SEPARATING COMMAND ERROR, CHAINING COST AND ONE-CALL ERROR IN THE ROLLOUT EVALUATION OF A VIDEO WORLD MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Action-conditioned video world models are usually evaluated by rolling them out under logged actions and plotting the error against horizon, and the growth of that error is attributed to errors compounding through the model's own outputs. The rollout score alone cannot support that attribution, because it has no reference: it does not show whether the evaluation gave the model the right commands, how much of the growth a single call to the same moment would already show, or whether the actions mattered. For models that take a time shift as input, three additional generations on the same items provide these references without retraining: a rollout with the commands recomputed as in training, one call to each scored horizon, and a rollout that repeats the first command. We applied them to the public evaluation of a released family of navigation world models (NWM; 49M to 1.01B parameters), on fresh test cohorts and with decision rules registered in advance. Correcting the released rollout code, which gives every call after the first a wrong command, lowers the 1B model's 4 FPS rollout LPIPS by 18.8% [16.9, 20.6]; in a post-hoc analysis the growth of the command error accounts for a third of the released curve's rise from 2 to 8 s. The release's authors attribute a crossover between their 1 FPS and 4 FPS rollouts to accumulated errors and loss of context. On the release's own 20-trajectory protocol, the released 1B checkpoint shows this crossover under the released commands only in the point estimates, and it disappears once the commands are corrected (post hoc; we cannot tell whether their figure used this code). With correct commands, the growth of the error that chaining adds over a single call accounts for between a sixth and two fifths of the rise from 2 to 8 s, and the two cohorts' estimates of that growth disagree beyond their intervals (post hoc). The larger share's post-hoc interval reaches about one half, so removing this chaining cost while leaving the one-call error unchanged would leave at least about half of the rise. Scale lowers the one-call error, although the released 1B checkpoint is worse than the 700M one on single calls (post hoc). The effect of scale on the chaining cost is unresolved on LPIPS; on DreamSim the cost decreases. We recommend reporting rollout scores together with these references.

1 INTRODUCTION

An action-conditioned video world model is evaluated by converting the logged actions of a recorded trajectory into a command for each model call, rolling the model out, and comparing the generated frames with the recorded video. The error grows with horizon, and the growth is usually attributed to errors compounding through the model's own outputs, since each generated frame becomes the context for the next call. The release audited here interprets its own rollout curves this way (Bar et al., 2025), and the same interpretation motivates the standard remedies for drift: training on the model's own outputs (Bengio et al., 2015; Huang et al., 2025), noise augmentation of the context (Valevski et al., 2025) and per-frame noise schedules (Chen et al., 2024; Ruhe et al., 2024).

A rollout score has no reference, so it cannot support this interpretation by itself. If the logged trajectory is converted into commands incorrectly, every call is mis-commanded, and the resulting error grows with the number of calls just as compounding error would; the score cannot tell the two

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

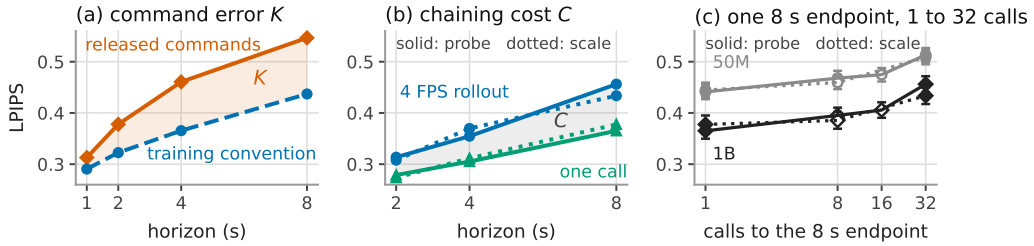


Figure 1: (a) The command error K is the band between the 1B model's 4 FPS rollout under the released commands and under the training convention; its growth from 2 to 8 s is a third of the released curve's rise (post hoc; confirmation cohort, 200 trajectories). (b) The chaining cost C is the band between the same model's corrected 4 FPS rollout and a single call to the same moment; its growth from 2 to 8 s is two fifths of the corrected rollout's rise on the probe cohort (solid lines, band) and a sixth on the scale cohort (dotted), 150 trajectories each (post hoc). (c) The 8 s endpoint reached with corrected commands in 1, 8, 16 (probe cohort only) and 32 calls by the 1B and 50M models on both cohorts, on a logarithmic axis; the 8- and 16-call rollouts (open markers) condition on generated frames 1 s and 0.5 s apart, a spacing the model never saw (Section 3.1). Bars are 95% trajectory-bootstrap intervals of each mean, not of paired differences; (a) and (b) show means only.

apart. Nor can it separate the cost of chaining calls from the difficulty of predicting further ahead in a single call, or show whether the actions mattered, since a model that ignored them and simply continued the video would still be scored, with nothing in the score to show it.

For models conditioned on both an action and a time shift, three extra generations on the same items supply the missing references without retraining (Section 3): one call to the same moment, a rollout with the commands recomputed as in training, and a rollout with the first command repeated. The first two split the released rollout's error into three terms, $E_{\text{released}} = J + C + K$: the one-call error, the chaining cost and the command error. The two command references share the rollout's first-call noise, and the one-call reference is paired with it by trajectory. None of the three is new as an idea; our contribution is to use them as references for the public evaluation of a released family of navigation world models at four sizes (Bar et al., 2025), with decision rules registered before each cohort was read.

The released rollout code tells every call after the first to move 1.25 waypoint units too far forward, and it expresses every command in the first frame's axes rather than the current frame's (Section 4). Recomputing the commands in the training convention lowers the 1B model's 4 FPS rollout LPIPS by 18.8% (registered claim H1). Post hoc, the growth of the command error from 2 to 8 s is a third of the released curve's rise (Figure 1a), and, for the released 1B checkpoint on the release's own 20-trajectory protocol, the crossover between 1 FPS and 4 FPS rollouts that the release attributes to accumulated errors and loss of context appears under the released commands only in the point estimates and is absent once the commands are corrected. That comparison also changes the context spacing, and we cannot tell whether the release's figure came from this code.

With correct commands, the growth of the chaining cost is two fifths of the corrected curve's rise from 2 to 8 s on one cohort and a sixth on the other (post hoc; Section 5.1, Figure 1b), so on the point estimates the one-call error accounts for the larger part of the rise. The larger share's post-hoc interval reaches about one half, so a remedy that removed the chaining cost and left the one-call error unchanged would leave at least about half of the rise. The command error, again post hoc and on a different cohort, grows with horizon by about as much as chaining does on one cohort and by more on the other.

The logged actions lower the 1B model's rollout error relative to a repeated first command (registered claim H6, MATERIAL; Section 5.2), so the rollouts use the actions; the repeated command is a gross perturbation, however, and the test leaves open whether the score resolves small action errors. From 50M to 1B the one-call error falls (registered claim H10), while the registered test of whether the chaining cost falls between the same two sizes is inconclusive (H9), and across all four sizes the chaining cost shows no detectable trend on LPIPS and a decrease on DreamSim (Section 6). Post

hoc, the released 1B checkpoint's one-call error exceeds the 700M checkpoint's, on the release's own single-step items and on a fresh cohort.

The study covers one released family and one dataset, with LPIPS as the registered metric, DreamSim alongside and one checkpoint per size; both metrics measure distance to the recorded future rather than frame quality, and the fresh cohorts stop at 8 s. Because the terms are defined as differences between generation arms, a rollout score without references cannot separate them in any family. The two command references need neither a time shift nor actions that compose; each costs one more rollout per item, and the one-call reference costs one call per item and scored horizon.

2 Setting and related work

The family. The models are the conditional diffusion transformers (Peebles & Xie, 2023) of Navigation World Models (NWM), released with weights and evaluation code (Bar et al., 2025) at four sizes: 49M, 196M, 687M and 1.01B parameters, counted from the released architecture (we write 50M, 200M, 700M and 1B below). Each model takes four consecutive context frames, an action and a time shift of up to 64 frames in either direction, and generates the frame at that shift. The action (x, y, θ) is the goal displacement in the current camera frame, in normalised waypoint units, and the heading change in radians; we refer to the action passed to a given call as its command. The same checkpoint can therefore predict a frame directly, in one call, or autoregressively, through a chain of shorter calls. We use the RECON navigation dataset of Shah et al. (2021), one of the datasets the release trains and evaluates on and the one its rollout figure uses, at 4 FPS with a waypoint unit of 0.25 m. We sample as released (ancestral sampling, 250 denoising steps, no guidance) and report LPIPS (Zhang et al., 2018), the registered primary metric, with DreamSim (Fu et al., 2023) alongside. According to the maintainers, the released checkpoints were retrained after face blurring. The four released checkpoint files we use, one per size, carry the same training-step count in their names (Appendix D). Every number in this paper describes the release, and averages over horizons are over 2, 4 and 8 s unless a horizon is named.

Related work. Action-conditioned video prediction and learned simulators go back at least to Oh et al. (2015) and Finn et al. (2016), and recent video-generative models cover games, robotics and driving (Yang et al., 2024; Hu et al., 2023; Wu et al., 2024; Bruce et al., 2024; Alonso et al., 2024). Compounding error in iterated prediction has been studied in direct-versus-iterated forecasting (Marcellino et al., 2006; Chevillon, 2007), where direct forecasts are in theory more robust to a misspecified one-step model, and in model-based reinforcement learning (Talvitie, 2014; Venkatraman et al., 2015; Hafner et al., 2019; Asadi et al., 2019; Janner et al., 2019; Lambert et al., 2021; 2022). The time-conditioned dynamics models of Lambert et al. (2021) are the closest analogue of our one-call reference. Training on the model's own outputs, one of the remedies for drift named in Section 1, was introduced for imitation and sequence models (Ross et al., 2011; Bengio et al., 2015; Ranzato et al., 2016; Lamb et al., 2016) and carried over to video diffusion (Huang et al., 2025). Memory mechanisms address forgetting (Xiao et al., 2025; Zhang & Agrawala, 2025). In language modelling, He et al. (2021) found that autoregressive distortions are less incremental than exposure bias predicts, and we ask the same question of a video world model. Harvey et al. (2022) compare sampling schemes that reach the same frames of a video through different sequences of model calls, the closest precedent for our comparison of rollouts with different numbers of calls, and Genie's Δ_tPSNR (Bruce et al., 2024) is a precedent for asking whether actions matter. The audit in Section 4 follows earlier re-measurements of published evaluations (Parmar et al., 2022; Engstrom et al., 2020; Agarwal et al., 2021; Musgrave et al., 2020). The release itself already reports single-call errors on its own items. What we add is the registered use of the three references as paired references, with single calls and rollouts on the same items and horizons and the command references sharing the rollout's first-call noise.

3 Method

3.1 The three references

Write $E(h, N)$ for the error at horizon h when the model reaches that moment in N calls of h/N each, with commands taken from the logged trajectory in the model's training convention. The first

reference, a single call to horizon h, gives the one-call error $J(h) = E(h, 1)$ (the release calls this single-step prediction), and we define the chaining cost C by

$$E(h, N) = J(h) + C(h, N), \qquad C(h, N) = E(h, N) - E(h, 1).$$

The decomposition holds by definition and does not identify a mechanism. Every rise and every share of a rise in this paper is measured from 2 s to 8 s, the shortest and longest horizons in the registered average (one-call frames at 1 s were generated but are not in it).

The second reference is the rollout $E(h, N)$ itself, with commands recomputed in the training convention, generated on the same items and with the same first-call noise as the released rollout. The difference $K(h, N) = E_{\text{released}}(h, N) - E(h, N)$ is then the command error, the effect of the released evaluation's command composition on the score, and it is zero at $N = 1$ by construction. The third reference repeats the first call's command at every later call, and its difference from E measures what the per-step actions are worth to the score.

Reading C as the cost of chaining assumes that the rollouts being compared differ only in the number of calls. This holds at 4 FPS: every shift is within the trained range, and every call conditions on four consecutive frames at the training rate, first real and then generated. The 2 FPS and 1 FPS rollouts condition on generated frames 0.5 s and 1 s apart, a context format the model never saw, so their C mixes chaining with a change of context spacing. We report these rollouts but base no claim about chaining on them, and the same applies to the 4 FPS versus 1 FPS comparisons in Section 4, which repeat the comparison the release reports. At 4 FPS, too, C contains the effect of fresh sampling noise at each later call and of the autoencoder round trip, in which every generated frame is decoded to pixels and encoded again before it serves as context. J is not an irreducible floor either: it comes from the same model asked for a long shift, so its growth with horizon includes whatever the model does worse at long shifts, as well as the uncertainty of a single recorded future. Finally, the one-call reference assumes that the logged actions compose into a single command that determines the target frame. This holds for planar displacement in a static scene; it does not hold for, say, a game controller.

3.2 Pairing and resampling

At a given rollout rate, the rollouts that differ only in their commands share the diffusion noise of the first call, so their first frames are bit-identical. A committed check verifies this on every generated set and recomputes every command from its label and the logged poses. Predictions that reach the same horizon in different numbers of calls are paired by trajectory (same context, same target), which removes between-trajectory variance but not sampling noise. Intervals come from resampling whole trajectories, never frames (20,000 draws, percentile intervals), and differences are computed before rounding. Each of the five fresh cohorts is a set of test trajectories drawn from a registered permutation, disjoint from every other cohort and from the index files shipped with the release, and read once under a plan committed before its first frame was generated. We call them confirmation (200 trajectories; H1 and H2), audit (150; the voided claims and the single-call command-following slopes), probe (150; H6 and chaining), dose (150; H8) and scale (150; H9, H10, chaining and the four sizes). The release's own item files are read as shipped, and Table 1 in Appendix D gives the details.

3.3 Registration

All claims adjudicated in this paper (H1, H2, H6, H8, H9, H10) were registered before the data behind them were generated, each with an estimand, a threshold, a control and a kill condition under which the test would be void. Appendix A lists them together with the registered claims that were voided and never run: two because no registered turn command lay within the model's training support (in-place turns over four or more frames are outside it), and four because the instrument they relied on failed its validation gate (Appendix C). Appendix A also lists the post-hoc analyses, which are labelled as post hoc wherever they appear. No familywise adjustment was registered, and a Bonferroni adjustment over the six claims changes none of the verdicts (Appendix G). H1, H6, H8 and H10 reached decisive verdicts: the released commands inflate the 1B model's rollout error materially (H1), the score registers gross action errors (H6), the forward offset alone is harmful in

the correct frame (H8), and the design detects scale in the one-call error (H10). H2 came out MIXED and H9 INCONCLUSIVE, and the readings of the curve that build on them are post hoc.

4 COMMAND ERROR IN THE PUBLIC EVALUATION

4.1 THREE DEFECTS

In the training convention, each action is the goal displacement in the frame of the call that executes it, in waypoint units, normalised with an affine map whose forward range is $[-2.5, 5]$. The released rollout evaluation instead expresses the cumulative waypoints in the first frame's axes, normalises them, differences the normalised values and sums the differences per stride. Because differencing after normalising removes the normalisation's intercept, every call after the first is told to move $\Delta X + 1.25$ waypoint units forward, where 1.25 is the midpoint of the forward range (Appendix B gives the arithmetic). The displacements also stay in the first frame's axes, so the commands become wrong as soon as the robot turns. Finally, the cumulative heading is wrapped before differencing, and a clip that crosses $\pm\pi$ receives one call whose heading command is off by 2π.

The defective composition affects every rollout, and every rollout-based comparison, produced with the released rollout evaluation code. It does not affect the released code's single-step evaluation, whose direct time-shift path sums from the first row and is correct, or the released planner, which we did not evaluate, at its default stride: the planner gives each call a correctly normalised command, although it scores a plan as a polyline in the first frame's axes. The release describes its actions as relative to the current location, which is the training convention, and the rollout code departs from that description. We believe this is an implementation error rather than a deliberate choice, and we are notifying the maintainers. Our measurements cannot show whether the numbers in the release's own report came from this code path.

4.2 EFFECT OF CORRECTING THE COMMANDS

On the confirmation cohort, the 1B model's 4 FPS rollout has a mean LPIPS over 2, 4 and 8 s of 0.4617 under the released commands and 0.3750 under the training convention. The relative reduction is 0.1879 [+0.1692, +0.2064], which H1's registered rule classifies as POSITIVE, and the absolute difference is 0.0867 [+0.0776, +0.0958] (interval post hoc). On DreamSim the relative reduction is 0.3269 [+0.2957, +0.3566]. At 1 FPS, where the same horizon takes a quarter as many calls, the LPIPS reduction is 0.1380 [+0.1198, +0.1564]. At 8 s the released and corrected rollouts score 0.5466 and 0.4370; Figure 1a gives every horizon.

The command error grows from 0.0554 [+0.0451, +0.0657] at 2 s to 0.1096 [+0.0953, +0.1241] at 8 s (intervals post hoc). Over that span the released curve rises by 0.1687 [+0.1539, +0.1836] and the corrected curve by 0.1145 [+0.1001, +0.1287], so the command error's growth of 0.0542 [+0.0383, +0.0703] is a third of the released curve's rise (0.3211 [0.2372, 0.4025]). For the 200M model its share of the rise is a quarter (0.2358 [0.1537, 0.3117]). These rises, growths and shares are post hoc, as is every such quantity in this paper.

The release compares its 4 FPS and 1 FPS rollouts and reports that the 1 FPS variant is better at first, but that "after 8 seconds, predictions degrade due to accumulated errors and loss of context and the 4 FPS becomes superior" (Bar et al., 2025). We repeated this comparison post hoc for the released 1B checkpoint on the release's own protocol and split, which has 150 items on 20 trajectories, so the intervals below are approximate and probably too narrow. On this protocol, correcting the commands lowers rollout LPIPS by a relative 0.1573 [+0.1276, +0.1874] at 4 FPS and 0.1389 at 1 FPS. Under the released commands the point estimates reproduce the crossover, although the interval does not exclude zero: at 16 s, 4 FPS minus 1 FPS is -0.0175 [-0.0392, +0.0031] (0.5491 against 0.5666). Under corrected commands the 1 FPS rollout is better at 16 s (4 FPS minus 1 FPS +0.0419 [+0.0161, +0.0696]; 0.4874 against 0.4456), so the crossover is absent; the interaction, the released difference minus the corrected one, is -0.0594 [-0.0902, -0.0297]. The comparison also changes the context spacing (Section 3.1), and we cannot decide whether the release's figures came from this code path: a co-author's answer on the release's issue tracker names the shipped rollout file, which the released code sends through this path, but not the code that produced the figure. We did not re-run the comparison with DIAMOND (Alonso et al., 2024) that appears in the same figure of the release.

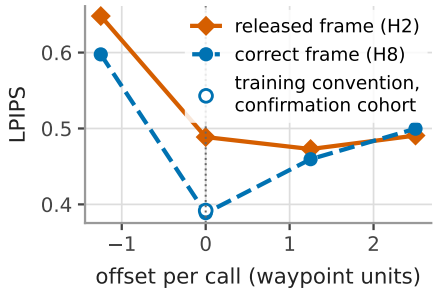


Figure 2: Zero offset is best in the correct frame convention and +1.25 in the released one. Mean 4 FPS rollout LPIPS of the 200M model against the forward offset added to every call after the first, with displacements in the first frame's axes (released convention; H2, confirmation cohort) or the current frame's (correct convention, used in training; H8, dose cohort). The released evaluation is the +1.25 point of the released-frame curve, and the open marker is the confirmation cohort's training-convention arm. The curves come from different cohorts, so their vertical gap is not a paired difference.

4.3 Forward-offset doses in the two frame conventions

Registered claim H2 applied four doses of the offset in the released frame convention with the 200M model: −1.25 units (N1), 0 (O, the offset removed), the released +1.25 (R) and +2.5 (D2). It predicted that the four differences N1 minus O, R minus O, D2 minus O and D2 minus R would all be positive (error lowest at zero offset and rising as the offset grows in either direction), and its verdict was MIXED. Two of the four differences went in the registered direction, D2 minus O was unresolved, and R minus O went the opposite way: the released arm scores lower than the arm with the offset removed (R minus O −0.0158 [−0.0238, −0.0077]; arm means 0.4729 and 0.4887). The negated offset is far worse (N1 minus O +0.1595 [+0.1483, +0.1705]), and the doubled offset is slightly worse than the released one (D2 minus R +0.0179 [+0.0105, +0.0253]). The two defects interact: with displacements in the first frame's axes, the commanded forward component shrinks as the robot turns away from its initial heading, and a forward surplus restores part of it. A post-hoc analysis of clips that do not turn over the horizon (absolute net heading change of at most 10 degrees and chord over path of at least 0.95, a rule fixed before the subset was formed) supports this account: on those clips the surplus is harmful, with R minus O +0.0178 [+0.0099, +0.0259] at 1 s (52 clips) and +0.0271 [+0.0121, +0.0417] at 2 s (37 clips).

Registered claim H8 put the same question to a fresh cohort with the frame convention correct (Figure 2). There the same +1.25-unit surplus raises the 200M model's rollout LPIPS from 0.3884 to 0.4595, by 0.0711 [+0.0607, +0.0814]. H8's estimand, the relative reduction that the logged commands give, is 0.1547 [+0.1332, +0.1757], which its registered rule classifies as HARMFUL. The 1B model, with five times as many parameters, pays about as much: the surplus costs it 0.0701 [+0.0598, +0.0802], and the logged commands' relative reduction is 0.1575 [+0.1361, +0.1785]. This comparison between sizes crosses the two GPU types. The best arm is zero offset in the correct frame convention and +1.25 in the released one. Both responses are asymmetric, so we report the ordering of the arms instead of a fitted optimum. The registered expectation that a quadratic fit would place the optimum within 0.3125 units of zero failed on both cohorts (Appendix E).

5 Chaining and the action channel

5.1 One endpoint reached in 1 to 32 calls

With commands in the training convention throughout, we ask the 1B and 50M models for the frame 8 s ahead in one call and for the same frame in 8, 16 and 32 calls (Figure 3 shows one clip); the 8- and 16-call rollouts change the context spacing, so we base no claim on their costs (Section 3.1). On the probe cohort the mean LPIPS at 8 s is 0.3650, 0.3951, 0.4056 and 0.4562 for the 1B model, and 0.4410, 0.4680, 0.4747 and 0.5133 for the 50M model. On the scale cohort, with 1, 8 and 32 calls, it is 0.3776, 0.3854 and 0.4336 for the 1B model and 0.4446, 0.4595 and 0.5087 for the 50M model (Figure 1c). On the registered average over 2, 4 and 8 s, the 4 FPS chaining cost of the 1B model is 0.0585 [+0.0497, +0.0672] on the probe cohort and 0.0496 [+0.0404, +0.0585] on the scale cohort; for the 50M model it is 0.0565 [+0.0483, +0.0648] and 0.0535 [+0.0448, +0.0619]. As a fraction of the one-call error it is 0.1545 [0.1238, 0.1859] (1B, scale cohort).



Figure 3: One clip's 8 s frame as the 1B model predicts it in 1, 8 and 32 calls (scale cohort), next to the recorded frame: the one-call frame invents a paved path, the 8- and 32-call frames invent a building, and LPIPS rises by 0.07 across the row. The clip is the one closest to the cohort medians among clips whose error rises from 1 to 8 to 32 calls, as the cohort's does. The 8-call rollout changes the context spacing (Section 3.1). The recorded frame is centre-cropped to a square for display. It is not the scored target, which goes through the model's resize and crop, so its differences from the generated frames in field of view and border come from the display.

At 8 s, the longest horizon of the fresh cohorts, the 1B model's 32-call cost is 0.0911 [+0.0752, +0.1070] on the probe cohort and 0.0560 [+0.0408, +0.0711] on the scale cohort. Post hoc, these are one fifth (0.1997 [0.1673, 0.2314]) and one eighth (0.1292 [0.0954, 0.1624]) of the rollout's error at that horizon. In a post-hoc check of the sixteen chaining quantities measured on both cohorts (two models, two rates, three horizons and their mean), this pair is the only one whose intervals do not overlap. The 50M model's 8 s costs are 0.0723 and 0.0641. The two 1B runs were made on different GPU types, but their 2 s and 4 s costs agree within their intervals (0.0348 against 0.0343, 0.0495 against 0.0584), and the 50M model, generated on one GPU type for both cohorts, agrees within its intervals at 8 s. Cohort variation, hardware numerics and chance cannot be separated after the fact, so we quote the 8 s cost as a range and do not rely on its exact value.

Measured against the rise of the curve instead of its level, chaining is not a small part of it (post hoc, LPIPS only). From 2 to 8 s the corrected 4 FPS rollout on the probe cohort rises by 0.1426 [+0.1262, +0.1589], of which the one-call error accounts for 0.0862 [+0.0696, +0.1029] and chaining for 0.0564 [+0.0382, +0.0745], or two fifths (0.3953 [0.2819, 0.5022]). On the scale cohort the rise is 0.1257 [+0.1090, +0.1423], and chaining accounts for 0.0217 [+0.0059, +0.0377] of it, a sixth (0.1725 [0.0491, 0.2873]). For the 50M model the shares are 0.2488 [0.1448, 0.3441] and 0.1706 [0.0558, 0.2763]. The cohorts disagree on the 1B model's chaining growth, as they do on its 8 s cost, although the shares overlap narrowly. On the point estimates the one-call error accounts for the larger part of the rise in every case, though the probe interval reaches one half. The command error of Section 4 grows with horizon by about as much as chaining does on one cohort and by more on the other (0.0542 against 0.0564 and 0.0217; the cohorts differ, so the comparison is not paired).

A post-hoc analysis of which clips have a high chaining cost finds no clear association with turning. The rank correlation between the per-clip 8 s cost and the accumulated absolute heading change is +0.033 and −0.134 for the 1B model on the two cohorts and +0.055 and +0.045 for the 50M model. The 1B model's least-turning quartile has a mean 8 s cost of 0.0832 on the probe cohort, against 0.0911 over all clips, and 0.0874 on the scale cohort, against 0.0560, so chaining is not confined to the clips that turn most.

5.2 Dependence on the per-step actions

On the probe cohort, replacing the logged per-step actions with the first call's command, repeated at every later call, raises the 1B model's rollout LPIPS from 0.3748 to 0.4951, by 0.1203 [+0.1083, +0.1325] (interval post hoc). H6's estimand, the relative reduction that the logged actions give, is 0.2431 [+0.2213, +0.2648], which its registered rule classifies as MATERIAL: the rollouts use the per-step actions. For the 50M model the relative reduction is 0.1403 [+0.1245, +0.1564], and the difference between the two reductions, +0.1028 [+0.0812, +0.1242], compares runs made on different GPU types. On the scale cohort the 200M and 700M models give relative reductions of 0.2325 [+0.2083, +0.2561] and 0.2705 [+0.2478, +0.2928]; we do not rank sizes across cohorts. The repeated command is a gross perturbation: dead-reckoned before any frame was generated, it ends 28.37 waypoint units from the logged position on a 34.03-unit path and 175.3 degrees off in

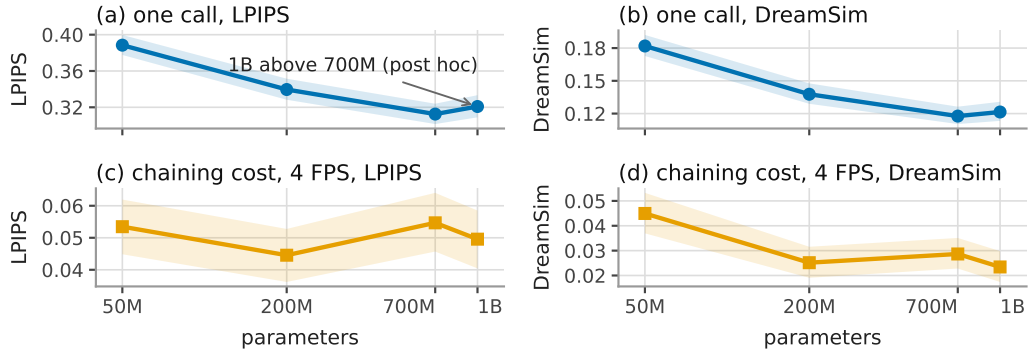


Figure 4: On the scale cohort, the one-call error (top) falls on LPIPS from 50M to 700M and is higher at 1B than at 700M (post hoc), while the 4 FPS chaining cost (bottom, on its own axis) shows no detectable trend on LPIPS (left) and a decrease on DreamSim (right). Values are means over 2, 4 and 8 s against the number of parameters, with 95% trajectory-bootstrap bands from one joint resampling; with one checkpoint per size, the bands cover evaluation variance only. The 700M and 1B bands overlap although their paired difference is resolved on LPIPS (post hoc; it is not resolved on DreamSim). The bottom row's axes span a narrower range than the top row's.

heading at 8 s (medians). With a perturbation this large, H6 can show that the score registers gross action errors, but not whether it resolves small ones.

6 Model size

6.1 One-call error

On the scale cohort the one-call error falls from 0.3884 [0.3777, 0.3997] at 50M to 0.3396 [0.3283, 0.3515] at 200M and 0.3125 [0.3015, 0.3240] at 700M, and rises to 0.3209 [0.3089, 0.3334] at 1B (Figure 4). The adjacent paired differences (post hoc) are −0.0489 [−0.0540, −0.0438], −0.0271 [−0.0325, −0.0217] and +0.0083 [+0.0034, +0.0137]. The registered positive control H10 confirms that the design detects scale: from 50M to 1B the difference is −0.0676 [−0.0746, −0.0607]. A log-linear fit to the four points gives a slope of −0.0553 [−0.0605, −0.0501] per decade of parameters; with one checkpoint per size, this interval also covers evaluation variance only.

The comparison between 700M and 1B is post hoc. The inversion was first seen on the release's own single-step items (500 items on 301 trajectories at 1, 2 and 4 s), where the 1B checkpoint's LPIPS exceeds the 700M checkpoint's by +0.0067 [+0.0042, +0.0092] and its DreamSim by +0.0026 [+0.0010, +0.0044]. It appears again on the fresh scale cohort (+0.0083 [+0.0034, +0.0137] LPIPS; DreamSim +0.0037 [−0.0001, +0.0083]). In 4 FPS rollouts the two sizes are not resolved (+0.0032 [−0.0055, +0.0118] LPIPS, −0.0015 [−0.0074, +0.0042] DreamSim). The contrast was selected from 24 size contrasts, with the release's items as the cohort that generated the hypothesis and the scale cohort as an unregistered replication: the direction was on record before the scale cohort's 700M scores were read, but the contrast was defined afterwards. Under a Bonferroni adjustment over the 24 contrasts, the LPIPS inversion survives on both cohorts and the DreamSim one on the release's items only (Appendix G). The comparison concerns single checkpoints whose file names carry the same training-step count, so it shows that the released 1B checkpoint is not the best of the released checkpoints on single steps but not that capacity hurts; no interval here covers run-to-run variance.

6.2 Chaining cost

H9 tested whether scale reduces the chaining cost by comparing the smallest and the largest size. The registered prediction, NOT-REDUCED, required the interval for the 1B-minus-50M difference to exclude a reduction of a fifth or more of the 50M model's cost on the probe cohort (0.0113); refutation required the interval to lie entirely below zero. The outcome is INCONCLUSIVE, −0.0039

[−0.0149, +0.0071]. The design was foreseeably underpowered: with 150 trajectories the interval's half-width is about the size of the bar, so a true zero had roughly even odds of confirming the prediction (post hoc). The interval also admits a reduction of up to 28% of the 50M model's cost on the scale cohort (0.0535). Across the four sizes the 4 FPS chaining cost is 0.0535 [0.0448, 0.0619], 0.0446 [0.0362, 0.0528], 0.0547 [0.0457, 0.0640] and 0.0496 [0.0404, 0.0585], with a log-linear slope of −0.0002 [−0.0083, +0.0080] per decade. No trend is detected, and a change of a fifth in either direction is not excluded. The slope is not the registered test and is no more precise than it; we report both. On DreamSim the cost is 0.0450, 0.0251, 0.0287 and 0.0234 (slope −0.0141 [−0.0210, −0.0071]; two-size difference −0.0215 [−0.0310, −0.0121]), a decrease that the per-size means place between 50M and 200M (post hoc; no chaining contrast between adjacent sizes was computed). From 200M to 1B the per-size intervals overlap on both metrics. A claim that scale reduces compounding error should therefore state the metric and the range of sizes it refers to.

7 Recommendations, limitations, conclusion

Recommendations for evaluation. (1) Generate the three references alongside every rollout, on the same items, with the command references sharing the rollout's first-call noise, and report at each horizon the command error, the chaining cost and the effect of the actions (which registers gross errors only), with the shares of the rise for the first two. A model without a time shift can use a separately trained direct model as its one-call reference, at the cost of confounding the one-call error with differences between the two models; we have not tested this. (2) Test the command composition before any frame is generated, independently of any score. A constant-velocity straight path must give the same command at every call; zero motion must be encoded as $n(0)$, the normalised value of zero displacement, not as zero in normalised units (Appendix B); forward motion after a quarter turn must map to the forward axis; and every command must decode back to the logged relative pose, with heading compared as a number and not modulo 2π. The last of these tests catches all three defects described here. Training and evaluation should share one function for constructing commands. (3) Report at least two perceptual metrics when a claim concerns chaining or scale (Section 6.2). (4) Resample trajectories rather than frames, state the number of trajectories next to every interval, and size the cohort for the smallest contrast of interest; Section 6.2 shows what 150 trajectories could not resolve.

Limitations. The study covers one family and one dataset, uses perceptual metrics, and has one checkpoint per size and one sample per item. Every scaling statement is therefore about particular checkpoints. The one-call error includes the irreducible uncertainty of a single recorded future, and both metrics score the distance to that future, not frame quality in the sense of video-generation benchmarks (Huang et al., 2024). With one sample per item, sample diversity and inaccuracy are not separated, and the shares of the rise are computed on LPIPS only. The one-call reference needs time-shift conditioning, which fixed-step world models lack. We cannot decide whether the release's own figures came from the defective code path (Section 4.1), and the released checkpoints, retrained after face blurring, need not be the ones behind them. H6 tests the score only against a gross action error. H9's design was underpowered, and the scaling conclusions depend on LPIPS being the registered metric. Post hoc, the 1B model's 8 s chaining cost differs between our two cohorts beyond its intervals. The two runs were on different GPU types, the instances no longer exist, and the three secondary contrasts that cross GPU types (listed with Table 1) were not re-run on one machine. The 2 FPS and 1 FPS rollouts, and the chaining cost itself, carry the confounds listed in Section 3.1. The fresh cohorts stop at 8 s, and the 16 s crossover rests on 20 trajectories. Nothing here measures heading or lateral error, although the frame defect concerns turns: the instrument for that failed its gate, and its claims were voided before any data were read. The registration's timestamps come from our own commit history. Control outcomes and the family's planner are not evaluated.

Conclusion. For the one released family we audited, registered tests show that correcting the public evaluation's command composition lowers the 1B model's 4 FPS rollout LPIPS by 18.8% and that the one-call error falls from 50M to 1B. Whether scale changes the chaining cost is unresolved on LPIPS, where the registered test was underpowered (post hoc); on DreamSim the cost decreases. The split of the corrected curve's rise between chaining and the one-call error is post hoc and differs between cohorts; on the point estimates the one-call error accounts for the larger part on both. We recommend reporting rollout scores together with the three references.

AI USE STATEMENT

In this work, we used generative AI tools (large language model assistants) to propose and refine hypotheses, including the registered claims; to design and give feedback on the methodology and the experiments; to help develop the framework of the three references; to derive the command arithmetic of Appendix B from the released code; to implement methods, including the analysis and experiment-orchestration code; to reformat the public dataset with the preprocessing its maintainers prescribe and to select the evaluation cohorts from it; and to interpret results. We have not used generative AI tools for translation; generating synthetic data sets, proving mathematical claims and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to suggest experimental parameters, to create and edit software code and figures, to search for, identify and summarise the literature, to brainstorm, to suggest the structure and title of the paper, to draft and edit the registration documents and the text of this paper, to format references, to review code and prose, and to summarise the experimental record. The authors made the decisions on what to register, run and report, and reviewed all AI-assisted work: the analysis code is accompanied by tests with planted violations, every number given to three or more decimals is checked mechanically against the record, and the authors checked every number, sign and claim against the record. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The anonymised supplement contains the registration with its dated amendments and its lines of the SHA-256 hash log, the split ledger, every cohort file, every scored row, check report and result file, the analysis code with its planted-violation tests, and the figure script, with a checksum for every file. Anonymisation replaced absolute paths, redacted two phrases of the registration and reworded a few strings that name the authors' other work; the supplement lists the files these edits touched, whose recorded checksums refer to the unedited originals. The originals and the commit history, whose hashes and timestamps the registration cites, will be released on de-anonymisation. Before the PDF is built, a script checks that every number given to three or more decimals in the text appears in the record.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in Atari. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Kavosh Asadi, Dipendra Misra, Seungchan Kim, and Michael L. Littman. Combating the compounding-error problem with a multi-step model. *arXiv preprint arXiv:1905.13320*, 2019.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

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

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Guillaume Chevillon. Direct multi-step estimation and forecasting. *Journal of Economic Surveys*, 21(4):746–785, 2007.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep RL: A case study on PPO and TRPO. In *International Conference on Learning Representations (ICLR)*, 2020.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning (ICML)*, 2019.

William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tianxing He, Jingzhao Zhang, Zhiming Zhou, and James Glass. Exposure bias versus self-recovery: Are distortions really incremental for autoregressive text generation? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

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

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Alex Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Nathan Lambert, Albert Wilcox, Howard Zhang, Kristofer S. J. Pister, and Roberto Calandra. Learning accurate long-term dynamics for model-based reinforcement learning. In *IEEE Conference on Decision and Control (CDC)*, 2021.

Nathan Lambert, Kristofer Pister, and Roberto Calandra. Investigating compounding prediction errors in learned dynamics models. *arXiv preprint arXiv:2203.09637*, 2022.

Massimiliano Marcellino, James H. Stock, and Mark W. Watson. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135 (1-2):499–526, 2006.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision (ECCV)*, 2020.

Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L. Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in Atari games. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2016.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. In *International Conference on Machine Learning (ICML)*, 2024.

Dhruv Shah, Benjamin Eysenbach, Gregory Kahn, Nicholas Rhinehart, and Sergey Levine. Rapid exploration for open-world navigation with latent goal models. In *Conference on Robot Learning (CoRL)*, 2021.

Erik Talvitie. Model regularization for stable sample rollouts. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations (ICLR)*, 2025.

Arun Venkatraman, Martial Hebert, and J. Andrew Bagnell. Improving multi-step prediction of learned time series models. In *AAAI Conference on Artificial Intelligence*, 2015.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. iVideoGPT: Interactive VideoGPTs are scalable world models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

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

Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

A REGISTERED CLAIMS AND VERDICTS

claim	cohort	estimand (units)	rule	control	verdict
H1	confirmation (200)	relative reduction of 1B 4 FPS rollout LPIPS, released to training convention	POSITIVE material iff lo > 0 and rho ≥ 0.10; NEGATIVE iff hi $<$ 0 or [lo, hi] inside ± 0.05	call-0 identity; command arithmetic; planted-row self-test	POSITIVE (material), +0.1879 [+0.1692, +0.2064]
H2	confirmation (200)	four paired LPIPS differences among offset doses, 200M, released frame	SUPPORTED iff all four in the registered direction; INSENSITIVE iff all inside a band; else MIXED	as H1	MIXED (R minus O opposite to registration; D2 minus O +0.0021 [-0.0098, +0.0140])
H3, H4	audit (150)	single-call turn response against a computed camera	(registered; support-gated)	training-support count per turn	VOID: no registered turn command on training support; never run
H3', H4', H5, H7	audit (150)	counterfactual turns and their composition against a computed camera	(registered; instrument-gated)	lens gate on held-out real pairs	VOID: gate failed (0.6726 / 0.4773 against [0.95, 1.05]); never run
S1	confirmation (200)	difference of H1's relative reduction, 1B minus 200M	secondary, no verdict	as H1	+0.0167 [-0.0030, +0.0367]; crosses the two GPU types
H6	probe (150)	relative reduction of 1B 4 FPS rollout LPIPS, first command repeated to logged actions	ACTION-BLIND iff inside ± 0.05; MATERIAL iff lo > 0 and rho ≥ 0.10; SMALL iff lo > 0; WORSE iff hi < 0	call-0 identity; command arithmetic	MATERIAL, +0.2431 [+0.2213, +0.2648]
H8	dose (150)	relative reduction of 200M 4 FPS rollout LPIPS from the +1.25-unit surplus arm to the logged commands, correct frame	HARMFUL iff lo $>$ 0 (material iff rho ≥ 0.10); COMPENSATES iff hi < 0; NEUTRAL iff inside ± 0.05	as H6	HARMFUL (material), +0.1547 [+0.1332, +0.1757]
H9	scale (150)	LPIPS difference of 4 FPS chaining cost, 1B minus 50M (registered prediction: NOT-REDUCED)	REDUCES iff hi $<$ 0; NOT-REDUCED iff lo > -0.0113; else INCONCLUSIVE	as H6; a failed check voids the model	INCONCLUSIVE, -0.0039 [-0.0149, +0.0071]
H10	scale (150)	LPIPS difference of one-call error, 1.01B minus 49M	IMPROVES iff hi $<$ 0	as H9	IMPROVES, -0.0676 [-0.0746, -0.0607]
readout gate	train-split real frames degraded to model grade (100 + 100 trajectories)	retrieval error in frames	median ≤ 1 and p90 ≤ 2	seeded violations	PASSED, medians 0.111 and 0.112

In the rules, rho is the point estimate of the relative reduction, one minus the ratio of the training-convention arm's mean to the other arm's mean; for H1, H6 and H8 the lower end of the interval also exceeds 0.10. Every row's threshold, control and kill condition was filed in a hashed registration, released with the code, before the corresponding rows were read. All but S1 were filed before their cohort was generated; S1's amendment was filed after the confirmation cohort was claimed

and before any of its rows were read. The post-hoc rule for the near-straight subset of Section 4.3 was filed after H2 had been seen. The four-size tables and slopes of Section 6 are registered secondary analyses without a verdict. The following are post hoc, carry no verdict, and are labelled as such wherever they appear: the near-straight subset; the rises, growths and shares of the rise in Sections 4.2 and 5.1 and the 8 s shares of the level; the reading of the DreamSim decrease as lying between 50M and 200M; the cross-cohort comparison and the turning correlation of Section 5.1; the per-horizon command-error intervals and the crossover intervals of Section 4.2; the medians and re-placement of Appendix E; the size contrasts of Section 6.1; the saturation of the forward response (Appendix E); the power statement for H9 (Section 6.2); the intervals on the absolute differences of H1 and H6; and the multiplicity check of Appendix G.

B THE COMMAND ARITHMETIC

The training convention takes the goal displacement (x, y) in the current frame in waypoint units and normalises it as $n(v) = 2\,(v - v_{\min})/(v_{\max} - v_{\min}) - 1$ with $v_{\min} = (-2.5, -4)$ and $v_{\max} = (5, 4)$, so that $n(x) = 4x/15 - 1/3$ and zero forward motion is encoded as $-1/3$. The released rollout evaluation expresses the cumulative waypoints in the first frame's axes, normalises them, prepends a row of zeros, takes differences, and sums the differences per stride. Row 0 is $n(X_1)$, which is correct. Every later row is $4\,\Delta X/15$, because the intercept cancels, and the model decodes it as $\Delta X + 1.25$: the offset is $(v_{\min} + v_{\max})/2$ on the forward axis and zero on the lateral axis, whose range is symmetric. On a straight path at one unit per frame, a call is told 2.25 units instead of 1 at stride 1, and 5.25 instead of 4 at stride 4; at 4 FPS this is an extra 0.3125 m per call. The cumulative heading is wrapped into $[-\pi, \pi)$ before differencing, so a clip that crosses $\pm\pi$ receives one call with a heading command off by 2π. The direct time-shift path sums from row 0 and is correct. In the training convention, a call from frame t to frame t' takes the logged position at t' expressed in the camera frame at t, divides it by the waypoint unit and applies n, and takes the heading change from t to t', wrapped after differencing; our training-convention arms construct every call this way. Our arms are the released composition; the same with the offset removed but the first-frame axes kept; the training convention; the released composition with the offset doubled or negated; the training convention with -1.25, $+1.25$ or $+2.5$ units added to each later call; and the training convention with the first call's command repeated at every later call. A committed check recomputes every command from its label and the logged poses and verifies that call 0 is bit-identical across the command variants at each rollout rate.

C INSTRUMENTS AND GATES

Two instruments were registered for reading where a generated frame's camera is. The computed-camera instrument fits an equidistant fisheye model and predicts the image effect of a heading change. Its gate required the ratio of recovered to logged heading change on held-out real pairs to lie in [0.95, 1.05]. It returned 0.6726 in the 5–10 degree bin and 0.4773 in the 10–20 degree bin (the 20–40 degree bin had no pairs), so it failed, and the claims that depended on it were declared void and never run. The retrieval readout compares a generated frame with the real frames from three frames before to 64 frames after the last context frame (the four context frames and the 64 that follow), using a perceptual distance and no camera model, and takes the index of the best match as the progress along the path. Its gate on real frames passed, with median errors of 0.111 and 0.112 frames and 90th percentiles of 0.286 and 0.280 on two sets of 100 trajectories, and no window was flagged. It is used in Appendix E.

D COHORTS

Table 1: cohorts. The adjudicated contrasts (H1, H2, H6, H8, H9, H10) compare arms generated on one machine. The secondary contrasts that cross GPU types (S1, H6's 1B-versus-50M difference, H8's 200M-versus-1B comparison) are named where they appear. The replication on the scale cohort is unregistered (Section 6.1).

cohort	trajectories	used for	generated on
confirmation	200	H1, H2, S1, the near-straight look, K's share of the rise, the command-following slope across H2's arms (Appendix E)	1B arms on the A100, 200M arms on the A30
audit	150	voided claims; single-call command-following slopes	A30
probe	150	H6, rollouts by number of calls, chaining, C's share of the rise	1B arms on the A100, 50M arms on the A30
dose	150	H8, progress along the path	200M arms on the A30, 1B arms on the A100
scale	150	H9, H10, four-size tables, size inversion (replication), rollouts by number of calls, chaining, C's share of the rise, action dependence at 200M and 700M, Figure 3	A30
release's items	500 on 301	single-step table, size inversion (hypothesis)	A100
release's rollout file	150 on 20	its own protocol, the crossover	A100

Eligible trajectories are test-split trajectories of at least 36 frames that appear in none of the release's own index files. Their names are sorted, permuted with a fixed seed and assigned in order: 200 to the confirmation cohort (H1, H2, S1), 150 to the audit cohort (voided claims, and the single-call command-following slopes of Appendix E), 150 to the probe cohort (H6), 150 to the dose cohort (H8) and 150 to the scale cohort (H9, H10). Each trajectory contributes one item, with the context ending at $t = \max(3, \lfloor (L - 33)/2 \rfloor)$ so that the 8 s target exists. Each later cohort was built only after the rebuilt permutation had reproduced every earlier committed cohort. The release's own single-step items (500 items on 301 trajectories, at 1, 2 and 4 s) and its rollout file (20 distinct trajectories, to 16 s) are read as shipped. Every cohort is claimed on a split ledger before its first frame is generated and is read once under a committed plan. Generation ran on one A30 and one A100-40GB instance, as Table 1 shows; the size inversion replicates from the A100 (the release's items) to the A30 (the scale cohort). The released checkpoint files are cdit_s_100000, cdit_b_100000, cdit_l_100000 and cdit_xl_100000; a fifth released file, cdit_xl_ego4d_200000, is not used.

E SECONDARY RESULTS AND FAILED EXPECTATIONS

Progress along the path. For the correctly commanded 200M rollouts of the dose cohort, the readout of Appendix C locates each generated frame along the true path. The registered ratio of executed to logged progress, averaged over 1, 2 and 4 s, is 1.164 [1.042, 1.316]; it is a mean of per-clip ratios and is inflated by clips with little true progress. The post-hoc medians of executed minus true progress are -0.08, -0.02 and +0.21 waypoint units at 1, 2 and 4 s. Scoring each frame against its best-matching real frame instead of the target lowers LPIPS by at most 0.06 (post hoc; from 0.2936 to 0.2579 at 1 s and from 0.3774 to 0.3209 at 4 s). Most of the error is therefore not in how far along the path the rollout is; heading and lateral position are not measured.

Command following. The registered slope of executed on commanded progress (readout of Appendix C, straight windows) ranges from 0.513 [0.393, 0.645] (50M) to 0.570 [0.435, 0.712] (1B) for scaled single-call commands to 1, 2 and 4 s on the audit cohort, and is 0.427 [0.323, 0.536] across H2's four arms at 1 and 2 s (200M, confirmation cohort). The response saturates above the logged command (post hoc), so the slope summarises the response and does not measure a gain. By horizon, the 1B model's action dependence (H6's relative reduction, Section 5.2) is 0.2810, 0.2939 and 0.1662 at 2, 4 and 8 s; at 8 s both arms are far from the logged frame.

Other secondary results and failed expectations. Per-horizon reductions, the DreamSim companions of every claim, the 2 FPS and 1 FPS rollouts, the chaining cost per size and horizon, and the action dependence of the two middle sizes are reported with intervals in the released record; the main text quotes those its argument uses. Several of our expectations, each recorded in the registration or the record before the relevant data were read, failed. A quadratic fit to the dose arms was expected to place the optimum within 0.3125 units of zero. It gave +1.3129 [+1.2362, +1.3978] in the released frame, where the frame defect makes a surplus help (Section 4.3), and +0.8482 [+0.8037, +0.8932]

in the correct frame, where the best arm is zero and the vertex is an artefact of the asymmetric response. The registration expected the action dependence to grow with horizon; it falls at 8 s. The 1 FPS rollout was expected to be a clean reference for chaining; it changes the context spacing. H9 was expected to come out NOT-REDUCED; it is inconclusive. The registration assumed that the next logged action is largely predictable from what the robot was already doing, which holds only in part: per 0.25 s step, the previous command leaves 0.583 of the forward command's variance and 0.375 of the heading command's unexplained, and at one command per second it leaves 0.937 of the heading command's. An early reading of the translation-following data, that "the model executes half of what it is told", was withdrawn in the record once the saturating response described above had been seen.

F COMPUTE

All generation used the released sampler at 250 steps on one A30 and one A100-40GB cloud instance, under a registered spending cap of about USD 45; the work used roughly 60 GPU-hours including development. Every scored row, provenance record, check report and result file is checksummed and listed in the released record. Apart from the three shown in Figure 3, the generated frames were not retained.

G MULTIPLICITY

The registration adjudicates each claim on its own 95% interval. Recomputed at a Bonferroni-adjusted level over the six adjudicated claims, no verdict changes: H1 +0.1879 [+0.1623, +0.2122]; H6 +0.2431 [+0.2136, +0.2730]; H8 +0.1547 [+0.1256, +0.1824]; H9 −0.0039 [−0.0185, +0.0109]; H10 −0.0676 [−0.0770, −0.0584]; H2's R minus O −0.0158 [−0.0265, −0.0048]. Over the 24 post-hoc size contrasts, the 1B-minus-700M single-step inversion is +0.0067 [+0.0030, +0.0108] (LPIPS) and +0.0026 [+0.0002, +0.0056] (DreamSim) on the release's items, and +0.0083 [+0.0007, +0.0171] (LPIPS) on the scale cohort; the scale-cohort DreamSim contrast and both rollout contrasts include zero.