A WORLD MODEL CAN PREDICT THE FUTURE WITHOUT KNOWING ITS CAUSE: INTERVENTIONAL EVALUATION OF ACTION-CONDITIONED WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A world model exists to answer a planner's counterfactual question: if the action were different, how would the future differ? Most current evaluations score one generated future for its fidelity, plausibility or memory. Causal response, whether the predicted future changes correctly when the action changes, is instead a property of *two* futures. An elementary proposition makes this exact: no score that sees only a rollout and its context can measure causal response, and neither can any functional of the distribution of such rollouts. A reference-based score such as PSNR certifies it only when the prediction error is small compared with the true effect; under an equal-error assumption, this fails in most games for the models we test. We therefore measure the response directly, replaying paired interventions on an exact emulator under common random numbers. We study three world models of different design on 26 Atari games and a recurrent state-space model on six DeepMind Control tasks. On Alien the diffusion model's PSNR is within a decibel of its PSNR on Pong, and its consistency score is higher. Yet by delay 20 its response to a changed action is world-sized and has near-zero alignment with the world's. The alignment contrast holds under two intervention designs and in all three architectures. For the diffusion model it holds in all 96 analysis settings over 7 seeds and at a doubled horizon, and for the token model in at least 92 of 96. Causal difficulty transfers between every pair of architectures ($\rho = 0.60$ to 0.83, all surviving Holm correction), although absolute levels differ, and blindness is a property of a model on a game under a context distribution. Two instruments meant to characterise the failure did not survive their own checks. A latent probe reports an encoded consequence that the same probe does not find once the action pathway is held fixed, and an automated object-level readout is not validated by a blinded audit with three annotators. We therefore report both as audited instruments, not as findings. Alignment's association with planning usefulness remains inconclusive. On Alien, where every model is blind, no model shows a detectable gain over a random policy, yet elsewhere models with low alignment under the diagnostic protocol can still improve returns.

1 INTRODUCTION

A world model is a learned simulator: given what an agent has seen and what it does next, it predicts what it will see. Its purpose is to answer a planner's counterfactual question, *if I took this action instead of that one, how would the future differ?* This use is why world models are trained inside agents (Ha & Schmidhuber, 2018; Hafner et al., 2021; Micheli et al., 2023; Alonso et al., 2024) and why large video models are described as world models for robotics (Bruce et al., 2024; NVIDIA et al., 2025). Yet the quality of their futures is judged one trajectory at a time: by fidelity to a reference video (Unterthiner et al., 2018; Zhang et al., 2018; Huang et al., 2024), by physical plausibility (Bansal et al., 2025; Motamed et al., 2026; Li et al., 2025) and, most recently, by memory (Zhang et al., 2026; Fang et al., 2026). Each of these evaluations takes a single generated trajectory and asks whether it is a good continuation.

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.

That design cannot answer the planner's question. Whether a model's future changes correctly when the action changes is a property of two trajectories that differ in one action. An action-blind score, which sees one trajectory and its context but not the supplied action, cannot in general identify what *would have happened* under a different action, so it cannot measure this property (Section 2). A model can therefore score as well as the environment on consistency, physical plausibility and memory while responding to its action in the wrong way, or not at all. To our knowledge, the paired response of trained pixel world models driven by an agent's actions has not been measured against an exact counterfactual.

This matters in practice. A planner that uses a model which ignores its actions cannot tell good actions from bad ones, however plausible the imagined videos look. Selecting world models by scores that cannot see this failure is blind to the property that the planner's question is about. Our evidence is confined to simulator-trained models, where exact counterfactual truth exists; for video world models of the real world, the question is posed here, not answered.

We measure the two-trajectory quantity directly. Two futures share the same context and the same actions except on a known window. The emulator is replayed under both, so the true consequence is exact for both arms, and the model imagines both with common random numbers. At every delay we compare the model's change between the arms with the world's, as a direction-aware *causal alignment* together with a magnitude and an error, both in pixel space and in the model's latent. The one-trajectory scores and Genie's ΔPSNR (Bruce et al., 2024) are computed on the same rollouts. We apply the measurement to three Atari world models (a pixel diffusion model, a discrete-token transformer and a stochastic-latent transformer that we train) and to a self-trained recurrent state-space model (RSSM) on six DeepMind Control (DMControl) tasks. It gives four results.

- **Scope, then an instrument.** No action-blind score of one rollout, nor of the law of rollouts, can measure causal response (Proposition 1). A reference-based score certifies it only when the prediction error is small compared with the true effect; under an equal-error assumption, this fails in 23, 24 and 25 of 26 games for our three models (Corollary 1). The instrument has exact ground truth for both arms, common random numbers, calibrated guards and gated statistics (Section 3).
- **Accuracy does not certify response, and causal difficulty transfers.** DIAMOND predicts Alien within a decibel of Pong and rates it more consistent, yet its delay-20 alignment is 0.002 on Alien against 0.86 on Pong. The alignment contrast holds in 96 of 96 analysis settings, under a second intervention design, at a doubled horizon for DIAMOND and in the two other architectures. The three architectures' rankings of the 26 games are correlated ($\rho = 0.71$, 0.60 and 0.83, all surviving Holm correction), and a game's verdict can change with the context distribution (Sections 4.1 and 4.2).
- **Where the change goes, and an audited object-level readout.** DIAMOND's low-alignment responses at pixel level are either nearly silent (3 of its 7 never-aligned cells, magnitude at most 0.06) or world-sized and, in three of the other four, mostly outside the world's support. In its cells with a three-arm alignment below 0.05 the median off-support fraction is 0.66, against 0.14 in cells at or above its tracked threshold. A third arm finds the change corresponding to the action taken in 22 of 24 tracked cells but in only 5 of 28 never-aligned ones, and not on Alien (Appendix G). An automated object readout labels 24 of the 25 assessable never-aligned cells as failing to improve on zero displacement. A blinded audit by three annotators, who agree with each other moderately to substantially, finds the locator under-detecting in every class and its displacement verdict only weakly related to theirs. We therefore report the readout as audited locator labels, not as a finding about the models, and the scores' association with those labels (Section 4.3) as an exploratory property of those labels. A latent probe without action-pathway controls reports a consequence that the same probe, with the action pathway held fixed, does not find (Section 4.4).
- **Usefulness, and what the data do not support.** In open-loop ranking with a frozen scorer, alignment's association with the usefulness of a model's futures is inconclusive, and prediction accuracy's is positive but inconclusive. In closed-loop planning on Alien, where all three models are blind, no model shows a detectable gain over a random policy, and elsewhere a positive gain does not require high alignment. A latency law, a consistency-versus-causality trade-off, a link to published returns and a contrast with ΔPSNR were tested and are not supported (Sections 4.5 and 4.6).

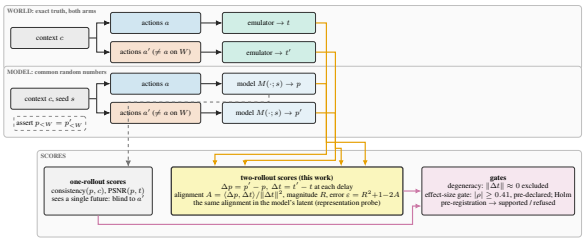

Figure 1: Paired-intervention measurement of causal response. One context c and two action sequences a, a' that differ only on a window W; the emulator gives exact futures t, t' (world lane), the model imagines both under the same seed s and must satisfy $p_{<W} = p'_{<W}$ (model lane). A one-rollout score (dashed grey) sees only p, the two-rollout score (orange) compares Δp with Δt at each delay, in pixel and latent space, and every statistic passes the gates (purple).

Figure 1 shows the measurement, and Appendix A places it among existing evaluations (Table 4) and reviews related work. The closest neighbours differ in what is intervened on and what is scored. ScratchWorld (Lin & Zhang, 2026) replays Scratch programs exactly on a pinned virtual machine, but it scores field-level state changes of language models that predict symbolic state. Genie's ΔPSNR uses two rollouts, but it measures how much the fidelity to the factual reference changes, not whether the change matches the world's. A recent position paper argues that world models should be judged by interventional action fidelity and decision usefulness (Yu et al., 2026); we provide the measurement.

2 Why no action-blind score of one rollout can measure causal response

We first make the planner's question precise. A world model is a map M from a context c, an action sequence a and its own randomness s (the seed) to a future $x = M(c, a; s)$. The environment E maps (c, a) to the true future, which an emulator replays exactly. A *one-rollout score* is a functional $S(x, c)$ of one generated future and its context, such as a consistency, roughness or realism score. A *distributional score* is a functional of the joint law of (c, x), such as FVD.

Definition 1 (paired causal response)**.** For a context c, two action sequences a, a' that differ only on a window W, a delay d counted from the start of W, and a seed s, the world's response is $\Delta t_d = E(c, a')_d - E(c, a)_d \in \mathbb{R}^D$ and the model's paired response is $\Delta p_d(s) = M(c, a'; s)_d - M(c, a; s)_d$, both arms imagined under the same s. Its *alignment*, *magnitude* and *normalised error* are $A_d = \langle \Delta p_d, \Delta t_d \rangle / \|\Delta t_d\|^2$, $R_d = \|\Delta p_d\| / \|\Delta t_d\|$ and $\varepsilon_d = \|\Delta p_d - \Delta t_d\|^2 / \|\Delta t_d\|^2$. They are defined on $\{\|\Delta t_d\| \geq \tau\}$ for a guard $\tau > 0$ and satisfy $\varepsilon_d = R_d^2 + 1 - 2A_d$ identically. Under common random numbers Δp_d is a function of (c, a, a', s) alone, so the response is defined under one coupling of the model's randomness. The seed sweeps of this paper vary the base action sequence a, and with it c and the true futures, on fixed emulator start states; they report variation over action sequences, not over s.

Proposition 1. *Fix an environment E and an evaluation distribution under which $c \sim \pi$, $a \mid c \sim \mu(\cdot \mid c)$ (actions may depend on the context, as they do under a policy), and the seed $s \sim \nu$ is a uniform variable independent of (c, a). There exist models M_1, M_2 such that $(c, M_1(c, a; s))$ and $(c, M_2(c, a; s))$ have the same joint law, so that every one-rollout score and every distributional score has the same distribution under both, while the causal response of M_1 equals the environment's for every (c, a, a', s) and that of M_2 is identically zero. The statement has content when the environment's response is nonzero with positive probability, as in every game here (assumptions in Appendix B).*

Proof sketch. Take $M_1 = E$ and $M_2(c, a; s) = E(c, \tilde{a}(c, s))$, where $\tilde{a}$ draws an action from $\mu(\cdot \mid c)$ using the seed, not the supplied action; $(c, M_2(c, a; s))$ then has the law of $(c, E(c, a))$, while under common random numbers the two arms of M_2 coincide (full proof in Appendix B).

The construction is not exotic: M_2 has learned the dynamics and substitutes an action of its own for the one it is given. A score that also sees the supplied action is outside the class, and so is one that

sees the true future of the same arm, such as PSNR(x, t); the following corollary bounds what the latter can certify.

Corollary 1. *For any model and any delay,* $\|\Delta p - \Delta t\| \le \|p' - t'\| + \|p - t\|$. *If both arms are predicted with error norm* e*, the root normalised response error* $\sqrt{\varepsilon} = \|\Delta p - \Delta t\|/\|\Delta t\|$ *is at most* $2e/\|\Delta t\|$*, a bound that does not certify* $\varepsilon < 1$ *whenever* $2e \ge \|\Delta t\|$.

The corollary is the triangle inequality, and its condition is sufficient, not necessary. Two arms with the same large error have a perfect response, so a bound above one shows only that accuracy does not certify the response, not that the response is wrong. At 30 dB the error norm of a 64×64 frame is about 7.0, while the true response at delay 20 in our suite has a median of about 8 units (2 to 35 across the 78 cells). Computed per game from the measured PSNR, the ratio $2e/\|\Delta t\|$ exceeds one in 23, 24 and 25 of 26 games for DIAMOND, IRIS and STORM (Appendix Q). At the accuracy of these models, PSNR seldom certifies a response error below the response itself. Algorithm 1 (Appendix B) summarises one episode. Alignment is the projection of the model's change onto the true change (1 for the true change, 0 for an unrelated one, negative for the wrong direction). The identity holds per episode pair, not for aggregated medians. Alignment alone cannot separate an absent response from a misdirected one; the triple can. Appendix B gives the estimators, the guard on $\|\Delta t_d\|$ and an off-support bound $|A_d| \le R_d\sqrt{1 - O_d}$.

3 Measurement

Models and metrics. We evaluate DIAMOND (pixel diffusion) and IRIS (transformer over discrete tokens), the two reference Atari-100k models with public weights, and STORM (a transformer over stochastic categorical latents), which releases no weights and which we train per game with its authors' configuration. All three run on the 26 games, and our own weak RSSM on six DMControl tasks is the testbed for the probe controls (Appendix C). Alongside our consistency proxy and PSNR against the actual future, we compute Genie's controllability ΔPSNR (Bruce et al., 2024), which asks whether the action *matters*, not whether the response is *correct*. ΔPSNR and alignment agree on this suite, which we read as convergent validity. Their rank correlation is 0.69 to 0.98 up to delay 8 over the eight games with ΔPSNR stored at every delay (Table 5), and delay-20 ΔPSNR exists for 60 of the 78 cells (Appendix M). A per-game inverse-dynamics proxy fails its own gate on seven of sixteen cells (Tables 5 and 6).

Protocol. Each episode yields a pair of rollouts that share a context and an action sequence except on an intervention window, eight steps by default. The emulator is replayed under both for exact ground truth, and the model imagines both with common random numbers. The instrument asserts that the two imagined futures coincide before the intervention; the recorded maximum difference is zero in every sweep file (tolerances in Appendix C). Each Atari harness also passed a registration test that this assertion cannot see: an action changed at step K must first change the frame at index $K{+}1$, as in the emulator. An earlier version of the IRIS and STORM harnesses failed the test by one step, and every IRIS and STORM number in the main text postdates the correction. Ratios are guarded by a floor calibrated from the observed true divergences. Unless a table says otherwise, PSNR, consistency and alignment are means over seven evaluation seeds of the per-seed median over 60 episode pairs; the seeds vary the action sequence on shared start states, and a $\pm$ gives the seed-to-seed standard deviation. Magnitude and error are medians over the pooled guarded pairs (Appendix B). The released models are one checkpoint per game, so their training-seed variation is measured only on the flagship pair (Alien and Pong, Section 4.2); for STORM, which we train, a second training seed on every game is reported in Section 4.2 (Appendix E).

Interventions and gates. A fixed intervention leaves 14 of 26 games without a visible effect. We therefore let the emulator choose, per game and without any model, the action offset and warm-up that maximise the true effect, and all 26 games then carry a live intervention. Every statistic passes a degeneracy gate: units whose true divergence is too small to divide by are excluded and counted. We report estimates with intervals and mark those that meet a pre-specified effect-size gate ($|\rho| = 0.41$ at $n = 26$), which is a decision rule fixed in advance, not a power statement; the 80% power minimum detectable effect is 0.53. We state the multiplicity correction for each test family. The predictions about newly measurable games were written from the emulator alone before any model was run on these interventions.

Table 1: The flagship pair under both interventions, three architectures (7 evaluation seeds $\times$ 60 episodes). PSNR of the un-intervened rollout against the actual future at delay 20; consistency, the one-rollout proxy; causal alignment at delays 1 and 20 (mean $\pm$ s.d. over seeds at delay 20); response magnitude R and normalised error $\varepsilon = R^2+1-2A$ at delay 20 (medians over the pooled guarded pairs, so the identity need not hold between the displayed aggregates). STORM has no fixed-intervention run.

	intervention	game	PSNR (dB)	consistency	align. d=1	align. d=20	R d=20	ε d=20
DIAMOND	fixed	Alien	34.4	0.876	0.41	0.002 ± 0.001	1.51	3.23
DIAMOND	fixed	Pong	35.2	0.842	0.56	0.859 ± 0.029	1.33	1.03
DIAMOND	emulator-chosen	Alien	30.2	0.853	0.18	0.003 ± 0.004	1.11	2.18
DIAMOND	emulator-chosen	Pong	35.2	0.843	0.52	0.848 ± 0.037	1.28	1.01
IRIS	fixed	Alien	32.9	0.852	0.51	0.025 ± 0.043	1.38	2.89
IRIS	fixed	Pong	34.4	0.867	0.31	0.315 ± 0.067	1.15	1.51
IRIS	emulator-chosen	Alien	29.8	0.840	0.07	0.001 ± 0.006	1.15	2.31
IRIS	emulator-chosen	Pong	34.4	0.867	0.26	0.217 ± 0.089	1.11	1.61
STORM	emulator-chosen	Alien	30.0	0.852	0.11	0.004	0.67	1.40
STORM	emulator-chosen	Pong	35.7	0.879	0.29	0.376	1.02	1.27

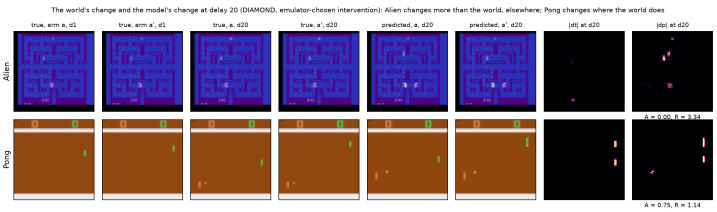

Figure 2: What zero alignment at world-sized magnitude looks like: one episode each of Alien and Pong (emulator-chosen intervention, DIAMOND, delay 20). Left to right: the two true arms at delays 1 and 20, the two predicted arms at delay 20, and the absolute pixel change of the world ($|\Delta t|$) and of the model ($|\Delta p|$). In Alien the world's change is the player's position and the model changes several other things by more (magnitude 3.34, alignment 0.00); in Pong the model's change coincides with the world's (alignment 0.75, magnitude 1.14).

4 Results

4.1 The flagship: matched accuracy, a world-sized response in the wrong place

The flagship pair shows the gap that Section 2 allows: two games with matched accuracy and opposite causal response (Table 1, Figures 2 and 4; all tests in Appendix P). Under the fixed intervention DIAMOND predicts Alien and Pong to within a decibel (34.4 against 35.2 dB) and rates Alien *more* consistent. Yet its alignment on Alien is 0.002, against 0.86 on Pong. The response is not absent: its magnitude is 1.51 times the world's, a world-sized change in an unrelated direction, and its normalised error is 3.23 against 1.03 on Pong. Under the emulator-chosen intervention Alien is predicted 5 dB below Pong, with the same verdict. IRIS reproduces the contrast at a lower level: 0.001 on Alien against 0.22 on Pong under the emulator-chosen intervention, and 0.025 against 0.31 under the fixed one, where an early response on Alien (0.51 at delay 1) is lost by delay 20. On Pong IRIS is directionally aligned but classed aligned-then-forgotten; its tracked threshold, 0.11, is by definition half of its emulator-chosen Pong value. STORM gives 0.004 against 0.38. Alien is one of the games at zero in every architecture, not uniquely the worst. Matching on accuracy does not change the verdict (rule and intervals in Appendix C): the 123 of 348 guarded Alien pairs inside Pong's PSNR interquartile band have alignment 0.002 for DIAMOND, and the 155 such IRIS pairs have 0.015. The causal half of the conclusion holds in 96 of 96 analysis settings (delay, guard floor, statistic, episode count) for DIAMOND. For IRIS it holds in 96 of 96 under the emulator-chosen intervention and 92 of 96 under the fixed one. The consistency inversion holds for DIAMOND in every setting under the fixed intervention and in 88 of 96 under the emulator-chosen one, and not for IRIS. Consistency therefore does not separate the blind game from the tracked one, and in one architecture it orders them the wrong way. For DIAMOND at a horizon of 48 steps, Alien is at zero from delay 20 through delay 44, and its early response (0.19 at delay 1) never reaches tracking, while Pong stays tracked at every delay with enough guarded pairs (Table 21).

Contexts from the trained policy. We repeated the measurement on six games (7 seeds, 30 episodes each) with the warm-up, context and base actions produced by each released agent's own

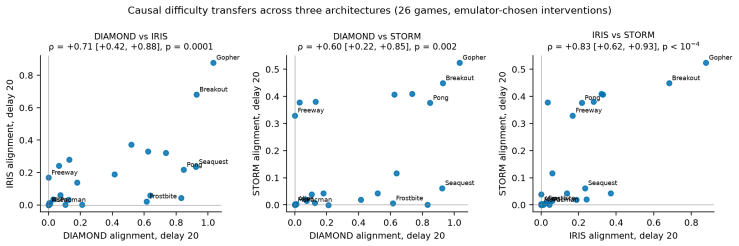

Figure 3: Cross-architecture agreement per game: causal alignment at delay 20 under the emulator-chosen interventions for every pair of the three architectures (26 games). Spearman ρ with bootstrap interval and permutation p; all three pairs survive Holm correction.

actor in the real emulator. We used the pure policy and, on Alien and Pong, also ten percent random actions (Appendix D). The flagship verdict holds, and two games' verdicts change. Alien stays blind (DIAMOND -0.005 and 0.015, IRIS 0.005), BankHeist stays at 0.000, and Pong retains directional alignment where measurable (IRIS 0.61 under the pure policy, DIAMOND 0.77 with random actions; levels, not the formal class). Freeway, never-aligned for DIAMOND (0.000) and aligned-then-forgotten for IRIS (0.17) under random contexts, reaches tracked levels under the policies' contexts (0.47 and 0.59). Breakout, tracked under random contexts, falls to 0.000 for DIAMOND at delay 20 on the pairs whose true arms still differ (IRIS 0.11, at its tracked threshold). Our classifications therefore depend on the context distribution, as those of any single-distribution evaluation do.

4.2 Causal difficulty transfers across architectures

The flagship is not a coincidence of one model: causal difficulty transfers across the three architectures (Figure 3). The rankings of the 26 games by causal alignment are correlated between every pair ($\rho = +0.71$, $+0.60$ and $+0.83$ for DIAMOND against IRIS, DIAMOND against STORM and IRIS against STORM, all surviving Holm correction; Kendall's $W = 0.81$; intervals in Appendix P). Across training seeds, a second STORM checkpoint per game ranks the games much as the first does ($\rho = 0.74$, 95% CI $[0.44, 0.90]$) and reproduces the flagship (Alien 0.003, Pong 0.51). A second training seed of DIAMOND and IRIS on the flagship pair gives the same verdict (Alien 0.004 and 0.003, Pong 0.83 and 0.25 at delay 20; Appendix D). Consistency rankings transfer at least as strongly and PSNR rankings comparably (Appendix P), so transfer is a property of the games, not of the causal score. Within every architecture, prediction accuracy is positively but uncertainly associated with alignment ($\rho = 0.38$, 0.34 and 0.47, none surviving Holm correction; Appendix P), and the flagship shows that this association does not certify the response. Table 24 classifies every cell by the level and persistence of its alignment (counts, thresholds and definitions in Table 7, Appendix E; sensitivity in Appendix P). Two failure modes recur: a response that is never aligned, and one that is aligned and then forgotten, with a measurable horizon (Appendix E). In the text, *blind* denotes a never-aligned cell; the pre-registered scorecard (Appendix T) uses its own criteria.

4.3 Where the change goes, and an audit of an object-level readout

Zero alignment at world-sized magnitude admits two readings: the change does not depend on which action the model was given, or it is specific to the action but lands in the wrong place. A third arm per episode tests the correspondence between action and consequence. In 25, 24 and 24 of 26 games, the model's response to the chosen action projects more on the true change under that action than its response to a wrong action does (one-sided binomial $p < 10^{-5}$ each). The comparison is Holm-significant in 17, 14 and 9 games, which include 22 of the 24 tracked cells but 5 of the 28 never-aligned ones, and not on Alien. The models' changes therefore correspond to the action taken more often than not. A low-alignment cell's change also lands mostly outside the support of the world's change: the median off-support fraction is 0.66 in DIAMOND's 10 cells with three-arm alignment $A_b < 0.05$, against 0.14 in its 11 with $A_b \geq T$, the tracked threshold of Table 7. The off-support bound makes this relation partly definitional (Appendix G). A pixel-level score cannot say whether a low alignment reflects a wrong consequence or a sprite drawn a few pixels off, so we built a readout in object coordinates. The actuated object is the set of pixels where the two true arms first differ. A template locator searches for it in the generated frames at delays 8 and 20

Table 2: Blinded human audit of the object-level readout: 59 panels (21 games, 43 cells) annotated independently by three authors, and the consensus labels (at least two decided and all who decided agree). Pairwise Cohen's kappa between annotators: 0.56 to 0.81 on presence and 0.46 to 0.75 on improvement (per pair in Appendix K). *No verdict*: the locator did not find the object; intervals for every kappa, the protocol, decision rules, contingency tables and the breakdown by class are in Appendix K.

	annotator 1	annotator 2	annotator 3	consensus
presence: panels decided / unsure	49 / 10	58 / 1	59 / 0	49 / 10
locator recall / precision vs annotator	0.43 / 0.86	0.39 / 0.80	0.41 / 0.69	0.40 / 0.77
kappa, locator presence vs annotator	0.00	-0.12	-0.09	-0.14
sees object: tracked, never-aligned (of decided)	15/15, 7/10	16/17, 14/17	14/17, 12/18	14/15, 10/13
displacement compared (undecided; no verdict)	20 (18; 21)	22 (5; 32)	23 (5; 31)	17 (21; 21)
agreement / kappa, locator vs annotator	0.70 / 0.44	0.59 / 0.20	0.57 / 0.14	0.65 / 0.31
locator improvement calls agreed	8 of 8	6 of 9	5 of 8	4 of 5
locator no-improvement calls agreed	6 of 12	7 of 13	8 of 15	7 of 12
no locator verdict (annotator sees improvement)	21 (9)	32 (20)	31 (14)	21 (11)

(the paper reports 20). For every episode in which it finds the object in both generated arms, the locator's estimate of the model's displacement of the object between the arms is compared with the true displacement, against the baseline of a zero displacement (up to 30 evaluated episodes per cell, 1,955 in all; Appendix K). Its audit comes first.

Audit of the readout. Three authors annotated the same 59 panels (21 games, 43 cells) independently, blind to each other's sheets and to every metric, under decision rules fixed before the first labels were read (Appendix K). Where both decided, they agree with each other moderately to substantially: pairwise Cohen's kappa is 0.56 to 0.81 on presence and 0.46 to 0.75 on improvement (Table 2; pairs and panel counts in Appendix K). Against them the template locator under-detects in every class, with recall 0.43, 0.39 and 0.41 at precision 0.86, 0.80 and 0.69 against the three annotators (Table 2). Its displacement verdict is only weakly related to theirs (kappa 0.44, 0.20 and 0.14; consensus counts in Table 2). We therefore report the readout's outputs as locator labels. They do not establish a model-level presence contrast, and the failure label below is not validated as a fact about the models.

The locator's labels. A colour locator misses 46%, 60% and 60% of the objects the template locator finds on the 78 cells (1,955 episodes), so Figure 7 (Appendix K) uses the template locator, whose self-check on true frames passes in 95% of episodes. It detects the object in both generated arms in a median 0.89 of episodes in tracked cells (BattleZone, whose template is degenerate, is the exception; Table 15), against 0.41 in never-aligned cells (pooled; per-architecture medians in Figure 7). A cell is labelled *no improvement on zero displacement* when the estimated error is strictly smaller than a zero displacement's in fewer than half of those episodes (ties count as no improvement; at least 3 such episodes per cell). Tracked cells (22 of 24 assessable) improve in a median 0.59 of detected episodes (median errors 5.7 against 12.2 pixels). Never-aligned cells (25 of 28 assessable) improve in 0.09 (6.9 against 3.9 pixels), and 24 of the 25 carry the label; the exception, IRIS UpNDown, sits at exactly half. The zero-displacement baseline is stronger in blind cells, where the world moves the object less, so the label is easier to earn there (Appendix K).

Which scores predict the locator's labels. On the 60 cells (26 games) where every score is defined, no score shows detectable discrimination of the non-detection label (alignment AUC 0.56 [0.34, 0.74], PSNR 0.62, the response error 0.60; two-sided game-level permutation $p > 0.1$). On the 55 cells (26 games) with enough detections, three two-rollout scores discriminate the no-improvement label: alignment 0.88 [0.76, 0.96] ($p < 0.001$), ΔPSNR 0.84 [0.65, 0.95] and the magnitude 0.78, against consistency 0.43 and PSNR 0.55. The response error's and the off-support fraction's estimates are uncertain, and 3 of the seven rows survive Holm correction within that family (stratified game-block permutation, Appendix P; Table 14). These are exploratory estimates for locator-derived labels on the evaluated cells. They do not show which score detects a semantic failure, and they do not show that alignment discriminates better than ΔPSNR. Predictors built from the emulator's own futures (Appendix O) show what the scores register by construction. Four wrong responses keep the correct predictor's PSNR (30.5 dB) and a consistency within 0.02 of it. ΔPSNR moves on three of them but stays near the correct predictor's value on the wrong-action one (3.49 against 3.63 dB), while alignment takes a distinct value on each (0.50 to -0.60).

4.4 WHERE DOES THE INTERVENTION GO? A REPRESENTATION PROBE

Pixel alignment says that the decoded future does not move correctly, but not whether the model's *representation* knows the consequence. We therefore apply the same estimator to the model's latent for its predicted frames against its latent for the true frames. For the RSSM on six DMControl tasks the answer first looks decisive: near the intervention the imagined latent aligns at 0.97 to 1.00 with the true-frame latent while the decoded pixels sit at zero, and far from it the alignment collapses (Figure 8, Appendix L). Two controls show otherwise. The posterior state receives the action as well as the frame, so the "true latent change" has two sources. Holding the action input at arm a and supplying the frames of arm b leaves the posterior almost unchanged near the intervention: the frames-only share of the true latent change is at most 9% at delays 1 to 3. Holding the frames and changing only the action reproduces the full change there, and it aligns with the imagined latent at 0.97 to 1.01 at delays 1 to 3 (Table 16). So the alignment near one was the action input's echo through the recurrent state. Where the visual consequence does enter the posterior (delays 16 and 20, and delay 8 on walker), the imagined latent's alignment with the full true latent change falls to 0.00 to 0.34 (one exception, Appendix L), and its frames-only alignment is at most 0.01. The uncontrolled probe therefore does not establish that the consequence is encoded, and the existing decoder does not render one from either state. These checkpoints are weak (18 to 26 dB), which makes the near-perfect uncontrolled reading the more telling.

For the token model, the sampled-token embedding is as blind as its pixels, but at the pre-sampling hidden state the uncontrolled probe again looks like decoding failure: on the two games whose pixels are blind at delay 20, the hidden alignment is 0.41 (Alien) at delay 1 and 0.26 (BankHeist) at delay 3. Under the same controls the transformer's state does encode the frames (with the action tokens held fixed; Appendix L). In the pixel-blind games, however, the imagined state tracks that visual change weakly or not at all, and the uncontrolled values are the action token's echo. STORM's state gives the same answer (Appendix L). A latent probe of an action-conditioned model therefore needs action-pathway controls to establish that it reports an encoded consequence rather than the action input's echo through the state. That echo is the world-model counterpart of causal confusion in imitation learning (de Haan et al., 2019). The controls diagnose this probe; they do not show that no other readout could recover the consequence.

4.5 WHAT THE DATA DO NOT SUPPORT

Several claims we set out to make are not supported (Appendix M). The latency of the true consequence is at most a weak correlate of alignment, not a law (Appendix F). Its correlation with alignment is negative in all six estimates of that appendix (-0.58 to -0.32, uncorrected p from 0.035 to 0.29 over games). Its sharpest pre-registered prediction, that Alien would be tracked under an immediate-effect intervention, was refuted for DIAMOND at the thresholds written with the predictions and left unscored for IRIS, whose Alien cell missed their accuracy gate; overall, 8 of 11 scored predictions were confirmed for DIAMOND and 9 of 11 for IRIS. A decorrelation of consistency from causal alignment is not established: none is detected within any architecture at $n = 26$ (all $p > 0.1$; Appendix M). No association is detected between causal alignment and published agent returns ($\rho = 0.20$, $p = 0.35$, $n = 24$; all designs unpowered, with minimum detectable $|\rho|$ at 80% power 0.55 at $n = 24$ and 0.73 at $n = 12$). Finally, the predicted contrast against ΔPSNR did not appear: the two agree at the game level, and the eight dissenting transformer cells are in Appendix M.

4.6 PLANNING WITH THE REWARD INFORMATION HELD FIXED

Does alignment predict how useful a model's futures are? We hold the contexts, the candidate actions and the true futures fixed and vary the world model, and one reward scorer per game, trained on emulator frame pairs and then frozen, scores every architecture. Scorer gates leave fifteen usable games per architecture. Eight candidates per context are ranked by scored predicted return and by the emulator's replayed return, and the same scorer applied to the true frames gives a true-frame reference that separates a weak scorer from a weak model (sample sizes, gates and STORM's matching in Appendix N). No score, ours included, passes the pre-declared gate of $\rho = 0.41$, which was fixed in advance for 45 cells (Appendix P). With games as the resampling and permutation unit, the model's ranking correlation relates to alignment at $\rho = 0.30$ ($p = 0.15$), to the response error

Table 3: Closed-loop planning, excerpt (all twelve games in Appendix N): mean emulator return over 60 episodes, bootstrap 95% interval; an asterisk marks a model whose interval of the difference from the random policy lies entirely above zero; the oracle is the emulator planning under the same frozen scorer. Every model is blind on Alien; CrazyClimber's alignments are 0.07, 0.24 and 0.02 (mean-of-seed-medians estimator, Tables 23 and 12; Table 17 uses pooled medians).

game	random	oracle	DIAMOND	IRIS	STORM
Alien	16.8 [13.7, 20.0]	31.7 [27.2, 36.2]*	18.3 [14.8, 22.0]	15.7 [12.7, 18.8]	17.0 [13.7, 20.5]
CrazyClimber	118.3 [100.0, 136.7]	166.7 [151.7, 181.7]*	171.7 [153.3, 190.0]*	180.0 [160.0, 200.0]*	161.7 [145.0, 178.3]*

at -0.21 and to prediction PSNR at 0.37 ($p = 0.087$). A post-hoc estimate on the 30 matched DIAMOND and IRIS cells, outside the pre-declared analysis, gives 0.43 for PSNR ($p = 0.066$); intervals are in Appendix N. The three pre-declared estimates are therefore inconclusive, and none is corrected for the three predictors tested. Within architectures ($n = 15$ each), only IRIS reaches $p < 0.05$, uncorrected, for alignment (0.66), ΔPSNR (0.63) and PSNR (0.57); Appendix N gives these and the cells that break the pattern.

Closed loop. Ranking fixed candidates is not planning, so we also closed the loop: random-shooting model-predictive control inside the emulator, against a random policy and an emulator-rollout oracle, on the twelve games whose scorer passes a stricter pre-declared gate (Table 3; design and per-game returns in Appendix N). The oracle's gain interval over the random policy lies above zero in 8 of 12 games. On those eight, the models' gain intervals lie above zero in 5, 3 and 4 games for DIAMOND, IRIS and STORM (uncorrected counts; STORM's timing in Appendix N), and the median oracle-normalised gain is 0.61, 0.25 and 0.40. On Alien no model shows a detectable gain: every model-minus-random interval includes zero (Appendix N). The gain's correlation with alignment and with PSNR rests on eight games per architecture (Appendix N) and is not interpreted. On CrazyClimber all three models are useful, DIAMOND and STORM at low diagnostic alignment (Table 3), so a horizon-12 planner can gain from such a model; the diagnostic protocol does not establish that the response is also low on the planner's own contexts. A shared scorer can also err differently on different generators. Accuracy alone does not decide usefulness: at the same accuracy IRIS ranks Frostbite at half the reference and Alien at chance (Appendix N).

5 Limitations

Exact counterfactual ground truth requires a simulator; ours are Atari and DMControl at 64×64. Training-seed variation is uneven. STORM has a second training seed on all 26 games, while DIAMOND and IRIS are released as one checkpoint per game, so their seven seeds vary evaluation, not training; a second training seed of each on the flagship pair reproduces the verdict (Section 4.2). The DMControl RSSM is our own and weak. Three blinded annotators find the object locator under-detecting in every class (presence recall 0.43, 0.39 and 0.41 against the three annotators) and its displacement verdict only weakly related to theirs (kappa 0.44, 0.20 and 0.14). The object-level readout is therefore reported as locator labels, not as presence or displacement facts about the models, and the metric comparison built on them is exploratory. Common random numbers couple the two arms' sampling and are not a property of the model. PSNR is compared between games, not against a static-prediction baseline (Appendix C).

6 Conclusion

When the action changes, does the model's future change the way the world's does? On three architectures and 26 games, causal blindness is common: 28 of 78 cells never align, and 20 align and then forget. The failure transfers across architectures and depends on the context distribution. In the 28 never-aligned cells the change is at least half the size of the world's in 19 and nearly silent in 6. Where these cells respond, the change is of the wrong kind, and in the diffusion model it lands mostly outside the world's support. An object-level readout of this change is not validated by three blinded annotators, and an uncontrolled latent probe reports a consequence that the controlled probe does not find. Alignment's link to planning usefulness is inconclusive: there is no detectable gain on Alien, where every model is blind, and there are gains at low alignment elsewhere. Consistency and accuracy alone do not certify the measured response.

Reproducibility statement

Every result in this paper is computed from stored per-episode result files. The anonymised code (the measurement library `ccaus/`, the sweep and analysis scripts and the extension experiments), the per-episode result files, the three human annotation sheets and the dated pre-registration files are available at `https://anonymous.4open.science/r/wm-causal-response-731E`. One pipeline (`scripts/rerun_reg/analyze_all.sh`) recomputes the main analysis outputs from the result files and runs the test suites, each extension experiment has its own evaluation script, and before submission an audit script checked that every decimal number in the paper body occurs in a result file. The measurement library ships with four test suites, including property tests swept over dimension, sparsity and scale, and a mutation suite that reproduces six faults, five of them among the eight of Appendix U, and checks that each is detected; the registration test of Appendix C is run separately. Pre-registered predictions and their scorecard are released as written, and the rules fixed before each analysis (the human-check decision rules, the consensus rule, the closed-loop game list) are stated where they apply. Checkpoints are the public DIAMOND and IRIS releases, our STORM checkpoints (two training seeds per game) and our second-seed DIAMOND and IRIS checkpoints for the flagship pair; our own checkpoints, the RSSM's code and checkpoints and the frozen reward scorers are released on acceptance, while the per-episode result files they produced are included; emulator ground truth is the Arcade Learning Environment (ALE) with `repeat_action_probability` $= 0$.

AI use statement

We employed a large language model (LLM) to assist in refining the language and enhancing the clarity of this manuscript, and to fix bugs in the software code. No experimental result, number, or reference originates from a language model: every number is generated by scripts from stored result files and every reference was verified against its source. The core scientific ideas, the experimental design, the results analysis and the final conclusions were conceived by the authors, who take responsibility for the final content of this work.

Ethics statement

This work evaluates simulator-trained world models (public releases and models we trained on the same simulators) and involves no deployment. Its only human data are the object-panel annotations made by three of the authors for the audit of Section 4.3; no external participants were recruited and no personal data were collected.

References

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Manuel Wüthrich, and Stefan Bauer. CausalWorld: A robotic manipulation benchmark for causal structure and transfer learning. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.04296.

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. doi: 10.52202/079017-1873. arXiv:2405.12399.

Hidehisa Arai, Keishi Ishihara, Tsubasa Takahashi, and Yu Yamaguchi. ACT-Bench: Towards action controllable world models for autonomous driving. *arXiv preprint arXiv:2412.05337*, 2024.

Mido Assran, Adrien Bardes, David Fan, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

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. arXiv:2406.03520.

Daniel M. Bear, Kevin Feigelis, Honglin Chen, Wanhee Lee, Rahul Venkatesh, Klemen Kotar, Alex Durango, and Daniel L. K. Yamins. Unifying (machine) vision via counterfactual world modeling. *arXiv preprint arXiv:2306.01828*, 2023.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The Arcade Learning Environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013. doi: 10.1613/jair.3912. arXiv:1207.4708.

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, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.15391.

Lars Buesing, Theophane Weber, Sébastien Racanière, S. M. Ali Eslami, Danilo Rezende, David P. Reichert, Fabio Viola, Frederic Besse, Karol Gregor, Demis Hassabis, and Daan Wierstra. Learning and querying fast generative models for reinforcement learning. *arXiv preprint arXiv:1802.03006*, 2018.

Silvia Chiappa, Sébastien Racanière, Daan Wierstra, and Shakir Mohamed. Recurrent environment simulators. In *International Conference on Learning Representations (ICLR)*, 2017. arXiv:1704.02254.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1905.11979.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, et al. WorldScore: A unified evaluation benchmark for world generation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. doi: 10.1109/ICCV51701.2025.02573. arXiv:2504.00983.

Jianjie Fang, Yingshan Lei, Qin Wan, Ziyou Wang, Yuchao Huang, Yongyan Xu, Baining Zhao, Weichen Zhang, Chen Gao, Xinlei Chen, and Yong Li. iWorld-Bench: A benchmark for interactive world models with a unified action generation framework. *arXiv preprint arXiv:2605.03941*, 2026.

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. AdaWorld: Learning adaptable world models with latent actions. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2503.18938.

David Ha and Jürgen Schmidhuber. World Models. *arXiv preprint arXiv:1803.10122*, 2018. Also as Recurrent World Models Facilitate Policy Evolution, NeurIPS 2018.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with discrete world models. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.02193.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2. arXiv:2301.04104 (Mastering Diverse Domains through World Models).

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, et al. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/CVPR52733.2024.02060. arXiv:2311.17982.

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, et al. Model-based reinforcement learning for Atari. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1903.00374.

Nan Rosemary Ke, Aniket Didolkar, Sarthak Mittal, Anirudh Goyal, Guillaume Lajoie, Stefan Bauer, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Christopher Pal. Systematic evaluation of causal discovery in visual model based reinforcement learning. In *NeurIPS Datasets and Benchmarks Track*, 2021. arXiv:2107.00848.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, et al. WorldModelBench: Judging video generation models as world models. In *NeurIPS Datasets and Benchmarks Track*, 2025. arXiv:2502.20694.

Yufeng Lin and Jialu Zhang. ScratchWorld: Evaluating if world models compute executable consequences. *arXiv preprint arXiv:2606.31689*, 2026.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning (ICML)*, 2022. arXiv:2202.03169.

Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the Arcade Learning Environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018. doi: 10.1613/jair.5699. arXiv:1709.06009.

Willi Menapace, Stéphane Lathuilière, Sergey Tulyakov, Aliaksandr Siarohin, and Elisa Ricci. Playable video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. doi: 10.1109/CVPR46437.2021.00993. arXiv:2101.12195.

Willi Menapace, Stéphane Lathuilière, Aliaksandr Siarohin, Christian Theobalt, Sergey Tulyakov, Vladislav Golyanik, and Elisa Ricci. Playable environments: Video manipulation in space and time. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. doi: 10.1109/CVPR52688.2022.00357. arXiv:2203.01914.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2209.00588.

Vincent Micheli, Eloi Alonso, and François Fleuret. Efficient world models with context-aware tokenization. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2406.19320.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026. doi: 10.1109/WACV61042.2026.00099. arXiv:2501.09038.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for Physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in Atari games. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. arXiv:1507.08750.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world models are happy with 100k interactions. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2303.07109.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954. arXiv:2102.11107.

Yiqing Shen, Aiza Maksutova, Chenjia Li, and Mathias Unberath. Counterfactual world models via digital twin-conditioned video diffusion. *arXiv preprint arXiv:2511.17481*, 2025.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. Short version as FVD: A new metric for video generation, ICLR 2019 Workshop on Deep Generative Models for Highly Structured Data.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2408.14837.

Zizhao Wang, Xuesu Xiao, Zifan Xu, Yuke Zhu, and Peter Stone. Causal dynamics learning for task-independent state abstraction. In *International Conference on Machine Learning (ICML)*, 2022. arXiv:2206.13452.

Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, Zhengzhong Liu, Eric P. Xing, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.06114.

Yang Yu, Shiyuan Zhang, Yifei Sheng, Haoxiang Ren, and Haoxin Lin. How should world models be evaluated for embodied decision-making? A decision-making-centric position. *arXiv preprint arXiv:2606.15032*, 2026.

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. doi: 10.1109/CVPR.2018.00068. arXiv:1801.03924.

Shengjun Zhang, Zhang Zhang, Simin Huang, Zhenyu Tang, Hanyang Wang, Chensheng Dai, Min Chen, Yifan Li, Yuxin Li, Yingjie Chen, Hao Liu, Chen Li, et al. MBench: A comprehensive benchmark on memory capability for video world models. *arXiv preprint arXiv:2606.00793*, 2026.

Weipu Zhang, Gang Wang, Jian Sun, Yetian Yuan, and Gao Huang. STORM: Efficient stochastic transformer based world models for reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.52202/075280-1182. arXiv:2310.09615.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2411.04983.

A Related work

Table 4: Evaluations of world models and video generators, compared on the properties that decide whether a score can measure causal response. ✓: the property holds, verified in the paper; –: verified absent; ○: not applicable to that evaluation; ?: not reported or not verified by us. *Action intervention*: a real agent action is changed on a known window. *Two rollouts*: outcomes under two intervention settings (two action sequences, or two environment settings) are compared. *Exact truth, both arms*: simulator ground truth for both settings. *No judge*: no vision-language model (VLM) or human rater in the score. *Pixel world models*: evaluates trained action-conditioned pixel or video world models. *Direction*: the score can report that the model's change under an action intervention points in the wrong direction relative to the exact counterfactual world response (which requires two arms and exact truth; a score that compares generated motion with a commanded direction, as iWorld-Bench's trajectory following does, is action-aware but has no counterfactual reference and is marked absent). *Latent probe*: the same score is applied inside the model's representation.

	action intervention	two rollouts	exact truth, both arms	no judge	pixel world models	direction	latent probe
FVD, LPIPS, PSNR (Unterthiner et al., 2018; Zhang et al., 2018)	–	–	○	✓	✓	–	–
VBench (Huang et al., 2024)	–	–	○	?	✓	–	–
MBench (Zhang et al., 2026)	–	–	○	–	✓	–	–
iWorld-Bench (Fang et al., 2026)	–	–	○	✓	✓	–	–
VideoPhy (Bansal et al., 2025)	–	–	○	–	✓	–	–
Physics-IQ (Motamed et al., 2026)	–	–	○	✓	✓	–	–
WorldModelBench (Li et al., 2025)	–	–	○	–	✓	–	–
Genie ΔPSNR (Bruce et al., 2024)	latent	✓	one arm	✓	✓	–	–
ScratchWorld (Lin & Zhang, 2026)	program	✓	✓	✓	– (LMs)	–	–
CausalWorld (Ahmed et al., 2021)	env. vars	✓	✓	✓	– (policies)	○	–
Ke et al. (2021)	env. vars	✓	✓	✓	– (struct.)	○	–
This work	✓	✓	✓	✓	✓	✓	✓

World models. Action-conditional prediction of Atari frames was posed by Oh et al. (2015) and extended to long horizons by Chiappa et al. (2017), whose models were judged by prediction error and by whether an agent could act through them; Buesing et al. (2018) made such rollouts the substrate for planning. Learned simulators for control in the modern sense go back to Ha & Schmidhuber (2018); the recurrent state-space models of the Dreamer line made them practical for pixels (Hafner et al., 2021; 2025), SimPLe established the Atari-100k protocol (Kaiser et al., 2020), and transformer world models followed (Micheli et al., 2023; Robine et al., 2023; Zhang et al., 2023; Micheli et al., 2024). Diffusion world models (Alonso et al., 2024; Valevski et al., 2025), large video world models (Bruce et al., 2024; Xiang et al., 2024; NVIDIA et al., 2025; Yang et al., 2024) and latent-space models used for planning (Zhou et al., 2025; Assran et al., 2025; Gao et al., 2025) extend the family; in every one of them the counterfactual question, how the future would differ under a different action, is the property the planner relies on. We evaluate DIAMOND, IRIS, STORM and a recurrent state-space model (RSSM) because the first two are the reference Atari-100k models with public weights, the third a transformer over stochastic latents that we train per game, the last our own DMControl model, and all have an emulator that can be replayed for exact counterfactual truth (Bellemare et al., 2013; Machado et al., 2018; Tassa et al., 2018).

Evaluation. Benchmarks for video world models score the plausibility and memory of a single generated trajectory, following the video-generation tradition of FVD and perceptual metrics (Unterthiner et al., 2018; Zhang et al., 2018; Huang et al., 2024) and its physics-focused successors (Bansal et al., 2025; Motamed et al., 2026; Li et al., 2025). MBench (Zhang et al., 2026) decomposes memory into entity, environment and *causal* consistency, the last asking whether the consequences of an event the model has seen continue to unfold; it is an observational, one-trajectory score, judged by rule-based metrics and a vision-language model (VLM) on real video, and contains no intervention. iWorld-Bench (Fang et al., 2026) evaluates fourteen models on generation

quality, trajectory following (the generated camera motion against the commanded direction, an action-aware but single-rollout score) and memory with computed metrics validated against human preference, again without paired interventions against an exact counterfactual. ScratchWorld (Lin & Zhang, 2026) is the closest in spirit: it observes that overlap scores reward copying persistent state, replays Scratch programs exactly on a pinned virtual machine to obtain replay-verified transitions and counterfactual outcomes, and reports intervention sensitivity among its diagnostics; its subjects are prompted language and reasoning models predicting symbolic Scratch state, and its primary score is a value-aware changed-field F_1, with the overlap-based full-state accuracy kept only as a diagnostic of the copying confound. Action conditioning of pixel and video world models has been evaluated before, in every evaluation we know of observationally, that is, from generated futures without an exact counterfactual reference: playable video generation (Menapace et al., 2021; 2022) scores whether learned discrete actions produce distinct and consistent motions in one generated video, the ancestor of Genie's ΔPSNR; ACT-Bench (Arai et al., 2024) scores whether a driving world model follows a commanded trajectory, read back by a learned estimator; WorldScore (Duan et al., 2025) includes a controllability dimension judged on one generated video, and AdaWorld (Gao et al., 2025) evaluates its latent actions by simulation quality and visual planning. Each asks whether the action matters, is followed, or is distinguishable in a single trajectory; none replays the same situation under a different action to obtain the exact counterfactual and compares the model's change with the world's. Counterfactual world modeling (Bear et al., 2023) shares a word with us and not an object: it queries a masked predictor with counterfactual prompts to extract structure from images, and evaluates nothing against a replayed truth; Shen et al. (2025) generates counterfactual videos from a digital twin rather than measuring a model's response. We evaluate trained pixel world models under action interventions with a vector response that has a direction, a magnitude and an error, and probe the model's latent.

Causality. The distinction between observing a trajectory and intervening on it is Pearl's (Pearl, 2009); causal representation learning asks when latent causes are identifiable from data (Schölkopf et al., 2021; Lippe et al., 2022), CausalWorld (Ahmed et al., 2021) evaluates manipulation agents under interventions on the task's causal variables (generalisation and robustness), and Ke et al. (2021) evaluate causal structure discovery in structured or latent state. Causal confusion in imitation learning (de Haan et al., 2019) is the policy-side analogue of what our latent probes find: a learner that reads the action input rather than its consequence. Causal dynamics learning (Wang et al., 2022) uses a learned causal graph of the dynamics for state abstraction. None of these measures whether a trained pixel world model's decoded future responds correctly to an intervention, which is the quantity built here.

B PROOFS, ESTIMATORS AND THE OFF-SUPPORT BOUND

Estimators. For a pair of rollouts with predicted futures p, p' and true futures t, t', at delay d after the intervention we report three quantities of the vector response: *causal alignment* $A_d = \langle \Delta p_d, \Delta t_d \rangle / \|\Delta t_d\|^2$, the projection of the model's change onto the true change (1 for the true change, above 1 for a change in the true direction that overshoots it, 0 for an unrelated change, negative for the wrong direction); *response magnitude* $R_d = \|\Delta p_d\| / \|\Delta t_d\|$; and the *normalised response error* $\varepsilon_d = \|\Delta p_d - \Delta t_d\|^2 / \|\Delta t_d\|^2 = R_d^2 + 1 - 2A_d$, which is 0 for a perfect response, 1 for no response, 2 for a world-sized response orthogonal to the truth, and 4 for a world-sized response in the opposite direction; the identity holds per episode pair, and the tables report either means over seeds of per-seed medians or medians over pooled pairs (each caption says which), so aggregated columns need not satisfy it. Alignment alone is a projection and cannot distinguish an absent response from a misdirected one; the triple can. All three divide by $\|\Delta t_d\|$ and are undefined when it is near zero, so every cell is guarded by a floor calibrated from the observed divergences and the number of excluded cells is reported. On the excluded cells, where the world's change is below the floor (small, not necessarily zero), we report the unnormalised *below-guard response* $\|\Delta p_d\|$, the change the model makes where the world barely changes; it is a change where the world did not change only when Δt_d is exactly zero, which the decoupling analysis of Appendix H measures separately. On the same rollouts we compute a consistency proxy (seam discontinuity plus roll-out roughness, blind to the intervention), PSNR of p against t, and ΔPSNR. A response can also be split, using the ground truth alone, by where it lands relative to the world's change:

Lemma 1 (off-support bound). *Let S be the set of coordinates on which Δt_d is nonzero and $O_d = \|\Delta p_{d,S^c}\|^2/\|\Delta p_d\|^2$ the fraction of the model's change energy outside it. Then $|A_d| \le R_d\sqrt{1-O_d}$.*

Proof. $|\langle \Delta p_d, \Delta t_d\rangle| = |\langle \Delta p_{d,S}, \Delta t_d\rangle| \le \|\Delta p_{d,S}\|\,\|\Delta t_d\| = \sqrt{1-O_d}\,\|\Delta p_d\|\,\|\Delta t_d\|$ by Cauchy-Schwarz; divide by $\|\Delta t_d\|^2$ (with $\Delta t_d \neq 0$ by the guard and O_d defined as 0 when $\Delta p_d = 0$). The bound is stated for the exact support. For a thresholded support S_τ (Appendix G keeps the coordinates where the world's change exceeds one percent of its maximum), let $q_\tau = \|\Delta t_{d,S_\tau^c}\|^2/\|\Delta t_d\|^2$ be the fraction of the world's change energy outside S_τ and O_τ the model's off-support fraction with respect to S_τ; splitting the inner product over S_τ and its complement and applying Cauchy-Schwarz to each part gives $|A_d| \le R_d\big[\sqrt{(1-O_\tau)(1-q_\tau)}+\sqrt{O_\tau q_\tau}\big]$, which reduces to the lemma at $q_\tau = 0$.

So a change that lands mostly outside the world's change can be well aligned only by being correspondingly larger than the world's: at world-sized magnitude, alignment is bounded by $\sqrt{1-O_d}$; Appendix G uses this to read the off-support fraction.

Proof of Proposition 1. Take $M_1(c,a;s) = E(c,a)$. Let $\tilde{a}(c,s)$ be a measurable function of the context and the seed, but not of a, whose conditional law given c is $\mu(\cdot \mid c)$ when $s \sim \nu$; since s is uniform on $[0,1]$, the conditional quantile transform of $\mu(\cdot \mid c)$ provides one (for finite action sequences it is the inverse of the conditional distribution function). Define $M_2(c,a;s) = E(c,\tilde{a}(c,s))$, which ignores a. Since s is independent of (c,a), the pair $(c,\tilde{a}(c,s))$ has the law of (c,a); hence $(c,E(c,\tilde{a}(c,s)))$ has the law of $(c,E(c,a))$, that is, $(c,M_2(c,a;s))$ and $(c,M_1(c,a;s))$ have the same joint law, and so does any functional of (c,x) or of its law. Under common random numbers the same s is supplied to both arms, so $M_2(c,a';s)-M_2(c,a;s) = E(c,\tilde{a}(c,s))-E(c,\tilde{a}(c,s)) = 0$ for every (c,a,a',s), whereas $M_1(c,a';s) - M_1(c,a;s) = E(c,a') - E(c,a)$. When actions are independent of the context, as in our measurement, $\tilde{a}$ depends on the seed alone.

Remarks on the assumptions. Three assumptions are implicit in Definition 1 and Proposition 1 and are stated here. (i) The proposition is non-trivial only when the environment's response $E(c,a') - E(c,a)$ is nonzero on a set of (c,a,a') of positive probability; every game in the suite has a nonzero true response at delay 20 under its chosen intervention (Appendix C). (ii) Exact replay requires the context to determine the emulator state: in our protocol the context is a reset seed, a warm-up and an action prefix, which do, whereas four visible frames alone need not (two states can share their last four frames), so the measurement is of the model given the frames and the replayed truth given the state. (iii) The model's response is defined under a fixed coupling s of its randomness; common random numbers make that coupling reproducible across the two arms and across runs, and the response under one coupling is not a functional of the model's conditional output distribution alone (two models with the same conditional law can differ in their coupled response). The independence of s from (c,a) assumed in Proposition 1 concerns the evaluation distribution under which one-rollout scores are compared; the paired measurement fixes s in each episode and does not rely on it.

Discussion of Corollary 1. The corollary is the triangle inequality. It gives a sufficient condition for $\varepsilon < 1$, not a necessary one (Section 2); a score that sees only the two error norms, such as PSNR, certifies $\varepsilon < 1$ only when $2e < \|\Delta t\|$, since errors of those norms in opposite directions reach the bound. For frames in $[-1,1]$ with $D = 12{,}288$ coordinates, a PSNR of P dB is an error norm $e = 2\sqrt{D}\,10^{-P/20}$, about 7.0 at 30 dB, while the true response $\|\Delta t\|$ at delay 20 ranges from about 2 to 35 across the 78 architecture-game cells (median 8; the pooled-median estimator of Table 22). Computed per game from our measurements (Appendix Q), the equal-error ratio $2e/\|\Delta t\|$ exceeds one in 23, 24 and 25 of 26 games for DIAMOND, IRIS and STORM (the cells nearest the threshold are DIAMOND Breakout and Gopher at 1.014 and 1.014; STORM Qbert at 0.982) and lies between 1.1 and 3.8 in the never-aligned cells. Where the intervened arm's error was also stored, the exact ratio $(\|p-t\| + \|p'-t'\|)/\|\Delta t\|$ exceeds one in 25 of 26 IRIS games and 26 of 26 STORM games, and in both DIAMOND flagship cells (Alien 2.82, Pong 2.05); it differs from the equal-error value by up to 0.87 (IRIS UpNDown, 1.26 against 2.13), so the equal-error ratio, whose two medians are taken over different pairs, is a summary under a stated assumption, not an approximation with a guaranteed accuracy, and the exact ratio is reported wherever it exists. At the accuracy of current world models, PSNR does not certify a response error below the response itself.

Algorithm 1 Paired-intervention measurement of causal response, one episode

Require: context c, actions a, window W, delays D, emulator E, model M, seed s
$a' \leftarrow a$; for $t \in W$: $a'_t \leftarrow$ a different action (offset chosen per game by the emulator)
$t \leftarrow E(c, a),\ t' \leftarrow E(c, a')$ *exact truth by replay*
$p \leftarrow M(c, a;\ s),\ p' \leftarrow M(c, a';\ s)$ *common random numbers*
assert $p_{<W} = p'_{<W}$ *a cause cannot act backwards*
for all $d \in D$ **do**
  $\Delta t \leftarrow t'_d - t_d;\ \Delta p \leftarrow p'_d - p_d$
  record the mean of the two per-arm consistency scores of $p_{\le d}$ and $p'_{\le d}$ against c, and PSNR(p_d, t_d) per-arm one-rollout scores
  if $\|\Delta t\|$ below the guard: record the below-guard response $\|\Delta p\|$ and **continue**
  record A_d, R_d, ε_d and ΔPSNR *two-rollout*
  record the same triple in latent space when the model exposes one
end for
return per-delay records (aggregated by bootstrap median, guarded, gated)

C Measurement protocol and models

Emulator and preprocessing. Atari games are run through the Arcade Learning Environment v5 interface with frame skip 4 and sticky actions disabled (`repeat_action_probability` $= 0$), so a replay from the same reset seed under the same actions reproduces the same frames bit for bit. Frames are resized to 64×64 by area interpolation and scaled to $[-1, 1]$, the convention of both released models; every metric is computed on these arrays with $D = 64{\cdot}64{\cdot}3 = 12{,}288$ coordinates. DMControl tasks are rendered at the same size through the RSSM's own pipeline.

Registration check. Each harness was checked for a one-step offset between the actions it consumes and the frames it returns, because the pre-window identity assertion cannot see one: the test changes the single action at index K after the context of a random action sequence, in three episodes of Pong and Breakout, runs the harness for six steps and requires the first differing predicted frame to be index $K{+}1$ (the test uses interventions with an immediate emulator effect on Pong and Breakout, so it does not depend on the emulator's own first difference, which is printed alongside and was $K{+}1$ in these runs). STORM's context of eight frames starts four steps earlier so that the intervention hits the same absolute emulator time where the game has a warm-up of at least four steps (19 games); in the seven games whose chosen warm-up is zero the shift is clipped and STORM's intervention lands four emulator steps later than for the four-frame models. An earlier version of the IRIS and STORM harnesses failed this check by one step (they responded at index K); every IRIS and STORM measurement in this paper was taken after the correction except the IRIS reward-head ranking of Appendix M, which was not rerun (the IRIS driver consumed actions from index C instead of $C{-}1$; the STORM harness decoded the frame after the next one); the test is part of the released code but is run separately from the sweeps, and its pass is recorded as prose, not as a stored artefact. The released test covers the three Atari harnesses; the DMControl RSSM pairs each observation with the action taken at it, as its training data does, and was checked by inspection of its harness rather than by the test.

Contexts and interventions. An episode is defined by a reset seed. After w random warm-up steps (identical in both arms, drawn from a generator seeded by the episode) the model receives a context of 4 frames (DIAMOND), 1 frame (IRIS, which is reset from that frame alone: its key-value cache is emptied and filled with the frame's 16 tokens, after which it accumulates the model's own sampled tokens; the three earlier context frames serve only as the consistency reference), 8 frames (STORM, whose per-layer cache receives one entry per context frame and is never truncated within the horizon) or 10 frames (RSSM). IRIS's cache reaches its 340-token limit once, at rollout step 19, and is rebuilt from the tokens of the frame the model itself predicted at step 18, identically in both arms, so delays 16 and 20 are produced after that rebuild. IRIS and STORM sample ancestrally at temperature one (no argmax); the RSSM's stochastic draw is replaced by its mean for evaluation. Actions are drawn uniformly for the context and the 24-step horizon (30 for the RSSM). The intervention replaces the action at steps 3 to 10 after the context (an 8-step window, $K = 3$) by the action shifted by an offset k modulo the action set, and leaves every other action unchanged, so the two arms differ only on the window. Delays are counted from the first step of the window: $d \in \{1, 2, 3, 5, 8, 12, 16, 20\}$,

extended to $\{28, 36, 44\}$ at horizon 48. For the RSSM the intervention is a single step at $K = 4$ after the context whose continuous action is set to the minimum of the action range in one arm and to the maximum in the other (no cyclic offset and no window), with delays $\{1, 2, 3, 5, 8, 12, 16, 20, 24\}$.

Common random numbers, and what the seeds vary. Both arms of an episode call the harness with the same seed, and the harness reseeds the torch generator once per rollout step from the episode index and the step (DIAMOND before each frame's denoising loop, IRIS before each frame's token draws, STORM once before the context encoding and before each step), so the two arms draw identical noise until the first differing action changes the logits. The emulator reset seed and the warm-up generator are functions of the episode index, so the sixty start states of a game and their warm-up actions are shared by all seven evaluation seeds and, for DIAMOND and IRIS, by both architectures (STORM shares the reset seeds, but its warm-up is four steps shorter where the game has one, clipped at zero in seven games, and its frames are max-pooled, so its contexts and true futures are matched within STORM only); the seven evaluation seeds vary the base action sequence, and through it the contexts and the true futures, and nothing else. Seed variation in this paper is therefore variation over action sequences on fixed start states, not a sweep over the model's sampling noise. The pre-window identity assertion compares the two arms' predicted frames at indices 0 to $K-1$ of every episode pair (tolerance 0 for DIAMOND and IRIS; 10^{-5} for STORM after cuDNN determinism was enforced and TF32 disabled, following a 5×10^{-7} discrepancy in a first GPU run) and aborts the run on failure; the recorded maximum pre-window difference is 0 in every result file (980, 420 and 252 seed-level entries for the three architectures). The true arms come from a deterministic emulator (frame-skip 4, sticky actions off) driven by identical actions up to the window, so their identity before it holds by construction. Common random numbers make the measured response reproducible under one coupling of the model's randomness; they do not make it a functional of the model's conditional distribution alone.

Below the guard is not no change. The guard excludes (pair, delay) cells whose true response lies below a data-calibrated floor; such cells are excluded from the three ratios and counted, but a true response below the floor is small, not zero. Only the decoupling analysis of Appendix H uses exact equality ($\Delta t = 0$) to define an inert pair, and no statement in the paper treats a below-guard cell as an unchanged world.

Choosing the intervention with the emulator. Under the fixed choice ($k = 3$, $w = 0$) fourteen games showed no true response at delay 20 in the sweeps (fifteen in the 20-episode search, which also returns zero for RoadRunner), because the intervention fell inside an idle window after reset or the offset was a no-op. The search (`scripts/emulator_intervention_search.py`) is model-blind and sequential: for each game it evaluates every non-zero cyclic offset $k \in \{1, \ldots, n_{\text{act}} - 1\}$ of the game's action set at $w = 0$ and keeps the one that maximises the median over 20 emulator-only episodes of the true response $\|\Delta t_{20}\|$ (the same quantity the sweeps record); it then evaluates $w \in \{30, 60, 120, 240\}$ at that offset and keeps the $w \in \{0, 30, 60, 120, 240\}$ with the largest median response, so $w = 0$ is retained only when it is also the maximum (seven games) and the offset is not re-searched at the chosen warm-up. Ties go to the smallest candidate; in the fourteen games where every offset gives a zero response at $w = 0$ the offset is therefore $k = 1$ by tie, and no game ended without a response at its chosen (k, w); $w = 240$ was never selected. The search episodes are the first 20 episodes of evaluation seed 42 (the same start states and action sequences), so 20 of the 420 evaluation pairs per cell (4.8%) were also used to choose the intervention; the other six seeds use new action sequences on the same start states. The resulting (k, w) per game was fixed before any model was run under it; the search file carries no timestamp field, and its modification time (4 September 2026, before every emulator-chosen model run) is the record.

Policy-context control. This arm tests policy-generated contexts and base trajectories with replayed interventions; it is not an evaluation of two independently adapting closed-loop policies. For Alien, Pong, BankHeist, Freeway, Breakout and Boxing, the warm-up, the context and the base action sequence are produced by the released agent's own actor (DIAMOND's and IRIS's actor-critic, shipped in the same checkpoint as the world model) acting in the real emulator, sampling from its policy with an epsilon-random action with probability 0 (and 0.1 for Alien and Pong); the intervention replaces the same eight-step window by the per-game emulator-chosen offset, both arms are replayed exactly (the policy's trajectory is asserted replayable), and the model imagines both under

common random numbers. Seven seeds, 30 episodes each. The reading rule was fixed before the runs: blindness persists if Alien stays below 0.05 and Pong stays tracked in both reference architectures, and the six-game Spearman correlation between the two context types is positive; either outcome is reported. That correlation is $+0.03$ for DIAMOND and $+0.83$ for IRIS.

Accuracy matching. The matched comparison of Section 4.1 uses the fixed intervention at delay 20. For each architecture, Pong's interquartile band of factual-arm PSNR is computed over all 420 Pong episode pairs (DIAMOND $[34.5, 37.1]$ dB, IRIS $[33.1, 36.0]$ dB); an Alien episode pair is matched when its own factual-arm PSNR lies inside that band (inclusive), among the 348 pairs above the guard. The unit is the episode pair, pooled over the seven evaluation seeds (the matched pairs come from every seed); the estimate is the median delay-20 alignment over the matched pairs with a percentile bootstrap interval from 2,000 resamples of the pairs; no seed-level clustering is applied and no test is run, the comparison being between the matched median and the all-pairs median of the same architecture. Under the emulator-chosen intervention the matched sets are too small to report (3 DIAMOND, 1 IRIS and 13 STORM pairs, each matched on its own architecture's Pong band). The band, rather than a symmetric tolerance, was chosen so that the matched Alien pairs sit inside Pong's typical accuracy; no other rule was tried. Matched estimates at delay 20 under the fixed intervention: DIAMOND, 123 of 348 guarded Alien pairs inside the band, median alignment 0.002 (95% bootstrap CI $[-0.006, 0.028]$); IRIS, 155 of 348 pairs, 0.015 ($[0.000, 0.062]$). PSNR and consistency are compared between games, not against a static-prediction baseline, which is not reported; alignment does not use PSNR, so the causal verdict does not depend on that baseline.

Model sampling. DIAMOND is sampled with 3 denoising steps from the released per-game checkpoints; IRIS samples one categorical token per position from its released per-game checkpoints; the RSSM is our own DreamerV3-style model with the stochastic path made deterministic for evaluation. In every case the generator is re-seeded at each step with a function of the episode seed and the step index, identically in both arms, so that two rollouts that receive the same actions produce the same frames and the pre-intervention identity assertion holds exactly (maximum pre-window difference 0 in every file).

Scores on the same rollouts. Consistency is $\exp(-(\text{seam} + \text{roughness}))$, where seam is the root-mean-square difference between the first predicted frame and the last context frame and roughness the mean root-mean-square second difference along the rollout, both divided by the standard deviation of the context; it never sees the action. PSNR is computed between the un-intervened prediction and its true frame at each delay with data range 2, capped at 100 dB. ΔPSNR follows Genie: the PSNR of the un-intervened prediction against its truth minus the PSNR of the intervened prediction against the same un-intervened truth, so it measures how much the future moves when the action changes, not whether it moves correctly. The representation probes apply the same alignment, magnitude and error to the model's own latent for its predicted frames and to the latent it assigns to the true frames.

The RSSM and its states. The DMControl model is our own recurrent state-space model in the DreamerV3 style, trained per task from pixels with one training seed (42) on an offline set of intervention-free episodes (ten episodes of 250 steps at action repeat 2 for cheetah_run and walker_walk, twenty for the other four tasks, i.e. about 2,500 and 5,000 environment steps), cut into length-50 subsequences with stride 25 (90 and 180 subsequences), for 100 epochs at batch size 32 (200 and 500 gradient steps; Adam, learning rate 3×10^{-4}, gradient norm clipped at 100; loss = reconstruction MSE + reward MSE + KL with weight 1 and one free nat), and evaluated from the epoch with the lowest mean training loss (there is no held-out split). Its state is the pair (h, z): $h \in \mathbb{R}^{256}$ is the deterministic recurrent state, updated by a GRU from the previous (z, a); $z \in \mathbb{R}^{32}$ is the stochastic state, drawn from the prior $p(z \mid h)$ during imagination and from the posterior $q(z \mid h, e)$ during observation, where e is the CNN embedding of the current frame. The feature vector used by the decoder, the reward head and every latent quantity in the paper is the concatenation $[h, z] \in \mathbb{R}^{288}$. The imagined latent z^{im} is the feature after `imagine_step` under the supplied action; the latent assigned to the true frames z^* is the feature after `observe_step` teacher-forced on the true frames with the same actions; the stochastic draw is made deterministic (its mean) for evaluation so that both sides are functions of their inputs. This is why the action enters z^*: h carries a through the GRU whether the frames differ or not, which is the confound the frames-only con-

trol removes. These checkpoints are weak world models: on the probe's episodes their imagined frames reach 18 to 26 dB against the truth with no decay across the 30-step horizon, close to a static prediction, and their predicted rewards are uncorrelated with the true ones (Spearman -0.08 to 0.26). Their pixel blindness is therefore expected, and nothing in the paper treats it as evidence about competent RSSMs.

The third architecture. STORM releases no Atari checkpoints, so we train one world model per game with the authors' code and configuration (100k environment steps, 2-layer 512-wide transformer; two training seeds, the second reported in Appendix E), on the same GPU class as the rest of the paper. Two facts differ from the released models and are stated for fairness: the authors trained with bfloat16 mixed precision on hardware that supports it, and our GPUs do not, so we train with float16 mixed precision and loss scaling with an fp16-safe attention mask, the only code changes; and STORM's training environment uses the emulator's default sticky actions and a max over the last two of four repeated frames, so its evaluation replays use the same frame pooling with sticky actions disabled for exact truth, and its eight-frame context starts four steps earlier so that the intervention hits the same emulator time as for the four-frame models (except in the seven games without a warm-up, where the shift is clipped at zero and the intervention lands four steps later). Evaluation is float32 for all three architectures. Appendix J lists every game. A fourth architecture, TWM (Robine et al., 2023), trained per game with its authors' code in our software environment, did not produce functioning agents (scores at the random level on Pong, Breakout and Freeway) and its imagined futures do not depend on the action, so it is not reported; the failure is in our training run, not a statement about the model.

Visual share. For a latent probe, the full true latent change is $\Delta z^* = z_b^* - z_a^*$ with each posterior computed from its own frames and actions; the frames-only change Δz_F^* recomputes z_b^* from arm b's frames with arm a's actions, and the actions-only change Δz_A^* recomputes it from arm a's frames with arm b's actions. The visual share is $\|\Delta z_F^*\|/\|\Delta z^*\|$; it is a ratio of norms, not an additive decomposition, and frames-only alignment is reported only where the visual share is at least 0.25. Shared-baseline null: for ten games (Alien, Pong, BankHeist, Boxing, Breakout, Freeway, Gopher, Seaquest, MsPacman, Frostbite), 40 episodes each, three rollouts under the same actions with seeds s, $s + 10^4$, $s + 2 \cdot 10^4$; the reported value is the median over episodes and games of the cosine between the two seed differences at delay 20, and the noise-to-response ratio is the median of $\|\Delta p_{\text{seed}}\|/\|\Delta p_b\|$.

Cost. The core sweep of one architecture over 26 games (7 seeds $\times$ 60 episodes, two arms, 24 steps, both interventions where applicable) costs 5 GPU-hours for DIAMOND and 5 for STORM on one V100, and 50 for IRIS, whose token-by-token sampling is the slow step; training the 26 STORM models took 162 GPU-hours. The measurement itself is therefore cheaper than training one of the models it evaluates.

Sample sizes. Atari: 26 games, 7 evaluation seeds, 60 episodes per seed; the emulator-chosen intervention for all three architectures, the fixed intervention for DIAMOND and IRIS. DMControl: 6 tasks, 4 seeds, 20 episodes per seed. Horizon-48 ablation: Alien and Pong, 7 seeds, 60 episodes. Pre-sampling probe: 6 games, 7 seeds, 60 episodes; its action-pathway controls and the STORM hidden-state probe: 6 games, 7 seeds, 30 episodes.

Models and metrics. We evaluate three Atari-100k world models on all 26 games: the two reference models with public weights, DIAMOND (pixel diffusion) and IRIS (transformer over discrete tokens), and STORM (a transformer over stochastic categorical latents), which releases no weights and which we train per game with its authors' configuration; and our own RSSM on six DMControl tasks, a weak model by the measures below, that we use as a second substrate and as the testbed for the representation-probe controls: three Atari model families and one DMControl model, two substrates. As a method-free evaluation paper there is no performance state of the art to beat; the relevant state of the art is the *metrics*. Alongside our consistency proxy (a one-rollout score) and PSNR against the actual future (a reference-based score), we compute Genie's controllability ΔPSNR (Bruce et al., 2024), the closest existing measure: it needs two rollouts, like ours, but asks whether the action *matters* to the prediction, not whether the response is *correct*. On this suite the two agree: rank correlation between ΔPSNR and causal alignment is 0.83 to 0.98 at every delay for

DIAMOND and 0.69 to 0.88 up to delay 8 for IRIS ($n = 8$ games each, those with ΔPSNR at every delay), and no game in either architecture shows a large ΔPSNR with near-zero alignment. We read this as convergent validity for the measurement rather than a contrast; alignment adds sign, per-delay decomposition and guarded denominators. MBench's "causal consistency" (Zhang et al., 2026) is an observational one-trajectory score and belongs to the class the proposition covers. iWorld-Bench's trajectory following (Fang et al., 2026) compares the generated camera motion with the commanded direction, so it is action-aware and outside the class of Proposition 1, like the same-arm scores of Section 2, but it is a single-rollout score without a counterfactual reference: it can verify that the generated motion follows the command, not that the change from a different command is the world's. We also train, per game, a small inverse-dynamics model on emulator transitions and read the action back from consecutive *generated* frames, the one-rollout "action consistency" proxy; the statistic in Table 6 applies that reading to both arms and compares them, a two-rollout use of the one-rollout instrument. It fails its own gate (held-out emulator accuracy at least twice the majority baseline) on seven of sixteen (architecture, game) cells, and where usable does not cleanly separate causally blind from tracked games; we report it as an unreliable baseline rather than a contrast. Tables 5 and 6 give both.

Table 5: Genie's controllability ΔPSNR (dB) against causal alignment on the same rollouts, at delays 1 and 20. Rank correlation between the two is 0.83 to 0.98 (DIAMOND) and 0.69 to 0.88 up to delay 8 (IRIS); no cell has ΔPSNR > 1 dB with alignment < 0.1.

	DIAMOND: ΔPSNR / align.		IRIS: ΔPSNR / align.	
game	d=1	d=20	d=1	d=20
Alien	+0.17 / 0.18	+0.10 / 0.00	+0.02 / 0.07	+0.03 / 0.00
Pong	+1.50 / 0.52	+1.57 / 0.85	+0.00 / 0.26	+0.26 / 0.22
BankHeist	+0.00 / 0.00	+0.00 / 0.00	+0.00 / 0.00	+0.00 / 0.00
Boxing	+1.57 / 0.65	+0.28 / 0.07	+0.47 / 0.26	+0.33 / 0.06
Breakout	+5.68 / 0.79	+4.11 / 0.93	+2.11 / 0.71	+1.28 / 0.68
Freeway	+0.11 / 0.03	+0.01 / 0.00	+0.39 / 0.44	+0.05 / 0.17
CrazyClimber	+0.73 / 0.39	+0.06 / 0.07	+0.03 / 0.08	+0.32 / 0.24
DemonAttack	+2.28 / 1.17	+1.52 / 0.64	+0.55 / 0.23	+0.38 / 0.06

Table 6: Inverse-dynamics "action consistency". The inverse-dynamics model (IDM) is a one-rollout instrument: it reads an action from consecutive generated frames of a single rollout. The statistic displayed here applies it to both arms and compares the two readings, so it is a two-rollout use of a one-rollout instrument and is not itself an action-blind score. Gate: held-out accuracy on emulator frames must be at least twice the majority baseline. The alignment column is measured on the IDM rollouts themselves (a separate run from Table 23). *arms differ*: fraction of episodes in which the IDM reads different actions from the two generated arms, at delay 1 (inside the window) and 20 (after it).

	game	IDM gate (acc / majority)	arms differ d=1	arms differ d=20	causal align. d=20
DIAMOND	Alien	fail (0.11/0.06)	–	–	–
DIAMOND	BankHeist	fail (0.05/0.06)	–	–	–
DIAMOND	Boxing	pass (0.32/0.06)	0.93	0.57	0.073
DIAMOND	Breakout	pass (0.59/0.25)	0.58	0.34	0.921
DIAMOND	CrazyClimber	fail (0.14/0.11)	–	–	–
DIAMOND	DemonAttack	pass (0.49/0.17)	0.94	0.12	0.637
DIAMOND	Freeway	pass (0.83/0.33)	0.57	0.28	0.000
DIAMOND	Pong	fail (0.18/0.17)	–	–	–
IRIS	Alien	fail (0.05/0.06)	–	–	–
IRIS	BankHeist	fail (0.05/0.06)	–	–	–
IRIS	Boxing	pass (0.33/0.06)	0.82	0.66	0.056
IRIS	Breakout	pass (0.60/0.25)	0.58	0.40	0.639
IRIS	CrazyClimber	fail (0.17/0.11)	–	–	–

continued on the next page

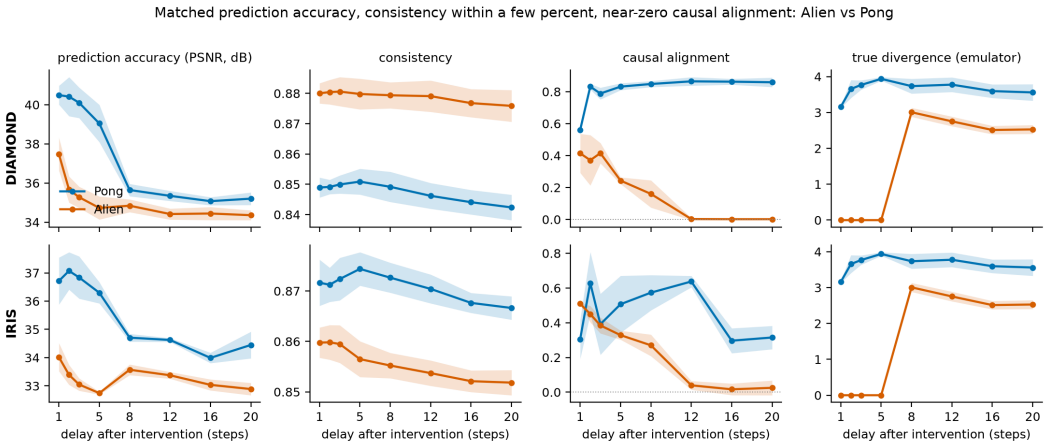

Figure 4: Alien vs Pong in DIAMOND (top) and IRIS (bottom) under the fixed intervention (offset 3, no warm-up; Table 1 gives both interventions): delay-20 prediction accuracy within 0.9 dB (DIAMOND) and 1.6 dB (IRIS), consistency within a few percent, and delay-20 alignment of 0.002 against 0.86 (DIAMOND) and 0.025 against 0.31 (IRIS). Right: the emulator's true effect in Alien is invisible for five steps, then appears.

	game	IDM gate (acc / majority)	arms differ d=1	arms differ d=20	causal align. d=20
IRIS	DemonAttack	pass (0.49/0.17)	0.67	0.31	0.059
IRIS	Freeway	pass (0.83/0.33)	0.82	0.21	0.168
IRIS	Pong	pass (0.41/0.17)	0.60	0.43	0.203

D THE FLAGSHIP PAIR OVER DELAY AND UNDER THE AGENTS' OWN CONTEXTS

Figure 4 plots prediction accuracy, consistency and causal alignment against delay for the flagship pair in the two released models; the policy-context control follows. IRIS conditions on a single frame (Appendix C); whether that difference contributes to its lower absolute alignment is not something these data can isolate, and we do not attribute it.

A second training seed of the released models. DIAMOND and IRIS are released as one checkpoint per game, so we trained a second seed of each on the flagship pair with the authors' code and default Atari-100k configuration (seed 2; DIAMOND 1,000 epochs, IRIS 600 epochs, 100k environment steps, one V100 per game) and measured it with the sweep of Section 3 (7 evaluation seeds $\times$ 60 episodes). The reading was fixed before training: the replicate reproduces the flagship if seed-2 Alien is below 0.05 and seed-2 Pong is above the architecture's tracked threshold at delay 20 (DIAMOND under both interventions, IRIS under the emulator-chosen one). Delay-20 alignment, seed 2 against the released checkpoint: DIAMOND Alien 0.004 against 0.003 (chosen) and 0.000 against 0.002 (fixed), Pong 0.83 against 0.85 and 0.85 against 0.86; IRIS Alien 0.003 against 0.001 and 0.009 against 0.025, Pong 0.25 against 0.22 and 0.31 against 0.31; prediction PSNR of the replicates is within 2.6 dB of the released checkpoints in every cell. Both replicates satisfy the pre-declared reading.

Contexts from the trained policy. Random actions are how the models' replay buffers begin, not how their agents act, so we repeated the measurement on six games with the warm-up, the context and the base action sequence produced by each released agent's own actor in the real emulator (Appendix C). The flagship survives: under its own policy DIAMOND's alignment on Alien is -0.005 (0.015 with ten percent random actions) and IRIS's 0.005, at 31.0 and 29.4 dB, BankHeist stays at 0.000 for both, and on Pong IRIS retains directional alignment at 0.61 under the pure policy and DIAMOND at 0.77 with ten percent random actions (levels under this context distribution; in the random-context taxonomy of Table 24 IRIS Pong is aligned-then-forgotten). The control also shows what an evaluation distribution can do: under the pure policy the two true Pong arms reconverge by

delay 20 in 192 of 210 pairs (180 of 210 with ten percent random actions; the trained policy pins the paddle), so DIAMOND's Pong value is quoted from the ten-percent-random condition and rests on its remaining 30 pairs; on Breakout the true arms reconverge in 147 of 210 pairs, and on the other 63 DIAMOND's delay-20 alignment is 0.000 in every seed, against 0.93 under random contexts, so DIAMOND's Breakout verdict changes too (IRIS falls from 0.68 to 0.11, at its tracked threshold); and Freeway changes verdict, never-aligned for DIAMOND (0.000) and aligned-then-forgotten for IRIS (0.17) under random contexts and at tracked levels under the policies' contexts (0.47 and 0.59). Causal blindness on the flagship is therefore not an artefact of random-action contexts, and a game's verdict can depend on the distribution of contexts it is measured on, a limitation of any single-distribution evaluation, ours included.

E CROSS-ARCHITECTURE TRANSFER AND THE TAXONOMY

Table 7 gives the class counts per architecture; the threshold rule, its sensitivity, the causal horizons and the consensus ranking follow. The second STORM training seed transfers to the other two families at $\rho = 0.50$ (DIAMOND) and 0.54 (IRIS); Table 8 gives both seeds' values for every game. The absolute levels differ: the median delay-20 alignment over the 26 games is 0.15 for DIAMOND, 0.06 for IRIS and 0.03 for STORM, and the two architectures of a pair assign the same class to 17, 15 and 16 of the 26 games (DIAMOND and IRIS, DIAMOND and STORM, IRIS and STORM).

Table 7: Taxonomy of causal failure across 78 (game, architecture) cells under emulator-chosen interventions. Rules applied in order, with A_{20} the delay-20 alignment and the peak the maximum over delays 1 to 8: *partial* if A_{20} or any of delays 1 to 8 is unmeasurable; *tracked* (directionally, not necessarily accurately) if $A_{20} \geq T$ and A_{20} is at least half the peak; *aligned-then-forgotten* if the peak is at least T and A_{20} is below half the peak; *never-aligned* if the peak is below T; otherwise (peak at least T, $A_{20} < T$, A_{20} at least half the peak) *partial*; the rules are exhaustive and mutually exclusive. T is 0.40 for DIAMOND, written with the pre-registered predictions, and half of the architecture's own Pong delay-20 alignment for IRIS and STORM (DIAMOND 0.40, IRIS 0.11, STORM 0.19).

	tracked	aligned-then-forgotten	never-aligned	partial
DIAMOND	11	6	7	2
IRIS	7	9	8	2
STORM	6	5	13	2

Table 8: STORM under two training seeds: delay-20 causal alignment per game under the emulator-chosen interventions (mean over 7 evaluation seeds of per-seed medians; the same offset and warm-up per game for both seeds), with the seed-1 taxonomy class. Spearman correlation between the two seeds' game rankings 0.74 (bootstrap 95% interval $[0.44, 0.90]$); the seed-2 taxonomy, with its own threshold of half seed-2 Pong (0.25), agrees with seed 1 on 24 of 26 games (seed 2: 5 tracked, 5 aligned-then-forgotten, 14 never-aligned, 2 partial); seed-2 flagship Alien 0.003, Pong 0.509; cross-family transfer with seed 2 in place of seed 1: DIAMOND 0.50 ($p = 0.013$), IRIS 0.54 ($p = 0.005$).

game	seed 1 A_{20}	seed 2 A_{20}	class (seed 1)
Alien	+0.004	+0.003	never-aligned
Amidar	+0.000	-0.000	never-aligned
Assault	+0.007	+0.024	aligned-then-forgotten
Asterix	+0.001	-0.000	never-aligned
BankHeist	-0.000	-0.000	never-aligned
BattleZone	+0.407	+0.462	tracked
Boxing	+0.015	+0.015	never-aligned
Breakout	+0.449	+0.495	tracked
ChopperCommand	+0.019	+0.007	never-aligned
CrazyClimber	+0.021	+0.001	never-aligned

continued on the next page

game	seed 1 A_{20}	seed 2 A_{20}	class (seed 1)
DemonAttack	+0.117	+0.142	aligned-then-forgotten
Freeway	+0.330	+0.402	tracked
Frostbite	+0.006	+0.016	never-aligned
Gopher	+0.524	+0.139	tracked
Hero	+0.043	+0.014	never-aligned
Jamesbond	+0.039	+0.007	aligned-then-forgotten
Kangaroo	+0.378	+0.182	partial
Krull	-0.000	+0.003	never-aligned
KungFuMaster	+0.042	+0.009	aligned-then-forgotten
MsPacman	+0.000	+0.000	never-aligned
Pong	+0.376	+0.509	tracked
PrivateEye	+0.410	+0.291	tracked
Qbert	+0.380	+0.001	partial
RoadRunner	+0.003	+0.019	never-aligned
Seaquest	+0.062	+0.016	aligned-then-forgotten
UpNDown	+0.000	+0.004	never-aligned

The tracked threshold is 0.40 for DIAMOND, written with the pre-registered predictions, and half of Pong's delay-20 alignment for the two lower-level models, 0.11 for IRIS and 0.19 for STORM; halving it leaves DIAMOND's counts unchanged and moves IRIS's by two cells, while doubling it reclassifies six of DIAMOND's tracked cells and moves STORM's aligned-then-forgotten count from five to two (Appendix P). The causal horizon of a game (the first of the stored delays 1, 2, 3, 5, 8 and 20, before or after the peak, at which its alignment is below half its delay-1-to-8 peak, computed for games whose peak is at least 0.05 and censored at 24 when the alignment never falls that far by delay 20 or the peak is below 0.05; a game with an unmeasurable early delay in either architecture is dropped, leaving $n = 24$ of 26) correlates at 0.24 and 0.20 between DIAMOND and each transformer (permutation $p = 0.252$ and 0.352) and at 0.73 between IRIS and STORM ($p = 0.0006$); this is a different rule from the per-cell horizon of Table 24, which is left undefined for never-aligned and partial cells. The consensus ranking of the three architectures puts BankHeist, MsPacman, Amidar, Alien, UpNDown and RoadRunner hardest and Gopher, Breakout, PrivateEye, BattleZone, Pong and Seaquest easiest.

F LATENCY AND THE TIMING OF THE RESPONSE

Latency was the first game-level correlate of blindness we found; this appendix gives the correlate under both intervention designs, the pre-registered test of it, and the timing of the response within an episode.

F.1 LATENCY OF THE CONSEQUENCE: A GAME-LEVEL CORRELATE

We asked which property of a game predicts blindness. Under the fixed intervention, the latency of the true consequence, the first delay at which the true change reaches a quarter of its delay-20 size, correlated with alignment at $\rho = -0.48$ over the 24 DIAMOND and IRIS cells (12 games; $p = 0.018$ over cells, 0.076 over games as blocks) and at -0.50 among the 15 well-predicted cells over 9 games ($p = 0.059$ and 0.102), negative at every threshold of a sweep. It generalised only in part: under the emulator-chosen interventions the correlation over all 26 games is -0.42 for DIAMOND ($p = 0.035$) and -0.40 for IRIS ($p = 0.045$), and among well-predicted games it is -0.32 for DIAMOND ($n = 13$, $p = 0.290$) and -0.58 for IRIS ($n = 11$, $p = 0.075$), and within a single game episodes with immediate effects are tracked no better than latent ones. We report latency as a game-level correlate under some designs, not a law.

A pre-registered test. The fixed intervention used above leaves 14 of 26 games without a visible effect because it falls inside an idle window after reset. Letting the emulator choose, per game and model-blind, the action offset and a random warm-up that maximise the true effect makes all 26 games measurable. Before running any model on these interventions we computed each game's latency from the emulator alone and wrote down the law's predictions: 17 games predicted TRACKED

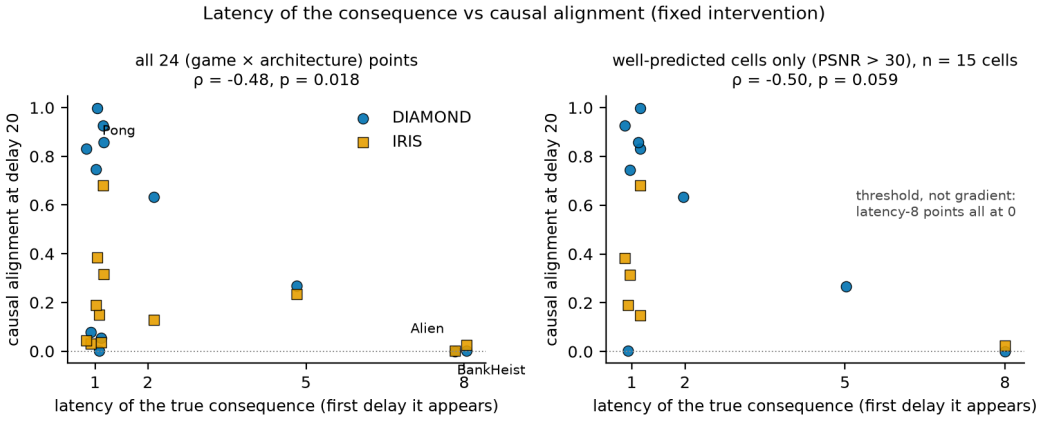

Figure 5: Latency of the true consequence vs causal alignment, fixed intervention (offset 3, no warm-up), delay 20. Left: all 24 architecture-game cells (12 games, DIAMOND and IRIS), $\rho = -0.48$, $p = 0.018$ permuting cells and 0.076 permuting games as blocks. Right: well-predicted cells (PSNR>30; 15 cells over 9 games), $\rho = -0.50$, $p = 0.059$ and 0.102. A threshold, not a gradient: every latency-8 point sits at zero. Under the emulator-chosen interventions the sign is the same in both architectures (-0.32 and -0.58 among well-predicted games) and significant in neither (§F.1).

(latency 1), 2 predicted BLIND (BankHeist, latency 12; Kangaroo, latency 5), 7 with no prediction (latency 2 to 3). Under its new intervention Alien's latency becomes 1, so the law predicts it will now be tracked, the sharpest test available, since Alien was causally blind under the original intervention in both reference architectures. The predictions were locked in a dated file before the runs; the confirmation thresholds (blind below 0.05; tracked at or above 0.40 for DIAMOND and 0.09 for IRIS, set at about half of each architecture's fixed-intervention Pong alignment as measured before the IRIS harness correction of Appendix C; PSNR above 30 dB) are those of the scoring script, written with the predictions, and are not part of the dated record. The outcome is mixed and we report it as such: for DIAMOND, 8 of 11 scored predictions were confirmed, including both BLIND calls (BankHeist 0.000, Kangaroo 0.029), but the sharpest prediction failed, Alien remains blind (0.003) under an immediate-effect intervention, and the TRACKED prediction also failed for Freeway (0.000, blind) and CrazyClimber (0.066, aligned-then-forgotten); for IRIS 9 of 11 were confirmed; Alien is not scored for IRIS because its PSNR (29.8 dB) falls below the 30 dB gate written with the predictions, although its alignment is 0.001. Together with the within-game null above, this supports reading causal blindness as a property of the (model, game, context distribution) cell rather than of the episode, and latency as a game-level correlate under some designs; a failed latency hypothesis does not show that latency plays no causal role or that episode-level variation is irrelevant, only that neither was detected here. What the test strengthens is the flagship: Alien is causally blind under both interventions, in both reference architectures (and under the chosen intervention in STORM), at prediction quality comparable to Pong.

F.2 THE RESPONSE IS IMMEDIATE AND WORLD-SIZED REGARDLESS OF THE WORLD

Why should latency matter? We measured the absolute model divergence $\|\Delta p\|$, defined even when the true divergence is zero. Three findings, all under the fixed intervention (offset 3, no warm-up; STORM has no fixed-intervention run). First, DIAMOND responds *immediately* to any action change in every episode (roughly world-sized where the world moves: ratio of median model change to median true change 0.75 for DIAMOND and 0.65 for IRIS in Pong at delay 1, per-pair median ratio 0.77 and 0.68); where the world did not move at delay 1 (the two true frames identical) DIAMOND moved anyway in every pair (Alien 277 of 277, Pong 22 of 22), while IRIS moved in 123 of 277 on Alien and 5 of 22 on Pong, so an early response is not itself a sign of blindness. Second, this comparison detects no episode-level association within Alien: episodes whose true effect was immediate are tracked at delay 20 no better than latent ones (median alignment 0.002 over 122 immediate pairs against 0.001 over 226 latent ones, $p = 0.86$). Third, in Pong the response stays aligned to delay 20; in Alien and BankHeist it does not, even after the real effect arrives. We therefore report latency as a game-level correlate of causal blindness and do not claim an episode-level mechanism. The Freeway control, again under the fixed intervention, adds a second failure mode:

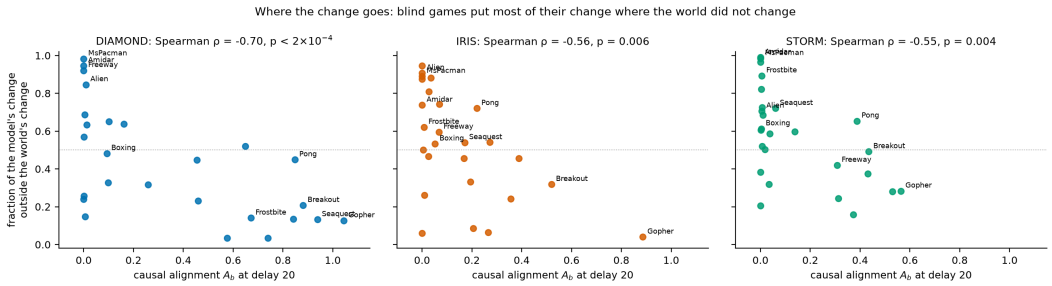

Figure 6: Where the model's change lands. For every game and all three architectures, the fraction of the model's change energy at delay 20 that falls outside the support of the world's change, against causal alignment. Blind games put most of their change where the world did not change; tracked games put most of it where the world did.

the world moves at delay 1 in 92% of episodes and DIAMOND tracks it at once (alignment 0.66 to 0.76 by the two estimators), then *forgets*, its response collapses from 1.80 to 0.27 while the true displacement persists at 1.8, and alignment reaches 0.002 by delay 20. Causal blindness thus arrives in two forms under this design: *never-aligned* (Alien, BankHeist) and *aligned-then-forgotten* (Freeway, Boxing); these are the behaviours under the fixed intervention, and the emulator-chosen taxonomy of Section 4.2 assigns its own classes (DIAMOND Freeway is never-aligned there, with a peak of 0.17, Table 24). Immediacy of the true effect buys initial tracking at most, not persistence, and does not guarantee even that. The per-episode files behind these numbers are released with the code.

Knowing when versus knowing what. A last question the same files answer is whether the model's change depends on whether the world's does. On pairs where the intervention was inert in the emulator (the two true frames identical at delay 20, $\Delta t = 0$ exactly; pairs with a small but nonzero true change lie below the guard and are excluded from the ratio, and active pairs are those above the guard with a nonzero true change; Appendix H) the two arms differ only in the action input the model received, so a model whose change were a context-independent echo of that input would change as much there as anywhere. The ratio of the model's median change on inert pairs to its median change on active pairs shows no detectable correlation with alignment in any architecture ($\rho = -0.29$, -0.11 and -0.15; $n = 19$, 18 and 18, minimum detectable $|\rho|$ 0.60 to 0.62; Appendix H). What it shows instead is that low-alignment cells split: in RoadRunner, Freeway and CrazyClimber (never-aligned, never-aligned and aligned-then-forgotten for DIAMOND), DIAMOND's change is small when the action is inert (ratios 0.01, 0.11, 0.11) and world-sized when it is not, in the wrong direction, while in Boxing (aligned-then-forgotten) it changes the same amount regardless (0.97), as does IRIS in Boxing and DemonAttack (aligned-then-forgotten, 0.89 and 0.83) and in ChopperCommand, a tracked cell (0.73). A small ratio is consistent with a change gated by the world's, but an action echo whose size depends on the context would produce the same pattern, so the ratio separates cells, not mechanisms, and does not by itself show that a model knows when an action matters.

G WHERE THE CHANGE GOES

Zero alignment at world-sized magnitude admits two readings: the model changes the future in a way that does not depend on which action it was given, or it changes the future in an action-specific way that lands in the wrong place. A third arm per episode tests the correspondence between action and consequence; it does not identify what the off-support change encodes. Besides the base action sequence a and the intervened one b, we replay a second intervention c, a different action on the same window whose true future differs from both, and imagine all three with the same seed (Appendix I). Two quantities decide. The cosine between the model's two responses, Δp_b and Δp_c, says whether the model produces the same change for different actions; its median over games is 0.46 to 0.47 for all three architectures, and the median cosine between the world's two responses is also 0.46. This coincidence does not show that the models' changes depend on the action to the world's degree: two differences that share a baseline have a cosine near one half even when the changes are unrelated, so the comparison needs a shared-baseline null (three rollouts under the same actions with different

seeds), which we report in Appendix I. What the three arms establish directly is that different actions produce different outputs. A permutation test of the correspondence itself is possible from the same three arms: per episode we compare the model's response to action b projected on the world's response to b (the correct pairing) with its response to the wrong action c projected on the same truth, and ask whether the correct pairing wins. At delay 20 it does in 25 of 26 games for DIAMOND, 24 for IRIS and 24 for STORM (binomial $p < 10^{-5}$ in each), with 17, 14 and 9 games individually significant after Holm correction. Across games, then, the models' changes correspond to the action taken more often than not; the next sentences say where. Whether the correspondence is present follows the taxonomy of Section 4.2, which is computed from the same alignment data and so corroborates rather than independently validates it: among the cells classified as tracked, 11 of 11 pass for DIAMOND, 6 of 7 for IRIS and 5 of 6 for STORM; among those classified never-aligned, 2 of 7, 2 of 8 and 1 of 13. Alien itself does not pass in any architecture (Holm $p = 0.09$, 1.00 and 0.97). The test measures correspondence with one true consequence; a blind model could still encode the action identity in a displaced sprite or texture and fail it, so the defensible statement is that correspondence with the true consequence was rarely detected in blind cells, not that those cells carry no action information. In the games with $A_b \geq 0.40$ (11, 2 and 4 games) the response to the wrong action projects on the true change at 0.48, 0.52 and 0.47 of the response to the right one for the three architectures, close to the cosine between the two true responses (0.49, 0.49 and 0.48); the projection depends on relative magnitudes as well as direction, so we read it as consistent with an action-specific response rather than as proof of one. What distinguishes a blind game is where its change lands (Figure 6). The fraction of the model's change energy outside the support of the world's change is 0.66 in DIAMOND's cells with $A_b < 0.05$ in the three-arm run (10 cells; range 0.15 to 0.98) against 0.14 in those with $A_b \geq T$ (11 cells; 0.03 to 0.52), 0.77 against 0.33 for IRIS and 0.65 against 0.33 for STORM; these descriptive sets differ from the taxonomy classes (over DIAMOND's seven never-aligned cells the median is 0.85). Across the 26 games it correlates with alignment at $\rho = -0.70$ ($p < 2 \times 10^{-4}$, 5,000 permutations) for DIAMOND, -0.56 ($p = 0.006$) for IRIS and -0.55 ($p = 0.004$) for STORM. Off-support energy and alignment are not independent: by Lemma 1, $|A| \leq R\sqrt{1-O}$ for the exact support, and for the thresholded support used here $|A| \leq R[\sqrt{(1-O)(1-q)} + \sqrt{Oq}]$ with q the fraction (not measured here) of the world's change energy below the threshold (Appendix B), so high off-support energy bounds alignment by construction and the correlation is partly definitional. The defensible statement is that most low-alignment responses place a large part of their change energy outside the pixels the intervention affects; whether that change encodes wrong consequences or sprite displacement and timing is not answered by the pixel analysis, and the object-level readout of Section 4.3 did not survive its human audit, so the question stays open; the off-support fraction does not discriminate the locator's labels either (Table 14). No PSNR association is detected: none between PSNR and off-support energy in any architecture ($\rho = -0.30$, -0.02 and -0.15, all $p > 0.1$), the size of the world's response is not detected as a predictor of which games are hard ($\rho = 0.10$ against the consensus difficulty), and the latency association depends on architecture and intervention design (Section F.1).

H DECOUPLING INDEX PER GAME

For each game under the emulator-chosen intervention, the median of the model's change $\|\Delta p_{20}\|$ on episode pairs where the intervention was inert in the emulator (the two true frames identical at that delay; pairs with a small nonzero true change are excluded from the ratio) divided by its median on pairs where it was not, with a bootstrap interval over pairs; games with fewer than ten pairs of either kind are omitted. A ratio near zero means the model changes only when the world does; a ratio near one means it changes the same amount either way.

Table 9: Decoupling index at delay 20 (emulator-chosen interventions): the median of the model's change on inert pairs (true change exactly zero) divided by its median on active pairs (true change above the guard and nonzero); pairs with a small nonzero true change below the guard belong to neither group, so inert, excluded and active pairs partition the 420 pairs of a cell. n: inert/active counts; games with fewer than ten pairs of either kind are omitted. A and R are the pooled medians over guarded pairs of Table 22, not the means of per-seed medians of Table 23.

model	game	A	R	ratio [95% CI]	$\|\Delta p\|$ inert	$\|\Delta p\|$ active	n inert/active
DIAMOND	BankHeist	+0.00	0.06	0.61 [0.27, 1.05]	1.03	1.68	56/364
DIAMOND	MsPacman	+0.00	0.03	0.97 [0.91, 1.05]	0.18	0.19	10/410
DIAMOND	Amidar	+0.00	0.03	0.96 [0.91, 1.02]	0.11	0.11	41/379
DIAMOND	RoadRunner	+0.00	0.85	0.01 [0.01, 0.01]	0.08	10.69	109/311
DIAMOND	Freeway	+0.00	1.01	0.11 [0.10, 0.12]	0.22	2.13	60/360
DIAMOND	Kangaroo	+0.00	1.03	0.96 [0.87, 1.06]	5.11	5.34	199/221
DIAMOND	UpNDown	+0.00	0.90	0.04 [0.03, 0.05]	0.50	12.26	192/228
DIAMOND	CrazyClimber	+0.05	0.81	0.11 [0.07, 0.21]	0.79	6.95	166/254
DIAMOND	Boxing	+0.07	0.96	0.97 [0.84, 1.08]	15.59	16.05	10/404
DIAMOND	Jamesbond	+0.08	0.75	0.34 [0.14, 0.49]	1.56	4.58	142/278
DIAMOND	Hero	+0.17	0.99	0.29 [0.06, 0.81]	2.14	7.46	25/393
DIAMOND	ChopperCommand	+0.41	0.94	0.09 [0.03, 0.13]	0.71	8.05	48/365
DIAMOND	KungFuMaster	+0.50	1.02	0.42 [0.15, 0.56]	3.91	9.25	74/346
DIAMOND	Frostbite	+0.60	1.00	0.03 [0.02, 0.03]	0.21	7.99	149/271
DIAMOND	DemonAttack	+0.63	1.53	0.73 [0.62, 0.82]	6.49	8.87	30/390
DIAMOND	Pong	+0.84	1.28	0.67 [0.62, 0.72]	3.17	4.74	140/271
DIAMOND	Seaquest	+0.93	1.12	0.02 [0.01, 0.02]	0.14	9.32	200/216
DIAMOND	Breakout	+0.95	1.06	0.02 [0.02, 0.02]	0.11	5.61	102/302
DIAMOND	Gopher	+1.04	1.17	0.34 [0.33, 0.39]	5.30	15.51	50/370
IRIS	Amidar	+0.00	0.39	0.23 [0.00, 20.79]	0.33	1.44	41/379
IRIS	RoadRunner	+0.00	0.76	0.00 [0.00, 0.00]	0.00	9.75	109/311
IRIS	BankHeist	+0.00	1.04	0.73 [0.05, 1.41]	27.11	37.24	56/364
IRIS	Jamesbond	+0.00	0.75	0.60 [0.47, 0.76]	1.90	3.19	142/278
IRIS	UpNDown	+0.01	0.77	0.09 [0.05, 0.14]	0.85	9.56	192/228
IRIS	Frostbite	+0.01	0.99	0.00 [0.00, 0.00]	0.00	7.72	149/271
IRIS	Kangaroo	+0.02	0.91	0.79 [0.70, 1.01]	3.92	4.94	199/221
IRIS	Boxing	+0.05	0.94	0.89 [0.71, 1.01]	13.98	15.68	10/404
IRIS	DemonAttack	+0.06	1.03	0.83 [0.67, 0.96]	5.05	6.12	30/390
IRIS	Hero	+0.14	0.97	0.22 [0.03, 0.85]	1.58	7.04	25/393
IRIS	Freeway	+0.16	1.02	0.63 [0.12, 0.93]	1.34	2.13	60/360
IRIS	CrazyClimber	+0.17	0.76	0.13 [0.09, 0.20]	0.79	6.27	166/254
IRIS	ChopperCommand	+0.18	0.90	0.73 [0.56, 0.98]	5.74	7.87	48/365
IRIS	Pong	+0.24	1.11	0.60 [0.56, 0.67]	2.52	4.17	140/271
IRIS	Seaquest	+0.26	1.01	0.09 [0.02, 0.15]	0.78	8.56	200/216
IRIS	KungFuMaster	+0.38	0.95	0.58 [0.47, 0.69]	4.98	8.61	74/346
IRIS	Breakout	+0.72	0.99	0.00 [0.00, 0.00]	0.00	5.24	102/302
IRIS	Gopher	+0.88	0.99	0.00 [0.00, 0.00]	0.00	13.82	50/370
STORM	BankHeist	-0.00	0.61	0.91 [0.14, 1.67]	15.63	17.26	83/334
STORM	Amidar	+0.00	0.00	1.00 [0.99, 1.00]	0.00	0.00	32/388
STORM	UpNDown	+0.00	0.29	0.00 [0.00, 0.07]	0.00	3.85	189/231
STORM	RoadRunner	+0.00	0.76	0.60 [0.47, 0.83]	6.01	10.07	104/316
STORM	Frostbite	+0.00	0.86	0.00 [0.00, 0.05]	0.00	6.31	147/273
STORM	CrazyClimber	+0.00	0.76	0.13 [0.07, 0.34]	0.80	6.17	154/266
STORM	Assault	+0.01	0.75	1.10 [0.87, 1.33]	7.16	6.52	12/408
STORM	ChopperCommand	+0.02	0.78	0.39 [0.25, 0.74]	2.78	7.13	45/373
STORM	KungFuMaster	+0.03	0.99	0.76 [0.55, 0.97]	7.08	9.37	69/349
STORM	Hero	+0.03	0.85	0.34 [0.13, 0.85]	2.62	7.69	28/391
STORM	Jamesbond	+0.04	0.86	0.76 [0.55, 1.01]	3.74	4.90	145/275

continued on the next page

model	game	A	R	ratio [95% CI]	$\|\Delta p\|$ inert	$\|\Delta p\|$ active	n inert/active
STORM	Seaquest	+0.05	0.66	0.14 [0.12, 0.16]	0.84	5.96	199/217
STORM	DemonAttack	+0.12	0.81	0.78 [0.60, 1.03]	4.76	6.08	24/395
STORM	Freeway	+0.33	0.91	0.00 [0.00, 0.00]	0.00	2.09	66/351
STORM	Kangaroo	+0.37	0.92	0.23 [0.16, 0.33]	1.18	5.21	224/196
STORM	Pong	+0.39	1.02	0.69 [0.57, 0.74]	2.88	4.17	162/253
STORM	Breakout	+0.44	0.96	0.35 [0.22, 0.48]	1.90	5.47	101/306
STORM	Gopher	+0.53	1.00	0.39 [0.25, 0.55]	5.83	15.05	42/378

I CONSEQUENCE SPECIFICITY PER CELL

Three arms per episode under the emulator-chosen intervention, 3 seeds and 40 episodes per game; the wrong action c is the first other offset whose true future differs from both a's and b's at delay 20; episodes without one are skipped; STORM's counts differ because its contexts and truths are matched within STORM only (Appendix C). Off-support uses the thresholded support of Appendix B (coordinates where the world's change exceeds one percent of its maximum).

Shared-baseline null. Under the same actions, three rollouts with different seeds give two differences that share the first rollout as baseline; their cosine is the value the geometry produces without any action change. The null gives a median cosine of 0.49 for DIAMOND, 0.50 for IRIS and 0.45 for STORM at delay 20 on the same ten games, against observed values of 0.47, 0.48 and 0.46: the observed cosine is what a shared baseline produces with no action change at all. The same runs show that changing the seed under a fixed action changes the future by 0.20 (DIAMOND), 0.93 (IRIS) and 0.98 (STORM) of the change produced by changing the action, which is why the measurement holds the seed fixed. For the two token models, whose sampled tokens diverge once their logits differ, the magnitude of a response after the window therefore includes sampling divergence set off by the action; alignment is affected only through the small projection of an unrelated change onto the true one.

Table 10: Consequence specificity per cell at delay 20 (three arms per episode: base a, the chosen action b, and a wrong action c whose true future differs from both). A_b: alignment of the response to b with the true change under b; $A_{c\to b}$: the response to the wrong action projected on the same truth; $\cos(\Delta p_b, \Delta p_c)$: similarity of the model's two responses; $\cos(\Delta t_b, \Delta t_c)$: similarity of the world's two responses; off-support: fraction of $\|\Delta p_b\|^2$ outside the support of Δt_b; PSNR of the intervened arm against its own truth. Values are per-cell medians over episodes (3 seeds $\times$ 40 episodes). Classes, applied in this order with the first matching rule winning (so a cell is never in two), with T the architecture's tracked threshold: *tracked-specific*: $A_b \geq T$ and $A_{c\to b} \leq A_b/2$; *tracked-unspecific*: $A_b \geq T$ otherwise; *blind-silent*: $A_b < 0.05$ and $R < 0.1$; *blind-agnostic*: $A_b < 0.05$ and $\cos(\Delta p_b, \Delta p_c) \geq 0.8$; *blind-misplaced*: $A_b < 0.05$ and off-support ≥ 0.9; *blind-other*: $A_b < 0.05$ otherwise; *partial*: $0.05 \leq A_b < T$. These descriptive classes are distinct from the Section 4.2 taxonomy.

model	game	n	A_b	$A_{c\to b}$	$\cos(\Delta p_b, \Delta p_c)$	$\cos(\Delta t_b, \Delta t_c)$	off-support	PSNR_b	class
DIAMOND	Amidar	106	-0.00	+0.00	+0.46	+0.48	0.95	30.8	blind-silent
DIAMOND	MsPacman	118	+0.00	+0.00	+0.49	+0.44	0.98	28.3	blind-silent
DIAMOND	Freeway	95	+0.00	+0.00	+0.39	+0.50	0.92	38.0	blind-misplaced
DIAMOND	BankHeist	99	+0.00	+0.00	+0.41	+0.45	0.24	11.1	blind-silent
DIAMOND	RoadRunner	93	+0.00	+0.00	+0.47	+0.46	0.57	25.3	blind-other
DIAMOND	UpNDown	72	+0.00	-0.00	+0.44	+0.41	0.26	23.0	blind-other
DIAMOND	Kangaroo	71	+0.00	+0.03	+0.38	+0.49	0.69	31.3	blind-other
DIAMOND	CrazyClimber	72	+0.01	+0.00	+0.41	+0.48	0.15	36.3	blind-other
DIAMOND	Alien	119	+0.01	+0.00	+0.46	+0.47	0.85	30.2	blind-other
DIAMOND	Qbert	119	+0.01	+0.00	+0.44	+0.41	0.63	27.0	blind-other
DIAMOND	Boxing	118	+0.09	+0.02	+0.47	+0.45	0.48	22.7	partial
DIAMOND	Jamesbond	79	+0.10	+0.00	+0.41	+0.40	0.33	31.6	partial

continued on the next page

model	game	n	A_b	$A_{c\to b}$	$\cos(\Delta p_b, \Delta p_c)$	$\cos(\Delta t_b, \Delta t_c)$	off-support	PSNR_b	class
DIAMOND	Assault	120	+0.10	+0.03	+0.47	+0.44	0.65	26.9	partial
DIAMOND	Krull	120	+0.16	+0.08	+0.48	+0.45	0.64	28.5	partial
DIAMOND	Hero	113	+0.26	+0.01	+0.43	+0.47	0.32	28.9	partial
DIAMOND	KungFuMaster	96	+0.45	+0.00	+0.43	+0.46	0.45	29.6	tracked-specific
DIAMOND	ChopperCommand	108	+0.46	+0.13	+0.50	+0.38	0.23	31.6	tracked-specific
DIAMOND	BattleZone	120	+0.58	+0.16	+0.38	+0.46	0.03	30.8	tracked-specific
DIAMOND	DemonAttack	110	+0.65	+0.40	+0.55	+0.52	0.52	30.3	tracked-unspecific
DIAMOND	Frostbite	77	+0.67	+0.19	+0.64	+0.56	0.14	34.3	tracked-specific
DIAMOND	PrivateEye	119	+0.74	+0.42	+0.52	+0.59	0.03	31.6	tracked-unspecific
DIAMOND	Asterix	120	+0.84	+0.41	+0.47	+0.49	0.14	32.8	tracked-specific
DIAMOND	Pong	76	+0.85	+0.43	+0.47	+0.50	0.45	35.0	tracked-unspecific
DIAMOND	Breakout	74	+0.88	+0.45	+0.51	+0.50	0.21	38.4	tracked-unspecific
DIAMOND	Seaquest	67	+0.94	+0.48	+0.47	+0.45	0.13	33.8	tracked-unspecific
DIAMOND	Gopher	104	+1.04	+0.49	+0.46	+0.47	0.13	29.8	tracked-specific
IRIS	Amidar	106	+0.00	+0.00	+0.67	+0.48	0.74	30.5	blind-other
IRIS	MsPacman	118	+0.00	+0.00	+0.89	+0.44	0.89	27.0	blind-silent
IRIS	BankHeist	99	+0.00	+0.00	+0.40	+0.45	0.06	15.0	blind-other
IRIS	Krull	120	+0.00	+0.00	+0.45	+0.45	0.95	30.4	blind-misplaced
IRIS	Alien	119	+0.00	-0.00	+0.47	+0.47	0.91	29.8	blind-misplaced
IRIS	Jamesbond	79	+0.00	+0.00	+0.38	+0.40	0.87	28.9	blind-other
IRIS	RoadRunner	93	+0.00	+0.00	+0.52	+0.46	0.50	25.7	blind-other
IRIS	Frostbite	77	+0.01	+0.00	+0.46	+0.56	0.62	27.7	blind-other
IRIS	UpNDown	72	+0.01	+0.00	+0.44	+0.41	0.26	22.8	blind-other
IRIS	Kangaroo	71	+0.03	+0.05	+0.46	+0.49	0.47	31.0	blind-other
IRIS	DemonAttack	110	+0.03	+0.02	+0.50	+0.52	0.81	30.3	blind-other
IRIS	Assault	120	+0.03	+0.00	+0.47	+0.44	0.88	24.6	blind-other
IRIS	Boxing	118	+0.05	+0.03	+0.46	+0.45	0.53	22.8	partial
IRIS	Freeway	95	+0.07	+0.00	+0.52	+0.50	0.60	33.4	partial
IRIS	Asterix	120	+0.07	+0.05	+0.50	+0.49	0.74	27.0	partial
IRIS	Hero	113	+0.17	+0.01	+0.35	+0.47	0.46	29.6	tracked-specific
IRIS	Seaquest	67	+0.17	+0.12	+0.48	+0.45	0.54	28.7	tracked-unspecific
IRIS	ChopperCommand	108	+0.19	+0.04	+0.41	+0.38	0.33	29.2	tracked-specific
IRIS	CrazyClimber	72	+0.21	+0.06	+0.37	+0.48	0.09	34.9	tracked-specific
IRIS	Pong	76	+0.22	+0.04	+0.49	+0.50	0.72	34.2	tracked-specific
IRIS	BattleZone	120	+0.26	+0.00	+0.43	+0.46	0.06	30.2	tracked-specific
IRIS	Qbert	119	+0.27	+0.00	+0.43	+0.41	0.54	27.3	tracked-specific
IRIS	PrivateEye	119	+0.36	+0.11	+0.60	+0.59	0.24	25.7	tracked-specific
IRIS	KungFuMaster	96	+0.39	+0.12	+0.46	+0.46	0.46	27.6	tracked-specific
IRIS	Breakout	74	+0.52	+0.29	+0.49	+0.50	0.32	35.8	tracked-unspecific
IRIS	Gopher	104	+0.89	+0.42	+0.48	+0.47	0.04	34.2	tracked-specific
STORM	Krull	120	-0.00	+0.00	+0.46	+0.49	0.97	30.5	blind-misplaced
STORM	BankHeist	92	-0.00	-0.00	+0.41	+0.42	0.20	14.7	blind-other
STORM	MsPacman	117	+0.00	-0.00	+0.50	+0.46	0.98	28.1	blind-silent
STORM	Amidar	115	+0.00	+0.00	+0.50	+0.45	0.99	30.5	blind-silent
STORM	UpNDown	62	+0.00	+0.00	+0.49	+0.39	0.38	23.4	blind-other
STORM	RoadRunner	90	+0.00	+0.01	+0.47	+0.57	0.61	24.6	blind-other
STORM	Boxing	119	+0.00	+0.02	+0.40	+0.42	0.61	22.9	blind-other
STORM	Asterix	118	+0.00	+0.00	+0.45	+0.46	0.82	27.6	blind-other
STORM	Frostbite	79	+0.00	-0.00	+0.59	+0.33	0.89	27.8	blind-other
STORM	Alien	119	+0.01	+0.01	+0.38	+0.46	0.71	30.0	blind-other
STORM	Assault	119	+0.01	+0.00	+0.49	+0.45	0.73	28.0	blind-other
STORM	CrazyClimber	65	+0.01	-0.00	+0.40	+0.25	0.52	33.4	blind-other
STORM	KungFuMaster	102	+0.01	+0.00	+0.45	+0.49	0.69	27.2	blind-other
STORM	Hero	113	+0.02	+0.00	+0.15	+0.44	0.50	29.1	blind-other
STORM	ChopperCommand	109	+0.03	+0.01	+0.46	+0.43	0.32	28.4	blind-other
STORM	Jamesbond	84	+0.04	+0.02	+0.44	+0.43	0.59	31.3	blind-other

continued on the next page

model	game	n	A_b	$A_{c\to b}$	$\cos(\Delta p_b, \Delta p_c)$	$\cos(\Delta t_b, \Delta t_c)$	off-support	PSNR_b	class
STORM	Seaquest	73	+0.06	+0.02	+0.45	+0.44	0.72	28.6	partial
STORM	DemonAttack	111	+0.14	+0.06	+0.52	+0.51	0.60	29.7	partial
STORM	Freeway	93	+0.31	+0.19	+0.47	+0.50	0.42	40.4	tracked-unspecific
STORM	Kangaroo	54	+0.31	+0.10	+0.49	+0.48	0.24	35.8	tracked-specific
STORM	BattleZone	120	+0.37	+0.11	+0.50	+0.47	0.16	31.0	tracked-specific
STORM	Pong	63	+0.39	+0.22	+0.51	+0.49	0.65	35.6	tracked-unspecific
STORM	Qbert	119	+0.43	+0.10	+0.35	+0.45	0.37	32.9	tracked-specific
STORM	Breakout	81	+0.43	+0.25	+0.52	+0.52	0.49	35.4	tracked-unspecific
STORM	PrivateEye	119	+0.53	+0.27	+0.63	+0.57	0.28	28.5	tracked-unspecific
STORM	Gopher	106	+0.56	+0.23	+0.44	+0.45	0.28	28.0	tracked-specific

Table 11: Action-consequence correspondence at delay 20: the model's response to the chosen action projected on the true change under that action (A_b), the response to a wrong action projected on the same truth ($A_{c\to b}$), the mean per-episode difference (the estimand of the test, which differs from the difference of the two medians in Table 10), and the one-sided sign-flip permutation p (20,000 flips; alternative: the correct pairing projects more) after Holm correction over the 26 games of each architecture. The null of the sign-flip test is that the per-episode difference is symmetric about zero, i.e. that the model's projection on the true change is exchangeable between the chosen and the wrong action; b and c are not randomised (c is the first other offset whose true future differs from both), so this is a test of that exchangeability, not a randomisation test, and exact emulator replay guarantees the two truths, not this assumption. The suite-level count in the text uses a one-sided binomial test of the number of games with a positive mean difference against one half.

model	game	n	A_b	$A_{c\to b}$	diff.	Holm p	pass
DIAMOND	Alien	119	+0.01	+0.00	+0.04	0.088	no
DIAMOND	Amidar	106	-0.00	+0.00	-0.01	0.909	no
DIAMOND	Assault	120	+0.10	+0.03	+0.09	0.006	yes
DIAMOND	Asterix	120	+0.84	+0.41	+0.37	0.001	yes
DIAMOND	BankHeist	99	+0.00	+0.00	+0.08	0.020	yes
DIAMOND	BattleZone	120	+0.58	+0.16	+0.32	0.001	yes
DIAMOND	Boxing	118	+0.09	+0.02	+0.04	0.151	no
DIAMOND	Breakout	74	+0.88	+0.45	+0.30	0.001	yes
DIAMOND	ChopperCommand	108	+0.46	+0.13	+0.22	0.001	yes
DIAMOND	CrazyClimber	72	+0.01	+0.00	+0.10	0.104	no
DIAMOND	DemonAttack	110	+0.65	+0.40	+0.25	0.001	yes
DIAMOND	Freeway	95	+0.00	+0.00	+0.01	0.909	no
DIAMOND	Frostbite	77	+0.67	+0.19	+0.24	0.001	yes
DIAMOND	Gopher	104	+1.04	+0.49	+0.55	0.001	yes
DIAMOND	Hero	113	+0.26	+0.01	+0.20	0.001	yes
DIAMOND	Jamesbond	79	+0.10	+0.00	+0.12	0.001	yes
DIAMOND	Kangaroo	71	+0.00	+0.03	+0.01	0.909	no
DIAMOND	Krull	120	+0.16	+0.08	+0.10	0.001	yes
DIAMOND	KungFuMaster	96	+0.45	+0.00	+0.26	0.001	yes
DIAMOND	MsPacman	118	+0.00	+0.00	+0.00	0.088	no
DIAMOND	Pong	76	+0.85	+0.43	+0.32	0.001	yes
DIAMOND	PrivateEye	119	+0.74	+0.42	+0.24	0.001	yes
DIAMOND	Qbert	119	+0.01	+0.00	+0.07	0.151	no
DIAMOND	RoadRunner	93	+0.00	+0.00	+0.09	0.005	yes
DIAMOND	Seaquest	67	+0.94	+0.48	+0.38	0.001	yes
DIAMOND	UpNDown	72	+0.00	-0.00	+0.07	0.147	no
IRIS	Alien	119	+0.00	-0.00	+0.01	1.000	no
IRIS	Amidar	106	+0.00	+0.00	-0.02	1.000	no
IRIS	Assault	120	+0.03	+0.00	+0.04	0.158	no

continued on the next page

model	game	n	A_b	$A_{c\to b}$	diff.	Holm p	pass
IRIS	Asterix	120	+0.07	+0.05	+0.02	0.113	no
IRIS	BankHeist	99	+0.00	+0.00	-0.06	1.000	no
IRIS	BattleZone	120	+0.26	+0.00	+0.22	0.001	yes
IRIS	Boxing	118	+0.05	+0.03	+0.04	0.384	no
IRIS	Breakout	74	+0.52	+0.29	+0.26	0.001	yes
IRIS	ChopperCommand	108	+0.19	+0.04	+0.10	0.009	yes
IRIS	CrazyClimber	72	+0.21	+0.06	+0.12	0.113	no
IRIS	DemonAttack	110	+0.03	+0.02	+0.06	0.007	yes
IRIS	Freeway	95	+0.07	+0.00	+0.16	0.001	yes
IRIS	Frostbite	77	+0.01	+0.00	+0.05	0.748	no
IRIS	Gopher	104	+0.89	+0.42	+0.41	0.001	yes
IRIS	Hero	113	+0.17	+0.01	+0.13	0.001	yes
IRIS	Jamesbond	79	+0.00	+0.00	+0.08	0.007	yes
IRIS	Kangaroo	71	+0.03	+0.05	+0.06	0.701	no
IRIS	Krull	120	+0.00	+0.00	+0.00	1.000	no
IRIS	KungFuMaster	96	+0.39	+0.12	+0.18	0.001	yes
IRIS	MsPacman	118	+0.00	+0.00	+0.01	0.772	no
IRIS	Pong	76	+0.22	+0.04	+0.14	0.009	yes
IRIS	PrivateEye	119	+0.36	+0.11	+0.11	0.004	yes
IRIS	Qbert	119	+0.27	+0.00	+0.20	0.001	yes
IRIS	RoadRunner	93	+0.00	+0.00	+0.06	0.021	yes
IRIS	Seaquest	67	+0.17	+0.12	+0.11	0.001	yes
IRIS	UpNDown	72	+0.01	+0.00	+0.05	0.214	no
STORM	Alien	119	+0.01	+0.01	+0.01	0.968	no
STORM	Amidar	115	+0.00	+0.00	+0.00	0.645	no
STORM	Assault	119	+0.01	+0.00	+0.02	0.968	no
STORM	Asterix	118	+0.00	+0.00	+0.01	0.968	no
STORM	BankHeist	92	-0.00	-0.00	-0.12	1.000	no
STORM	BattleZone	120	+0.37	+0.11	+0.21	0.001	yes
STORM	Boxing	119	+0.00	+0.02	+0.01	0.968	no
STORM	Breakout	81	+0.43	+0.25	+0.26	0.001	yes
STORM	ChopperCommand	109	+0.03	+0.01	+0.08	0.058	no
STORM	CrazyClimber	65	+0.01	-0.00	+0.10	0.058	no
STORM	DemonAttack	111	+0.14	+0.06	+0.06	0.007	yes
STORM	Freeway	93	+0.31	+0.19	+0.09	0.016	yes
STORM	Frostbite	79	+0.00	-0.00	+0.10	0.058	no
STORM	Gopher	106	+0.56	+0.23	+0.28	0.001	yes
STORM	Hero	113	+0.02	+0.00	+0.13	0.007	yes
STORM	Jamesbond	84	+0.04	+0.02	+0.05	0.268	no
STORM	Kangaroo	54	+0.31	+0.10	+0.19	0.040	yes
STORM	Krull	120	-0.00	+0.00	+0.00	0.968	no
STORM	KungFuMaster	102	+0.01	+0.00	+0.06	0.302	no
STORM	MsPacman	117	+0.00	-0.00	-0.00	1.000	no
STORM	Pong	63	+0.39	+0.22	+0.14	0.059	no
STORM	PrivateEye	119	+0.53	+0.27	+0.15	0.001	yes
STORM	Qbert	119	+0.43	+0.10	+0.27	0.001	yes
STORM	RoadRunner	90	+0.00	+0.01	+0.02	0.968	no
STORM	Seaquest	73	+0.06	+0.02	+0.03	0.409	no
STORM	UpNDown	62	+0.00	+0.00	+0.02	0.968	no

J THE THIRD ARCHITECTURE ON ALL GAMES

Table 12: The third architecture, STORM, on all 26 games under the emulator-chosen interventions at delay 20 (7 seeds × 60 episodes): PSNR, consistency, alignment A, magnitude R, error ε, taxonomy class and peak alignment over delays 1 to 8. Classes use the rule of Table 7 with T = half of STORM's own Pong delay-20 alignment (0.19; DIAMOND 0.40 written with the pre-registered predictions, IRIS 0.11). This table is STORM's first training seed; the second seed is summarised in Appendix E. R and ε are pooled medians over guarded pairs (the estimator of Table 22); PSNR, consistency and A are means of per-seed medians.

game	PSNR	consist.	A	R	ε	class	peak
Gopher	26.6	0.748	+0.524	1.00	0.94	tracked	0.70
Breakout	36.6	0.846	+0.449	0.96	0.97	tracked	0.75
PrivateEye	28.4	0.796	+0.410	0.93	0.97	tracked	0.57
BattleZone	31.1	0.859	+0.407	0.86	0.87	tracked	0.64
Qbert	32.5	0.829	+0.380	0.89	0.92	partial	–
Kangaroo	36.1	0.924	+0.378	0.92	0.97	partial	–
Pong	35.7	0.879	+0.376	1.02	1.27	tracked	0.54
Freeway	41.1	0.818	+0.330	0.91	1.00	tracked	0.60
DemonAttack	29.5	0.567	+0.117	0.81	1.30	aligned-then-forgotten	0.66
Seaquest	31.7	0.851	+0.062	0.66	1.14	aligned-then-forgotten	0.43
Hero	29.5	0.684	+0.043	0.85	1.09	never-aligned	0.17
KungFuMaster	27.0	0.778	+0.042	0.99	1.58	aligned-then-forgotten	0.32
Jamesbond	31.8	0.674	+0.039	0.86	1.37	aligned-then-forgotten	0.26
CrazyClimber	32.8	0.738	+0.021	0.76	1.01	never-aligned	0.13
ChopperCommand	28.8	0.882	+0.019	0.78	1.29	never-aligned	0.12
Boxing	23.2	0.625	+0.015	0.80	1.56	never-aligned	0.12
Assault	27.8	0.338	+0.007	0.75	1.54	aligned-then-forgotten	0.22
Frostbite	29.4	0.887	+0.006	0.86	1.68	never-aligned	0.04
Alien	30.0	0.852	+0.004	0.67	1.40	never-aligned	0.11
RoadRunner	25.1	0.830	+0.003	0.76	1.43	never-aligned	0.01
Asterix	27.7	0.443	+0.001	0.28	1.07	never-aligned	0.18
UpNDown	25.6	0.699	+0.000	0.29	1.02	never-aligned	0.01
Amidar	30.5	0.959	+0.000	0.00	1.00	never-aligned	0.00
MsPacman	28.1	0.870	+0.000	0.00	1.00	never-aligned	0.00
BankHeist	14.5	0.449	-0.000	0.61	1.18	never-aligned	0.00
Krull	30.2	0.789	-0.000	0.57	1.34	never-aligned	0.00

K OBJECT-LEVEL VALIDATION

Displacement endpoint. For every episode in which the template locator finds the actuated object in both generated arms, the locator's estimate of the model's displacement of the object between the arms is compared with the true displacement; the endpoint is the fraction of such episodes in which the estimated displacement error, in pixels, is smaller than the error of a zero displacement (which equals the true displacement's norm), together with the two median errors. A cell is assessable on this endpoint when at least three episodes have the object in both arms. Hits are stored at tolerances of 2, 4 and 8 pixels; the paper reports 8. An *eligible* episode is one whose true displacement exceeds 8 pixels, so that a zero displacement would fail an 8-pixel tolerance; accuracy conditional on eligibility is also tabulated. A stricter endpoint, the fraction of all episodes with a hit within 8 pixels that a zero displacement would not have passed, counts ineligible episodes as failures and is therefore bounded above by one minus the zero-pass rate even for a perfect predictor; Table 15 lists it for comparison. The game-level permutation test in Table 14 permutes the object labels across games with every architecture's cell of a game moving together (5,000 permutations) and reports the two-sided p, the fraction of permuted labelings whose AUC lies at least as far from one half as the observed AUC; the bootstrap resamples games the same way; the tabulated p-values are uncorrected, and the Holm correction within each seven-row family is reported in Section 4.3. A cell's label on the displacement endpoint is *no improvement on zero displacement* when the model's error is strictly smaller than the zero displacement's in fewer than half of the episodes with the object in

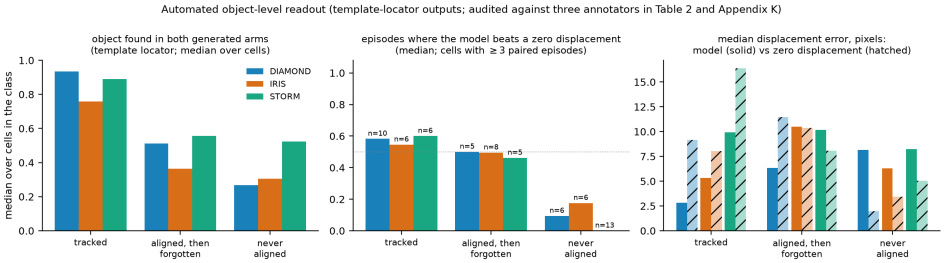

Figure 7: Automated object-level readout of the taxonomy (template-locator outputs, audited against three annotators in Table 2 and Appendix K). The actuated object is the set of pixels where the two true arms first differ; a normalised-cross-correlation template locator searches for it in the generated frames (up to 30 episodes per cell; Appendix K). Left: fraction of evaluated episodes in which the locator finds the object in both generated arms. Middle: among episodes with the object found in both arms, the fraction in which the estimated displacement error is strictly smaller than a zero displacement's, in cells with at least 3 such episodes (the number of such cells is printed above each bar). Right: median estimated displacement error in pixels (solid) and the zero displacement's (hatched) on the same episodes. Bars are medians over the cells of a class per architecture.

both arms; ties count as no improvement, so the label does not assert that the estimated displacement is worse than zero, only that it fails to beat it. On Alien the template finds the player in 0.53, 0.43, 0.39 of episodes and the improvement fraction is 0.19, 0.23, 0.00 for the three architectures; on Pong the locator finds the paddle for all three models and DIAMOND's estimated displacement improves on the baseline in 0.84 of detected episodes. Human check: see the paragraph and table below. The zero-displacement baseline is stronger where the world moves the object less (median zero error 3.9 px in never-aligned cells against 12.2 in tracked), so the label is easier to earn in blind cells; restricted to episodes whose true displacement exceeds 8 px, the within-tolerance accuracy is 0.37 in tracked cells, 0.33 in aligned-then-forgotten and 0.00 in never-aligned ones (Table 15).

The two locators. Both arms are rolled in the real emulator at the native 210×160 resolution. The first delay $d_0 \in \{1, \ldots, 20\}$ at which any pixel of the two true frames differs by more than 30 (L1 over RGB) defines the actuated object as the set of differing pixels at d_0 (no dilation); an episode without such a delay is skipped. A *lane* is the union over delays 1 to 20 of those difference masks, dilated once by a 9×9 kernel, and every search is restricted to it. A colour signature is the most frequent RGB value among the object's pixels that differ from the frame's background colour by more than 30 (episodes with fewer than four such pixels are skipped). The *colour locator* marks pixels within L1 distance 40 of the signature inside the lane and returns their centroid; the object counts as present in a generated frame when at least a quarter as many pixels match as in the corresponding true frame. The *template locator* cuts, from the true frame of the same arm at the same delay, the bounding box (padded by 2 pixels) of the largest connected component of the signature-colour mask, and matches this RGB patch by zero-mean normalised cross-correlation (OpenCV TM_CCOEFF_NORMED) inside the lane of the generated frame after nearest-neighbour upsampling to the native 210×160; the object counts as found when the peak correlation is at least 0.60 in both generated arms (fixed before the runs), the estimated position being the centre of the peak. The matcher is invariant to brightness and contrast, not to colour. The true displacement is the difference of the colour-locator centroids in the two true frames; the estimated displacement is the difference of the template centres in the two generated frames. The self-check matches the arm-b template back into the true frame it was cut from and passes when the peak is at least 0.9; it fails in 107 of 1,955 delay-20 episodes, of which 80 are BattleZone episodes whose template spans the full frame width and is rejected as degenerate, so BattleZone's near-zero detection rates (0.03, 0.03 and 0.23) are a property of the locator, not of the models. Episodes are indexed 0 to 29 per cell; every random quantity (emulator reset, warm-up, base actions, model seed) is a function of the episode index, so the first six episodes of the 30-episode run are the six of the earlier run from which the human panels were drawn, verified identical record by record. Delays 8 and 20 are evaluated; the paper reports 20.

Human check. Annotated independently by three authors after the analysis was fixed, each seeing no metric value and no other annotator's sheet, with the rules fixed before the first labels were

read: a panel enters the presence comparison when both arms are labelled yes or no, human present meaning yes in both; it enters the displacement comparison when the labels yield a verdict, direction no meaning no improvement and direction yes with magnitude about right or too little meaning improvement, while direction unsure, or direction yes with magnitude too much or unsure, is undecided; one panel, the first in hashed order, was used to explain the questions before labelling and is excluded from every sheet (59 panels remain). The human rule is a qualitative proxy for the pixel endpoint, not the same quantity: a prediction in the right direction that overshoots can still have a smaller error than a zero displacement (15 pixels predicted for a true 10 has error 5), and the magnitude labels do not say how far an overshoot goes, so overshooting panels are undecided rather than failures (3, 3 and 2 panels for the annotators, the locator finding the object in 1, 2 and 0 of them). Consensus labels, whose rule was fixed before the second sheet was integrated, exist where at least two annotators decided and all who decided agree; any disagreement is undecided (9 panels on presence, 16 on improvement). Agreement between annotators over the panels both decided, for the pairs 1 and 2, 1 and 3, 2 and 3: presence 0.96, 0.90 and 0.86 (Cohen's kappa 0.81, 0.65 and 0.56; 49, 49 and 58 panels), improvement 0.82, 0.87 and 0.73 (kappa 0.63, 0.75 and 0.46; 38, 39 and 51 panels); the annotators were unsure on 10, 1 and 0 panels on presence and undecided on 18, 5 and 5 on displacement. Presence: the template locator's recall against the annotators is 0.43, 0.39 and 0.41 at precision 0.86, 0.80 and 0.69 (consensus recall 0.40, precision 0.77; colour locator recall 0.31, 0.31 and 0.36), and its presence label has kappa 0.00, -0.12 and -0.09 with theirs; the annotators see the object in 15 of 15, 16 of 17 and 14 of 17 decided tracked panels, 11 of 13, 11 of 13 and 10 of 13 aligned-then-forgotten, 7 of 10, 14 of 17 and 12 of 18 never-aligned and 9 of 11, 10 of 11 and 8 of 11 partial ones, so these locator rates do not establish a model-level presence contrast. Displacement: on the panels each annotator and the locator could both judge (20, 22 and 23), agreement is 0.70, 0.59 and 0.57 (kappa 0.44, 0.20 and 0.14), the locator's improvement calls agreeing on 8 of 8, 6 of 9 and 5 of 8 and its no-improvement calls on 6 of 12, 7 of 13 and 8 of 15; on the 17 consensus panels with a locator verdict the counts are tp 4, fn 5, fp 1, tn 7 (agreement 0.65), and the locator has no verdict on 21 of the 38 consensus panels, 11 of which the annotators judge as improving. On the 4 consensus never-aligned panels with a locator verdict the annotators see improvement in 0 and the locator in 1, and the annotators see right-direction movement in 1, 4 and 3 never-aligned panels the locator did not detect. Table 13 gives the breakdown by class. Reading: the annotators agree with each other moderately to substantially; the locator under-detects in every class and its displacement verdict is only weakly related to human judgment; the object-level labels, and the metric comparison built on them (Table 14), are properties of the automated readout that the human check does not validate. Sampling and blinding: the panels were drawn by a seeded stratified rule (up to four per architecture and taxonomy class, plus episode 0 of Alien, Pong, Breakout, Asterix and BankHeist for every architecture; the 63 panels so drawn were capped at the first 60 in hashed-name order, which dropped three STORM panels), cover 21 games and 43 of the 78 cells (20 DIAMOND, 21 IRIS, 18 STORM panels), and were presented in the order of their hashed file names; each panel shows the two true frames, the two generated frames and the true frame at the first difference with the actuated object boxed, with no architecture, game name, class or score; the game is recognisable from its frames, the architecture is not; all three annotators are authors and knew the hypothesis. Contingency counts by pair (presence: both present, first absent and second present, first present and second absent, both absent; improvement: both yes, first yes and second no, first no and second yes, both no): annotators 1 and 2: 42, 2, 0, 5 and 17, 4, 3, 14; annotators 1 and 3: 38, 1, 4, 6 and 17, 4, 1, 17; annotators 2 and 3: 43, 0, 8, 7 and 20, 10, 4, 17. Game-clustered bootstrap intervals (the resampling unit is the game; 2,000 replicates, replicates with an undefined kappa dropped): kappa between annotators [0.46, 1.00], [0.28, 0.90] and [0.32, 0.77] on presence and [0.32, 0.87], [0.51, 0.93] and [0.24, 0.71] on improvement; kappa of the locator's displacement verdict with the annotators [0.17, 0.71], [-0.09, 0.53] and [-0.31, 0.47] and with the consensus [-0.20, 0.62]; of its presence label with the annotators [-0.23, 0.25], [-0.36, 0.05] and [-0.36, 0.13]. With 20, 22 and 23 comparable panels these intervals are wide, and the disagreements listed above carry more information than a verbal grade of the kappa values.

Table 13: Blinded human check of the object readout by 3 annotators on the same 59 panels (one example panel excluded), by taxonomy class of the panel's cell; each annotator evaluated on their own labels, then on the consensus labels (a label exists when at least two annotators decided and all who decided agree; any disagreement counts as undecided). Cohen's kappa between annotators on the derived labels over the panels both decided: annotators 1-2: presence 0.81 over 49 panels, improvement 0.63 over 38; annotators 1-3: presence 0.65 over 49 panels, improvement 0.75 over 39; annotators 2-3: presence 0.56 over 58 panels, improvement 0.46 over 51. *present / absent / unsure*: presence label (present = seen in both generated arms); *recall*: fraction of the panels labelled present in which the template locator found the object; *compared*: panels on which both the annotator (direction no; or direction yes with magnitude about right or too little; other combinations undecided) and the locator (object found with a displacement) could judge whether the predicted displacement improves on zero, the human rule being a qualitative proxy for the pixel endpoint; *human* and *locator*: how many of those each judged as improving; *agree*: panels on which the two verdicts coincide; *missing*: panels the annotator could judge but the locator could not (in brackets, how many of those the annotator judged as improving).

labels	class	panels	present / absent / unsure	recall	compared	human	locator	agree	missing (improving)
annotator 1	tracked	17	15 / 0 / 2	0.80	10	10	6	6	3 (2)
	aligned-then-forgotten	13	11 / 2 / 0	0.36	6	3	2	5	3 (3)
	never-aligned	18	7 / 3 / 8	0.14	3	0	0	3	7 (1)
	partial	11	9 / 2 / 0	0.11	1	1	0	0	8 (3)
annotator 2	tracked	17	16 / 1 / 0	0.75	11	8	5	6	4 (3)
	aligned-then-forgotten	13	11 / 2 / 0	0.36	6	2	2	4	7 (6)
	never-aligned	18	14 / 3 / 1	0.21	4	1	2	3	12 (4)
	partial	11	10 / 1 / 0	0.10	1	1	0	0	9 (7)
annotator 3	tracked	17	14 / 3 / 0	0.79	11	9	5	7	4 (2)
	aligned-then-forgotten	13	10 / 3 / 0	0.30	5	1	1	3	6 (6)
	never-aligned	18	12 / 6 / 0	0.25	6	1	2	3	12 (3)
	partial	11	8 / 3 / 0	0.12	1	1	0	0	9 (3)
consensus	tracked	17	14 / 1 / 2	0.79	8	7	4	5	3 (2)
	aligned-then-forgotten	13	10 / 2 / 1	0.30	4	1	0	3	5 (5)
	never-aligned	18	10 / 3 / 5	0.20	4	0	1	3	9 (2)
	partial	11	8 / 1 / 2	0.12	1	1	0	0	4 (2)

Table 14: Which score predicts the locator's labels? AUC, pooled over three architectures, for separating cells by two automated labels derived from the template locator (Appendix K; these labels are locator outputs, which the human audit of Table 2 does not validate as facts about the models): column 2, the object found in both generated frames at delay 20 in fewer than half of the episodes (29 of 60 cells where every score is defined); column 3, the estimated displacement error strictly smaller than a zero displacement's in fewer than half of the episodes with the object found in both arms, ties counting as no improvement (35 of 55 cells with at least three such episodes); every score is oriented so that an AUC above one half points the right way. Intervals: game-clustered bootstrap (games resampled, every architecture's cell of a game together); p: two-sided game-level permutation test (the fraction of 5000 game-permuted labelings whose AUC lies at least as far from one half as the observed AUC), uncorrected; the two labels are treated as separate seven-row families for Holm correction. Column 4 uses the stricter strict-hit label (a hit within 8 pixels that a zero displacement would not have passed, below one half of all episodes; 59 of 60 cells), which counts ineligible episodes as failures and is listed for comparison; the last two columns give the paired game-clustered difference between the off-support fraction's AUC and the row's for the two automated labels.

score	not found [CI], p	no improvement on zero [CI], p	strict hit [CI], p	O minus row, not found	O minus row, no improvement
consistency (one rollout)	0.50 [0.32, 0.68], 0.979	0.43 [0.23, 0.61], 0.448	0.71 [0.55, 0.82], 0.624	-0.05 [-0.31, +0.20]	+0.21 [-0.07, +0.50]
PSNR (reference, one arm)	0.62 [0.41, 0.81], 0.239	0.55 [0.30, 0.77], 0.677	0.85 [0.71, 0.95], 0.747	-0.17 [-0.44, +0.10]	+0.08 [-0.22, +0.38]
dPSNR (two rollouts, factual reference)	0.58 [0.39, 0.76], 0.394	0.84 [0.65, 0.95], < 0.001	0.85 [0.71, 0.97], 0.253	-0.13 [-0.33, +0.05]	-0.20 [-0.39, +0.01]
alignment A (this work)	0.56 [0.34, 0.74], 0.548	0.88 [0.76, 0.96], < 0.001	0.97 [0.92, 1.00], 0.240	-0.11 [-0.30, +0.08]	-0.24 [-0.42, -0.08]
magnitude R (this work)	0.49 [0.30, 0.67], 0.950	0.78 [0.64, 0.90], 0.007	0.97 [0.92, 1.00], 0.244	-0.04 [-0.27, +0.20]	-0.14 [-0.39, +0.11]
error eps (this work)	0.60 [0.43, 0.75], 0.273	0.62 [0.39, 0.80], 0.228	0.95 [0.87, 1.00], 0.367	-0.15 [-0.35, +0.06]	+0.01 [-0.18, +0.24]
off-support fraction O (this work)	0.45 [0.26, 0.63], 0.632	0.64 [0.43, 0.82], 0.238	0.66 [0.52, 0.81], 0.752	–	–

Table 15: Object-level readout per cell at delay 20 (locator outputs; the human audit is in Table 13), with the displacement endpoint of Appendix K. The actuated object is the set of pixels where the two true arms first differ. *colour*, *template*: fraction of episodes in which the dominant-colour locator and the normalised-cross-correlation template locator (Appendix K) find the object in both generated arms; n: episodes evaluated; *disp.*: median true displacement of the object between the arms in pixels; *z.pass*: fraction of episodes in which a zero displacement is already within 8 pixels of the true one (on such episodes an 8-pixel tolerance cannot distinguish a correct displacement from a zero one, so they are ineligible for the tolerance-based accuracy); *elig.*: eligible episodes; *acc.*: fraction of eligible episodes with the object found in both arms in which the estimated displacement is within 8 pixels of the true one (– when no eligible episode has the object); *pairs*: episodes with the object found in both arms; *m.err*, *z.err*: median displacement error in pixels of the estimated displacement and of a zero displacement over those episodes; *better*: fraction of those episodes in which the estimated error is smaller than the zero displacement's; *strict*: the stricter hit score (hits a zero displacement would not pass, over all episodes), listed for comparison; *swap*: $2\langle \Delta p_{bc}, \Delta t_{bc} \rangle / \|\Delta t_{bc}\|^2$, the only column taken from the earlier 40-episode readout (every other column, including *colour*, is from the 30-episode audit run). Cells with fewer than three paired episodes are not assessable on the displacement endpoint; – in m.err, z.err and better marks a cell without a paired episode.

model	game	class	colour	template	n	disp.	z.pass	elig.	acc.	pairs	m.err	z.err	better	strict	swap
DIAMOND	Assault	aligned-then-forgotten	0.38	0.83	24	15.1	0.33	16	0.50	20	6.3	11.5	0.60	0.25	+0.31

continued on the next page

model	game	class	colour	template	n	disp.	z.pass	elig.	acc.	pairs	m.err	z.err	better	strict	swap
DIAMOND	Boxing	aligned-then-forgotten	1.00	0.33	30	25.0	0.17	25	0.14	10	17.3	21.7	0.50	0.03	+0.16
DIAMOND	CrazyClimber	aligned-then-forgotten	0.59	0.69	29	0.2	0.97	1	–	20	1.8	0.1	0.00	0.00	+0.00
DIAMOND	Hero	aligned-then-forgotten	0.14	0.31	29	8.6	0.45	16	1.00	9	5.4	8.0	0.44	0.10	+0.62
DIAMOND	Jamesbond	aligned-then-forgotten	0.12	0.25	8	4.4	0.88	1	0.00	2	33.4	14.2	0.00	0.00	+0.00
DIAMOND	Krull	aligned-then-forgotten	0.04	0.88	24	15.0	0.08	22	0.42	21	8.9	15.5	0.62	0.33	+0.39
DIAMOND	Alien	never-aligned	0.07	0.53	30	0.9	0.77	7	0.00	16	11.1	2.2	0.19	0.00	+0.01
DIAMOND	Amidar	never-aligned	0.00	0.88	25	11.0	0.40	15	0.00	22	12.5	12.0	0.00	0.00	+0.00
DIAMOND	BankHeist	never-aligned	0.92	0.25	24	3.4	0.88	3	–	6	5.1	5.1	0.00	0.00	-0.00
DIAMOND	Freeway	never-aligned	0.30	0.17	30	6.0	0.67	10	0.00	5	3.0	0.0	0.40	0.00	+0.00
DIAMOND	MsPacman	never-aligned	0.00	0.03	29	24.0	0.14	25	0.00	1	26.8	26.8	0.00	0.00	+0.00
DIAMOND	RoadRunner	never-aligned	0.47	0.27	15	3.7	0.60	6	–	4	12.7	1.7	0.00	0.00	+0.25
DIAMOND	UpNDown	never-aligned	0.90	0.35	20	3.4	0.70	6	0.00	7	2.4	0.8	0.29	0.00	+0.02
DIAMOND	Kangaroo	partial	0.18	0.25	28	8.1	0.50	14	1.00	7	3.5	6.0	0.43	0.11	+0.00
DIAMOND	Qbert	partial	0.30	0.10	30	36.9	0.00	30	0.00	3	33.2	27.2	0.33	0.00	-0.00
DIAMOND	Asterix	tracked	0.00	0.60	30	8.2	0.47	16	0.00	18	11.5	7.8	0.00	0.00	+1.55
DIAMOND	BattleZone	tracked	1.00	0.03	30	3.2	0.83	5	0.00	1	76.2	8.9	0.00	0.00	+1.28
DIAMOND	Breakout	tracked	0.00	0.96	28	21.0	0.29	20	0.37	27	6.3	18.9	0.59	0.25	+1.96
DIAMOND	ChopperCommand	tracked	0.96	0.96	25	6.5	0.60	10	0.56	24	8.7	6.2	0.38	0.20	+0.61
DIAMOND	DemonAttack	tracked	0.73	0.27	26	8.0	0.69	8	1.00	7	1.0	4.0	0.57	0.08	+1.55
DIAMOND	Frostbite	tracked	0.80	0.80	20	3.1	0.75	5	0.00	16	2.5	1.1	0.19	0.00	+1.15
DIAMOND	Gopher	tracked	0.93	0.93	30	12.0	0.43	17	1.00	28	0.0	10.5	0.89	0.50	+2.06
DIAMOND	KungFuMaster	tracked	0.47	0.47	30	2.5	0.97	1	–	14	2.2	3.7	0.43	0.00	+0.99
DIAMOND	Pong	tracked	1.00	0.96	26	26.0	0.19	21	1.00	25	3.0	26.0	0.84	0.77	+1.79
DIAMOND	PrivateEye	tracked	0.77	0.93	30	24.8	0.23	23	0.38	28	11.7	31.8	0.68	0.27	+1.34
DIAMOND	Seaquest	tracked	0.82	1.00	11	13.8	0.09	10	0.90	11	2.6	13.8	0.91	0.82	+1.79
IRIS	Assault	aligned-then-forgotten	0.08	0.92	24	15.1	0.33	16	0.36	22	10.7	13.2	0.55	0.21	+0.07
IRIS	Asterix	aligned-then-forgotten	0.00	0.57	30	8.2	0.47	16	0.00	17	10.3	8.0	0.00	0.00	+0.14
IRIS	Boxing	aligned-then-forgotten	1.00	0.30	30	25.0	0.17	25	0.00	9	23.2	20.9	0.56	0.00	+0.12
IRIS	DemonAttack	aligned-then-forgotten	0.00	0.08	26	8.0	0.69	8	–	2	5.5	8.0	0.50	0.00	+0.18
IRIS	Freeway	aligned-then-forgotten	0.07	0.30	30	6.0	0.67	10	0.67	9	3.0	4.0	0.44	0.07	+0.56
IRIS	Frostbite	aligned-then-forgotten	0.65	0.65	20	3.1	0.75	5	0.00	13	10.8	1.0	0.15	0.00	+0.33
IRIS	Hero	aligned-then-forgotten	0.03	0.10	29	8.6	0.45	16	1.00	3	8.0	8.0	0.33	0.03	+0.59
IRIS	Pong	aligned-then-forgotten	0.77	0.88	26	26.0	0.19	21	0.33	23	19.0	26.0	0.57	0.23	+0.61
IRIS	Seaquest	aligned-then-forgotten	0.00	0.36	11	13.8	0.09	10	0.75	4	6.9	12.7	0.75	0.27	+0.30
IRIS	Alien	never-aligned	0.00	0.43	30	0.9	0.77	7	0.00	13	2.2	0.9	0.23	0.00	+0.00
IRIS	Amidar	never-aligned	0.00	0.92	25	11.0	0.40	15	0.20	23	11.0	12.0	0.22	0.12	+0.00
IRIS	BankHeist	never-aligned	0.71	0.21	24	3.4	0.88	3	–	5	6.6	5.8	0.00	0.00	+0.00
IRIS	Jamesbond	never-aligned	0.12	0.12	8	4.4	0.88	1	–	1	13.1	2.4	0.00	0.00	-0.00
IRIS	Krull	never-aligned	0.04	0.62	24	15.0	0.08	22	0.07	15	23.3	12.0	0.13	0.04	-0.00
IRIS	MsPacman	never-aligned	0.14	0.00	29	24.0	0.14	25	–	0	–	–	–	0.00	+0.00
IRIS	RoadRunner	never-aligned	0.33	0.20	15	3.7	0.60	6	–	3	6.0	0.9	0.00	0.00	+0.05
IRIS	UpNDown	never-aligned	0.90	0.40	20	3.4	0.70	6	1.00	8	1.5	1.1	0.50	0.10	-0.01
IRIS	Kangaroo	partial	0.00	0.11	28	8.1	0.50	14	–	3	1.2	2.1	0.33	0.00	+0.16
IRIS	Qbert	partial	0.17	0.03	30	36.9	0.00	30	1.00	1	4.8	42.3	1.00	0.03	+0.76
IRIS	BattleZone	tracked	1.00	0.03	30	3.2	0.83	5	0.00	1	79.2	8.9	0.00	0.00	+0.67
IRIS	Breakout	tracked	0.00	0.96	28	21.0	0.29	20	0.26	27	8.6	23.0	0.59	0.18	+0.93
IRIS	ChopperCommand	tracked	0.80	0.92	25	6.5	0.60	10	0.30	23	21.1	7.1	0.26	0.12	+0.54
IRIS	CrazyClimber	tracked	0.55	0.76	29	0.2	0.97	1	0.00	22	0.1	0.1	0.05	0.00	+0.06
IRIS	Gopher	tracked	0.40	0.90	30	12.0	0.43	17	1.00	27	0.0	9.0	0.85	0.47	+1.78
IRIS	KungFuMaster	tracked	0.07	0.33	30	2.5	0.97	1	1.00	10	2.0	3.5	0.50	0.03	+0.81
IRIS	PrivateEye	tracked	0.47	0.73	30	24.8	0.23	23	0.38	22	8.9	19.8	0.59	0.20	+0.05
STORM	Assault	aligned-then-forgotten	0.00	0.73	22	8.0	0.55	10	0.14	16	34.3	8.0	0.19	0.05	+0.07
STORM	DemonAttack	aligned-then-forgotten	0.12	0.54	24	12.0	0.38	15	0.71	13	2.0	8.1	0.46	0.21	+0.07
STORM	Jamesbond	aligned-then-forgotten	0.00	0.56	9	12.8	0.44	5	0.00	5	11.2	12.8	0.60	0.00	+0.28
STORM	KungFuMaster	aligned-then-forgotten	0.07	0.23	30	4.0	0.77	7	–	7	2.4	2.4	0.00	0.00	+0.09
STORM	Seaquest	aligned-then-forgotten	0.00	0.71	17	14.2	0.06	16	0.25	12	10.2	14.3	0.75	0.18	+0.01

continued on the next page

model	game	class	colour	template	n	disp.	z.pass	elig.	acc.	pairs	m.err	z.err	better	strict	swap
STORM	Alien	never-aligned	0.00	0.39	28	1.1	0.82	5	–	11	0.0	0.0	0.00	0.00	-0.00
STORM	Amidar	never-aligned	0.00	0.96	28	6.0	0.64	10	0.00	27	6.0	6.0	0.33	0.00	+0.00
STORM	Asterix	never-aligned	0.00	0.80	30	8.2	0.47	16	0.00	24	8.2	7.9	0.00	0.00	+0.00
STORM	BankHeist	never-aligned	0.65	0.41	17	3.7	0.76	4	0.00	7	119.8	3.9	0.00	0.00	-0.01
STORM	Boxing	never-aligned	0.20	0.50	30	18.7	0.23	23	0.09	15	16.7	15.6	0.27	0.03	+0.06
STORM	ChopperCommand	never-aligned	0.76	0.88	25	11.4	0.36	16	0.31	22	16.6	9.4	0.45	0.16	+0.09
STORM	CrazyClimber	never-aligned	0.63	0.59	27	0.2	0.96	1	0.00	16	1.0	0.1	0.12	0.00	+0.01
STORM	Frostbite	never-aligned	0.61	0.83	18	0.2	0.83	3	0.00	15	2.3	0.2	0.00	0.00	+0.03
STORM	Hero	never-aligned	0.07	0.27	30	6.5	0.60	12	0.00	8	11.5	5.0	0.25	0.00	+0.06
STORM	Krull	never-aligned	0.14	0.52	21	8.9	0.43	12	0.00	11	6.9	6.9	0.00	0.00	+0.00
STORM	MsPacman	never-aligned	0.00	0.13	30	19.0	0.20	24	0.00	4	40.9	37.4	0.00	0.00	+0.00
STORM	RoadRunner	never-aligned	0.15	0.23	13	11.6	0.38	8	0.00	3	31.3	2.7	0.00	0.00	+0.10
STORM	UpNDown	never-aligned	0.94	0.69	16	4.6	0.81	3	–	11	3.7	3.7	0.09	0.00	+0.00
STORM	Kangaroo	partial	0.00	0.28	25	2.3	0.72	7	–	7	1.0	0.0	0.14	0.00	+1.01
STORM	Qbert	partial	0.00	0.14	28	39.8	0.00	28	0.75	4	2.1	48.0	1.00	0.11	+0.96
STORM	BattleZone	tracked	0.97	0.23	30	4.0	0.87	4	0.00	7	15.0	15.0	0.43	0.00	+0.61
STORM	Breakout	tracked	0.00	0.88	25	17.7	0.40	15	0.21	22	8.4	17.7	0.59	0.12	+0.69
STORM	Freeway	tracked	0.08	0.42	26	8.0	0.54	12	0.50	11	5.1	4.8	0.45	0.04	+0.69
STORM	Gopher	tracked	0.73	1.00	30	14.0	0.47	16	0.75	30	4.0	14.0	0.87	0.40	+1.29
STORM	Pong	tracked	0.79	0.96	24	22.5	0.29	17	0.31	23	13.0	23.0	0.61	0.21	+1.16
STORM	PrivateEye	tracked	0.67	0.90	30	18.0	0.30	21	0.35	27	11.5	20.0	0.70	0.23	+0.74

L REPRESENTATION PROBE: CONTROLS AND THE DMCONTROL ARM

This appendix gives the two action-pathway controls of Section 4.4 with every value, the probe figure, the full controls table for the RSSM, IRIS and STORM, and the DMControl arm. The two action-pathway controls show that this alignment does not come from the visual consequence. The posterior state receives the action as well as the frame, so the "true latent change" between arms has two sources: the action input and the visual consequence. Holding the action input at arm a and supplying the frames of arm b leaves the posterior almost unchanged near the intervention: the frames-only change is at most 9% of the full change at delays 1 to 3 on every task; at delay 8 it is 2% and 4% on cartpole and cheetah, 23% to 41% on hopper, reacher and finger, and 81% on walker. Holding the frames and changing only the action reproduces the full change near the intervention and aligns with the imagined latent at 0.89 to 1.06 at every tabulated delay. The alignment near one was therefore the action input echoing through the recurrent state on both sides of the comparison, and required no visual evidence at all. Where the visual consequence does enter the posterior (a visual share of at least one half: delay 8 on walker and delays 16 and 20 on every task), the imagined latent no longer tracks the true latent change (alignment 0.00 to 0.34, with one exception, cartpole at delay 16, 0.68 at a share of 0.54), and its frames-only alignment is at most 0.01. Decoding the posterior states of the two true arms never renders the true pixel change (alignment 0.00 at every delay, magnitude at most 0.21 of the true change), so the existing decoder does not render the true consequence from the posterior either. The supported statement is therefore narrow: the uncontrolled probe does not establish that the consequence is encoded, the action pathway explains most of its apparent success, and the existing decoder does not render a consequence from either the imagined or the true posterior state. Whether some other readout could recover the consequence from the state is not tested here; a held-out readout to simulator state with shuffled targets would test it. These checkpoints are weak (18 to 26 dB, Appendix C), so their blindness is unsurprising; the point is that an uncontrolled probe certified a near-perfect latent for a model this weak.

For the token model the sampled-token embedding is as blind as its pixels (Alien 0.02 against 0.00 at delay 20), as one expects, since the frame is a fixed function of the sampled tokens, although a fixed decoder does not force the two alignments to coincide. At the pre-sampling hidden state the uncontrolled probe again looks like a decoding failure: on the two games whose pixels are blind at delay 20 the hidden alignment is 0.41 (Alien) at delay 1 and 0.26 (BankHeist) at delay 3, where its true hidden change first clears the guard among the table's delays, and 0.15 to 0.17 at delay 8, against pixel alignment near zero (Table 16). The same two controls give a different picture from

the RSSM's. Here the true hidden change is mostly visual: with the action tokens held fixed, the frames alone account for 0.70 to 1.00 of it from delay 3 on, so the transformer's state does encode the frames. But in the pixel-blind games the imagined state tracks that visual change weakly or not at all: frames-only alignment is 0.06 and 0.05 on BankHeist at delays 3 and 8 and 0.01 at delay 20, and on Alien 0.14 at delay 1 falling to 0.07 by delay 8 and 0.00 by delay 20, the trace of an early response that does not persist; the uncontrolled values of 0.4 to 0.5 were, once more, the action token's echo (actions-only alignment 0.34 to 0.57 at delays 1 to 8). In the four games whose pixels are not blind at delay 20 (Boxing, Breakout and Freeway, which the random-context taxonomy classes as aligned-then-forgotten, tracked and aligned-then-forgotten, and Pong) the state does track the visual change, at 0.30 to 0.76 for the first three at delays 1 to 3 and at 0.12 to 0.26 on Pong at delays 1 to 8, near their pixel alignment for the first three and at about half of it on Pong. For the token model, therefore, the blindness is not confined to the sampler: the tested readouts of its state do not recover the consequence either, with a modest sampling loss visible where tracking exists. STORM's transformer state, probed with the same controls on the same six games, gives the same answer: it encodes the frames (visual share 0.50 at delay 3 on Alien and 0.80 at delay 8 on BankHeist, rising to one by delay 20), does not track their change in the blind games (frames-only alignment 0.01 to 0.11 on Alien and BankHeist), and tracks it at about the pixel level in Breakout, Freeway and Pong, while in Boxing, never-aligned at delay 20 although its pixels respond at delays 1 to 8, it stays about as low as in the blind games (Table 16). The lesson is the same for the three probed models, the RSSM, IRIS and STORM, and for both latent readouts: a latent probe of an action-conditioned model requires action-pathway controls to establish its interpretation, since without them it can certify an encoded consequence it has not shown to be there; the state's shortcut through the action input is the world-model counterpart of causal confusion in imitation learning (de Haan et al., 2019). Two qualifications hold throughout: the hybrid frame-and-action inputs of the controls can lie outside the model's training distribution, and the visual share is a norm ratio that depends on the representation's coordinates; the controls therefore diagnose the interpretation of these probes and are not a theorem about the absence of information from the state.

Table 16: Controls for the representation probe. *latent* A: alignment of the imagined latent change with the change of the latent assigned to the true frames (the probe). *share*: fraction of that true latent change that remains when the action input is held fixed and only the frames differ (visual evidence); frames-only alignment is reported only where the share is at least 0.25. *actions-only*: alignment with the change produced by the action input alone (– where that change has zero norm and the alignment is undefined, as for IRIS at delay 20, where the visual share is 1.00). *decoded*: pixel alignment of the decoded true posterior states' difference with the true pixel change (RSSM). For the RSSM and STORM, near the intervention the share is near zero and actions-only alignment equals the probe, so the probe measured the action's echo; for IRIS the share at delay 1 ranges from 0.10 to 0.68 across its games and the frames-only column decides.

model	task / game	d	pixel A	latent A	share	frames-only A	actions-only A	decoded A
RSSM	cartpole swingup	1	0.000	0.98	0.00	–	0.98	0.00
RSSM	cartpole swingup	3	0.000	0.98	0.00	–	0.98	0.00
RSSM	cartpole swingup	8	-0.000	0.96	0.02	–	0.96	-0.00
RSSM	cartpole swingup	16	-0.000	0.68	0.54	0.01	0.96	-0.00
RSSM	cartpole swingup	20	-0.000	0.06	0.97	0.00	0.95	-0.00
RSSM	cheetah run	1	0.000	0.99	0.00	–	0.99	0.00
RSSM	cheetah run	3	0.000	0.99	0.00	–	0.99	0.00
RSSM	cheetah run	8	0.000	0.99	0.04	–	0.99	0.00
RSSM	cheetah run	16	-0.000	0.30	0.83	-0.00	1.00	-0.00
RSSM	cheetah run	20	-0.000	0.01	0.99	0.00	1.02	0.00
RSSM	walker walk	1	0.000	1.00	0.02	–	1.01	0.00
RSSM	walker walk	3	0.000	1.00	0.09	–	1.01	0.00
RSSM	walker walk	8	-0.000	0.34	0.81	0.00	1.01	0.00
RSSM	walker walk	16	0.000	0.00	1.00	0.00	1.03	0.00
RSSM	walker walk	20	-0.000	-0.00	1.00	-0.00	1.06	0.00
RSSM	reacher easy	1	-0.000	0.99	0.00	–	0.99	0.00

continued on the next page

model	task / game	d	pixel A	latent A	share	frames-only A	actions-only A	decoded A
RSSM	reacher easy	3	-0.000	1.00	0.02	–	1.00	0.00
RSSM	reacher easy	8	0.000	0.92	0.32	0.02	1.02	0.00
RSSM	reacher easy	16	0.000	0.01	1.00	0.00	1.03	0.00
RSSM	reacher easy	20	0.000	0.00	1.00	0.00	1.03	0.00
RSSM	finger spin	1	0.000	0.98	0.01	–	0.98	0.00
RSSM	finger spin	3	0.000	0.98	0.07	–	0.98	0.00
RSSM	finger spin	8	0.000	0.79	0.41	0.03	0.98	0.00
RSSM	finger spin	16	-0.000	0.00	1.00	0.00	0.97	-0.00
RSSM	finger spin	20	-0.000	0.00	1.00	0.00	0.96	-0.00
RSSM	hopper hop	1	-0.000	0.97	0.00	–	0.97	-0.00
RSSM	hopper hop	3	-0.000	0.97	0.02	–	0.97	-0.00
RSSM	hopper hop	8	0.000	0.91	0.23	–	0.96	0.00
RSSM	hopper hop	16	0.000	0.01	0.99	0.00	0.92	0.00
RSSM	hopper hop	20	0.000	0.00	1.00	-0.00	0.89	-0.00
IRIS	Alien	1	0.077	0.41	0.66	0.14	0.55	–
IRIS	Alien	3	0.058	0.24	0.83	0.09	0.44	–
IRIS	Alien	8	0.057	0.15	0.92	0.07	0.40	–
IRIS	Alien	20	0.000	0.00	1.00	0.00	–	–
IRIS	BankHeist	1	0.000	0.47	0.10	–	0.57	–
IRIS	BankHeist	3	0.000	0.26	0.70	0.06	0.49	–
IRIS	BankHeist	8	0.001	0.17	0.95	0.05	0.34	–
IRIS	BankHeist	20	0.000	0.01	1.00	0.01	–	–
IRIS	Boxing	1	0.261	0.53	0.60	0.30	0.64	–
IRIS	Boxing	3	0.328	0.43	0.80	0.33	0.53	–
IRIS	Boxing	8	0.222	0.24	0.94	0.17	0.39	–
IRIS	Boxing	20	0.051	0.04	1.00	0.04	–	–
IRIS	Breakout	1	0.718	0.81	0.48	0.61	0.85	–
IRIS	Breakout	3	0.898	0.80	0.82	0.58	0.73	–
IRIS	Breakout	8	0.903	0.78	0.91	0.59	0.63	–
IRIS	Breakout	20	0.670	0.46	1.00	0.46	–	–
IRIS	Freeway	1	0.448	0.76	0.65	0.76	1.03	–
IRIS	Freeway	3	0.609	0.76	0.91	0.54	0.68	–
IRIS	Freeway	8	0.400	0.66	0.97	0.57	0.50	–
IRIS	Freeway	20	0.161	0.38	1.00	0.38	–	–
IRIS	Pong	1	0.325	0.39	0.68	0.12	0.45	–
IRIS	Pong	3	0.420	0.36	0.90	0.18	0.48	–
IRIS	Pong	8	0.517	0.38	0.94	0.26	0.40	–
IRIS	Pong	20	0.229	0.17	1.00	0.17	–	–
STORM	Alien	1	0.118	0.83	0.00	–	0.83	–
STORM	Alien	3	0.063	0.63	0.50	0.11	0.81	–
STORM	Alien	8	0.053	0.36	0.81	0.05	0.71	–
STORM	Alien	20	0.004	0.02	1.00	0.02	0.29	–
STORM	BankHeist	1	0.000	0.77	0.00	–	0.77	–
STORM	BankHeist	3	0.000	0.54	0.34	0.05	0.67	–
STORM	BankHeist	8	-0.000	0.39	0.80	0.04	0.62	–
STORM	BankHeist	20	-0.000	0.04	0.99	0.01	0.34	–
STORM	Boxing	1	0.073	0.60	0.00	–	0.60	–
STORM	Boxing	3	0.082	0.43	0.66	0.13	0.54	–
STORM	Boxing	8	0.072	0.23	0.91	0.09	0.39	–
STORM	Boxing	20	0.015	0.03	1.00	0.03	0.10	–
STORM	Breakout	1	0.601	0.88	0.00	–	0.88	–
STORM	Breakout	3	0.709	0.82	0.85	0.52	0.69	–
STORM	Breakout	8	0.658	0.71	0.91	0.48	0.56	–
STORM	Breakout	20	0.416	0.41	1.00	0.40	0.21	–
STORM	Freeway	1	0.382	0.81	0.00	–	0.81	–
STORM	Freeway	3	0.557	0.78	0.85	0.37	0.54	–

continued on the next page

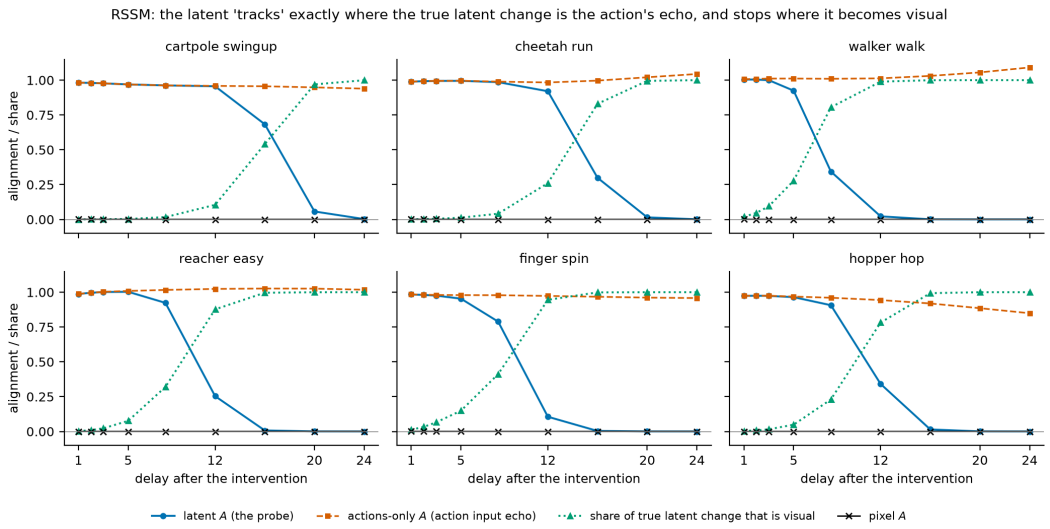

Figure 8: RSSM representation probe with its controls, six DMControl tasks (4 evaluation seeds $\times$ 20 episodes). The probe (latent A, blue) reports alignment near one at delays 1 to 3 and collapses between delays 5 and 20 depending on the task. The actions-only control (orange) stays near one at every delay (0.85 to 1.09 over the plotted delays 1 to 24; 0.89 to 1.06 at the delays of Table 16), matching the probe where the probe is high and remaining after it has collapsed, and the share of the true latent change that is visual (green) is near zero during the early near-one alignment and rises to one where the probe collapses. Pixel alignment (black) is zero throughout. The probe measured the action input's echo through the recurrent state, not an encoded consequence.

model	task / game	d	pixel A	latent A	share	frames-only A	actions-only A	decoded A
STORM	Freeway	8	0.529	0.63	0.88	0.38	0.52	–
STORM	Freeway	20	0.288	0.49	1.00	0.46	0.57	–
STORM	Pong	1	0.248	0.87	0.00	–	0.87	–
STORM	Pong	3	0.417	0.75	0.70	0.54	0.79	–
STORM	Pong	8	0.513	0.69	0.85	0.52	0.62	–
STORM	Pong	20	0.362	0.45	0.99	0.42	0.50	–

L.1 DMControl: an RSSM whose pixels never track

This arm is a weak, self-trained model (Appendix C), included as a second substrate and, above all, as the testbed for the representation-probe controls of Section 4.4; its blindness is not offered as a finding about recurrent state-space models. On six DMControl tasks an RSSM world model shows the separation in its starkest form: consistency is flat across the horizon while causal alignment is within ± 0.002 at delay one and 0.000 at delay 24 on every task and seed, although the simulator's own divergence persists (it falls by a tenth to a fifth on cheetah and on some finger seeds) or grows (ten- to twelve-fold on cartpole). The RSSM moves in response to the action, its response magnitude is nonzero, but in a direction uncorrelated with the world's; and, as Section 4.4 showed, an uncontrolled probe of its latent reports tracking that the action-pathway controls take away.

M Tested negatives and the reward-head ranking design

Three claims we set out to make are not supported, and we list them because the gates that refused them are part of the method. A decorrelation between consistency and causal alignment across games is not established: at $n = 26$ no association between the two is detected within any architecture ($\rho = -0.32$, 0.07 and 0.13, none significant), which excludes neither a weak negative relation nor a weak positive one. No association between causal alignment and the published Atari-100k returns of agents trained inside these models is detected, in any of three designs (DIAMOND and IRIS on the games measurable under the fixed intervention; every design unpowered). And the contrast

we predicted against Genie's ΔPSNR did not appear; the two agree. Two further readings are not supported at $n = 26$: that consistency does not transfer across architectures (it does, Appendix P) and that prediction accuracy predicts causal alignment within IRIS (the association is not detected, Appendix P).

What ΔPSNR cannot see. Agreement at the game level does not make the two scores interchangeable. ΔPSNR asks whether the future moves when the action changes; alignment asks whether it moves as the world does. Searching the 60 cells that have a delay-20 ΔPSNR (every IRIS and STORM cell and the eight DIAMOND games of Table 5) for a positive ΔPSNR (bootstrap interval above zero) beside an alignment below 0.05 finds no such cell for DIAMOND, two for IRIS (Asterix, ΔPSNR $+0.26$ dB at alignment 0.044; Jamesbond, $+0.02$ dB at 0.001) and six for STORM (Boxing, ChopperCommand, Hero, Jamesbond and KungFuMaster at $+0.18$ to $+0.35$ dB with alignment 0.015 to 0.043, and CrazyClimber, whose ΔPSNR of 0.00 dB passes the interval rule by a hair, $[0.00, 0.16]$, and is negligible at the displayed precision although its magnitude is 0.76): in the two IRIS and five STORM cells with a measurable ΔPSNR the transformers' futures move with the action in a direction unrelated to the world's. This is an exploratory search over 60 cells with an uncorrected interval-based selection rule, reported as a description of where the two scores part, not as a tested claim. Appendix O shows the same blind spot on constructed failures.

Ranking candidate actions with the model's own reward head. The published-returns analysis above cannot isolate the world model, so we ran a controlled version: fixed contexts, eight candidate action sequences, the same model seed for every candidate, the candidates ranked by the model's own predicted return (DIAMOND's reward model, IRIS's reward head) and by the emulator's replayed return. With random candidates and a 20-step horizon, four of ten games produced no context in which the emulator's returns differed, and the rest give a null: over the six informative games, ranking quality follows causal alignment at $\rho = 0.43$ for DIAMOND ($p = 0.381$) and 0.03 for IRIS ($p = 1.00$), against a minimum detectable $|\rho|$ of 0.92 at this size. The IRIS values of this design were measured before the harness correction of Appendix C and were not rerun. The pattern inside the numbers explains why this design cannot answer the question: DIAMOND ranks Seaquest's candidates at chance ($\rho = 0.00$) while its pixel response there is 0.93, and Frostbite and Breakout, aligned at 0.60 and 0.95, are ranked near the reference in Section 4.6's design. Lengthening the horizon to 40 steps and holding an action for half the candidates makes eight games informative per architecture and gives $\rho = 0.19$ for DIAMOND and 0.59 for IRIS ($n = 8$ each, minimum detectable 0.85), with Freeway ranked with zero top-choice regret by both models ($\rho = 0.59$ and 1.00) although DIAMOND's response to the intervention is zero under both the brief and the sustained design (IRIS's is 0.16 under the brief one). Neither six-game value is interpretable at this size. A model's own reward head is a second learned component, and its quality need not track whether the dynamics respond to the action, so a ranking test that uses it measures the head as much as the dynamics. Section 4.6 holds the reward information fixed instead.

N PLANNING: OPEN LOOP AND CLOSED LOOP

This appendix gives the open-loop design, the open-loop estimates with their intervals, every open-loop cell, and the closed-loop results on all twelve games. No score, ours included, shows an association that passes the pre-declared gate, and we report the estimates with their intervals rather than as absences. Over the 45 cells, with intervals from a game-clustered bootstrap that keeps every architecture's cell of a resampled game together and p-values from a game-clustered block permutation, the model's ranking correlation relates to the brief-intervention alignment at $\rho = 0.30$ (95% CI $[-0.09, 0.62]$, $p = 0.15$), to the response error at -0.21 ($[-0.57, 0.24]$), and to prediction PSNR at 0.37 ($[-0.03, 0.66]$, $p = 0.087$); PSNR's interval includes zero and its permutation test does not reach 0.05, so accuracy's positive estimate is inconclusive here (the pre-declared gate asked for 0.41, computed for 45 independent cells). Within architectures ($n = 15$ each, minimum detectable $|\rho|$ 0.67) the alignment estimate is -0.01 ($[-0.50, 0.50]$) for DIAMOND, 0.66 ($[0.23, 0.89]$, $p = 0.006$) for IRIS and 0.37 ($[-0.22, 0.80]$) for STORM, and IRIS's PSNR and ΔPSNR estimates are 0.57 ($[0.06, 0.90]$, $p = 0.030$) and 0.63 ($[0.23, 0.84]$, $p = 0.015$), so the pre-declared rule (alignment's interval excluding zero while PSNR's includes it) is met in no architecture. Contexts, candidates and true futures are identical for DIAMOND and IRIS, which share episode seeds and emulator preprocessing; STORM's emulator replays use its frame pooling and start four steps ear-

lier where the game has a warm-up (in Assault, Asterix, Boxing and Breakout, which have none, the candidates instead start four steps later), so its cells are matched within STORM and only approximately across architectures. On the context-matched DIAMOND and IRIS cells alone ($n = 30$ in 15 games), the estimates are $\rho = 0.29$ for alignment ($[-0.15, 0.68]$, $p = 0.22$), -0.33 for the error and 0.43 for PSNR ($[-0.09, 0.75]$, $p = 0.066$). Grouping cells by response points the same way without settling it: the 16 cells with alignment below 0.05 and a usable true-frame reference (above 0.2) have a median ranking correlation of 0.26 (the raw correlation with the emulator's ranking, not the ratio to the reference), 6 of them at or below chance, and the 6 with alignment at or above 0.40 a median of 0.60, 4 of them near the reference (one-sided $p = 0.024$ by an unclustered permutation of cell labels, 5,000 draws, one exploratory test among several). The cells that break the expected pattern are instructive on both sides. DIAMOND's response on Asterix, DemonAttack and Seaquest is aligned at 0.83, 0.63 and 0.93, yet its futures rank candidates at or below chance on the first two and at a third of the reference on the last; Frostbite and Breakout, aligned at 0.60 and 0.95, rank at the reference. Conversely Freeway, UpNDown, CrazyClimber and Qbert, whose alignment is at most 0.11, rank at 0.75 to 1.00 of the reference. We tested the natural explanation for the latter, that a sustained action produces a response a brief one does not, by re-measuring alignment with the intervention held for forty steps: Freeway's alignment stays at 0.01 at delay 20 and 0.00 at delay 44, so the explanation fails for Freeway, the only one of the four measured with a sustained intervention. A model can be blind by this measure, which demands the right change at each pixel and each delay, and still order coarse outcomes correctly; and a model can produce the right change and still rank candidates badly: these experiments do not separate errors in reward-relevant prediction from scorer errors.

Table 17: Ranking candidate action sequences with the reward information held fixed. A reward scorer trained on emulator frame pairs (independent of every model) scores each model's predicted futures under eight candidate sequences per context; candidates are ranked by scored predicted return and by the emulator's replayed return. ρ is the Spearman correlation between the two rankings, median over informative contexts with a bootstrap interval; the true-frame reference applies the same scorer to the true futures (a reference, not an upper bound: a generated future can be easier for the scorer than the true one); rel. is the model's ρ relative to that reference; regret is the normalised loss of the model's top choice; hit@2 the fraction of contexts in which the model's top choice is among the emulator's best two. The eleven games whose rewards are too sparse for the scorer are listed with the reason in the open-loop design paragraph of Appendix N; rel. is omitted (–) where the true-frame reference does not exceed 0.2.

model	game	A	informative	ρ model	ρ true-frame ref.	rel.	regret	hit@2
DIAMOND	Alien	+0.00	60/60	$+0.06\ [-0.06, +0.22]$	$+0.69$	0.08	1.07	0.35
DIAMOND	Amidar	+0.00	58/60	$+0.30\ [+0.16, +0.44]$	$+0.77$	0.39	1.14	0.55
DIAMOND	Assault	+0.12	57/60	$+0.13\ [+0.03, +0.25]$	$+0.00$	–	0.00	0.63
DIAMOND	Asterix	+0.83	60/60	$-0.10\ [-0.20, +0.06]$	$+0.82$	-0.12	1.33	0.25
DIAMOND	Boxing	+0.07	60/60	$+0.19\ [+0.05, +0.32]$	$+0.52$	0.37	0.89	0.55
DIAMOND	Breakout	+0.95	28/60	$+0.61\ [+0.58, +0.76]$	$+0.61$	1.00	0.00	0.96
DIAMOND	ChopperCommand	+0.41	37/60	$+0.00\ [-0.08, +0.08]$	-0.08	–	1.14	0.70
DIAMOND	CrazyClimber	+0.05	60/60	$+0.85\ [+0.79, +0.87]$	$+0.92$	0.93	0.00	0.88
DIAMOND	DemonAttack	+0.63	57/60	$+0.08\ [-0.12, +0.22]$	$+0.19$	–	1.14	0.49
DIAMOND	Freeway	+0.00	20/60	$+0.59\ [+0.58, +0.77]$	$+0.59$	1.00	0.00	1.00
DIAMOND	Frostbite	+0.60	36/60	$+0.84\ [+0.76, +0.87]$	$+0.90$	0.93	0.00	0.89
DIAMOND	MsPacman	+0.00	60/60	$+0.06\ [-0.12, +0.19]$	-0.33	–	1.01	0.30
DIAMOND	Qbert	+0.11	60/60	$+0.63\ [+0.55, +0.76]$	$+0.83$	0.75	0.38	0.60
DIAMOND	Seaquest	+0.93	39/60	$+0.25\ [+0.13, +0.44]$	$+0.85$	0.30	1.14	0.51
DIAMOND	UpNDown	+0.00	41/60	$+0.59\ [+0.33, +0.73]$	$+0.79$	0.75	0.00	0.76
IRIS	Alien	+0.00	60/60	$+0.05\ [-0.11, +0.15]$	$+0.69$	0.07	1.14	0.42
IRIS	Amidar	+0.00	58/60	$+0.19\ [+0.01, +0.34]$	$+0.77$	0.24	1.22	0.33
IRIS	Assault	+0.03	57/60	$-0.05\ [-0.22, +0.11]$	$+0.00$	–	1.33	0.37
IRIS	Asterix	+0.04	60/60	$+0.04\ [-0.16, +0.14]$	$+0.82$	0.05	1.33	0.35
IRIS	Boxing	+0.05	60/60	$+0.19\ [+0.05, +0.36]$	$+0.52$	0.36	0.78	0.63

continued on the next page

model	game	A	informative	ρ model	ρ true-frame ref.	rel.	regret	hit@2
IRIS	Breakout	+0.72	28/60	$+0.61$ $[+0.58, +0.76]$	$+0.61$	0.99	0.00	0.96
IRIS	ChopperCommand	+0.18	37/60	$+0.17$ $[-0.17, +0.28]$	-0.08	–	1.14	0.70
IRIS	CrazyClimber	+0.17	60/60	$+0.87$ $[+0.85, +0.89]$	$+0.92$	0.94	0.00	0.95
IRIS	DemonAttack	+0.06	57/60	$+0.06$ $[-0.14, +0.16]$	$+0.19$	–	1.14	0.56
IRIS	Freeway	+0.16	20/60	$+0.59$ $[+0.58, +0.77]$	$+0.59$	1.00	0.00	1.00
IRIS	Frostbite	+0.01	36/60	$+0.44$ $[+0.25, +0.54]$	$+0.90$	0.49	0.25	0.56
IRIS	MsPacman	+0.00	60/60	-0.13 $[-0.28, -0.07]$	-0.33	–	1.13	0.22
IRIS	Qbert	+0.27	60/60	$+0.73$ $[+0.66, +0.76]$	$+0.83$	0.88	0.00	0.73
IRIS	Seaquest	+0.26	39/60	$+0.38$ $[+0.06, +0.44]$	$+0.85$	0.44	0.00	0.56
IRIS	UpNDown	+0.01	41/60	$+0.15$ $[-0.09, +0.43]$	$+0.79$	0.20	1.00	0.54
STORM	Alien	+0.00	60/60	$+0.10$ $[+0.00, +0.30]$	$+0.72$	0.14	1.00	0.40
STORM	Amidar	+0.00	60/60	-0.22 $[-0.36, -0.05]$	$+0.87$	-0.25	1.27	0.40
STORM	Assault	+0.01	56/60	$+0.00$ $[-0.07, +0.11]$	$+0.00$	–	1.23	0.46
STORM	Asterix	+0.00	60/60	$+0.26$ $[+0.18, +0.35]$	$+0.78$	0.33	0.80	0.50
STORM	Boxing	+0.02	60/60	$+0.27$ $[+0.15, +0.43]$	$+0.44$	0.61	0.67	0.50
STORM	Breakout	+0.44	42/60	$+0.59$ $[+0.58, +0.76]$	$+0.67$	0.88	0.00	0.95
STORM	ChopperCommand	+0.02	38/60	-0.04 $[-0.16, +0.13]$	-0.08	–	1.14	0.61
STORM	CrazyClimber	+0.00	60/60	$+0.84$ $[+0.79, +0.87]$	$+0.91$	0.92	0.00	0.80
STORM	DemonAttack	+0.12	56/60	$+0.08$ $[-0.02, +0.28]$	$+0.40$	0.21	1.14	0.50
STORM	Freeway	+0.33	17/60	$+0.08$ $[-0.25, +0.58]$	$+0.59$	0.14	1.14	0.65
STORM	Frostbite	+0.00	37/60	$+0.31$ $[+0.17, +0.40]$	$+0.91$	0.34	0.62	0.73
STORM	MsPacman	-0.00	60/60	-0.05 $[-0.27, +0.11]$	-0.30	–	1.07	0.33
STORM	Qbert	+0.35	59/60	$+0.58$ $[+0.42, +0.69]$	$+0.81$	0.72	0.00	0.73
STORM	Seaquest	+0.05	45/60	$+0.39$ $[+0.28, +0.44]$	$+0.85$	0.47	1.00	0.60
STORM	UpNDown	+0.00	41/60	$+0.41$ $[+0.08, +0.62]$	$+0.77$	0.54	0.89	0.59

Open-loop design. The controlled version of the planning question holds the contexts, the candidate actions and the true futures fixed and varies the world model; the reward scorer is trained once per game on emulator frame pairs, independent of every model, and frozen, so that every architecture is scored by the same instrument. The scorer is a small convolutional classifier over a pair of consecutive 64×64 frames (six input channels; three stride-2 convolutions, a 256-unit layer and a three-way output) trained to predict the sign of the emulator's per-step reward (negative, zero, positive) with cross-entropy and inverse-frequency class weights (Adam, learning rate 10^{-3}, batch 128, 8 epochs) on 600 emulator episodes per game: after the game's warm-up, four random context actions and one action held for 60 steps, giving 64 frame pairs per episode and 38,400 pairs per game; 80% of the pairs train and 20% are held out, the split being over pairs rather than episodes, so the held-out accuracy is an optimistic estimate of generalisation to new episodes. A game's scorer is trained once and frozen (the DIAMOND run trained it; IRIS and STORM loaded it). The gate is a balanced held-out accuracy of at least 0.5 and at least 50 rewarded pairs; eleven of the 26 games fail it because the held and random actions produce almost no reward within the 64-step window: BankHeist (0 rewarded of 38,400 pairs); BattleZone (0 rewarded of 38,400 pairs); Gopher (0 rewarded of 38,400 pairs); Hero (3 rewarded of 38,400 pairs); Jamesbond (0 rewarded of 38,400 pairs); Kangaroo (0 rewarded of 38,400 pairs); Krull (32 rewarded of 38,400 pairs; balanced accuracy 0.50); KungFuMaster (0 rewarded of 38,400 pairs); Pong (0 rewarded of 8,800 pairs); PrivateEye (0 rewarded of 38,400 pairs); RoadRunner (40 rewarded of 38,400 pairs) (Pong's evidence is from an earlier 200-episode run of the same scorer design in which only alternate episodes hold an action, 44 steps per episode). For each of 20 contexts per seed and 3 seeds (60 contexts per cell; the contexts are the first four frames of an emulator rollout after the warm-up, from reset seeds shared across seeds and across DIAMOND and IRIS) we draw eight candidate sequences of 40 scored actions (41 are drawn; the last is neither imagined nor scored), four holding one action and four random, imagine each with the same model seed (common random numbers across candidates), and score the predicted frames; the scored return of a sequence is the sum over consecutive frame pairs of the scorer's expected reward sign, the emulator's return is the sum of reward signs of the replayed candidate, and a context is informative when the eight emulator returns are not all equal. Per informative context we record the Spearman correlation between scored and emulator returns for the model's frames and for the true frames (the true-frame reference, which separates a weak scorer from a weak model; a reference, not an upper bound, since a generated future can be easier for the scorer than the true one), the normalised regret

$(\max_k g_k - g_{k^*})/(\max_k g_k - \bar{g})$ with k^* the model's top candidate (undefined when all returns tie), and hit@2, whether the model's top candidate has one of the two largest emulator returns; per cell the median of ρ and of the regret over informative contexts with a bootstrap interval and the fraction of informative contexts for hit@2, cells needing at least ten informative contexts; *rel.* in Table 17 is the ratio of the model's median correlation to the reference's where the reference exceeds 0.2. Fifteen games are usable for each of the three architectures.

Table 18: Closed-loop planning inside the emulator: random-shooting model-predictive control (MPC; $K = 8$ candidates, horizon 12, 30 decision steps, 60 episodes per cell) with each world model imagining the candidates and one frozen per-game scorer reading their returns. Cells give the mean emulator return with a bootstrap interval; an asterisk marks a model whose bootstrap interval of the difference from the random policy (episodes resampled independently in each condition) lies entirely above zero (a positive gain; intervals entirely below zero are not marked). The oracle is an emulator-rollout oracle: it plans with the emulator itself under the same frozen scorer, not an optimal-return oracle.

game	random	oracle	DIAMOND	IRIS	STORM
Alien	16.8 [13.7, 20.0]	31.7 [27.2, 36.2]*	18.3 [14.8, 22.0]	15.7 [12.7, 18.8]	17.0 [13.7, 20.5]
Amidar	0.2 [0.0, 0.4]	0.2 [0.0, 0.5]	0.2 [0.1, 0.5]	0.4 [0.2, 0.7]	0.1 [0.0, 0.2]
Asterix	110.0 [100.0, 120.0]	133.3 [122.5, 144.2]*	134.2 [122.5, 145.8]*	111.7 [100.8, 122.5]	120.0 [111.7, 130.0]
Boxing	-0.0 [-0.2, 0.2]	0.6 [0.4, 0.8]*	0.0 [-0.2, 0.3]	-0.1 [-0.3, 0.2]	0.1 [-0.1, 0.3]
Breakout	0.0 [0.0, 0.0]	0.1 [0.1, 0.2]*	0.1 [0.1, 0.2]*	0.2 [0.1, 0.3]*	0.3 [0.2, 0.4]*
CrazyClimber	118.3 [100.0, 136.7]	166.7 [151.7, 181.7]*	171.7 [153.3, 190.0]*	180.0 [160.0, 200.0]*	161.7 [145.0, 178.3]*
Freeway	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
Frostbite	7.8 [5.2, 10.8]	17.0 [13.3, 20.8]*	13.7 [10.3, 17.2]*	11.5 [8.3, 14.8]	9.0 [6.2, 12.0]
MsPacman	38.7 [35.3, 42.0]	31.8 [29.7, 34.2]	35.7 [32.5, 39.0]	34.8 [32.2, 37.5]	39.8 [36.2, 43.5]
Qbert	22.1 [15.0, 29.2]	43.8 [39.6, 47.5]*	34.6 [27.9, 40.8]*	39.2 [33.3, 45.0]*	58.8 [45.8, 75.4]*
Seaquest	6.0 [3.7, 8.3]	16.7 [13.7, 19.7]*	7.3 [5.0, 10.0]	7.0 [4.7, 9.3]	10.0 [7.3, 12.7]*
UpNDown	72.5 [37.8, 115.3]	82.7 [44.8, 124.5]	74.8 [40.3, 115.5]	57.8 [27.2, 94.8]	66.3 [33.2, 107.3]

Closed-loop design. Random-shooting model-predictive control inside the emulator: an episode is reset from its seed (the same twenty reset seeds for every seed, condition and architecture), the game's warm-up and three random context actions are applied (seven for STORM, whose eight-frame context is filled the same way with no warm-up shift, so its first decision falls four emulator steps later), and at each of 30 decision steps eight random candidate sequences of 12 actions are drawn, each candidate is imagined from the last four real frames (eight for STORM) with a model seed shared by the eight candidates, scored by the frozen scorer on the imagined frames, and the first action of the best-scoring candidate is executed in the emulator; the context is always the last real frames, never imagined ones. The random policy executes the first action of the first candidate; the emulator-rollout oracle scores each candidate's true future obtained by cloning and restoring the emulator state, under the same scorer, candidates and horizon (an oracle for the scorer, not for the return). Returns are means over 3 seeds $\times$ 20 episodes with a 4,000-resample percentile bootstrap; the difference from the random policy has its own bootstrap interval (episodes resampled independently in each condition) and a model is counted as useful in a game when that interval lies entirely above zero; the oracle-normalised gain $G = (\bar{R}_{\mathrm{model}} - \bar{R}_{\mathrm{random}})/(\bar{R}_{\mathrm{oracle}} - \bar{R}_{\mathrm{random}})$ is defined per game where the oracle's gain interval lies above zero (8 of the 12 games; on MsPacman the oracle's return lies below the random policy's, -6.8 $[-10.8, -2.8]$, the scorer being anti-correlated with the emulator's return there, true-frame reference -0.33 in Table 17), the reported summary is the median of G over those eight games, and the correlations of gain with alignment and PSNR are computed on them. The random and oracle controls follow the DIAMOND and IRIS timing (first decision after three context actions, a four-frame context); STORM's first decision falls four emulator steps later, so its gains are measured against controls that act earlier and its closed-loop counts are comparable within STORM and only approximately with the other two architectures. On Alien the model-minus-random differences are +1.5 [-3.3, 6.5], -1.2 [-5.7, 3.2] and +0.2 [-4.5, 5.0] return points for DIAMOND, IRIS and STORM, each interval including zero, against the oracle's +14.8 [9.3, 20.5]. On the eight games the gain's Spearman correlation with the delay-20 alignment of Table 23 is 0.26, 0.76 and 0.71, and with PSNR 0.74, 0.60 and 0.69 (DIAMOND, IRIS, STORM; game-permutation p 0.54, 0.04 and 0.06 for alignment and 0.05, 0.14 and 0.07 for PSNR, uncor-

rected); at $n = 8$ the minimum detectable $|\rho|$ at 80% power is 0.85, so these estimates are reported and not interpreted. The twelve games are those of the fifteen open-loop games whose frozen scorer reached a balanced held-out accuracy of at least 0.8 on emulator frame pairs in the DIAMOND and IRIS preprocessing (Assault, ChopperCommand and DemonAttack fall below it), a list fixed before the closed-loop runs; the same scorers score lower on STORM's max-pooled frames, so the gate was applied once, not per architecture.

What the open loop does and does not show. Across the 45 cells (15 games), alignment's and the response error's associations with ranking quality are inconclusive and prediction accuracy's is positive but inconclusive (estimates and intervals above); no estimate of the pre-declared 45-cell analysis reaches the gate (the post-hoc estimate on the context-matched cells does), and none shows that a stronger association does not exist. Intervals are game-clustered bootstrap intervals and p-values come from a game-clustered block permutation (a random bijection of the games, every cell taking the ranking quality of the same-position cell of the game it is mapped to), two-sided. What survives is the diagnosis: IRIS predicts Alien and Frostbite at the same accuracy (29.8 and 29.5 dB) and ranks the first at chance and the second at 0.49 of the reference, so accuracy does not say which failure is present and the response measurement does (never aligned against aligned and then forgotten). Whether a response score can be made to predict usefulness by measuring the response on the reward-relevant coordinates is a question we leave open.

O CONSTRUCTED FAILURES: WHICH SCORE DETECTS WHICH

Table 19: Which score detects which failure. Predictors built from the emulator's own futures (8 games, up to 30 episodes each, delay 20, median over games of the per-game median); every realistic predictor carries the same 30 dB noise in both arms. PSNR gives every failure the value it gives the correct response and consistency lies within 0.02 of it; ΔPSNR assigns zero to the action-ignoring predictor and does not order the rest by correctness (the wrong-action predictor lies within 0.14 dB of the correct one); the response triple separates every failure construction from the correct response, while shared additive noise leaves the correct triple unchanged (the noiseless and noisy correct rows coincide). All constructions of an episode share one noise draw.

predictor	PSNR	consist.	ΔPSNR	A	R	ε
perfect (noiseless)	100.0	0.767	+69.93	+1.00	1.00	0.00
correct	30.5	0.620	+3.63	+1.00	1.00	0.00
action-ignoring	30.5	0.620	+0.00	+0.00	0.00	1.00
wrong action	30.5	0.619	+3.49	+0.50	0.99	0.97
delayed by 2	30.5	0.616	+2.73	+0.37	0.89	1.04
reversed	30.5	0.609	+1.60	-0.60	0.69	2.67
orthogonal (world-sized)	30.5	0.601	+1.61	+0.02	0.69	1.33
correct + irrelevant motion	30.5	0.616	+4.17	+0.96	1.16	0.47
half-size correct	30.5	0.623	+1.27	+0.50	0.50	0.25

To separate what each score can see from what these particular models happen to do, we build predictors from the emulator's own futures and score them with every metric in the paper (Table 19). Every realistic predictor carries the same 30 dB noise in both arms, the accuracy of the models we evaluate. A predictor that ignores the action, one that responds to the wrong action, one that responds two steps late, one that reverses the true change, one that changes the future by a comparable amount in an orthogonal direction, and one that adds irrelevant motion to the correct change all receive the same factual PSNR as the correct response (30.5 dB, because that score sees one arm) and a consistency within 0.02 of its 0.62 (0.60 to 0.62). ΔPSNR assigns zero to the action-ignoring predictor and values between 1.60 and 4.17 dB to the rest, but its ordering does not follow correctness: the wrong-action predictor scores within 0.14 dB of the correct one (3.49 against 3.63), and the reversed and orthogonal responses, which are deliberately incorrect, score 1.60 and 1.61 dB, close to the 1.27 of a correct response of half the size. Alignment, magnitude and error take distinct values on these constructions and characterise them (the reversed row differs from the ideal $A = -1$,

$R = 1, \varepsilon = 4$ because the reversed frames are clipped to the pixel range and carry the shared noise): reversed gives negative alignment (-0.60) and the largest error, orthogonal gives alignment near zero (0.02) at a magnitude comparable to the truth, wrong action gives half alignment (0.50) at full magnitude, delayed reduces alignment to 0.37 with the magnitude intact, irrelevant motion gives high alignment with magnitude above one, and half-size gives half alignment at half magnitude. All constructions of an episode share one noise draw and the table gives medians without intervals, so it is a set of worked examples of what each score registers, not a statistical comparison, and it says nothing about which score is useful for planning. The two IRIS cells with a positive ΔPSNR at an alignment below 0.05 listed in Appendix M (Asterix and Jamesbond) are the natural counterpart of the orthogonal row.

P STATISTICAL PROTOCOL, TESTS AND ABLATIONS

Every reported statistic passes through the same pipeline; this appendix states it so that each number in the paper can be recomputed from the released per-episode files.

Table 20: Statistical summary. Every correlation is a Spearman rank correlation with a Monte Carlo permutation p; *gate* means $|\rho|$ exceeds the pre-specified effect-size gate (0.41 at $n = 26$), a decision rule fixed in advance and not a statement of statistical power (the 80% power minimum detectable effect is given in Appendix P); the open-loop planning rows use the pre-declared planning gate of 0.41, which for 45 cells is the 80% power minimum detectable effect rather than the $p < 0.05$ threshold, so the column does not apply one rule to every row; *survives* refers to Holm correction within the row's family: the three pairwise transfer correlations of each quantity form one family, the six within-architecture correlations another; rows without a family show a dash. Cross-architecture rows use the emulator-chosen interventions at delay 20. Rows marked pooled combine cells of several architectures that share games; their tests are game-clustered where stated and unclustered where stated.

claim	test	n	estimate	p	gate	survives
Cross-architecture agreement, causal alignment (DIAMOND vs IRIS)	Spearman, 20,000 permutations	26	+0.71 [+0.42, +0.88]	0.0001	yes	yes
Cross-architecture agreement, causal alignment (DIAMOND vs STORM)	Spearman, 20,000 permutations	26	+0.60 [+0.22, +0.85]	0.0017	yes	yes
Cross-architecture agreement, causal alignment (IRIS vs STORM)	Spearman, 20,000 permutations	26	+0.83 [+0.62, +0.93]	$< 10^{-4}$	yes	yes
Cross-architecture agreement, consistency (three pairs)	Spearman, permutation	26	+0.72, +0.79, +0.91	all $< 10^{-4}$	yes	yes
Cross-architecture agreement, PSNR (three pairs)	Spearman, permutation	26	+0.69, +0.66, +0.73	≤ 0.0004	yes	yes
PSNR $\rightarrow$ alignment, DIAMOND	Spearman, permutation	26	+0.38	0.0546	no	no
PSNR $\rightarrow$ alignment, IRIS	Spearman, permutation	26	+0.34	0.0848	no	no
PSNR $\rightarrow$ alignment, STORM	Spearman, permutation	26	+0.47	0.0170	yes	no
Consistency $\rightarrow$ alignment, DIAMOND	Spearman, permutation	26	-0.32	0.1161	no	no
Consistency $\rightarrow$ alignment, IRIS	Spearman, permutation	26	+0.07	0.7472	no	no
Consistency $\rightarrow$ alignment, STORM	Spearman, permutation	26	+0.13	0.5153	no	no
Latency $\rightarrow$ alignment, DIAMOND (PSNR>30, chosen intervention)	Spearman, permutation	13	-0.32	0.2900	no	–
Latency $\rightarrow$ alignment, IRIS (PSNR>30, chosen intervention)	Spearman, permutation	11	-0.58	0.0746	no	–
Alignment $\rightarrow$ published agent return, pooled over architectures (24 cells, games shared)	Spearman, 20,000 permutations of cells (unclustered)	24	+0.20	0.3518	no	–
Open-loop ranking quality $\rightarrow$ alignment / PSNR, pooled (45 cells in 15 games)	Spearman; game-clustered bootstrap and block permutation	45	+0.30 [-0.09, +0.62] / +0.37 [-0.03, +0.66]	0.151 / 0.087	no / no	–

continued on the next page

claim	test	n	estimate	p	gate	survives
Open-loop ranking quality → PSNR, context-matched DIAMOND and IRIS cells (30 cells in 15 games; post hoc)	Spearman; game-clustered bootstrap and block permutation	30	+0.43 [-0.09, +0.75]	0.066	– (post hoc)	–
Closed loop: model gain over the random policy, per game (bootstrap interval of the difference entirely above zero)	episode bootstrap; cellwise counts, uncorrected	12 games	5 / 3 / 4 of 8 oracle-informative games	–	–	–
Correspondence (third arm): games whose mean per-episode difference $A_b - A_{c\to b}$ is positive at delay 20, DIAMOND / IRIS / STORM	one-sided binomial over games; per-cell sign-flip test, Holm over 26 games	26	25 of 26 / 24 of 26 / 24 of 26	$< 10^{-5}$ / $< 10^{-5}$ / $< 10^{-5}$	–	17 cells / 14 cells / 9 cells
STORM training seed 1 vs seed 2, per-game delay-20 alignment	Spearman, permutation; game bootstrap	26	+0.74 [+0.44, +0.90]	$< 10^{-4}$	yes	–
Kendall's W of the three architectures' game rankings by alignment (summarises the three pairwise correlations)	descriptive	26	0.81	–	–	–
Human audit: annotators 1 and 2, presence / improvement (Cohen's κ)	game-clustered bootstrap (21 games)	49 / 38 panels	0.81 [0.46, 1.00] / 0.63 [0.32, 0.87]	–	–	–
Human audit: annotators 1 and 3, presence / improvement (Cohen's κ)	game-clustered bootstrap (21 games)	49 / 39 panels	0.65 [0.28, 0.90] / 0.75 [0.51, 0.93]	–	–	–
Human audit: annotators 2 and 3, presence / improvement (Cohen's κ)	game-clustered bootstrap (21 games)	58 / 51 panels	0.56 [0.32, 0.77] / 0.46 [0.24, 0.71]	–	–	–
Human audit: locator displacement verdict vs annotator 1 / annotator 2 / annotator 3 / consensus (κ)	game-clustered bootstrap	20 / 22 / 23 / 17 panels	0.44 / 0.20 / 0.14 / 0.31	–	–	–
Alignment → object not found / no improvement on zero displacement (template labels, AUC; n = 60 / 55)	game-clustered bootstrap, two-sided game-level permutation	60 / 55	0.56 / 0.88	0.5481 / 0.0004	–	–
Off-support fraction → object not found / no improvement on zero displacement (template labels, AUC)	game-clustered bootstrap, two-sided game-level permutation	60 / 55	0.45 / 0.64	0.6323 / 0.2376	no	–
Flagship, Alien vs Pong per episode (DIAMOND, fixed intervention, delay 20, pairs above the guard)	Cliff's δ (Pong minus Alien)	348 vs 322	$+0.92$	–	–	–
Pre-registered predictions, DIAMOND	confirmed / refuted (thresholds fixed in advance)	11 scored	8 / 3	–	–	mixed
Pre-registered predictions, IRIS	confirmed / refuted (thresholds fixed in advance)	11 scored	9 / 2	–	–	mixed

Table 21: Ablations and sensitivity. Top: intervention duration (DIAMOND, 26 games × 7 seeds × 60 episodes at each) and intervention design. Bottom: the flagship conclusion swept over delay, guard floor, statistic and episode count (96 settings) for both released models under both interventions; evaluation-seed stability of the transfer; and the rollout horizon.

factor	levels	measurable games	rank inversions
intervention held for 1 step	fixed offset	8/26	1
intervention held for 4 steps	fixed offset	10/26	16
intervention held for 8 steps	fixed offset	12/26	29
intervention held for 16 steps	fixed offset	15/26	46
intervention design	fixed offset → emulator-chosen offset + warm-up	$12/26 \to 26/26$	–
sensitivity of the flagship	96 settings, DIAMOND, chosen intervention	causal half 96/96	consistency inversion 88/96

continued on the next page

factor	levels	measurable games	rank inversions
sensitivity of the flagship	96 settings, DIAMOND, fixed intervention	causal half 96/96	consistency inversion 96/96
sensitivity of the flagship	96 settings, IRIS, chosen intervention	causal half 96/96	consistency inversion 0/96
sensitivity of the flagship	96 settings, IRIS, fixed intervention	causal half 92/96	consistency inversion 0/96
evaluation seeds	7 (42, 7, 128, 256, 512, 1024, 2024)	DIAMOND vs IRIS ρ per seed +0.58 to +0.77	mean 0.68 ± 0.06
evaluation seeds	same seeds	DIAMOND vs STORM 0.49, IRIS vs STORM 0.67 (mean per-seed ρ)	–
rollout horizon	$24 \rightarrow 48$ steps, DIAMOND, chosen intervention	Alien 0.00 (d20) $\rightarrow$ 0.00 (d44); Pong 0.85 (d20; its later delays have fewer than 30 guarded pairs per seed and are not scored)	d$\leq$20 identical per episode (common random numbers)

Guard. Alignment, magnitude and error divide by the true response $\|\Delta t_d\|$. For each evaluation seed of each (game, architecture, intervention) file the guard floor is 0.25 times the median of $\|\Delta t_d\|$ over that seed's 60 pairs and all delays (seven floors per file; analyses that pool seeds use their mean); a (pair, delay) cell whose true response lies below the floor is excluded from the three ratios and counted. Exclusion counts are stored beside every aggregate (for example 39 of 660 cells for seed 42 of DIAMOND Alien at horizon 48, 21 to 39 across seeds). Consistency and PSNR are never guarded, since neither divides by the true response. Excluded cells are not discarded: on them we report the below-guard response $\|\Delta p_d\|$; a true response below the floor is small, not zero, and only exact zeros count as inert in Appendix H.

Aggregation. Within a file, the reported statistic at each delay is the median over the 60 episode pairs of a seed, with a 95% bootstrap interval from 2000 resamples of the pairs; across the 7 evaluation seeds we report the mean of the per-seed medians, and where a $\pm$ appears it is the standard deviation of the seven per-seed medians. The seven evaluation seeds vary the base action sequence only (emulator start states, warm-up actions and the model's sampling stream are functions of the episode index and are shared across seeds, Appendix C); the released checkpoints fix the training seed; training variability is measured by the STORM second-seed replicate on all 26 games (Section 4.2) and by a second training seed of DIAMOND and IRIS on the flagship pair (Appendix D).

Correlations and tests. Game-level correlations are Spearman's ρ over $n = 26$ games (one value per game, the delay-20 mean over seeds). Their p-values come from Monte Carlo permutation tests of the game labels: $20{,}000$ permutations for the transfer, within-architecture, latency, seed-replication and published-return tests, $5{,}000$ for the off-support, horizon, action-dependence and decoupling correlations, and $2{,}000$ game-clustered block permutations for the planning correlations; a p-value smaller than one permutation is written as a bound ($p < 10^{-4}$ or $p < 2 \times 10^{-4}$), never as 0. Spearman correlations use average ranks for ties, and a constant score is assigned zero correlation; the per-context planning correlations, whose emulator returns are small integers with frequent ties, are computed the same way from the stored candidate returns. Confidence intervals for ρ are bootstrap percentile intervals over games. Holm families are: for each of alignment, consistency and PSNR, the three pairwise transfer correlations of Section 4.2 (three tests each); the six within-architecture accuracy and consistency correlations (one family, corrected by hand: the smallest p, 0.017, times six is 0.10, so none survives); the pre-registered predictions are scored individually against thresholds fixed before the runs. Cliff's δ compares per-episode alignments of two games at one delay.

Power. Two numbers are reported. The pre-specified effect-size gate at $n = 26$ is $|\rho| = 0.41$; the minimum detectable effect at 80% power (two-sided $\alpha = 0.05$, Fisher z) is $|\rho| = 0.53$, and at $n = 8$ it is 0.85. The first number is the pre-specified effect-size gate of `ccaus/validate.py` (the *gate* column of Table 20): a decision rule fixed in advance, which refuses to mark a correlation as supported when $|\rho|$ lies below it, and not a statement of power; the second number is the power statement, and a null result is worded as "not detected" with that minimum detectable effect stated, never as absence. The planning gate of 0.41 is the 80% power minimum detectable effect at $n = 45$ (Fisher z), numerically equal to the $n = 26$ gate by coincidence; it was fixed for 45 independent

cells before the experiment and is kept as a decision rule although the cells sit in 15 game clusters, so the gate column of Table 20 does not apply one rule to every row. Failure to pass the gate is not read as evidence of absence.

Dependence. Observations pooled over (game, architecture) share games, so the taxonomy counts are descriptive and every suite-level correlation is computed within one architecture. Three analyses pool cells across architectures: the open-loop planning correlations (45 cells in 15 games; game-clustered bootstrap intervals and a game-clustered block permutation test, in which a random bijection of the games is drawn and every cell takes the outcome of the same-position cell of the game it is mapped to), the AUC comparison of Table 14 (60 and 55 cells over 26 games; the same game-clustered bootstrap, and a stratified block permutation in which games are grouped by the set of architectures they contribute, a random bijection of the games is drawn within each group, and every cell takes the label of the same architecture's cell in the game it is mapped to, so the permuted labels are a permutation of the observed labels within every architecture), and the published-return analysis (24 cells; an unclustered permutation of cells, reported as unpowered and descriptive). Table 14's two labels are corrected for multiplicity as separate seven-row families. The closed-loop counts of games in which a model's gain interval lies above zero are uncorrected descriptive counts. The 96 sensitivity settings of Section 4.1 re-analyse the same episodes under different analysis choices (delay in $\{8, 12, 16, 20\}$, guard fraction in $\{0.10, 0.25, 0.50, 1.00\}$ of the median true response, per-seed median or mean, and the first 15, 30 or 60 episodes of each seed: $4 \times 4 \times 2 \times 3 = 96$); they are a robustness sweep, not independent replications, and are reported as a count of settings in which the ordering holds.

Taxonomy threshold sensitivity. Counts of tracked / aligned-then-forgotten / never-aligned / partial cells when the tracked threshold T is halved or doubled. DIAMOND: $T/2$ gives 11/6/7/2; T gives 11/6/7/2; $2T$ gives 5/3/13/5. IRIS: $T/2$ gives 7/11/6/2; T gives 7/9/8/2; $2T$ gives 6/8/9/3. STORM: $T/2$ gives 6/11/7/2; T gives 6/5/13/2; $2T$ gives 5/2/16/3. The transfer correlations of Section 4.2 are computed on the continuous alignment values and do not depend on T.

Pre-registration. The intervention search and the latency predictions were written to files before the corresponding model runs (`emulator_intervention_search.json`, whose record is its modification time, Appendix C; `preregistered_latency_prediction.LOCKED_20260904T1619.json`, dated in its name); the scorecard (`prediction_scorecard.json`) scores those predictions after the runs, and the thresholds it uses (blind below 0.05; tracked at or above 0.40 for DIAMOND and 0.09 for IRIS; PSNR above 30 dB) are those of `scripts/score_predictions.py`, written with the predictions from the fixed-intervention Pong values; the dated registration file holds the predictions, not the thresholds. The scorecard's IRIS threshold was fixed before the registration correction and differs from the taxonomy threshold of 0.108 derived after it; no scored IRIS prediction has an alignment between the two values, so rescoring at 0.108 changes no verdict. The horizon-48 ablation and the pre-sampling probe were analysed under rules written into the results file before their outputs existed.

Q THE PSNR BOUND ON THE RESPONSE, PER GAME

Corollary 1 bounds the root normalised response error $\sqrt{\varepsilon}$ by $2e/\|\Delta t\|$ when both arms are predicted with error norm e. Table 22 computes that ratio from the measured factual-arm PSNR for every architecture and game, assuming the intervened arm is predicted as well as the factual one, and, wherever the intervened arm's error was stored (every IRIS and STORM cell and the two DIAMOND flagship cells), the exact ratio $(\|p - t\| + \|p' - t'\|)/\|\Delta t\|$; the delay-20 values carry three decimals so that threshold crossings can be checked (the equal-error ratio exceeds one in 23, 24 and 25 of 26 games; the exact ratio in 25 of 26 IRIS and 26 of 26 STORM games; the two differ by up to 0.87). A ratio above one means the PSNR level permits a response error larger than the true response, so the factual accuracy does not certify the response; the same table gives the measured response triple at delay 20 under the emulator-chosen interventions. The triple in Table 22 is a median over the pooled episode pairs above the guard (7 seeds $\times$ 60 episodes); the equal-error ratio is a ratio of medians, $2e(\text{median PSNR})/\text{median}\,\|\Delta t\|$, the PSNR median over all pairs and the $\|\Delta t\|$ median over the guarded pairs, and the exact two-arm ratio is the median over those pairs of the

per-pair ratio $(\|p-t\|+\|p'-t'\|)/\|\Delta t\|$, whereas Table 23 reports the mean over seeds of per-seed medians, so the two tables' alignments are different estimators of the same cell and need not agree.

Table 22: Per architecture and game under the emulator-chosen interventions: the equal-error bound ratio $2e/\|\Delta t\|$ at delays 1 and 20 (three decimals at delay 20 so that threshold crossings can be checked), the exact two-arm ratio $(\|p-t\|+\|p'-t'\|)/\|\Delta t\|$ at delay 20 where the intervened arm's error was stored (every IRIS and STORM cell; the two DIAMOND flagship cells), and the measured response triple at delay 20. A ratio above one means that the factual error does not certify $\varepsilon < 1$; – marks a ratio not computed (the exact ratio where the intervened arm's error was not stored; delay 1 where no pair passes the guard).

model	game	$2e/\|\Delta t\|$ d=1	d=20	exact d=20	A	R	ε
DIAMOND	Alien	3.12	2.792	2.815	+0.00	1.11	2.18
DIAMOND	Amidar	3.63	3.015	–	+0.00	0.03	1.00
DIAMOND	Assault	2.46	2.431	–	+0.12	1.11	1.80
DIAMOND	Asterix	1.49	1.097	–	+0.83	1.10	0.52
DIAMOND	BankHeist	132.75	3.763	–	+0.00	0.06	1.00
DIAMOND	BattleZone	2.69	1.229	–	+0.63	0.94	0.66
DIAMOND	Boxing	2.41	1.863	–	+0.07	0.96	1.72
DIAMOND	Breakout	0.80	1.014	–	+0.95	1.06	0.34
DIAMOND	ChopperCommand	1.75	1.374	–	+0.41	0.94	1.00
DIAMOND	CrazyClimber	1.56	0.807	–	+0.05	0.81	1.00
DIAMOND	DemonAttack	2.75	2.522	–	+0.63	1.53	1.96
DIAMOND	Freeway	2.49	2.623	–	+0.00	1.01	1.43
DIAMOND	Frostbite	1.34	0.732	–	+0.60	1.00	0.78
DIAMOND	Gopher	1.66	1.014	–	+1.04	1.17	0.26
DIAMOND	Hero	1.45	2.236	–	+0.17	0.99	1.18
DIAMOND	Jamesbond	5.49	2.392	–	+0.08	0.75	1.03
DIAMOND	Kangaroo	–	2.072	–	+0.00	1.03	1.74
DIAMOND	Krull	2.40	3.172	–	+0.21	1.19	1.93
DIAMOND	KungFuMaster	1.88	1.778	–	+0.50	1.02	1.01
DIAMOND	MsPacman	2.76	3.202	–	+0.00	0.03	1.00
DIAMOND	Pong	1.36	1.939	2.053	+0.84	1.28	1.01
DIAMOND	PrivateEye	1.39	0.945	–	+0.74	0.95	0.30
DIAMOND	Qbert	–	1.862	–	+0.11	1.00	1.47
DIAMOND	RoadRunner	1.49	1.633	–	+0.00	0.85	1.01
DIAMOND	Seaquest	0.64	1.630	–	+0.93	1.12	0.45
DIAMOND	UpNDown	3.21	1.097	–	+0.00	0.90	1.58
IRIS	Alien	4.06	2.916	2.931	+0.00	1.15	2.31
IRIS	Amidar	4.37	3.155	3.283	+0.00	0.39	1.00
IRIS	Assault	3.74	3.345	2.991	+0.03	1.29	2.65
IRIS	Asterix	2.82	2.187	2.144	+0.04	0.89	1.64
IRIS	BankHeist	100.52	2.296	2.669	+0.00	1.04	1.80
IRIS	BattleZone	3.59	1.528	1.535	+0.37	0.85	1.00
IRIS	Boxing	2.91	1.881	1.899	+0.05	0.94	1.75
IRIS	Breakout	1.17	1.222	1.510	+0.72	0.99	0.71
IRIS	ChopperCommand	2.68	1.688	1.705	+0.18	0.90	1.31
IRIS	CrazyClimber	3.92	0.909	1.386	+0.17	0.76	0.98
IRIS	DemonAttack	3.09	2.363	2.274	+0.06	1.03	1.70
IRIS	Freeway	4.80	4.532	4.722	+0.16	1.02	1.42
IRIS	Frostbite	5.64	1.753	2.268	+0.01	0.99	1.74
IRIS	Gopher	0.89	0.598	0.723	+0.88	0.99	0.19
IRIS	Hero	2.48	1.950	1.925	+0.14	0.97	1.30
IRIS	Jamesbond	6.75	2.602	2.337	+0.00	0.75	1.15
IRIS	Kangaroo	–	2.272	2.308	+0.02	0.91	1.45
IRIS	Krull	3.25	2.515	2.816	+0.00	0.98	1.94
IRIS	KungFuMaster	2.64	2.211	2.239	+0.38	0.95	1.10

continued on the next page

model	game	$2e/\|\Delta t\|$ d=1	d=20	exact d=20	A	R	ε
IRIS	MsPacman	3.43	3.742	3.726	+0.00	0.00	1.00
IRIS	Pong	2.08	2.072	2.419	+0.24	1.11	1.61
IRIS	PrivateEye	3.42	1.576	1.359	+0.32	0.92	1.01
IRIS	Qbert	–	1.694	1.848	+0.27	0.92	1.14
IRIS	RoadRunner	2.22	1.608	1.709	+0.00	0.76	1.00
IRIS	Seaquest	1.23	2.105	1.850	+0.26	1.01	1.30
IRIS	UpNDown	4.28	1.257	2.125	+0.01	0.77	1.31
STORM	Alien	2.51	2.458	2.418	+0.00	0.67	1.40
STORM	Amidar	4.11	2.737	2.642	+0.00	0.00	1.00
STORM	Assault	2.06	2.037	1.972	+0.01	0.75	1.54
STORM	Asterix	1.67	1.765	1.808	+0.00	0.28	1.07
STORM	BankHeist	30.09	2.312	2.335	-0.00	0.61	1.18
STORM	BattleZone	2.07	1.356	1.355	+0.41	0.86	0.87
STORM	Boxing	3.26	1.887	1.892	+0.02	0.80	1.56
STORM	Breakout	0.76	1.187	1.415	+0.44	0.96	0.97
STORM	ChopperCommand	3.01	1.877	1.852	+0.02	0.78	1.29
STORM	CrazyClimber	2.46	1.105	1.786	+0.00	0.76	1.01
STORM	DemonAttack	2.12	2.041	2.027	+0.12	0.81	1.30
STORM	Freeway	1.88	1.738	1.813	+0.33	0.91	1.00
STORM	Frostbite	3.83	2.297	2.591	+0.00	0.86	1.68
STORM	Gopher	1.17	1.417	1.322	+0.53	1.00	0.94
STORM	Hero	2.40	1.640	1.640	+0.03	0.85	1.09
STORM	Jamesbond	4.59	1.720	1.912	+0.04	0.86	1.37
STORM	Kangaroo	–	1.215	1.278	+0.37	0.92	0.97
STORM	Krull	2.47	2.092	2.319	-0.00	0.57	1.34
STORM	KungFuMaster	2.28	2.297	2.274	+0.03	0.99	1.58
STORM	MsPacman	2.95	2.782	2.770	-0.00	0.00	1.00
STORM	Pong	1.68	1.735	1.949	+0.39	1.02	1.27
STORM	PrivateEye	2.24	1.286	1.256	+0.40	0.93	0.97
STORM	Qbert	–	0.982	1.261	+0.35	0.89	0.92
STORM	RoadRunner	2.45	1.820	1.870	+0.00	0.76	1.43
STORM	Seaquest	1.19	1.269	1.797	+0.05	0.66	1.14
STORM	UpNDown	5.99	1.585	2.066	+0.00	0.29	1.02

R ALL GAMES UNDER THE EMULATOR-CHOSEN INTERVENTIONS

Table 23: All 26 games under the emulator-chosen interventions at delay 20: PSNR, consistency and A are, for each cell, the mean over the 7 evaluation seeds of the per-seed median over 60 episode pairs (the aggregation of Appendix P); R and ε are pooled medians over guarded pairs (the estimator of Table 22). True divergence is the emulator's, identical for DIAMOND and IRIS (STORM's replays use its own frame pooling, Table 12).

		DIAMOND					IRIS				
game	true div.	PSNR	consist.	A	R	ε	PSNR	consist.	A	R	ε
Gopher	13.50	29.9	0.762	1.033	1.17	0.26	34.3	0.760	0.878	0.99	0.19
Breakout	4.70	38.0	0.842	0.928	1.06	0.34	36.2	0.836	0.681	0.99	0.71
Seaquest	3.05	30.2	0.798	0.924	1.12	0.45	28.2	0.813	0.235	1.01	1.30
Pong	3.13	35.2	0.843	0.848	1.28	1.01	34.4	0.867	0.217	1.11	1.61
Asterix	9.11	32.9	0.403	0.831	1.10	0.52	27.0	0.446	0.044	0.89	1.64
PrivateEye	13.96	30.6	0.739	0.736	0.95	0.30	26.0	0.730	0.321	0.92	1.01
DemonAttack	5.52	29.9	0.524	0.637	1.53	1.96	30.5	0.621	0.059	1.03	1.70
BattleZone	8.95	32.0	0.851	0.625	0.94	0.66	30.2	0.843	0.330	0.85	1.00
Frostbite	4.69	37.0	0.890	0.616	1.00	0.78	29.5	0.854	0.022	0.99	1.74
KungFuMaster	7.47	29.5	0.767	0.518	1.02	1.01	27.6	0.760	0.371	0.95	1.10

continued on the next page

		DIAMOND					IRIS				
game	true div.	PSNR	consist.	A	R	ε	PSNR	consist.	A	R	ε
ChopperCommand	8.20	31.2	0.861	0.415	0.94	1.00	29.3	0.857	0.189	0.90	1.31
Krull	5.30	28.5	0.764	0.210	1.19	1.93	30.4	0.782	0.002	0.98	1.94
Hero	7.03	28.9	0.776	0.180	0.99	1.18	30.1	0.729	0.137	0.97	1.30
Qbert	9.79	27.5	0.813	0.130	1.00	1.47	28.3	0.792	0.280	0.92	1.14
Assault	8.45	26.7	0.263	0.125	1.11	1.80	24.2	0.331	0.032	1.29	2.65
Jamesbond	3.20	29.6	0.695	0.105	0.75	1.03	28.9	0.708	0.001	0.75	1.15
Boxing	16.91	22.9	0.545	0.074	0.96	1.72	22.8	0.548	0.060	0.94	1.75
CrazyClimber	2.50	35.9	0.798	0.066	0.81	1.00	35.3	0.766	0.242	0.76	0.98
Kangaroo	2.11	31.9	0.934	0.029	1.03	1.74	31.3	0.868	0.035	0.91	1.45
RoadRunner	11.59	26.5	0.835	0.009	0.85	1.01	26.7	0.869	0.005	0.76	1.00
UpNDown	3.95	28.3	0.731	0.005	0.90	1.58	27.1	0.757	0.013	0.77	1.31
Alien	4.86	30.2	0.853	0.003	1.11	2.18	29.8	0.840	0.001	1.15	2.31
Freeway	2.06	38.1	0.788	0.000	1.01	1.43	33.3	0.792	0.168	1.02	1.42
Amidar	4.21	30.8	0.982	0.000	0.03	1.00	30.2	0.798	0.000	0.39	1.00
MsPacman	5.37	28.3	0.852	0.000	0.03	1.00	26.9	0.804	0.000	0.00	1.00
BankHeist	27.19	11.4	0.947	0.000	0.06	1.00	15.3	0.506	0.000	1.04	1.80

S ALL 78 CELLS: MODE AND CAUSAL HORIZON

Table 24: All 78 (architecture, game) cells under emulator-chosen interventions: mode, peak causal alignment over delays 1 to 8, alignment at delay 20, and causal horizon (the first delay after delay 1 at which the alignment is below half its delay-1-to-8 peak, whether before or after the peak, so a cell whose alignment is still rising at delay 2 can carry a horizon of 2; > 20 if it is never below; undefined for never-aligned and partial cells). Classes as in Table 7: *partial* is a cell with an unmeasurable delay or with a peak above T and a delay-20 alignment between half the peak and T.

	game	mode	peak align.	align. d=20	ε d=20	horizon
DIAMOND	Krull	aligned-then-forgotten	0.88	0.210	1.93	16
DIAMOND	Hero	aligned-then-forgotten	0.88	0.180	1.18	16
DIAMOND	Assault	aligned-then-forgotten	0.61	0.125	1.80	12
DIAMOND	Jamesbond	aligned-then-forgotten	1.06	0.105	1.03	8
DIAMOND	Boxing	aligned-then-forgotten	0.65	0.074	1.72	12
DIAMOND	CrazyClimber	aligned-then-forgotten	0.51	0.066	1.00	16
DIAMOND	RoadRunner	never-aligned	0.04	0.009	1.01	–
DIAMOND	UpNDown	never-aligned	0.06	0.005	1.58	–
DIAMOND	Alien	never-aligned	0.18	0.003	2.18	–
DIAMOND	Freeway	never-aligned	0.17	0.000	1.43	–
DIAMOND	Amidar	never-aligned	0.00	0.000	1.00	–
DIAMOND	MsPacman	never-aligned	0.00	0.000	1.00	–
DIAMOND	BankHeist	never-aligned	0.01	0.000	1.00	–
DIAMOND	Qbert	partial	–	0.130	1.47	–
DIAMOND	Kangaroo	partial	–	0.029	1.74	–
DIAMOND	Gopher	tracked	1.06	1.033	0.26	> 20
DIAMOND	Breakout	tracked	1.00	0.928	0.34	> 20
DIAMOND	Seaquest	tracked	1.05	0.924	0.45	> 20
DIAMOND	Pong	tracked	0.83	0.848	1.01	> 20
DIAMOND	Asterix	tracked	1.11	0.831	0.52	> 20
DIAMOND	PrivateEye	tracked	0.74	0.736	0.30	> 20
DIAMOND	DemonAttack	tracked	1.17	0.637	1.96	> 20
DIAMOND	BattleZone	tracked	0.68	0.625	0.66	> 20
DIAMOND	Frostbite	tracked	0.90	0.616	0.78	> 20
DIAMOND	KungFuMaster	tracked	0.95	0.518	1.01	> 20

continued on the next page

	game	mode	peak align.	align. d=20	ε d=20	horizon
DIAMOND	ChopperCommand	tracked	0.68	0.415	1.00	> 20
IRIS	Seaquest	aligned-then-forgotten	0.76	0.235	1.30	16
IRIS	Pong	aligned-then-forgotten	0.53	0.217	1.61	16
IRIS	Freeway	aligned-then-forgotten	0.60	0.168	1.42	12
IRIS	Hero	aligned-then-forgotten	0.34	0.137	1.30	16
IRIS	Boxing	aligned-then-forgotten	0.34	0.060	1.75	12
IRIS	DemonAttack	aligned-then-forgotten	0.26	0.059	1.70	12
IRIS	Asterix	aligned-then-forgotten	0.40	0.044	1.64	12
IRIS	Assault	aligned-then-forgotten	0.19	0.032	2.65	12
IRIS	Frostbite	aligned-then-forgotten	0.24	0.022	1.74	16
IRIS	UpNDown	never-aligned	0.01	0.013	1.31	–
IRIS	RoadRunner	never-aligned	0.02	0.005	1.00	–
IRIS	Krull	never-aligned	0.01	0.002	1.94	–
IRIS	Jamesbond	never-aligned	0.09	0.001	1.15	–
IRIS	Alien	never-aligned	0.07	0.001	2.31	–
IRIS	BankHeist	never-aligned	0.01	0.000	1.80	–
IRIS	Amidar	never-aligned	0.00	0.000	1.00	–
IRIS	MsPacman	never-aligned	0.00	0.000	1.00	–
IRIS	Qbert	partial	–	0.280	1.14	–
IRIS	Kangaroo	partial	–	0.035	1.45	–
IRIS	Gopher	tracked	0.89	0.878	0.19	> 20
IRIS	Breakout	tracked	0.90	0.681	0.71	> 20
IRIS	KungFuMaster	tracked	0.70	0.371	1.10	> 20
IRIS	BattleZone	tracked	0.43	0.330	1.00	> 20
IRIS	PrivateEye	tracked	0.41	0.321	1.01	2
IRIS	CrazyClimber	tracked	0.41	0.242	0.98	2
IRIS	ChopperCommand	tracked	0.33	0.189	1.31	> 20
STORM	DemonAttack	aligned-then-forgotten	0.66	0.117	1.30	8
STORM	Seaquest	aligned-then-forgotten	0.43	0.062	1.14	16
STORM	KungFuMaster	aligned-then-forgotten	0.32	0.042	1.58	12
STORM	Jamesbond	aligned-then-forgotten	0.26	0.039	1.37	12
STORM	Assault	aligned-then-forgotten	0.22	0.007	1.54	8
STORM	Hero	never-aligned	0.17	0.043	1.09	–
STORM	CrazyClimber	never-aligned	0.13	0.021	1.01	–
STORM	ChopperCommand	never-aligned	0.12	0.019	1.29	–
STORM	Boxing	never-aligned	0.12	0.015	1.56	–
STORM	Frostbite	never-aligned	0.04	0.006	1.68	–
STORM	Alien	never-aligned	0.11	0.004	1.40	–
STORM	RoadRunner	never-aligned	0.01	0.003	1.43	–
STORM	Asterix	never-aligned	0.18	0.001	1.07	–
STORM	UpNDown	never-aligned	0.01	0.000	1.02	–
STORM	Amidar	never-aligned	0.00	0.000	1.00	–
STORM	MsPacman	never-aligned	0.00	0.000	1.00	–
STORM	BankHeist	never-aligned	0.00	-0.000	1.18	–
STORM	Krull	never-aligned	0.00	-0.000	1.34	–
STORM	Qbert	partial	–	0.380	0.92	–
STORM	Kangaroo	partial	–	0.378	0.97	–
STORM	Gopher	tracked	0.70	0.524	0.94	> 20
STORM	Breakout	tracked	0.75	0.449	0.97	> 20
STORM	PrivateEye	tracked	0.57	0.410	0.97	2
STORM	BattleZone	tracked	0.64	0.407	0.87	> 20
STORM	Pong	tracked	0.54	0.376	1.27	> 20
STORM	Freeway	tracked	0.60	0.330	1.00	> 20

T PRE-REGISTERED SCORECARD

Table 25: Pre-registered predictions (written from the emulator before any model run) and their outcome. TRACKED is scored only where PSNR > 30.

	game	latency	prediction	align. d=20	PSNR	verdict
DIAMOND	Kangaroo	5	BLIND	0.029	31.9	CONFIRMED
DIAMOND	BankHeist	12	BLIND	0.000	11.4	CONFIRMED
DIAMOND	Gopher	1	TRACKED	1.033	29.9	n/a (PSNR$\leq$30)
DIAMOND	Breakout	1	TRACKED	0.928	38.0	CONFIRMED
DIAMOND	Seaquest	1	TRACKED	0.924	30.2	CONFIRMED
DIAMOND	Pong	1	TRACKED	0.848	35.2	CONFIRMED
DIAMOND	Asterix	1	TRACKED	0.831	32.9	CONFIRMED
DIAMOND	DemonAttack	1	TRACKED	0.637	29.9	n/a (PSNR$\leq$30)
DIAMOND	BattleZone	1	TRACKED	0.625	32.0	CONFIRMED
DIAMOND	KungFuMaster	1	TRACKED	0.518	29.5	n/a (PSNR$\leq$30)
DIAMOND	ChopperCommand	1	TRACKED	0.415	31.2	CONFIRMED
DIAMOND	Krull	1	TRACKED	0.210	28.5	n/a (PSNR$\leq$30)
DIAMOND	Hero	1	TRACKED	0.180	28.9	n/a (PSNR$\leq$30)
DIAMOND	Assault	1	TRACKED	0.125	26.7	n/a (PSNR$\leq$30)
DIAMOND	Boxing	1	TRACKED	0.074	22.9	n/a (PSNR$\leq$30)
DIAMOND	CrazyClimber	1	TRACKED	0.066	35.9	REFUTED
DIAMOND	Alien	1	TRACKED	0.003	30.2	REFUTED
DIAMOND	Freeway	1	TRACKED	0.000	38.1	REFUTED
DIAMOND	MsPacman	1	TRACKED	0.000	28.3	n/a (PSNR$\leq$30)
IRIS	Kangaroo	5	BLIND	0.035	31.3	CONFIRMED
IRIS	BankHeist	12	BLIND	0.000	15.3	CONFIRMED
IRIS	Gopher	1	TRACKED	0.878	34.3	CONFIRMED
IRIS	Breakout	1	TRACKED	0.681	36.2	CONFIRMED
IRIS	KungFuMaster	1	TRACKED	0.371	27.6	n/a (PSNR$\leq$30)
IRIS	BattleZone	1	TRACKED	0.330	30.2	CONFIRMED
IRIS	CrazyClimber	1	TRACKED	0.242	35.3	CONFIRMED
IRIS	Seaquest	1	TRACKED	0.235	28.2	n/a (PSNR$\leq$30)
IRIS	Pong	1	TRACKED	0.217	34.4	CONFIRMED
IRIS	ChopperCommand	1	TRACKED	0.189	29.3	n/a (PSNR$\leq$30)
IRIS	Freeway	1	TRACKED	0.168	33.3	CONFIRMED
IRIS	Hero	1	TRACKED	0.137	30.1	CONFIRMED
IRIS	Boxing	1	TRACKED	0.060	22.8	n/a (PSNR$\leq$30)
IRIS	DemonAttack	1	TRACKED	0.059	30.5	REFUTED
IRIS	Asterix	1	TRACKED	0.044	27.0	n/a (PSNR$\leq$30)
IRIS	Assault	1	TRACKED	0.032	24.2	n/a (PSNR$\leq$30)
IRIS	Krull	1	TRACKED	0.002	30.4	REFUTED
IRIS	Alien	1	TRACKED	0.001	29.8	n/a (PSNR$\leq$30)
IRIS	MsPacman	1	TRACKED	0.000	26.9	n/a (PSNR$\leq$30)

U TESTED INVARIANTS OF THE INSTRUMENT

Each invariant below was violated during development by a fault that produced a plausible number, not a crash; the released mutation suite reproduces faults (2), (3), (4), (6), (7) and one unlisted dataset-gate fault and checks that each is detected; a unit test of the measurement suite covers fault (5); faults (1) and (8) are documented, not reintroduced. (1) A namedtuple observation made a determinism check compare two scalars and report exact truth. (2) Stochastic sampling in both the posterior and the prior of the RSSM turned a null intervention into a 15/255 response. (3) A consistency score using a norm over 12,288 pixels returned e^{-55} for every model. (4) A guard floor derived from an assumed scaling law rejected every cell, so a sweep measured nothing. (5) A mean of a ratio estimator followed a few near-zero denominators. (6) A relative decay rule with a zero baseline reported the strongest result as no separation. (7) A quantile threshold on sparse difference images matched every pixel. (8) A max/min ratio in the reporting layer printed 9.9×10^8. The

one-step registration offset of the IRIS and STORM harnesses is covered by the separate registration test.