of the photographed first frames. The number of videos of every model in every experiment is given in Table D.1. The subsets are three local-edit tracks (Table A.3) with their base, bump and dip, and shared-front tracks 0, 3, 4, 5, 7 and 11. They include the 60, 15 and 0° ramp pairs, the 0.85, 1.05 and 1.20 barriers and the $\kappa = 0, +8, -4$ chutes. Of the tray interventions they run the 30° and 60° shifted trays of both groups, one two-tray scene per angle, and the 10 mm-above near miss with its hit control, and of the prompt pairs four (two regular and two irregular layouts).

E.2. Prompts

Within a family the prompt is fixed, and {colour} stands for the ball colour of the scene. The prompts call the deflectors paddles and the launch ramp a launch track (Fig. E.1).

E.3. Seeds and repeatability

Every open-weight model generates each scene with seeds 0, 1 and 2, and the seed drives the first noise source of the workflow. Secondary noise sources keep the fixed values of the packaged workflows. A rerun with the same seed reproduces the decoded frames of LTX-2.5 (distilled), LTX-2.3, LTX-2.5 and MiniMax-H3 exactly, and 33 of 36 Cosmos 3 Nano reruns are identical files. Wan 2.2 (A14B) reruns differ in 0.2% of the pixels by at most 7 grey levels, which leaves the measured trajectory unchanged, so the sampler floor of the local edits is zero for every open-weight model except LTX-2.3 (0.15% in dv and 0.03% in dt). The closed models either ignore the seed (Seedance) or do not reproduce it (Wan 3.0), so their three generations per scene are replicates, and paired designs compare arm means across independent samples. A twelve-scene pilot set (seeds 0 and 1) supplies the image-to-video versions that the regeneration probe compares with.

F. Results by question

This part gives the per-model results behind each statement of the main paper, in the order of its questions. How the drawn geometry modulates motion (main paper, Sec. 4.1) is

Table F.1. Drawn geometry on readable videos. Track: intact videos, median amplitude ratio of the drawn undulation, median on-track fraction and kinematic gain. Deflectors: intact videos, median change of the deflector angle with its 95th percentile in brackets, and median shift.

model	track			deflectors			
	intact	ampl.	on track	kin.	intact	angle (°)	shift (mm)
★ Physical oracle		1	1	1		0	0
▲ LTX-2.3	104/120	0.998	0.975	1.08	1/72	0.11 (0.28)	0.29
▼ LTX-2.5 dist.	120/120	1.004	1.000	1.12	54/72	0.07 (0.19)	0.24
● LTX-2.5	99/120	0.999	1.000	1.04	4/72	–	–
◆ Cosmos 3 Nano	106/120	0.999	1.000	1.00	48/72	0.04 (0.16)	0.15
◆ MiniMax-H3	120/120	1.001	1.000	1.02	63/72	0.13 (0.28)	0.25
■ Wan 2.2	115/120	0.999	1.000	1.00	58/72	0.05 (0.17)	0.20
○ Wan 3.0	120/120	1.001	1.000	1.03	63/72	0.06 (0.16)	0.18
◀ Seedance 2.0	120/120	1.001	1.000	0.99	52/72	0.07 (0.22)	0.27
▶ Seedance mini	120/120	1.002	1.000	1.05	62/72	0.16 (0.37)	0.27

covered by Sec. F.1 and the contacts of Sec. F.2. Whether the modulation forms a local dynamics (Sec. 4.2) is covered by Secs. F.2 to F.4, what organizes the global trajectory (Sec. 4.3) by Secs. F.5 and F.6, and photographed first frames and regeneration (Sec. 4.4) by Secs. F.7 and F.8. Table 1 of the main paper holds one number per reading, and the tables below give each reading with its interval, floor and denominator. Unless stated otherwise, values are medians over readable, estimable videos with 95% intervals from a bootstrap over scenes (Sec. H), and n is the number of videos.

F.1. Response on the track family

Every model keeps the drawn track. The ball’s speed rises and falls with the drawn undulation, at 0.54 to 1.24 times the physical gain at 3 cycles per metre and ahead of the physical phase in eight models, while the mean slope adds little acceleration.

Geometry. On the readable videos the drawn undulation keeps its amplitude (per-model median ratio 0.998–1.004, per model and level 0.993–1.006) and the ball stays on the track (median on-track fraction 0.98–1.00, Table F.1). In the deflector family the deflectors keep their drawn angle (median change 0.04–0.16°, 95th percentile at most 0.37°) and position (median shift 0.15–0.29 mm). The kinematic gain of 0.99–1.12 shows the ball riding the drawn rails, so what follows concerns how the speed depends on the geometry.

Gain and phase. At $f = 3\text{ m}^{-1}$ the gain of LTX-2.3, both LTX-2.5 models and Wan 3.0 includes the physical value and that of the other five models lies below it (0.54–0.73). At $f = 4$ every gain lies below 1 (0.49–0.88, Table F.2), and over all levels and broadband components the gain is 0.44–1.24. At $f = 3$ eight models lead the physical

Tracks

track family

used in

☒ track main set
☒ local edits
☒ shared fronts
☒ clip length
☒ photographed tracks
☒ regenerated tracks

names the outcome

☒ yes, the stop at the end
☐ no

A realistic tabletop physics demonstration, filmed in real time as one continuous shot with a locked-off camera. A small {colour} ball rests at the top of a narrow white track mounted on a matte black board with small black-and-white checker markers in its corners. The ball starts rolling from rest and follows the track under gravity until it reaches the stop at the end. The track, the board and the camera remain completely rigid and stationary. The scene only contains the described setup and action.

Deflectors, named outcome

deflector family

used in

☒ deflector main set
☒ clip length
☒ photographed deflectors
☒ regenerated deflectors

names the outcome

☒ yes, at rest in the tray
☐ no

A realistic tabletop physics demonstration, filmed in real time as one continuous shot with a locked-off camera. A small {colour} ball rests at the top of a short white launch track mounted on a matte black board with small black-and-white checker markers in its corners; white flat paddles are fixed to the board below it and a white tray sits at the bottom. The ball starts rolling from rest, leaves the track, falls under gravity and bounces off the paddles until it comes to rest in the tray. The paddles, the track, the tray, the board and the camera remain completely rigid and stationary. The scene only contains the described setup and action.

Deflectors, neutral

interventions

used in

☒ prompt pairs
☒ shifted trays
☒ two trays
☒ no tray
☒ near misses
☒ permuted layouts

names the outcome

☐ yes
☒ no

A realistic tabletop physics demonstration, filmed in real time as one continuous shot with a locked-off camera. A small {colour} ball rests at the top of a short white launch track mounted on a matte black board with small black-and-white checker markers in its corners; white parts are fixed to the board below it. The ball starts rolling from rest, leaves the track and moves under gravity. The white parts, the track, the board and the camera remain completely rigid and stationary. The scene only contains the described setup and action.

Ramp pairs and barriers

one phrase changed

used in

☒ ramp pairs
☒ barriers

names the outcome

☐ yes
☒ no

in the track prompt:
~~follows the track under gravity until it reaches the stop at the end~~
→
moves along the track under gravity

Chutes

chutes

used in

☒ chutes

names the outcome

☐ yes
☒ no

A realistic tabletop physics demonstration, filmed in real time as one continuous shot with a locked-off camera. A small {colour} ball rests at the top of a short white track mounted on a matte black board with small black-and-white checker markers in its corners; a white tray is fixed to the board below the end of the track. The ball starts rolling from rest, leaves the end of the track and falls under gravity. The track, the tray, the board and the camera remain completely rigid and stationary. The scene only contains the described setup and action.

Structural negative prompt

guidance above 1

text, watermark, subtitles, people, hands, extra objects, camera motion, zoom, cuts, cartoon, painting, blurry

Figure E.1. **Prompts.** One card per prompt: where it is used, whether it names the outcome, and its text, with the colour slot highlighted.

phase by $15\text{--}41^\circ$, at least twice the per-video floor of $6.0\text{--}7.5^\circ$ (Wan 2.2 (A14B), with the smallest lead of $+15^\circ$, sits at 2.1 times its floor), and MiniMax-H3 responds in phase (-3° , $[-10, +6]$). At $f = 4$ the lead is $+2$ to $+45^\circ$, with Wan 2.2 (A14B) the in-phase model there. The broadband components give the same picture with wider intervals. The response repeats across seeds (phase coherence across the three seeds of a scene $0.67\text{--}0.96$). A second harmonic accompanies it: $2f/1f$ is $0.18\text{--}0.71$ at $f = 3$ and $f = 4$, against $0.02\text{--}0.05$ for the physical oracle.

Mean slope, relaxation and tempo. On the straight track the trend ratio is -0.10 to 0.65 against 1 for the stated

law, below 0.3 in seven models and 0.62 and 0.65 in the two Seedance models (Table F.3). On the undulating levels it is at most 0.42 (-0.33 to 0.31 at $f = 3$). On the straight level the exponent p is $0.51\text{--}1.62$ against 2: LTX-2.5 (distilled) moves at nearly constant speed and the two Seedance models come closest to a uniform acceleration. The relaxation rate is positive everywhere ($0.48\text{--}7.2\text{ s}^{-1}$ on the straight level, $1.3\text{--}14.6\text{ s}^{-1}$ on the undulating levels and $2.0\text{--}12.3\text{ s}^{-1}$ at $f = 3$), where the stated law has none, so the speed returns toward a preferred value within $0.07\text{--}0.74\text{ s}$. The balanced accuracy of $0.60\text{--}0.84$ at $f = 3$ lies above the monotone baseline ($0.60\text{--}0.68$) and below the physical oracle ($0.90\text{--}1.00$), and Wan 2.2 (A14B) sits at the

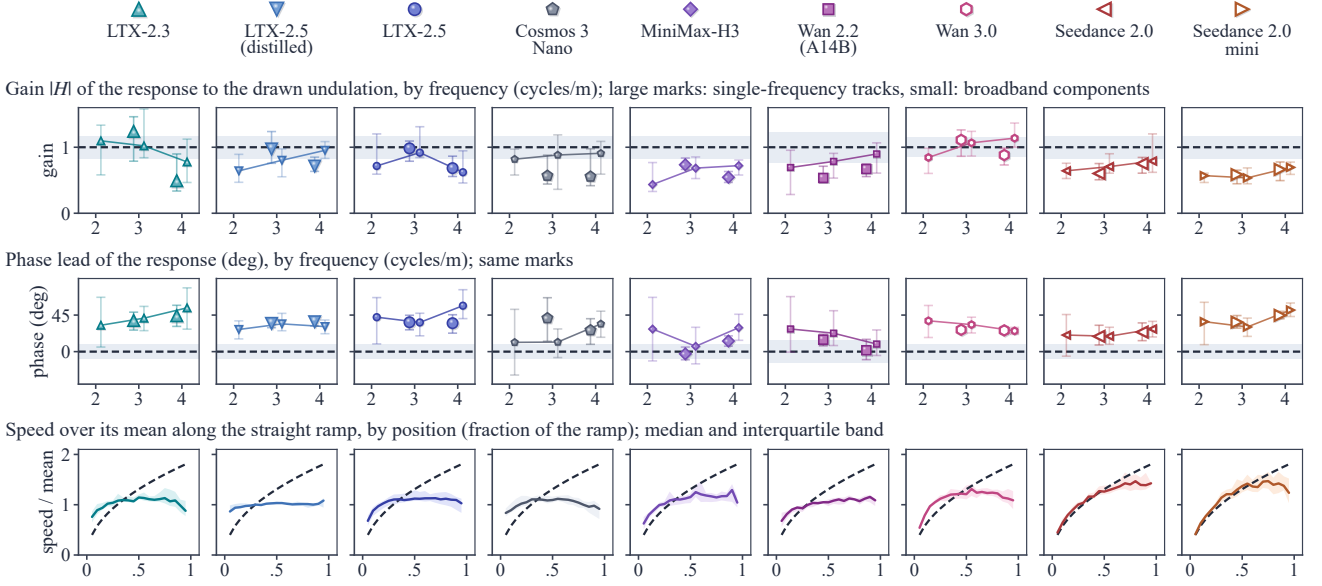


Table F.2. Response on the track family. Top: the first three rows of Fig. G.4, gain and phase by spatial frequency (large marks single-frequency tracks, small marks broadband components, physical value dashed, floor grey) and the speed along the straight ramp. Bottom: the complex gain, gain $|H|$ and phase (degrees, positive = speed leads the drawn height), for the single-frequency levels and the three broadband components, with 95% intervals below each median. n : estimable videos at $f = 3$, $f = 4$ and broadband.

model	n	gain $ H $					phase ($^{\circ}$)				
		$f = 3$	$f = 4$	bb 2	bb 3	bb 4	$f = 3$	$f = 4$	bb 2	bb 3	bb 4
★ Physical oracle		1	1	1	1	1	0	0	0	0	0
▲ LTX-2.3	24, 20, 24	1.24 [0.79, 1.46]	0.49 [0.34, 0.90]	1.10 [0.58, 1.33]	1.02 [0.84, 1.58]	0.78 [0.47, 1.12]	+38 [+27, +49]	+44 [+31, +57]	+32 [+6, +67]	+41 [+26, +56]	+54 [+28, +79]
▼ LTX-2.5 (distilled)	25, 23, 27	0.97 [0.76, 1.24]	0.71 [0.63, 0.85]	0.64 [0.47, 0.89]	0.80 [0.55, 0.98]	0.95 [0.83, 1.09]	+35 [+30, +41]	+37 [+32, +42]	+27 [+16, +38]	+34 [+26, +47]	+31 [+22, +39]
● LTX-2.5	24, 12, 15	0.98 [0.79, 1.09]	0.68 [0.56, 0.86]	0.72 [0.59, 1.20]	0.92 [0.84, 1.31]	0.62 [0.46, 0.95]	+36 [+27, +45]	+35 [+23, +45]	+42 [+10, +67]	+36 [+19, +47]	+57 [+41, +76]
◆ Cosmos 3 Nano	21, 19, 16	0.57 [0.44, 0.88]	0.56 [0.42, 0.98]	0.82 [0.58, 0.97]	0.88 [0.36, 1.19]	0.91 [0.59, 1.09]	+41 [+13, +66]	+26 [+9, +41]	+11 [−29, +52]	+12 [−7, +28]	+34 [+18, +50]
◆ MiniMax-H3	30, 30, 30	0.73 [0.66, 0.84]	0.54 [0.47, 0.63]	0.44 [0.33, 0.77]	0.68 [0.53, 0.85]	0.72 [0.58, 0.81]	−3 [−10, +6]	+13 [+7, +19]	+28 [−11, +67]	+7 [−15, +30]	+30 [+14, +46]
■ Wan 2.2 (A14B)	28, 30, 26	0.54 [0.46, 0.71]	0.67 [0.56, 0.75]	0.69 [0.28, 0.95]	0.78 [0.53, 0.91]	0.90 [0.61, 1.07]	+15 [+7, +21]	+2 [−10, +15]	+28 [−1, +68]	+23 [+8, +50]	+9 [−5, +27]
○ Wan 3.0	30, 30, 30	1.11 [0.86, 1.26]	0.88 [0.73, 0.99]	0.84 [0.60, 0.98]	1.07 [0.87, 1.24]	1.14 [1.00, 1.37]	+27 [+23, +30]	+26 [+21, +31]	+38 [+17, +57]	+33 [+23, +42]	+26 [+21, +30]
◁ Seedance 2.0	30, 30, 30	0.60 [0.51, 0.75]	0.75 [0.61, 0.85]	0.64 [0.53, 0.76]	0.71 [0.61, 0.90]	0.79 [0.62, 1.20]	+19 [+8, +32]	+24 [+14, +35]	+20 [−6, +46]	+19 [+9, +30]	+28 [+18, +37]
▷ Seedance 2.0 mini	30, 30, 30	0.59 [0.45, 0.65]	0.66 [0.49, 0.78]	0.57 [0.47, 0.61]	0.54 [0.45, 0.69]	0.70 [0.59, 0.78]	+36 [+29, +44]	+45 [+40, +50]	+37 [+9, +61]	+30 [+19, +41]	+51 [+43, +60]

baseline. All models are slower than the simulated motion ($\alpha = 1.5\text{--}3.2$ at $f = 3$).

Mechanism fits. Each model’s fingerprint was compared with ten candidate dynamics: the stated law, spatial smoothing, temporal smoothing, a response lag, a constant-speed cruise (relaxation at rate b toward a preferred speed with gravity kept), a first-order law in which speed follows the local slope, a curvature slowdown, the cruise with a Gaussian temporal smoothing of width τ , a relaxation with time constant τ toward a slope-dependent target speed $v^* - k\beta_0 W(s)$ without a physical drive, and a relaxation toward a saturated physical speed. Each candidate ran on a parameter grid along the drawn tracks of four scenes per level at the

model’s tempo, first-motion time and frame rate, through the scene camera with 1 px noise and the same estimator (14,400 simulated motions), and 30 components per model (gain, phase, $2f/1f$, gravities, b , g_w and accuracy per level, and p , b and trend ratio on the straight level) enter χ^2 with σ^2 the squared floor plus the squared half-width of the observed interval. The lowest AIC picks the best candidate, and a candidate suffices when $\chi^2/n \leq 2$ with no component beyond 3σ . Two families describe the nine models most economically (Table F.4). Relaxation to a slope-dependent target speed within 0.2–0.4 s ($k = 0.6$) fits Cosmos 3 Nano, LTX-2.3, LTX-2.5 and Wan 2.2 (A14B). Cruise with temporal smoothing, a relaxation toward a constant speed at 0.5–3.5 s^{-1} (time constant 0.3–2 s) followed by a smoothing of

Table F.3. Dynamics on the track family: tempo, trend ratio, relaxation rate b , straight-ramp exponent p and balanced accuracy (the physical oracle of the same scenes reads 0.90–1.00), with 95% intervals below each median. n : estimable videos at $f = 3$ and on the straight level.

model	n	tempo	trend ratio		relaxation rate b (s^{-1})		exponent	accuracy
		$\alpha, f = 3$	straight	$f = 3$	straight	$f = 3$	p , straight	$f = 3$
★ Physical oracle		1	1	1	0	0	2	0.90–1.00
▲ LTX-2.3	24, 27	2.05 [1.70, 2.61]	0.15 [0.09, 0.17]	−0.33 [−0.59, −0.30]	4.02 [2.79, 5.07]	12.33 [9.10, 16.15]	1.02 [0.95, 1.09]	0.67 [0.64, 0.72]
▼ LTX-2.5 (distilled)	25, 30	1.55 [1.36, 1.65]	−0.10 [−0.18, −0.06]	0.16 [0.01, 0.27]	7.22 [4.52, 19.36]	5.29 [3.02, 10.05]	0.51 [0.42, 0.73]	0.84 [0.83, 0.87]
● LTX-2.5	24, 26	1.63 [1.27, 1.73]	0.02 [0.02, 0.14]	0.23 [0.11, 0.33]	3.29 [1.93, 7.65]	4.93 [2.96, 6.09]	0.95 [0.76, 1.03]	0.75 [0.68, 0.81]
◆ Cosmos 3 Nano	21, 27	1.53 [1.16, 1.75]	−0.02 [−0.07, 0.04]	0.04 [−0.10, 0.20]	6.73 [4.83, 22.26]	4.99 [2.47, 8.16]	0.91 [0.86, 1.01]	0.71 [0.62, 0.76]
◆ MiniMax-H3	30, 30	2.66 [2.44, 2.87]	0.22 [0.21, 0.25]	0.19 [0.16, 0.26]	1.39 [0.96, 1.61]	2.10 [1.73, 2.23]	1.19 [1.17, 1.21]	0.79 [0.74, 0.82]
■ Wan 2.2 (A14B)	28, 29	3.17 [2.82, 3.32]	0.22 [0.06, 0.27]	0.04 [−0.02, 0.11]	1.79 [1.26, 3.54]	2.80 [2.57, 3.39]	1.14 [1.03, 1.19]	0.60 [0.57, 0.63]
○ Wan 3.0	30, 30	2.32 [2.04, 2.45]	0.25 [0.22, 0.35]	−0.02 [−0.04, 0.04]	1.84 [1.27, 2.11]	5.67 [5.07, 6.96]	1.23 [1.19, 1.25]	0.77 [0.74, 0.79]
◀ Seedance 2.0	30, 30	2.21 [2.06, 2.30]	0.62 [0.54, 0.80]	0.31 [0.20, 0.47]	0.62 [0.46, 0.77]	1.97 [1.38, 2.95]	1.54 [1.47, 1.72]	0.74 [0.70, 0.76]
▶ Seedance 2.0 mini	30, 30	1.98 [1.79, 2.28]	0.65 [0.63, 0.88]	0.07 [0.02, 0.25]	0.48 [0.26, 0.74]	4.62 [3.03, 6.29]	1.62 [1.54, 1.79]	0.67 [0.64, 0.70]

0.1–0.2 s, fits LTX-2.5 (distilled), MiniMax-H3, Seedance 2.0, Seedance 2.0 mini and Wan 3.0. The time-rescaled stated law suffices for none of the nine (χ^2/n 6.5–27.1) because it keeps the acceleration along the mean slope, and the first-order law scores worst (46–8768). One mechanism suffices for five models, and the largest deviations share one shape. On undulating tracks the relaxation rate of MiniMax-H3 and the two Seedance models is 1.6–3.7 times the rate that the fitted mechanism carries over from the straight level, Wan 3.0 relaxes on the straight level at half the carried rate, and the near-constant speed of LTX-2.5 (distilled) on the straight level ($p = 0.51$) is produced by no candidate. The pull toward the preferred speed thus changes with the shape of the track, and no single preferred-speed rule reproduces these models across the levels of the family. The two families lie within 0.1–0.2 in χ^2/n for Cosmos 3 Nano and LTX-2.5, and the shared-front tracks place both with the smoothing family (Sec. F.3), so their label is tentative. Sec. F.4 carries both families to local edits, ramp pairs and barriers.

F.2. Contacts and composition on the deflector family

At the first contact the closed models bounce, MiniMax-H3 bounces in part and three open-weight models slide along the deflector. On three deflectors the physical order of contacts appears in at most a quarter of any model’s videos, each contact errs on its own, and most models still bring the ball into the named tray.

First contact. Pooled over all 24 deflector scenes, the closed models reach an apparent normal restitution of 0.67–0.74 against the prescribed 0.85 (band 0.80–0.90) and leave 16–35° from the simulator’s exit for the measured incoming direction (Table F.5). MiniMax-H3 reads 0.45 and 23°, and Cosmos 3 Nano, LTX-2.5 (distilled) and Wan 2.2 (A14B)

slide (0.10–0.21, exits 57–78° away). On the single deflector MiniMax-H3 bounces as much as the closed models (0.66). There the exit direction turns with the deflector angle at 0.44–2.07° per degree against the physical 2.06, and the interval excludes 0 only for MiniMax-H3 and Wan 2.2 (A14B). LTX-2.3 has one estimable contact and is left uninterpreted.

Composition. With three deflectors (regular and irregular layouts pooled, 16 scenes) the physical order appears in 12–24% of the trackable videos of the closed models and in none of the open-weight models’ videos, while exact renders reproduce it every time (Table F.6). In the two Seedance models the exit is off by 19–31° at the first, second and third contact alike, and in Wan 3.0 the residual grows from 10° at the first contact to 22 and 28° at the second and third. MiniMax-H3, Wan 2.2 (A14B) and Cosmos 3 Nano seldom reach a second contact (2 of 42, 8 of 23 and 4 of 11 first contacts). A contact law fitted on each model’s single-deflector contacts and then frozen improves the three-deflector prediction for Seedance 2.0 only (26.4 to 22.1° on regular and 30.9 to 24.8° on irregular layouts). A consistent material of the wrong kind would be absorbed by this calibration, so the contact errors of the other models depend on the layout.

Destination against route. Under the prompt that names the outcome the ball reaches the tray in 98–100% of the trackable three-deflector videos of seven models whatever the route, and in 47% of those of Wan 2.2 (A14B) (Table F.6). The geometry leaves little room for a wrong route. The ensemble rests in the tray in 81–98% of its members on these layouts, and with any one deflector removed in 0–30% (medians 2% on regular and 13% on irregular layouts). Continued by the stated law from the state in which the

Table F.4. Mechanism fits on the track family: best candidate by AIC with its parameters, largest deviation $|z|$, whether one mechanism suffices ($\checkmark$ yes, $\times$ no), and χ^2/n ($n = 30$) of every candidate at its best grid point, with the best candidate tinted. Law: the time-rescaled stated law. Cruise+s.: cruise with temporal smoothing. Target: relaxation to a slope-dependent target speed. Sat.: saturating relaxation.

model	best	parameters	$ z _{\max}$	suff.	law	spatial	temporal	χ^2/n of each candidate						
								lag	cruise	1st order	curvature	cruise+s.	target	sat.
▲ LTX-2.3	target	τ 0.2 s, k 0.6	2.4	$\checkmark$	18.4	18.4	4.7	18.7	1.9	168.5	16.1	1.7	1.6	2.5
▼ LTX-2.5 (distilled)	cruise+s.	b 3.5 s $^{-1}$, τ 0.1 s	3.3	$\times$	14.8	14.4	7.1	18.2	2.1	191.9	13.6	1.3	1.4	4.6
● LTX-2.5	target	τ 0.2 s, k 0.6	2.5	$\checkmark$	11.5	10.9	3.7	12.9	2.4	133.8	10.5	1.2	1.0	2.0
◆ Cosmos 3 Nano	target	τ 0.2 s, k 0.6	1.7	$\checkmark$	14.5	14.2	7.1	14.6	1.0	45.9	13.8	0.8	0.7	1.5
◆ MiniMax-H3	cruise+s.	b 1 s $^{-1}$, τ 0.2 s	3.2	$\times$	25.2	23.8	7.6	24.3	6.4	6355.5	24.4	1.4	1.8	2.3
■ Wan 2.2 (A14B)	target	τ 0.4 s, k 0.6	1.6	$\checkmark$	13.4	13.2	6.8	13.5	2.1	1493.5	12.2	1.2	0.7	1.6
○ Wan 3.0	cruise+s.	b 3.5 s $^{-1}$, τ 0.1 s	3.7	$\times$	27.1	27.3	12.5	29.9	2.6	2375.6	26.6	1.9	3.0	7.5
◀ Seedance 2.0	cruise+s.	b 0.5 s $^{-1}$, τ 0.2 s	2.4	$\checkmark$	6.5	5.9	1.6	6.8	3.5	8768.3	5.9	0.9	1.6	1.8
▶ Seedance 2.0 mini	cruise+s.	b 0.5 s $^{-1}$, τ 0.2 s	3.1	$\times$	12.1	11.8	3.0	14.1	8.1	3812.2	10.6	2.2	4.7	5.4

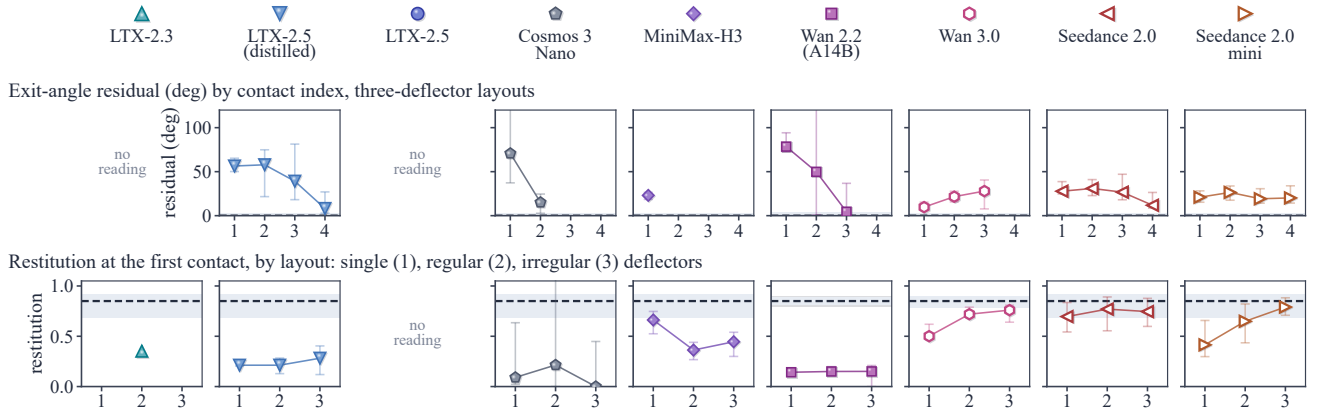


Table F.5. Contact dynamics on the deflector family. Top: exit-angle residual by contact index and first-contact restitution by layout. Bottom: restitution, exit residual and angle-response slope by model, with 95% intervals where estimated and denominators. Grey bands show the physical restitution range in the table and the measurement floors in the charts; floor values are reported in Table C.6.

model	all scenes	first-contact restitution e_n		residual ($^\circ$)			exit slope
		one deflector	three deflectors	all scenes	one deflector	one deflector	
★ Physical oracle	0.85 (0.80–0.90)	0.85	0.85	0	0	2.06	
▲ LTX-2.3	0.35 (1)	—	0.35 (1)	17	—	— (0)	
▼ LTX-2.5 (distilled)	0.21 [0.19, 0.27] (46)	0.21 (15)	0.23 (31)	57 [43, 65]	58	0.69 [−0.86, 3.01] (15)	
◆ Cosmos 3 Nano	0.10 [0.00, 0.41] (15)	0.09 (4)	0.10 (11)	68 [40, 83]	61	4.60 [4.19, 5.02] (4)	
◆ MiniMax-H3	0.45 [0.36, 0.54] (61)	0.66 (19)	0.41 (42)	23 [18, 29]	23	1.04 [0.44, 1.59] (19)	
■ Wan 2.2 (A14B)	0.15 [0.12, 0.17] (42)	0.14 (19)	0.15 (23)	78 [70, 92]	70	2.07 [0.58, 4.02] (19)	
○ Wan 3.0	0.67 [0.56, 0.75] (62)	0.50 (21)	0.75 (41)	16 [10, 22]	29	1.23 [−0.01, 2.79] (21)	
◀ Seedance 2.0	0.74 [0.60, 0.81] (52)	0.70 (18)	0.75 (34)	35 [28, 46]	47	0.60 [−0.26, 1.89] (18)	
▶ Seedance 2.0 mini	0.67 [0.56, 0.78] (59)	0.41 (17)	0.78 (42)	21 [16, 28]	21	0.44 [−0.11, 1.61] (17)	

generated ball leaves its last correct contact (or the launch ramp), the ball reaches the tray in 11–84% of the wrong-route videos of six of these models, from 11% for MiniMax-H3 to 84% for Wan 3.0 (Table F.7). The models exceed it by +16 to +87 percentage points with intervals that exclude 0, LTX-2.3 by +8 on 13 videos, and rescaling the exit speed by the model’s apparent gravity leaves +29 to +43 points.

Wan 2.2 (A14B) goes the other way: its ball slides along the first deflector and often stops short of the tray. This table covers all named-prompt videos, and Sec. F.5 pairs them with the neutral prompt.

Table F.6. Composition on the three-deflector layouts (regular and irregular pooled, prompt naming the tray). Exact order: share of trackable videos whose merged contact sequence equals the physical one, with 95% interval and counts, drawn on 0 to 100% with the physical 100% as a tick. Residual at contact $k = 1, 2, 3$ with the number of contacts. Calibrated law: median residual under the stated law and under the law fitted on the model’s single-deflector contacts, on regular and irregular layouts. Tray: share of trackable videos that end in the tray. Tinted: the one model for which the calibrated law improves the prediction. Floor of the residual 2.0–4.2° (90th percentile).

model	exact order share	residual at contact k (°)			stated → calibrated law (°)		tray
		$k = 1$	$k = 2$	$k = 3$	regular	irregular	
★ Physical oracle	100%	0	0	0	0	0	81–98% of members
▲ LTX-2.3	0% (0/13)	17 (1)	–	–	–	–	100% (13)
▼ LTX-2.5 (distilled)	0% (0/40)	56 (31)	58 (25)	39 (18)	47.0 → 47.4	50.5 → 51.9	100% (40)
◆ Cosmos 3 Nano	0% (0/34)	71 (11)	15 (4)	20 (1)	37.3 → 72.3	70.8 → 38.3	100% (34)
◆ MiniMax-H3	0% (0/46)	23 (42)	27 (2)	–	24.7 → 24.7	21.3 → 21.0	98% (46)
■ Wan 2.2 (A14B)	0% (0/36)	78 (23)	50 (8)	4 (4)	77.6 → 80.5	68.2 → 72.6	47% (36)
○ Wan 3.0	12% [2, 23] (5/42)	10 (41)	22 (39)	28 (14)	17.4 → 18.9	16.8 → 18.0	100% (42)
◁ Seedance 2.0	24% [11, 38] (8/34)	28 (34)	31 (27)	27 (20)	26.4 → 22.1	30.9 → 24.8	100% (34)
▷ Seedance 2.0 mini	14% [5, 26] (6/43)	21 (42)	26 (35)	19 (26)	23.4 → 23.7	19.2 → 19.6	100% (43)

Table F.7. Destination counterfactual on the three-deflector layouts, all named-prompt videos. Simulator: tray arrival of the stated law rolled forward from the model’s exit state after its last correct contact (videos used / wrong-route videos). Difference: model minus simulator in percentage points with 95% interval, then with the exit speed rescaled by the launch tempo and by the apparent gravity (videos used).

model	simulator	model minus simulator (points)		
		difference	tempo-scaled	gravity-scaled
▲ LTX-2.3	92% (13/13)	+8 [+0, +21]	+33 (3)	+33 (3)
▼ LTX-2.5 dist.	68% (40/40)	+32 [+16, +50]	+42 (24)	+31 (35)
◆ Cosmos 3 Nano	65% (34/34)	+35 [+24, +47]	+40 (15)	+39 (23)
◆ MiniMax-H3	11% (46/46)	+87 [+72, +100]	+43 (46)	+36 (11)
■ Wan 2.2	59% (27/36)	–22 [–44, –3]	–8 (24)	+29 (7)
○ Wan 3.0	84% (37/37)	+16 [+5, +29]	+14 (37)	+29 (28)
◁ Seedance 2.0	50% (26/26)	+50 [+33, +67]	+68 (25)	+31 (26)
▷ Seedance mini	30% (37/37)	+70 [+55, +85]	+76 (37)	+43 (37)

F.3. Locality: local edits and shared-front tracks

Only MiniMax-H3 changes speed at an edit in the physical direction while the motion before it stays unchanged, by a fifth of the physical change at the bump. In three open-weight models an edit ahead of the ball changes its path before the ball arrives beyond the null edit, and on shared-front tracks most models follow the drawn back-section phase with a lead of their own.

Local edits, open-weight models. All arms of a track share the seed, so edit-induced differences can be read relative to the rerun and null-edit floors (Table F.8). MiniMax-H3 slows by 2.4% at a bump (simulator –11.4%) and speeds up by 1.0% at a dip after normalizing by the pre-edit tempo (2.4% raw, simulator +7.8%), with the time before

the edit changed by at most 0.7%. Its arrival shift is 0.42 of the physical one, and its arrival is 0.7–0.8% late after both edits. LTX-2.5 reaches either edit 3–6% sooner and slows at the dip (–6.1%, [–8.6, –0.8]), where the simulated ball speeds up, while its reading at the bump (–2.1%, [–10.6, +7.5]) includes zero. Wan 2.2 (A14B) slows by 29% at the bump, 2.6 times the physical size, from before the bump on (time to the edit +15.3%), and all ten of its pairs come from the one seed that reaches the stop within the clip. The null edit of LTX-2.5 (distilled) changes dv by +4.1% [+2.1, +5.3] and is flagged, so its readings are not specific to the edit, and LTX-2.3 and Cosmos 3 Nano have too few pairs for a size reading. Before the ball reaches it, a bump or dip changes the path 1.7 to 4.0 times as much as the null edit of the same size, and the intervals separate from the null for LTX-2.3 and LTX-2.5 (bump and dip) and MiniMax-H3 (dip).

Local edits, closed models. Without seed control the arms are independent samples that also differ in tempo. On the three closed-model tracks Seedance 2.0 and Seedance 2.0 mini speed up at the dip by +9.3% [+3.4, +29.1] and +12.3% [+5.6, +13.1] (simulator +7.4%), in the physical direction and by at least the physical amount, with the time before the dip changed by –3.2 and +0.6%. Their slowdown at the bump (–5.0% [–12.4, +0.5] and –1.5% [–1.8, +4.9], simulator –10.6%) has intervals that include zero. Wan 3.0 has no reading, since its arms differ in the time to the edit by up to 14%.

Shared-front tracks. LTX-2.5 (distilled), LTX-2.5 and Wan 3.0 follow the drawn back-section phase with slopes whose intervals contain 1, Seedance 2.0 mini follows it with a slope of 0.80, and Seedance 2.0 and LTX-2.3 have slopes near 1 with intervals that also contain 0 (Table F.9). Their

Table F.8. Local edits, open-weight models (percent unless marked), pairs of the same track and seed. dv : speed change at the edit relative to the base video, normalized by the speed before the edit, with the unnormalized (raw) change at the dip in grey. Arrival: the normalized arrival shift of the bump over its physical value (1.07–1.10 on exact renders). T_{pre} : change of the time before the edit, bump / dip. Path: change of the path before the edit over that under the null edit, bump / dip, with * where the interval separates from the null. Null: null-edit dv , with † marking a flag. Floors: rerun 0 for every model except LTX-2.3 (0.15%), measurement 1.0% (dv) and 0.4% (dt).

model	pairs	dv bump	dv dip (raw)	arrival / physical	T_{pre}	path / null	null dv
★ Physical oracle		−11.0 to −11.7	+7.7 to +8.0	1	0, 0	—	0
▲ LTX-2.3	10, 16	+4.1 [0.0, +11.9]	−1.1 (−3.9) [−8.9, +3.7]	−1.57 [−4.67, 0.33]	−0.2, −0.2	2.1*, 2.2*	+2.5
▼ LTX-2.5 (distilled)	30, 31	−3.9 [−13.9, +0.9]	+1.9 (−1.4) [−2.2, +6.3]	0.34 [−1.76, 2.74]	−5.2, −3.5	2.8, 2.1	+4.1 †
● LTX-2.5	19, 27	−2.1 [−10.6, +7.5]	−6.1 (−4.4) [−8.6, −0.8]	2.93 [0.48, 5.11]	−3.1, −5.8	3.6*, 4.0*	−0.5
◆ Cosmos 3 Nano	11, 11	−2.6 [−27.0, +7.4]	−0.1 (−1.9) [−7.4, +3.0]	4.26 [2.22, 6.63]	−0.9, −0.5	3.0, 2.2	−1.2
◆ MiniMax-H3	35, 36	−2.4 [−3.6, −0.7]	+1.0 (+2.4) [−0.3, +2.9]	0.42 [0.08, 0.78]	−0.7, −0.1	2.0, 1.7*	−0.7
■ Wan 2.2 (A14B)	10, 9	−29.3 [−37.2, −12.6]	−3.9 (−0.8) [−32.8, −0.7]	1.90 [−0.15, 4.08]	+15.3, −0.9	1.1, 2.0	−8.2

leads have the sign and size of the track-family phase, and LTX-2.5 (distilled), Seedance 2.0 mini and LTX-2.3 lead by 17–25° more. MiniMax-H3 does not follow the back section as a whole (slope −0.11): the exploratory decomposition splits its response into a part locked to the drawn phase (0.41 of the physical amplitude, leading by 17°) and a larger part with the same phase on every track (0.55). Such a fixed part arises only with temporal smoothing, which is why the shared-front tracks place MiniMax-H3, Cosmos 3 Nano, LTX-2.5 and Seedance 2.0 with the smoothing family. In Wan 3.0, Seedance 2.0 mini and LTX-2.5 the lead grows from the front to the back section, where the ball is faster, by 9–22°, while both fitted families predict that it shrinks. Wan 2.2 (A14B) shows no measurable back-section response. The twelve front sections are drawn identically, and the motion over them spreads by 4.8–27.6 mm across tracks against a floor of 1.6–1.7 mm, the same kind of change that a null edit produces.

F.4. Energy: ramp pairs, barriers and chutes

The generated ball climbs above its release height, crosses barriers higher than its release point at every height tested, and keeps a crest speed that stays flat as the barrier rises. The fitted preferred-speed rules of Sec. F.1 reproduce none of the models on the track, the local edit and the ramp pair together.

Ramp pairs. Of the 72 readable ascents, 63 rise above the release height (61 by more than 10%, against a single-video floor of 0.055–0.091), where the simulated ball turns around at 0.93–0.98 (Table F.10). Thirty-seven reach the top stop, 12 are still climbing at the end of the clip, 10 stop on the slope and 13 turn back, 7 of those above the release height. Seven balls leave the track at the top and keep rising, 10–365 mm above the top stop (two in LTX-2.5 (distilled), three in LTX-2.5, two in Wan 2.2 (A14B)). The mean height ratio is 1.15–2.36 across models against 0.96 for the physical

oracle (0.97 for the two angles of the closed models). The 18 unreadable climbs also carry a height ratio and are left out of every count. The ratio rises with the angle because a steep ascent ends at a higher stop and a shallow one outlasts the clip, and the physical dependence on the angle (0.929 to 0.980) is below the resolution at model tempo.

Barriers. No model shows a threshold (Table F.11). The ensemble crosses with probability 1.00, 1.00, 0.42, 0.16, 0 and 0 at hump heights of 0.70, 0.85, 0.95, 0.971, 1.05 and 1.20 of the release drop. Six models cross every barrier they were given (the closed models ran three heights), including the two that stand 5% and 20% above the release drop, and across models the ball crosses 43 of the 49 impassable humps it reaches and fails at 7 of the 72 passable ones. Over the humps higher than the release drop the crossing ball of eight models keeps 0.33–0.87 of its speed at the foot, independent of the height, where the physical oracle falls from 0.53 through 0.35, 0.15 and 0.03 to a turn-around. The ball slows on the hump by an amount unrelated to the height it climbs. The outcome of Wan 2.2 (A14B) follows the seed: seed 0 crosses all six barriers, seed 1 turns back low on all six, seed 2 stops on the hump or does not reach it.

Chutes. The four chutes release the ball in the same state, and the stated law lands it at the same point whatever the drawn end curvature. Only Wan 3.0 lands where the drawn curvature points (Table F.12): its landing moves by −2.95 mm per unit of κ (permutation $p = 0.009$), about 35 mm from $\kappa = -4$ to $+8$, and its exit direction steepens with κ (−2.36° per unit, $p = 0.014$). The ball leaves the tip carrying the direction the chute had before its end. MiniMax-H3 also steepens its exit with κ ($p = 0.02$), but its landing order stays within the spread between chutes ($p = 0.10$). The other models show no ordering by κ ($p \geq 0.13$), and LTX-2.3, LTX-2.5 (distilled), Wan 2.2 (A14B) and Cosmos 3 Nano have 4–6 readable videos. The

Table F.9. Shared-front tracks. Slope of the back-section phase on the drawn back phase (1: follows the drawing, 0: continues the front) with its resultant length R , drawn on -0.5 to 1.5 with the physical 1 as a tick, and the track-family phase at $f = 3$ for comparison. Exploratory: locked part $|G|$ and its lead, fixed part $|C|$ (in brackets the physical oracle at the model’s tempo), front-section lead and its change to the back section. Front: spread of the motion over the identical front sections in mm (floor 1.6–1.7). Medians with 95% intervals. n : estimable videos.

model	n		back-section phase slope		track phase	exploratory					front (mm)
			slope	R		$ G $	lead of G	$ C $	front lead	back – front	
★ Physical oracle			1.00		0	1	0	0	0	0	0
▲ LTX-2.3	30		1.06 [−0.21, 1.18]	0.62	+38	0.90 [0.49, 1.29]	+59 [+37, +80]	0.16 [0.06, 0.64] (0.07)	+54	+12 [−23, +73]	19.5
▼ LTX-2.5 (distilled)	29		1.05 [0.97, 1.19]	0.83	+35	0.80 [0.63, 1.10]	+60 [+51, +70]	0.14 [0.05, 0.34] (0.22)	+55	−4 [−25, +23]	27.6
● LTX-2.5	34		0.95 [0.87, 1.01]	0.85	+36	0.86 [0.67, 1.04]	+27 [+19, +37]	0.32 [0.23, 0.45] (0.09)	+19	+9 [−1, +17]	14.4
● Cosmos 3 Nano	21		−0.28 [−0.50, 1.50]	0.52	+41	0.51 [0.36, 0.80]	+23 [+1, +50]	0.48 [0.30, 0.75] (0.17)	+39	−6 [−48, +68]	11.3
◆ MiniMax-H3	36		−0.11 [−0.24, 0.06]	0.68	−3	0.41 [0.35, 0.47]	+17 [+5, +26]	0.55 [0.47, 0.62] (0.04)	+63	−22 [−86, +21]	9.3
■ Wan 2.2 (A14B)	32		−0.50 [−0.50, 1.50]	0.30	+15	0.09 [0.03, 0.38]	—	0.05 [0.02, 0.35] (0.09)	+102	—	4.8
○ Wan 3.0	18		1.07 [0.97, 1.18]	0.98	+27	0.79 [0.67, 0.92]	+32 [+24, +39]	0.19 [0.08, 0.35] (0.01)	+18	+17 [+5, +28]	9.6
◁ Seedance 2.0	18		0.92 [0.00, 1.50]	0.69	+19	0.67 [0.37, 1.02]	+28 [+12, +43]	0.30 [0.09, 0.64] (0.08)	+21	+9 [−145, +103]	25.1
▷ Seedance 2.0 mini	18		0.80 [0.71, 0.89]	0.88	+36	0.54 [0.44, 0.67]	+53 [+36, +67]	0.17 [0.05, 0.31] (0.04)	+30	+22 [+3, +43]	17.5

Seedance models let the ball pause at the tip and then drop, so their fitted exit direction is not reported.

The fitted mechanisms beyond the track family. Each model’s two fitted mechanisms from Sec. F.1 were run on the local-edit, shared-front, ramp and barrier geometries at the model’s tempo and read by the same estimator (Table F.13, exploratory). On the local edits both families predict a local response in the physical direction and an unchanged motion before the edit. Only MiniMax-H3 has that form, at a size near the slope-target prediction, and the pre-edit changes of LTX-2.5 and Wan 2.2 (A14B) come from neither family. On ramps and barriers the two families separate through gravity. The slope-target law has a target speed with a positive floor, so its ball slows on an ascent without stopping and crosses every barrier. The smoothed cruise keeps gravity, so at the fitted rates its ball turns back low on the 60° ascent and crosses only the lowest barriers, and it still does so when the preferred speed is raised by half. All five models placed in the smoothed-cruise family on the track family behave like the slope-target law on ramps and barriers, and LTX-2.3 and LTX-2.5, placed with the slope target, agree with it there and depart from both families at the local edits. No single preferred-speed rule thus reproduces one model’s behaviour on the undulating track, at the local edit and on the ramp pair together. Three observations lie outside both families: the ball leaving the track at the top, the direction carried out of a curved chute, and outcomes set by the seed. The mechanism fits are the most economical description of the track response, and locality and energy call for something beyond them.

F.5. Outcome: prompt pairs and tray interventions

The named outcome strongly controls the destination: with the same first frame and seed and a prompt that leaves the outcome unnamed, tray arrival falls in all eight models tested, the steering beyond the ball’s own exit state shrinks to at most 27 points, and a drawn tray by itself barely attracts the ball.

Prompt pairs. This paragraph and Table F.14 use the pairs that are trackable under both prompts, while Table F.7 covers all named-prompt videos. Arrival falls by 33 to 96 percentage points. The unpaired arrival over all trackable videos of each prompt shows the same drop, and so do the pairs in which the ball is visible at the end of both clips (there Wan 3.0 shrinks to -14 points, $[-30, 0]$). The neutral prompt also slows the motion of seven models (tempo 1.22–2.35 times that of the named prompt, with LTX-2.3 unchanged on its three pairs), and part of the drop is motion that is still under way at the end of the clip. On the pairs whose neutral motion is complete (the ball has arrived, is at rest, or would have arrived at its own launch tempo), the drop stays large for Cosmos 3 Nano, LTX-2.3, Wan 2.2 (A14B) and Seedance 2.0 mini and becomes small for LTX-2.5 (distilled) and Wan 3.0. MiniMax-H3 shows that the named outcome also controls when the motion ends. At its own launch tempo its ball could not reach the tray within 5 s in 43 of 46 named-prompt videos, and 42 of them end in the tray. With the neutral prompt 40 of 45 are still moving at the end. The steering beyond the ball’s exit state, $+43$ to $+75$ points for MiniMax-H3 and the Seedance models with the named prompt, falls to -18 , 0 and $+11$ points, and the negative value comes from motions still under way at the

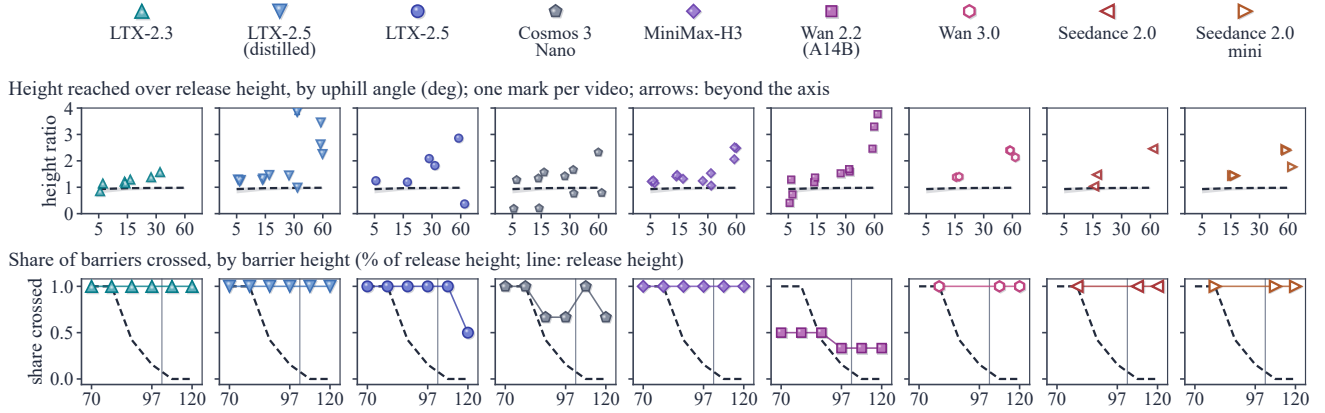


Table F.10. Energy. Top: rows 6 and 7 of Fig. G.4, the height reached over the release height by uphill angle (one mark per video) and the share of barriers crossed by height. Bottom: ramp pairs, readable climbs. Mean height ratio over ascents with 95% interval (drawn on 0 to 4 with the physical 0.962 as a tick), ascents above the release height as dots, and per angle the median ratio over one block per climb: ■ top stop, ■ still climbing, ■ stopped on the slope, ■ turned back (top-stop ratio in the header). Physical oracle: nominal turning height with the 64-member band. Last column: flat ascents reaching the end stop. The closed models ran 60, 15 and 0°.

model	mean height ratio		above release		per uphill angle (top-stop ratio)				0°
					60° (2.52)	30° (1.61)	15° (1.46)	5° (1.32)	
★ Physical oracle		0.962 (closed 0.973)	0	0.980 (0.92–1.00)	0.975 (0.90–1.00)	0.965 (0.87–1.00)	0.929 (0.76–1.00)		reaches
▲ LTX-2.3	— ●—	1.23 [1.15, 1.30]	●●●●●○	—	1.48 ■	1.21 ■	1.00 ■		○○○
▼ LTX-2.5 (distilled)	— ●—	2.36 [1.54, 3.70]	●●●●●●●○	2.62 ■	1.44 ■	1.33 ■	1.27 ■		●○○
● LTX-2.5	— ●—	1.60 [1.14, 2.06]	●●●●○	1.61 ■	1.95 ■	1.20 ■	1.25 ■		●○
◆ Cosmos 3 Nano	— ●—	1.15 [0.80, 1.49]	●●●●●○○○	1.56 ■	1.42 ■	1.34 ■	0.73 ■		●
◆ MiniMax-H3	— ●—	1.56 [1.48, 1.64]	●●●●●●●●	2.48 ■	1.24 ■	1.41 ■	1.22 ■		●○○
■ Wan 2.2 (A14B)	— ●—	1.76 [1.55, 1.95]	●●●●●●●○	3.30 ■	1.60 ■	1.28 ■	0.73 ■		●○○
○ Wan 3.0	— ●—	1.94 [1.84, 2.00]	●●●●	2.40 ■	not run	1.39 ■	not run		●●●
◀ Seedance 2.0	— ●—	1.66 [1.51, 1.80]	●●●	2.46 ■	not run	1.25 ■	not run		○○○
▶ Seedance 2.0 mini	— ●—	1.82 [1.61, 1.94]	●●●●●	2.40 ■	not run	1.44 ■	not run		●●○

end of the clip. Under the neutral prompt no model steers by more than +27 points (LTX-2.5 (distilled), $[-3, +52]$). Wan 3.0 keeps the most of its destination: it reaches the tray in 67% of its neutral videos and steers by +22 points ($[0, +43]$) without a named outcome. The route stays wrong under both prompts: the exact order appears in one named-prompt video each of the two Seedance models and in none of the others.

Shifted tray. With the neutral prompt only the drawn tray can attract the ball. The capture contrast is 0.42 for Seedance 2.0 and 0.10 for Seedance 2.0 mini with intervals that exclude zero, far below 1, and its interval includes zero for Wan 3.0 and MiniMax-H3 (Table F.15). The high arrival in the beyond group comes from the models’ own short

landings: the ball leaves the deflector slowly and lands 125–154 mm on the incoming side of the physical point already in the in-tray group, and the beyond trays sit exactly there. The landing then moves a further 26–74 mm toward the tray. MiniMax-H3 drops the ball almost vertically below the deflector in both groups. Cosmos 3 Nano, LTX-2.3, LTX-2.5 (distilled) and Wan 2.2 (A14B) leave one or two readable pairs (a second ball, struck deflectors, or balls that have not landed by the end).

Two trays and no tray. When a second, identical tray sits under the launch lip, the stated law rolled from the models’ own slow and steep exit states would mostly send the ball into it, and the models put it there in 0–3 videos (Table F.16). Their balls end elsewhere: MiniMax-H3 is still

Table F.11. Barriers. Per height ratio: crossings among the readable videos as dots (filled: crossed) over the median crest-speed ratio. Tinted: the barriers the stated law does not cross. Forbidden: crossings of the barriers the stated law does not cross (1.05, 1.20), with 0.95 and 0.971 counted by their nominal outcome. Missed: failures on barriers the stated law crosses. 50%: logistic crossing point. The closed models ran 0.85, 1.05 and 1.20.

model	barrier height over release drop						forbidden	missed	50%
	0.70	0.85	0.95	0.971	1.05	1.20			
★ Physical oracle	pass, 0.53	pass, 0.35	27/64, 0.15	10/64, 0.03	turn back	turn back	0	0	0.971–1.05
▲ LTX-2.3	0.10	0.52	0.59	0.54	0.45	0.41	6/6	0/12	none
▼ LTX-2.5 (distilled)	0.63	0.51	0.52	0.52	0.61	0.61	6/6	0/12	none
● LTX-2.5	0.16	0.47	0.16	0.15	0.47	0.87	3/4	0/8	1.20
◆ Cosmos 3 Nano	0.55	0.70	0.91	0.82	0.52	0.40	4/5	2/11	1.32 (extrap.)
◆ MiniMax-H3	0.82	0.62	0.58	0.56	0.56	0.68	6/6	0/12	none
■ Wan 2.2 (A14B)	0.83	0.78	0.87	0.70	0.96	1.07	2/6	5/9	by seed
○ Wan 3.0	not run	0.27	not run	not run	0.33	0.33	4/4	0/3	none
◀ Seedance 2.0	not run	0.46	not run	not run	0.53	0.45	6/6	0/2	none
▶ Seedance 2.0 mini	not run	0.42	not run	not run	0.46	0.43	6/6	0/3	none

Table F.12. Chutes. Landing slope (mm per unit κ in m^{-1}) with bootstrap interval and permutation p , mean landing offset in mm from the physical landing (720 mm), and the exit-direction slope ($^\circ$ per unit κ , with p). n : readable videos. Tinted: the model whose landing follows the drawn curvature.

model	n	landing slope	p	offset	exit slope (p)
★ Physical oracle		0		0	0
▲ LTX-2.3	4	-0.99 [-1.73, -0.25]	0.50	+3.9	+1.40 (0.51)
▼ LTX-2.5 dist.	6	+1.91 [0.87, 2.95]	0.13	-59.7	+0.07 (1.0)
◆ Cosmos 3 Nano	6	+2.45 [-1.16, 6.07]	0.50	-40.2	-0.63 (0.83)
◆ MiniMax-H3	12	-1.13 [-2.51, -0.19]	0.10	-59.3	-0.79 (0.020)
■ Wan 2.2	5	+4.02 [1.38, 6.66]	0.58	-3.9	+1.54 (0.08)
○ Wan 3.0	8	-2.95 [-4.18, -1.67]	0.009	-14.6	-2.36 (0.014)
◀ Seedance 2.0	8	+1.80 [-2.89, 6.00]	0.59	-2.8	–
▶ Seedance mini	9	+0.13 [-3.38, 3.37]	0.95	-3.6	–

moving at the end of all nine such videos, Wan 2.2 (A14B) rests on the first deflector or on the board, and Cosmos 3 Nano often leaves the ball on the launch ramp. Wan 3.0 puts four of six balls in the physical tray, while its own exit state would send five of six into the second one: its destination agrees with the scene’s physical outcome, consistent with its steering under the neutral prompt. With the tray removed, no model stops the ball where the tray stood, and only one Cosmos 3 Nano video draws the tray back.

Near misses and permuted layouts. A deflector 10 mm from the path is touched by every closed model in at least one of three seeds: Wan 3.0 in 3 of 3, Seedance 2.0 in 2 and Seedance 2.0 mini in 1, each a real rebound off the deflector end visible in all frames around the closest point. Wan 2.2 (A14B) touches the near-miss deflector in 3 of 7 videos (one per placement) and MiniMax-H3 in 1 of 9, whose path passes 70–134 mm below the deflectors placed above it. With three seeds these counts establish that the models’ paths come within 10 mm of the deflector end. On permuted layouts no trackable video reproduces either order (0 of 61), so the permutation reads at the floor. The first deflector alone is hit in 0/11 to 8/17 videos per model.

F.6. Timing: clip length

With the same scene, first frame, prompt and seed, the clip length controls the pace of the motion: from 3 to 8 s the same event stretches nearly in proportion, events fall at fixed fractions of the clip, and the extra time is spent at rest or on a repeat of the event. Both prompts of this experiment name the outcome.

Tempo and event times. On the tracks the duration exponent is 0.85–1.04 in three models (Table F.17), where a fixed physical rate gives 0 and a motion fitted to the clip gives 1. The intervals of MiniMax-H3 (1.04, [0.95, 1.14]) and LTX-2.3 (0.85, [0.53, 1.04]) include 1, and that of LTX-2.5 (distilled) (0.91, [0.90, 0.92]) lies just below it. MiniMax-H3 reaches the end stop at 90–93% of the clip at every

Table F.13. Observed readings against the two fitted mechanisms: the observed value in dark, then the predictions of the slope target and of the smoothed cruise in grey. Bump and dip: dv in percent. $|C|$: fixed part on shared-front tracks. Ramps: mean height ratio and the outcome on the 60° ascent (top: reaches the top stop, turn: turns back at the given ratio). Barriers: crossings (the predictions count the six heights). In brackets the family of the track-family fit.

model (family)	bump dv	dip dv	observed / slope target / smoothed cruise			
			fixed $ C $	ramps: mean	60° ascent	barriers crossed
▲ LTX-2.3 (target)	+4.1 / -3.5 / -2.7	-1.1 / +2.3 / +1.8	0.16 / 0.05 / 0.34	1.23 / 1.81 / 1.18	- / top / turn 0.36	18/18 / 6 / 6
▼ LTX-2.5 (distilled) (cruise)	-3.9 / -5.5 / -8.0	+1.9 / +3.0 / +3.3	0.14 / 0.02 / 0.35	2.36 / 1.81 / 1.02	top / top / turn 0.53	18/18 / 6 / 1
● LTX-2.5 (target)	-2.1 / -3.4 / -6.9	-6.1 / +1.8 / +1.9	0.32 / 0.02 / 0.28	1.60 / 1.81 / 1.01	top, turn / top / turn 0.46	11/12 / 6 / 4
◆ Cosmos 3 Nano (target)	-2.6 / -2.6 / -4.2	-0.1 / +1.9 / +0.3	0.48 / 0.01 / 0.52	1.15 / 1.81 / 1.78	top, turn / top / top	13/16 / 6 / 6
◆ MiniMax-H3 (cruise)	-2.4 / -4.4 / -9.3	+1.0 / +2.9 / +5.6	0.55 / 0.06 / 0.40	1.56 / 1.68 / 0.71	top / top / turn 0.60	18/18 / 6 / 1
■ Wan 2.2 (A14B) (target)	-29.3 / - / -	-3.9 / - / -	0.05 / 0.10 / 0.16	1.76 / 1.66 / 0.80	top / top / turn 0.53	by seed / 6 / 1
○ Wan 3.0 (cruise)	-7.0 / -5.3 / -7.3	-2.7 / +3.6 / +3.6	0.19 / 0.01 / 0.11	1.94 / 1.76 / 1.00	top / top / turn 0.61	7/7 / 6 / 3
◁ Seedance 2.0 (cruise)	-5.0 / -7.2 / -8.3	+9.3 / +4.9 / +4.9	0.30 / 0.04 / 0.48	1.66 / 1.74 / 0.72	top / top / turn 0.52	8/8 / 6 / 2
▷ Seedance 2.0 mini (cruise)	-1.5 / -3.5 / -8.5	+12.3 / +2.2 / +4.4	0.17 / 0.04 / 0.47	1.82 / 1.74 / 0.73	top / top / turn 0.51	9/9 / 6 / 2

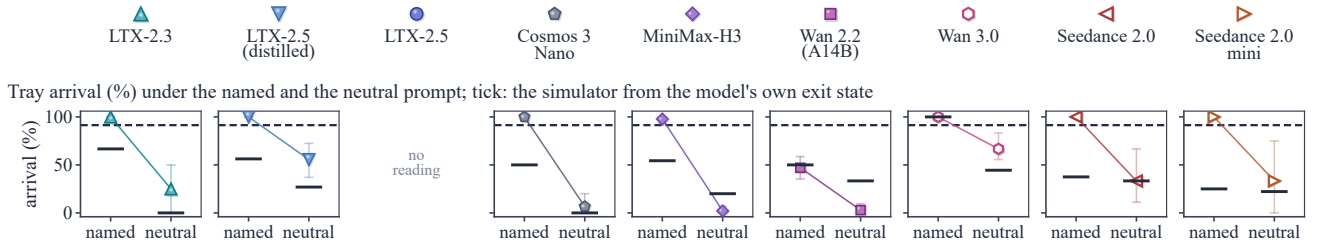


Table F.14. Effect of naming the destination. Top: tray arrival under named and neutral prompts. Bottom: matched prompt pairs, reporting arrival change, steering relative to the physical rollout from the model's exit state, tempo change and the status of neutral-prompt misses. Intervals are 95% scene-bootstrap intervals.

		tray arrival			steering			
model	pairs	neutral / named	difference	completed	neutral	named	tempo	neutral misses
★ Physical oracle		81–98% of members						
▲ LTX-2.3	8	25% / 100%	−75 [−100, −50]	−67 [−100, −33] (6)	+25 [0, +50] (8)	+33 [0, +50] (3)	1.01 ⁽³⁾	1 / 5 / 0
▼ LTX-2.5 (distilled)	27	56% / 100%	−44 [−63, −28]	−17 [−36, 0] (18)	+27 [−3, +52] (26)	+44 [+20, +69] (16)	2.35 ⁽¹⁵⁾	7 / 5 / 0
◆ Cosmos 3 Nano	15	7% / 100%	−93 [−100, −80]	−92 [−100, −75] (12)	+0 [−50, +50] (6)	+50 [+17, +83] (6)	1.38 ⁽⁵⁾	1 / 9 / 4
◆ MiniMax-H3	46	2% / 98%	−96 [−100, −89]	− ⁽²⁾	−18 [−30, −7] (45)	+43 [+31, +56] (46)	1.22 ⁽⁴⁵⁾	40 / 5 / 0
■ Wan 2.2 (A14B)	34	3% / 47%	−44 [−58, −28]	−42 [−75, −8] (12)	−20 [−40, 0] (15)	−14 [−39, +17] (22)	1.41 ⁽¹²⁾	20 / 8 / 5
○ Wan 3.0	9	67% / 100%	−33 [−44, −17]	−14 [−38, 0] (7)	+22 [0, +43] (9)	+0 [0, 0] (9)	1.25 ⁽⁹⁾	1 / 0 / 2
◀ Seedance 2.0	9	33% / 100%	−67 [−89, −33]	− ⁽³⁾	+0 [−27, +27] (9)	+62 [+40, +100] (8)	1.26 ⁽⁹⁾	5 / 0 / 1
▶ Seedance 2.0 mini	9	33% / 100%	−67 [−100, −25]	−40 [−100, 0] (5)	+11 [−45, +75] (9)	+75 [+50, +100] (8)	1.38 ⁽⁹⁾	1 / 2 / 3

Arrival: named (filled) to neutral (open) on 0 to 100%, over the band of physical arrival. Completed: pairs whose neutral motion is complete. Neutral misses: ■ still moving / ■ at rest / □ out of view or not visible at the end. Brackets: pairs or videos.

length, and its arrival times scale by 1.61 and 0.59 for clip ratios of 1.6 and 0.6. LTX-2.3 arrives at 55–59% of the clip and LTX-2.5 (distilled) at 47–52%, and both then rest. Wan 2.2 (A14B) runs twice as fast in 3 s as in 5 s (tempo ratio 0.49, [0.38, 0.53], 10 units) and keeps its 5 s pace at 8 s (0.85, [0.70, 1.12], 7 units). A single exponent would average these two opposite steps, so the two ratios describe its

tracks, and in 5 s only one of its twelve track videos reaches the stop. On deflectors the first tray contact gives γ of 1.06 for MiniMax-H3, 0.76 for Wan 2.2 (A14B) and 0.44 for LTX-2.5 (distilled). The tempo differences between lengths are 11–53 times the tempo floor.

Table F.15. Shifted tray (neutral prompt). Capture contrast over pairs of the same base scene and seed (0: ignores the tray, 1: follows its centre), drawn on -0.3 to 1 with the physical 0 as a tick, tray arrival in the in-tray and beyond groups (physical oracle 100% and 0%), and the median landing offset from the physical point in the two groups (mm, positive toward the deflector).

model	pairs	capture contrast	in	beyond	offset (mm)
★ Physical oracle		0	100%	0%	0/0
◆ MiniMax-H3	12	 -0.11 $[-0.26, 0.04]$	0%	17%	+317/+305
○ Wan 3.0	6	 0.41 $[-0.16, 0.45]$	83%	67%	+146/+215
◀ Seedance 2.0	6	 0.42 $[0.21, 0.54]$	83%	83%	+125/+199
▷ Seedance 2.0 mini	6	 0.10 $[0.06, 0.41]$	50%	100%	+154/+180

Table F.16. Two trays (neutral prompt). Final position of trackable videos, ■ physical tray / ■ second tray / □ elsewhere, and the share of rollouts of the stated law from the model’s own exit state that end in the second tray, unscaled and scaled by the launch tempo. Physical oracle: at least 63/64 in the physical tray, 0/64 touch the second.

model	n	final position physical / second / elsewhere	rollout to second
★ Physical oracle	64	$\geq 63 / 0 / -$	
▲ LTX-2.3	1	 $0 / 1 / 0$	1/1
▼ LTX-2.5 (distilled)	10	 $3 / 3 / 4$	10/10; 8/10
◆ Cosmos 3 Nano	6	 $1 / 0 / 5$	–
◆ MiniMax-H3	12	 $0 / 3 / 9$	10/12; 6/11
■ Wan 2.2 (A14B)	12	 $0 / 3 / 9$	7/9; 8/8
○ Wan 3.0	6	 $4 / 0 / 2$	4/6; 5/6
◀ Seedance 2.0	6	 $1 / 1 / 4$	5/5; 4/6
▷ Seedance 2.0 mini	6	 $3 / 2 / 1$	5/6; 1/6

Track fingerprints across lengths. The undulation response stretches with the motion (Table F.18). Apart from the two readable 8 s videos of LTX-2.3 and the phase of Wan 2.2 (A14B), the changes of gain and phase between lengths stay within twice their floors and those of accuracy within three times, so the pattern of the track family holds at every length. The relaxation rate, fitted in video time, changes most (13–49 times its floor). Converted to physical time (exploratory), it is the same at 5 and 8 s for both LTX models. At 3 s the converted rate and the trend ratio depart from the longer clips, which may come from the truncation of short clips, for which no floor exists.

Deflectors across lengths and the time after the event. MiniMax-H3 and LTX-2.5 (distilled) already bring every

settled ball into the tray at 3 s, and longer clips bring no more balls to the tray and no better order (exact order at 3, 5 and 8 s in 5/9, 4/10 and 6/11 videos of MiniMax-H3 and in 2/10, 5/10 and 3/9 of LTX-2.5 (distilled)). Wan 2.2 (A14B) lands 12/12 in 3 s and 12/12 in 8 s, and in 7 of the 8 s videos the ball then returns to the launch point and the event starts again. In the extra time of 8 s clips the LTX models leave the ball at rest for 3.5–3.8 s on tracks, LTX-2.5 (distilled) lifts the ball back onto a deflector and drops it into the tray a second time in 4 of 9 deflector videos, Wan 2.2 (A14B) loses the ball after the stop in 4 of 7 track videos, and MiniMax-H3 has about 0.5 s left after the event.

F.7. Photographed first frames

Replacing the rendered first frame with a photograph of the built scene leaves the readings of six models within their seed spread, and each of the three changes can be localized to a specific failure mode.

Decision rule photograph against control

For each model and configuration the photograph row is paired with its control, the rendered twin for open-weight models and the model’s own reading on the nominal scene for closed models. A reading changes with the visual domain only when three conditions hold together:

interval	the interval of the mean paired difference over configurations excludes zero,
seed spread	the pooled difference exceeds the seed spread $2s\sqrt{1/n_1 + 1/n_2}$,
floor	and most pairs exceed the floor of the photograph rows.

For a closed model a difference is attributed to the photograph only when it leaves the range produced by changing only the control (the twin minus nominal differences of the open-weight models). The floors are tight (gain 0.04–0.31, phase 1.5–25°, residual 0.5–4.6°) and the seed spread is usually larger, so the decisions rest mainly on the seed spread.

Results. Cosmos 3 Nano, LTX-2.3 on tracks, LTX-2.5 (distilled), Wan 2.2 (A14B), Seedance 2.0 mini and Wan 3.0 keep their readings (Table F.19). The tempo and gain differences of Wan 3.0 exceed its seed spread but lie within the range that the change of control alone produces (tempo +0.04 to +1.01, gain +0.02 to +0.38), and its clip duration stays the same (+0.05 s). Three changes remain. LTX-2.5 pushes the camera in on the photographs (fiducial drift in 9 of 12 videos against 0 of 12 twins), so most of its photograph rows are not estimable. The twin with the same measured geometry is readable in 11 of 12, implicating photographic appearance rather than the measured geometry. MiniMax-H3 errs by 17.5° more at contacts on photographs, with the same order and the same tray arrival. Seedance 2.0 shows a gain higher by 0.66, beyond both its seed spread and the range of the weaker control. LTX-2.3

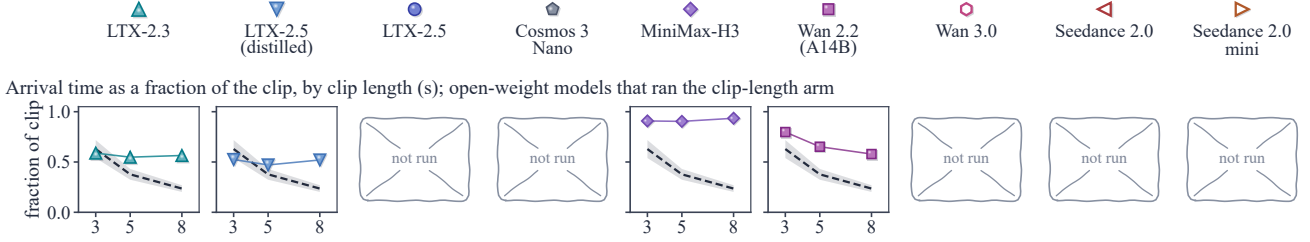


Table F.17. Clip length. Top: row 9 of Fig. G.4, the arrival at the end stop as a fraction of the clip by clip length. Bottom: tempo at 3 / 5 / 8 s (drawn on 0 to 4.5), duration exponent γ with interval and scene-seed units (drawn on -0.1 to 1.2 with the physical 0 as a tick), the arrival at the end stop as a share of the clip over readable videos (videos), and the paired ratio of arrival times 8 s over 5 s and 3 s over 5 s (physical oracle 1, a motion fitted to the clip 1.6 and 0.6). Wan 2.2 (A14B): the tempo ratios 3 s over 5 s and 8 s over 5 s in place of γ .

model	α at 3 / 5 / 8 s	γ , tracks	arrival at the end stop (share of clip)			arrival-time ratio	
			3 s	5 s	8 s	8/5	3/5
★ Physical oracle	equal at all lengths	0				1	1
▲ LTX-2.3	1.05 / 1.53 / 2.64	0.85 [0.53, 1.04] (11)	59% (4)	55% (8)	56% (3)	1.92	0.44 [0.38, 1.50]
▼ LTX-2.5 (distilled)	0.88 / 1.42 / 2.36	0.91 [0.90, 0.92] (12)	52% (9)	47% (9)	52% (11)	1.69 [1.31, 1.95]	0.75 [0.50, 0.85]
◆ MiniMax-H3	1.48 / 2.37 / 4.09	1.04 [0.95, 1.14] (12)	91% (12)	90% (10)	93% (10)	1.61 [1.53, 1.65]	0.59 [0.57, 0.60]
■ Wan 2.2 (A14B)	1.29 / 2.76 / 2.64	tempo 0.49; 0.85	80% (7)	65% (11)	58% (5)	—	0.91

Table F.18. Track readings at 3 / 5 / 8 s: gain, phase ($^\circ$), relaxation rate b (s^{-1} , video time) with the largest length difference over its floor in brackets, and $b\alpha$ (physical time, exploratory). Each line runs over the three lengths on a common scale per column ($|H|$ 0 to 2, phase -20 to $+50^\circ$, b 0 to 10, $b\alpha$ 0 to 40).

model	$ H $	phase	b	$b\alpha$
▲ LTX-2.3	0.71/0.90/1.82 (8.8)	+40/+39/+14 (1.0)	2.9/3.9/2.9 (13)	2.6/8.8/9.0
▼ LTX-2.5 (distilled)	0.45/0.70/0.52 (1.7)	+40/+34/+49 (1.4)	3.5/4.1/3.5 (33)	3.0/8.2/8.6
◆ MiniMax-H3	0.35/0.47/0.45 (1.2)	+7/-5/-1 (1.8)	5.1/1.5/1.4 (49)	8.3/4.0/5.7
■ Wan 2.2 (A14B)	0.71/0.46/0.68 (0.9)	-16/+17/-7 (3.5)	5.4/3.4/9.9 (49)	5.1/8.2/37.8

treats the deflectors as loose objects in every domain and adds new objects on photographs (Sec. D.3), which leaves it almost no trackable deflector videos (0 of 12 photographs, 1 of 9 twins).

The patterns of the main sets hold on photographs for every readable model. On three-deflector layouts the physical order appears in 0 or 1 of 6 photograph videos per model, and every trackable ball ends in the tray except with Wan 2.2 (A14B) (8 of 12, fewer on twins and nominal scenes). On tracks the relaxation rate is positive for every model and domain ($1\text{--}46\text{ s}^{-1}$), the median phase leads by $8\text{--}76^\circ$ for the eight models other than Cosmos 3 Nano, whose phase scatters with a seed spread of 59° , the median gain is $0.59\text{--}1.36$, and the trend ratio stays between -0.28 and 0.49 , far from 1.

F.8. Regeneration probe

Given a motion that is drawn correctly and wrong in its dynamics, the two open-weight models keep its timing, add

back their own response to the drawn shape, and change the motion where a drawn contact is violated. What they put in its place moves toward their own version of the scene and barely toward the stated law.

The lowest level copies. At the lowest noise level (LTX-2.5 (distilled) 0.7, Wan 2.2 (A14B) 0.90) the output stays within 4–15% of the distance between source and physics from the source (41% for the transparent deflector under Wan 2.2 (A14B)), and λ is near zero for every version (Table F.20). Even so the regeneration adds 17–21 mm of deviation to an exact source, ten times the 1.8 mm floor, so at this level $\lambda \approx 0$ reads as copying and the tendencies are read at the higher levels.

What is kept and what is replaced. The versions that break a drawn contact are pulled away from the source: the ball meets the transparent deflector from LTX 0.8 on

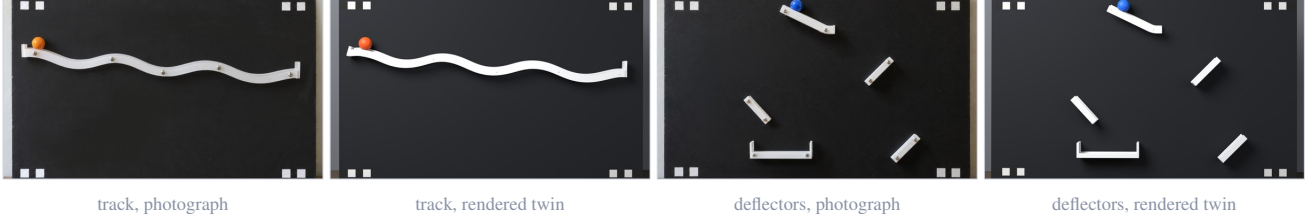


Table F.19. Photographed first frames. Top: a photographed first frame and its rendered twin for a track and for a deflector layout. Bottom: mean difference photograph minus control over configurations, with 95% intervals below. Tracks: share of estimable videos, tempo, gain, phase ($^{\circ}$). Deflectors: share of trackable videos, exact order, tray arrival, largest residual ($^{\circ}$). *, tinted: a change under the three-part rule. Control: rendered twin (open-weight models) or the nominal scene (closed models).

model	control	tracks				deflectors			
		estimable	α	$ H $	phase	trackable	order	tray	residual
▲ LTX-2.3	twin	+0.08 [-0.25, +0.50]	-0.12 [-0.36, +0.31]	+0.90 [+0.78, +1.28]	+49 [+6, +64]	-0.08 [-0.25, +0.00]	—	—	—
▼ LTX-2.5 (distilled)	twin	-0.25 [-0.75, +0.00]	+0.30 [+0.09, +0.62]	-0.27 [-0.31, -0.24]	+2 [-18, +23]	+0.08 [+0.00, +0.25]	-0.33 [-0.75, +0.00]	+0.00	+18.2 [+16.5, +19.6]
● LTX-2.5	twin	-0.42* [-0.67, -0.17]	+0.06 [+0.05, +0.07]	+0.17 [+0.14, +0.19]	+1 [-1, +4]	not run			
◆ Cosmos 3 Nano	twin	-0.17 [-0.33, +0.00]	-0.16 [-0.51, +0.04]	-0.25 [-0.74, +0.08]	-31 [-101, +110]	+0.08 [-0.17, +0.33]	+0.08 [+0.00, +0.25]	+0.00	-46 (1 config.)
◆ MiniMax-H3	twin	+0.00	-0.10 [-0.20, +0.02]	-0.01 [-0.09, +0.13]	+3 [-31, +13]	+0.17 [+0.00, +0.33]	+0.00	+0.00	+17.5* [+2.8, +32.2]
■ Wan 2.2 (A14B)	twin	+0.00 [-0.33, +0.33]	-0.04 [-0.18, +0.14]	-0.22 [-1.01, +0.06]	-16 [-55, +28]	+0.08 [+0.00, +0.25]	-0.17 [-0.50, +0.00]	+0.29 [+0.08, +0.54]	+28 [-25, +81]
○ Wan 3.0	nominal	+0.00	+0.26 [+0.13, +0.43]	+0.35 [+0.22, +0.40]	+20 [-7, +33]	+0.17 [+0.00, +0.50]	-0.17 [-0.50, +0.00]	+0.00	+10.9 [-3.5, +32.7]
◀ Seedance 2.0	nominal	+0.00	-0.02 [-0.27, +0.24]	+0.66* [+0.30, +0.85]	+15 [-10, +24]	+0.08 [+0.00, +0.25]	-0.04 [-0.38, +0.25]	+0.00	-7.7 [-14.7, -2.4]
▶ Seedance 2.0 mini	nominal	-0.08 [-0.25, +0.00]	+0.16 [-0.28, +0.63]	+0.02 [-0.14, +0.41]	-35 [-60, +20]	+0.08 [+0.00, +0.25]	-0.17 [-0.33, +0.00]	+0.00	+8.2 [-12.1, +31.2]

($\lambda = 1.01$) and from Wan 0.94 on (λ 1.09–1.12), and the restitution-0.3 source is pulled to $\lambda = 0.32$ at LTX 0.9 and replaced at Wan 0.94 and 0.95 (λ 0.99 and 1.40). The replacement keeps the dynamics wrong. The order stays wrong in every one of these videos and the first-contact restitution stays at 0.16–0.45 against 0.85 (Table F.21). Split along the two directions (exploratory, seeds 0 and 1, each output against the model’s own image-to-video video of the same scene and seed), the Wan 2.2 (A14B) regenerations of the restitution-0.3 source move 0.66 of the way toward that video and 0.10 toward physics (medians of 10 videos at 0.94 and 0.95). Per level the split is 0.40 and 0.23 at 0.94 (4 videos) and 0.73 and 0.08 at 0.95 (6 videos). A λ above 1 is the model’s own version overshooting physics along the source-to-physics direction. The versions that keep every drawn contact are accepted. Their λ stays below 0.1 except at LTX’s highest level (slowed source 0.24, $+90^{\circ}$ source 0.10), where the output drifts toward the model’s own version. The slowed motion keeps its tempo (2.01–2.10 under Wan and 1.81–1.90 under LTX, source 2.0), the constant-speed motion keeps its timing ($\lambda \leq 0.02$) and regains a response to the drawn undulation (gain 0.27–0.62 against 0.02 in the source), and the $+90^{\circ}$ phase returns toward each model’s own lead on the track family (LTX $+30^{\circ}$ at 0.9, Wan $+6$ and $+13^{\circ}$ at 0.94 and 0.95), away from the physical 0° .

From the exact source, and the switch of Wan 2.2 (A14B). Started from the exact source, both models drift away from the physical motion as the noise rises, by about 105–124 mm at the highest level (Table F.22). LTX-2.5 (distilled) drifts toward its own image-to-video version (κ 0.25–0.29, the same for the same and for another seed, so the direction is the model’s general rendering of the scene), and at 0.9 half of its seed-0 and seed-1 outputs across all versions have become that version (21 of 42). Most drift of Wan 2.2 (A14B) points elsewhere (κ 0.04–0.08). Its response to the noise level is a switch: the deviation stays at 36–40% of its own image-to-video deviation from 0.92 to 0.94 and jumps to 98% at 0.95, where 16 of 21 deflector outputs become its own version while the track outputs mostly still follow the source (2 of 23). The level of the switch depends on the scene, 0.94 for one irregular layout and 0.95 for another.

G. Case grids and per-model charts

The grids show the simulator and every model on one scene of each level and of the main interventions, and the charts give every model’s readings of every experiment on one page. Each grid cell is a stroboscopic composite: the ball at equal time steps, from light to full as time passes, on the board of the video, with the simulated path dashed and a small tick at each physical contact. Cells are chosen by rule. On a track level the cell shows the model’s estimable

Table F.20. Regeneration: λ per source version and noise level (0: stays on the source, 1: pulled onto physics), median with 95% interval and readable videos. The tint deepens with λ , from 0 to 1.2 and above (0  1.2).

	▼ LTX-2.5 (distilled)			■ Wan 2.2 (A14B)		
source version	0.7	0.8	0.9	0.90	0.94	0.95
<i>keeps every drawn contact</i>						
constant speed	0.01 [0.00, 0.02] (9)	0.00 [0.00, 0.01] (9)	0.00 [-0.06, 0.03] (9)	0.01 [0.00, 0.02] (9)	0.02 [0.01, 0.02] (6)	0.02 [-0.02, 0.47] (6)
phase +90°	0.00 [-0.02, 0.01] (9)	0.02 [0.01, 0.03] (9)	0.10 [0.04, 0.13] (9)	0.01 [0.01, 0.02] (9)	0.07 [0.03, 0.14] (7)	0.08 [0.04, 0.14] (8)
slowed 2×	-0.00 [-0.01, 0.00] (18)	0.06 [0.01, 0.19] (18)	0.24 [0.18, 0.41] (15)	0.00 [0.00, 0.01] (18)	0.00 [-0.01, 0.02] (17)	0.01 [-0.03, 0.02] (16)
<i>breaks a drawn contact</i>						
restitution 0.3	0.01 [-0.02, 0.02] (9)	0.03 [-0.05, 0.14] (9)	0.32 [0.17, 0.92] (7)	0.01 [0.01, 0.09] (8)	0.99 [0.11, 1.70] (6)	1.40 [0.52, 1.76] (9)
transparent deflector	0.07 [0.07, 0.14] (5)	1.01 [0.99, 1.28] (5)	1.01 [0.97, 1.11] (4)	0.28 [0.25, 0.62] (5)	1.09 [1.01, 1.38] (6)	1.12 [1.02, 1.35] (8)

Table F.21. The defining reading of each version after regeneration, at the three levels of each model, against the source and the physical oracle.

version: reading	source	oracle	▼ LTX 0.7/0.8/0.9	■ Wan 0.90/0.94/0.95
slowed: α	2.0	1.0	1.86/1.90/1.81	2.01/2.04/2.10
constant speed: $ H $	0.02	1.0	0.47/0.60/0.62	0.27/0.50/0.52
phase +90°: phase	+86	0	+61/+37/+30	+70/+6/+13
restitution 0.3: e_n	0.30	0.85	0.28/0.28/0.45	0.16/0.35/0.29
transparent: exact order	0	1	0/0/0	0/0/0

Table F.22. Exact sources: distance from the physical motion (mm, drawn on 0 to 220 with the physical motion at 0) and the component κ toward the model's own image-to-video version with the same and with another seed. Last column: outputs of all versions that became the model's own version ($\kappa \geq 0.5$), tracks and deflectors, one bar each.

level	d_{phys}	κ same	κ other	own version
▼ LTX 0.7	 17 [10, 31]	0.01	0.01	 0/24; 0/23
▼ LTX 0.8	 76 [33, 94]	0.03	0.04	 3/24; 6/22
▼ LTX 0.9	 124 [109, 175]	0.25 [0.21, 0.66]	0.29 [0.03, 0.32]	 12/24; 9/18
■ Wan 0.90	 21 [14, 42]	0.00	0.00	 0/24; 2/18
■ Wan 0.94	 77 [50, 104]	0.03	0.03	 1/21; 9/19
■ Wan 0.95	 105 [41, 211]	0.04 [0.00, 0.13]	0.08 [0.02, 0.52]	 2/23; 16/21

video of the scene nearest its median reading of the level (the ramp exponent on the straight level, gain and phase at 3 cycles per metre on the others). On a deflector level it shows a video with the model's usual tray outcome and the first-contact restitution nearest its median, and in the interventions one seed (for the closed models, one generation index) that is trackable in every arm with the model's usual outcome. A model with no readable video of a scene shows its most typical flagged video, tagged as an altered scene. Videos in which the camera moves are compensated frame

by frame with the measured board homography. Fig. 4 of the main paper shows four of the models on five of these scenes.

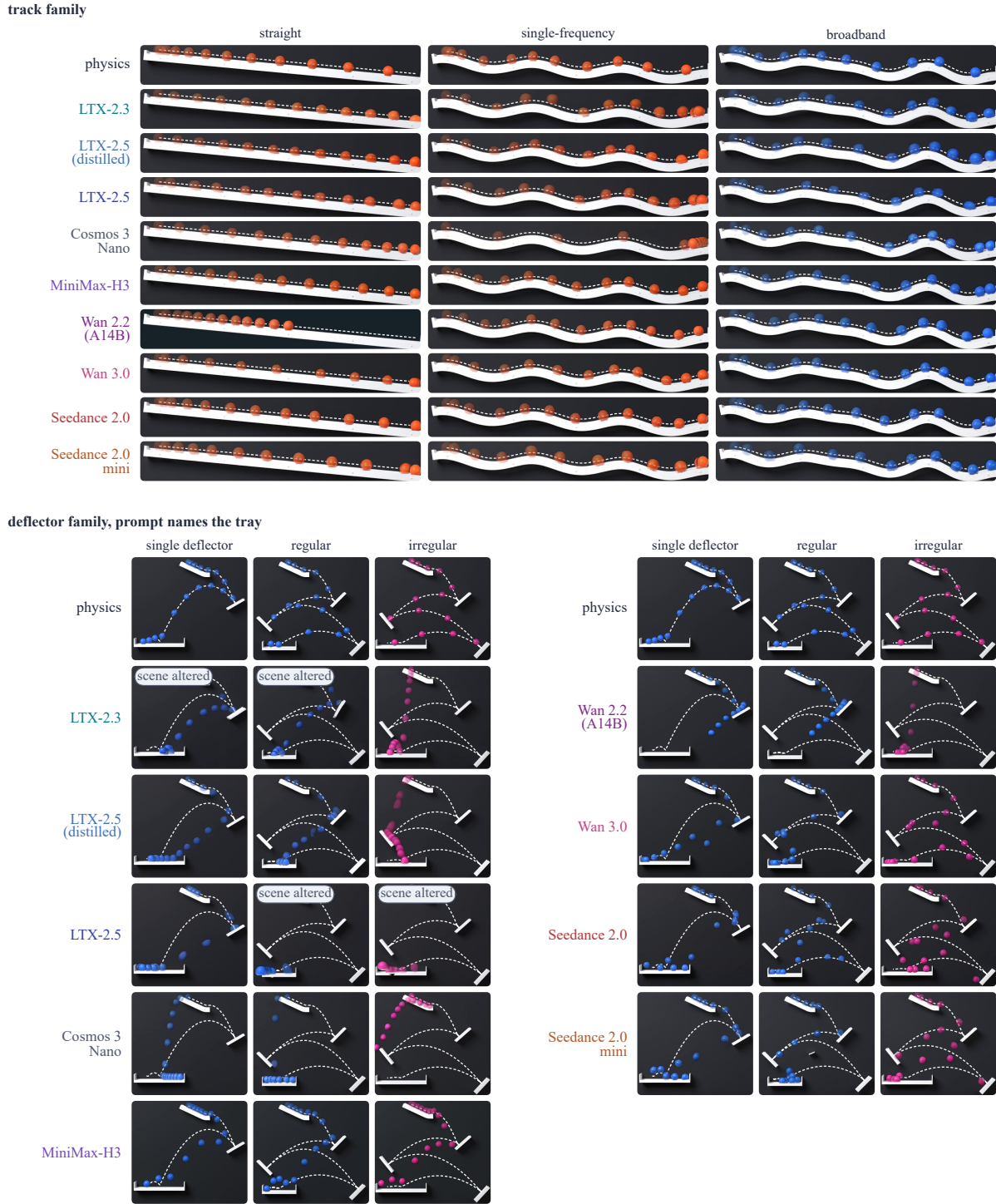


Figure G.1. **Main geometry families, all models.** Top: representative straight, single-frequency and broadband track scenes. Bottom: single-, regular- and irregular-deflector scenes under the named-outcome prompt. Dashed curves show physical trajectories; tagged cells mark models with no readable video of that scene.

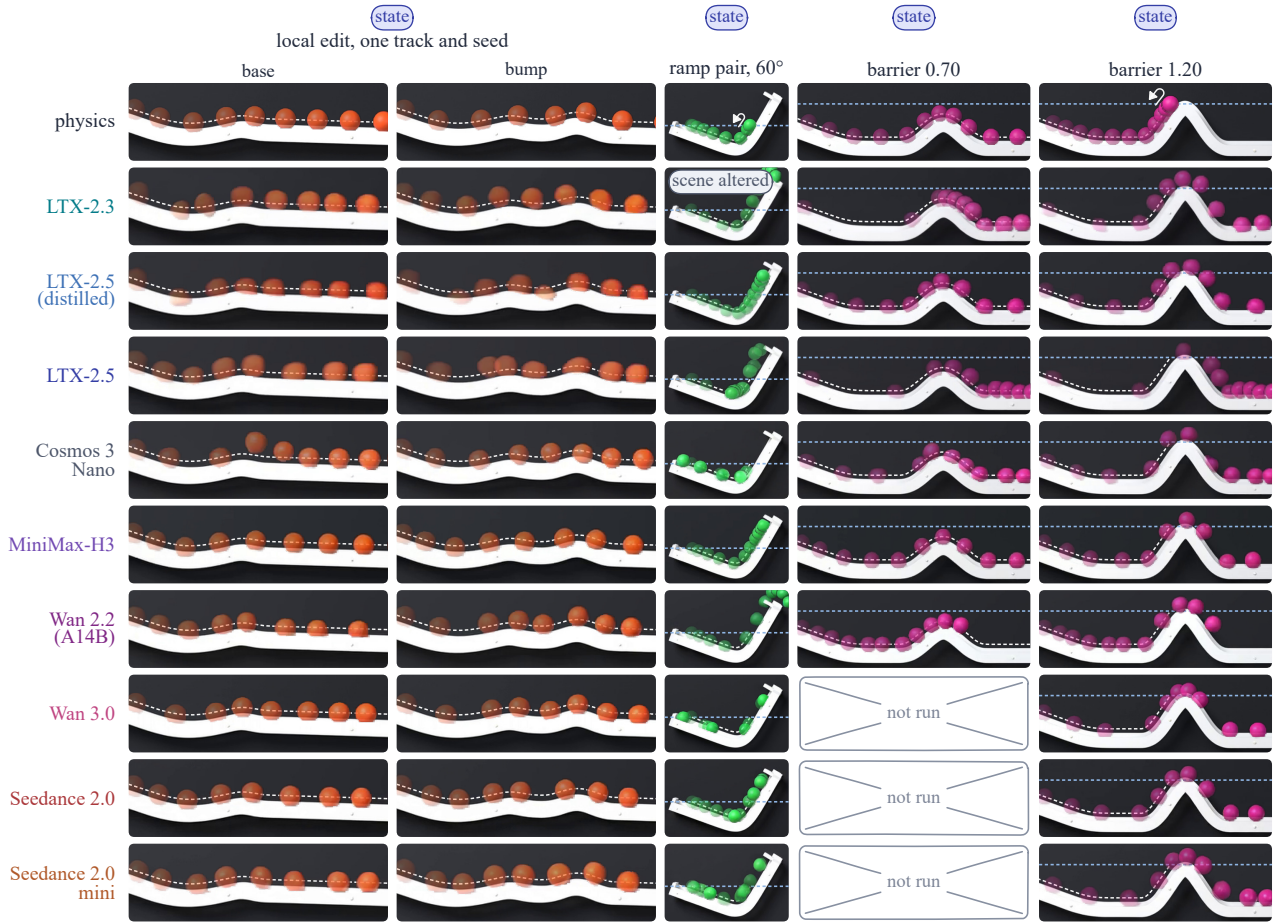


Figure G.2. **Interventions on state, all models.** Columns: the local edit (base and bump) on the one track that all nine models ran, the ramp pair with a 60° ascent, and the barriers at 0.70 and 1.20 of the release drop (the closed models ran 0.85, 1.05 and 1.20). Rows: the simulator and the nine models. The straight dashed line marks the release height and the arrow the physical turning point.

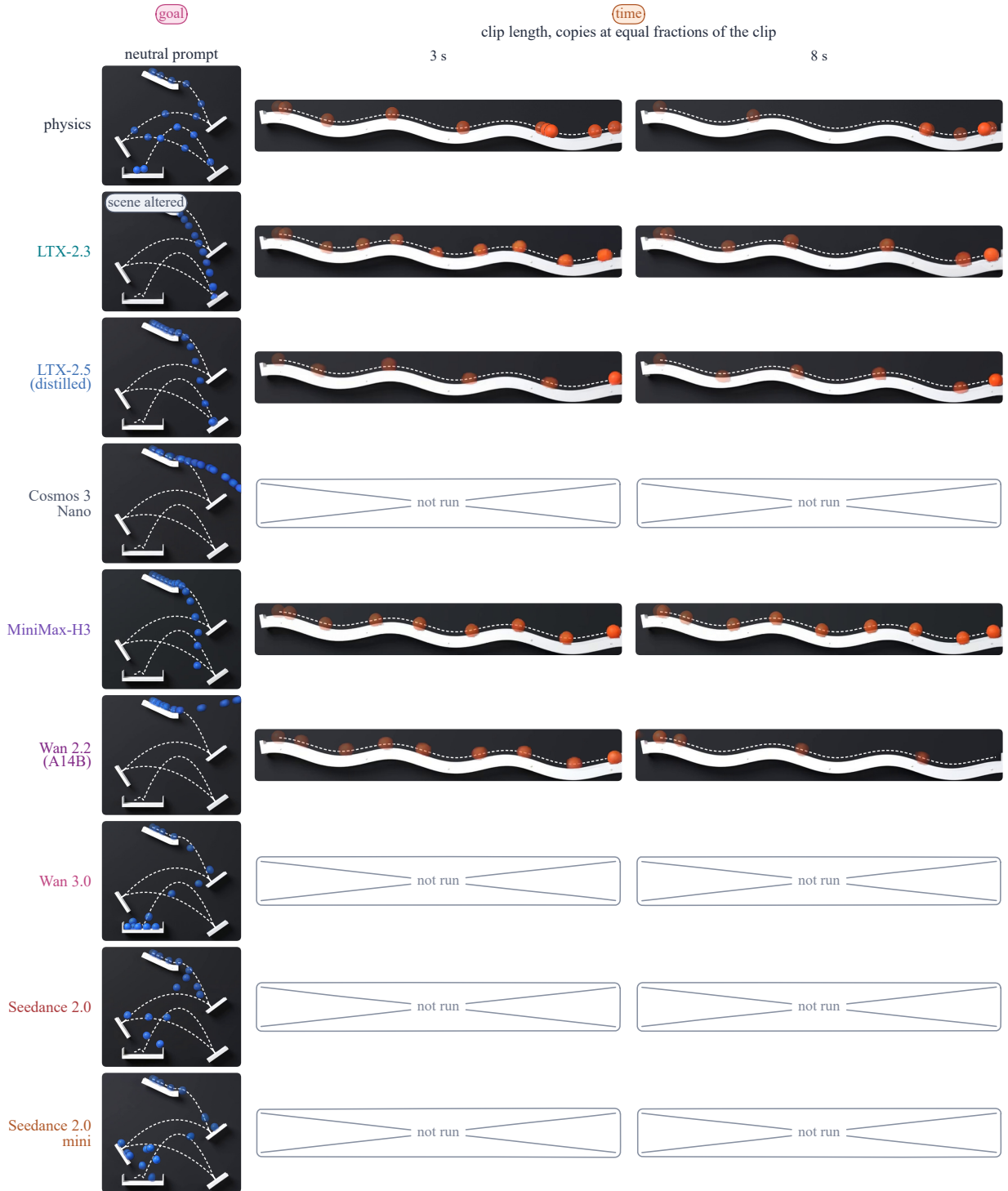


Figure G.3. **Interventions on goal and time.** Left: an irregular three-deflector scene under the prompt that leaves the outcome unnamed, from the prompt pairs, for the simulator and the eight models that ran them. Right: one single-frequency track at 3 and 8 s for the four open-weight models of the clip-length arm, with the ball drawn at equal fractions of each clip, so that a motion scaled to the clip gives the same spacing at both lengths. The tagged cell (LTX-2.3) shows an altered scene.

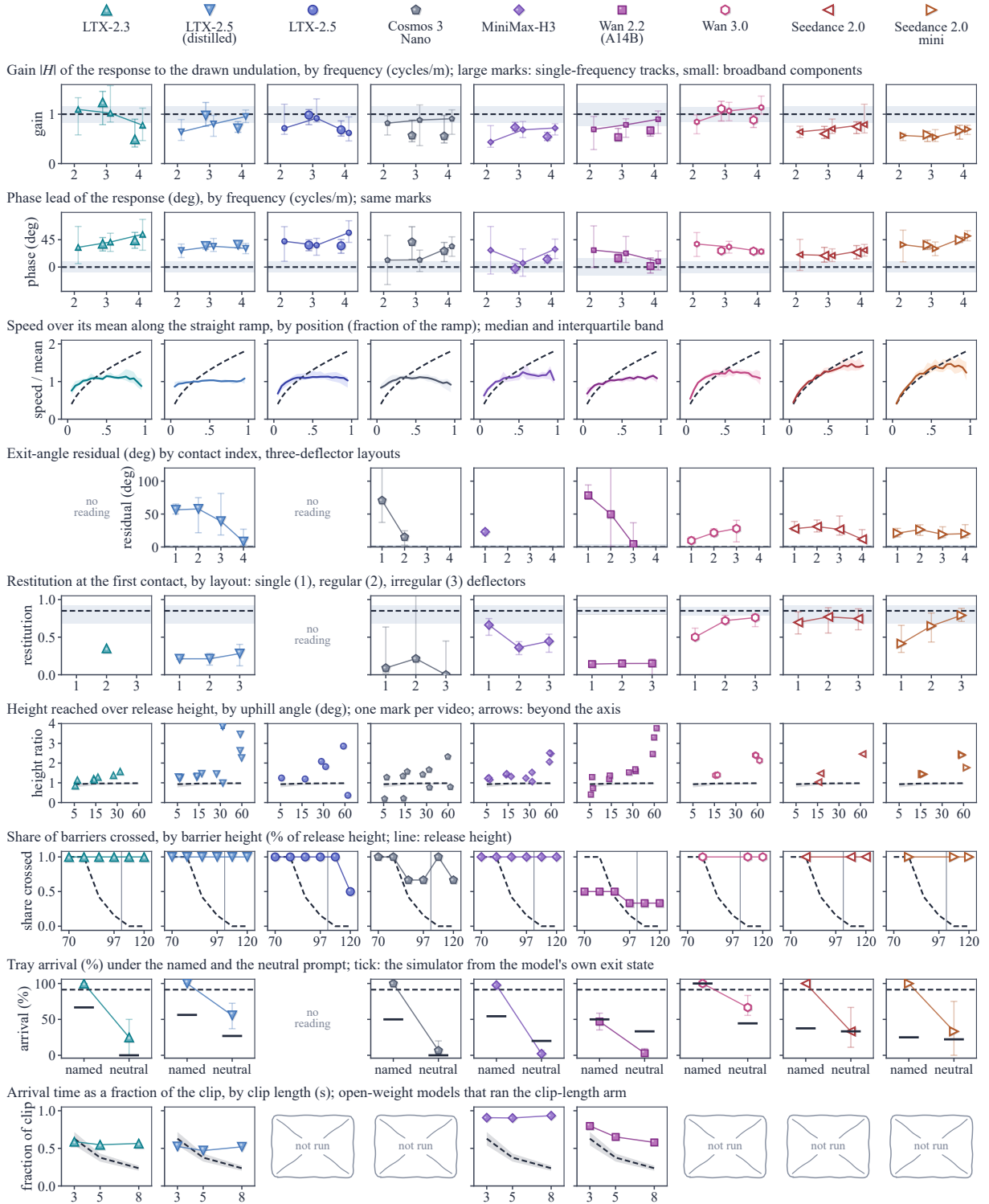


Figure G.4. **Per-model readings of every experiment.** One column per model and one row per reading, named above each row, with the physical value dashed. In the first two rows the grey band is the largest floor of the three undulating track levels at the model's frame rate. In the last row the physical band is the track-family transit time (1.62–2.15 s) divided by the clip length, dashed at its midpoint.



Figure G.5. **Recurring integrity failures.** Each pair shows the first frame and a later frame of a representative flagged video, the one nearest the median change among the flagged videos of the model shown. The later frame is the frame of largest change; for the shifted track it is shown as the amplified difference from the first frame, and for the duplicated ball it is a frame with both balls. Dashed outlines mark where the part was in the first frame. The examples show a rotating deflector, tilting tray, rotating launch ramp, shifted track, camera dolly, and ball duplication; counts and, where measured, magnitudes are reported in Table D.2.

H. Statistics

The unit of every analysis is a scene with its seeds, and every interval comes from resampling scenes. Paired designs remove what two videos share, permutation tests answer ordering questions, and a reading counts as a property of a model only when its interval clears the floor at the model’s frame rate and output size.

H.1. Units and intervals

Per-video readings are summarized by their median over the estimable videos of a cell, phases by their circular mean with the resultant length R , and binary outcomes (exact order, tray arrival) by their share. Intervals are 95% percentile intervals from 2000 bootstrap resamples of scenes, in which the seeds of a scene are drawn together, so that the spread between scenes, the larger source of variation, enters every interval. Within a video, the gain and phase come from a regression whose coefficients carry a Newey–West covariance. Ratios of coefficients have Fieller intervals, and the gain, phase and $2f/1f$ intervals come from 4000 draws of the coefficients. Every table carries its denominators (Sec. D.1), so that small cells can be read as such. Closed models contribute three to six scenes to each intervention family, and their readings are stated as intervals over those scenes.

H.2. Paired designs

Same seed. For the open-weight models the arms of an intervention share scene and seed, so each pair differs only in the changed element. Local edits pair every arm with the base video of the same track and seed and take the median over pairs, with a bootstrap over tracks. The rerun arm gives the sampler’s own floor, which is zero for every open-weight model except LTX-2.3. Prompt pairs match scene, first frame and seed, and differences in arrival are paired. The steering compares the two prompts through their means per scene, since the videos with a wrong route differ between prompts. Capture contrasts pair the in-tray and beyond scenes of the same base and seed, with a bootstrap over base scenes. Clip-length differences (3 minus 5 s, 8 minus 5 s) are paired by scene and seed, and the duration exponent is the least-squares slope of $\log \alpha$ on $\log T_{\text{clip}}$ after removing the mean of each scene and seed, with a bootstrap over scenes.

Independent samples. Closed models do not reproduce a seed, so an intervention compares the mean of an arm with the mean of the base over the same scenes, and the difference carries the variation between samples, including their tempo. Photographed first frames compare each photograph row with its control row per configuration, with a bootstrap over configurations and the seed-spread test of Sec. F.7, in

which s pools the seed deviations of the model around the centres of all its rows. The destination counterfactual compares the model’s arrival with the arrival of the stated law from the same exit states, with a bootstrap over scenes.

H.3. Permutation and stratified resampling

On chutes the four curvatures of one seed are exchangeable units for an open-weight model, so the p -value of the landing slope comes from permuting the curvature labels within each seed (across all videos for closed models). The interval in Table F.12 resamples seeds within each curvature and leaves out the spread between chutes, so it is narrower, and the permutation p carries the decision. On ramp pairs and barriers the intervals resample seeds within each scene and keep the design of angles and heights.

H.4. Reading intervals against floors

A reading is interpreted as a property of a model when its interval lies outside the floor at the model’s frame rate and size (Sec. C.4). Single-pair floors of the local edits (2.4–5.6% in dv) are of the size of the dip effect itself, so these readings rest on medians over about 36 pairs, whose error on the simulated motion stays within a few tenths of a percent. The null edit is flagged when the interval of its median difference from the base lies wholly beyond the larger of the rerun floor and the measurement floor (1.0% for dv , 0.4% for dt), a rule that flags none of 120 noise-only runs and all 120 runs in which the null arm carries the bump response. Clip-length differences are read as the larger of $|3s - 5s|$ and $|8s - 5s|$ over its floor. Readings marked exploratory (the locked and fixed parts of the shared-front response, the front and back leads, the relaxation rate in physical time, the decomposition of regenerated motion, and the mechanism predictions on other families) were added after the design was fixed. They inform the interpretation and carry no claim on their own.