DOES CALIBRATED SURPRISE DETECT PHYSICAL VIOLATIONS? AUDITING CONFORMAL MONITORS FOR LATENT WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Prediction error in a latent world model, or *surprise*, can flag an unexpected transition. What makes that alarm evidence of a physical inconsistency? We introduce VoE-Bench, a violation-of-expectation benchmark combining an audit of 29 checkpoints in three environments with focused tests of reference sensitivity, physical selectivity, and decision utility. In TwoRoom, changing the clean (unedited, default-rendering) reference moves alarms on the same 1,000 benignly recoloured episodes from 0.5% to 82.4%, although both nominal audits meet tolerance. Fresh references (9,600 episodes) confirm this sensitivity and reveal discrepancies from a point forecast frozen before scoring. In PushT, an action-aware multilayer perceptron (MLP) detects more action mismatches than its action-free counterpart under separate calibration. At approximately 5% empirical nominal false-positive rate, the original head has no resolved advantage over the action-aware MLP sharing its encoder and inputs in three-frame onset-window detection (-2.7 percentage points; 97.5% interval $[-7.8, +6.3]$). Under constant recolouring, mean teleport ordering above an appearance-only twin is 0.510 when the teleported object is recoloured and 0.878 when the other object is (other-minus-same difference 0.368; 97.5% interval $[0.332, 0.405]$). TwoRoom readouts recover position despite unselective residuals, and later PushT residuals reveal action evidence masked by an appearance transient. Separate planning tests resolve no success gain and expose the cost of mismatched screening references. The audit distinguishes reference sensitivity, appearance-dependent ordering, and temporal masking, identifying what must be checked before interpreting or acting on an alarm.

1 INTRODUCTION

Latent world models predict the consequences of actions in a learned representation space. Joint-embedding predictive architectures (JEPAs) support planning through these predictions without reconstructing every pixel (Sobal et al., 2025; Zhou et al., 2025; Maes et al., 2026). The discrepancy between predicted and observed representations, which we call *surprise*, also supplies a monitoring signal. A sudden displacement or motion inconsistent with a command should be surprising to a model of familiar dynamics. This motivates using prediction error to assess physical consistency (Garrido et al., 2025; Maes et al., 2026) and execution failures (Domberg & Schildbach, 2025; Ward et al., 2026). Yet an alarm alone does not identify which aspect of a transition is unexpected.

A controlled comparison makes this ambiguity concrete. A conformal monitor chooses its threshold from a *clean reference*: nominal episodes with no intervention and default rendering. In TwoRoom, a two-room navigation task, we calibrated the same model and score on two independently collected 600-episode references, produced by successive versions of our data generator. Both thresholds met our held-out audit tolerance at target false-alarm level $\alpha = 0.05$. Nevertheless, they alarmed on 0.5% and 82.4% of the *same* 1,000 benignly recoloured episodes, whose states and actions were unchanged (Section 3.2). The references placed the threshold on opposite sides of most recoloured scores. A subsequent study with a common collection protocol tests how much reference sampling alone can vary this response.

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.

We ask *when calibrated surprise separates a physical inconsistency from unfamiliar appearance, and which controls justify that interpretation*. Three properties require distinct evidence. *Nominal calibration* bounds alarms under exchangeable reference and query episodes, whose joint distribution is unchanged by relabelling their roles, and averages over reference draws (Shafer & Vovk, 2007); an independent audit checks the rate of one realized threshold. *Physical selectivity* asks whether a physical edit scores above a matched benign change; *decision utility* asks whether acting on alarms improves outcomes at an acceptable computational cost. Neither follows from calibration, whose nominal bound also need not hold for recoloured episodes.

We develop VoE-Bench, a *violation-of-expectation* benchmark, to test these distinctions using simulator interventions and appearance-matched controls. An *action mismatch* changes the executed action while retaining the command supplied to the monitor: the inconsistency is relative to that command, even when the motion remains physically lawful. A broad audit covers 29 checkpoints in three environments; focused studies in TwoRoom and PushT, a planar object-pushing task, then examine the reference, predictor, appearance, and temporal statistic. Our contributions are:

1. **A diagnosis of reference sensitivity.** We relate benign-alarm variability to the nominal tail and shifted-score concentration, then test an order-statistic point forecast frozen before scoring 9,600 new reference episodes.
2. **Controlled tests of physical evidence.** The action-aware PushT MLP detects more mismatches than its action-free counterpart under separate calibration. At approximately 5% empirical nominal FPR, the original head has no resolved timely-detection advantage over an equal-input MLP. Crossed PushT edits reveal appearance-dependent teleport ordering; TwoRoom readouts recover position despite unselective residuals; temporal controls expose masked PushT action evidence.
3. **An audit connecting alarms to decisions.** We combine independent nominal audits and matched physical controls with planning comparisons using fixed candidate sets or per-call optimizer budgets. These comparisons quantify decision outcomes and computational costs; the tested planning rules show no resolved success gain.

2 Background and Audit Protocol

A *parent* fixes an initial simulator state and commanded-action sequence. All its edited variants and appearance-matched unedited *twins* share a data role and stay together during resampling. Nominal episodes use default (*clean*) rendering; *benign* variants change appearance only. A *monitor* combines a score with a calibrated alarm threshold.

2.1 Latent world models and surprise

A frozen encoder f maps observation o_t to a d-dimensional latent z_t. A frozen predictor g uses h observations and commanded-action chunks a_t to predict the next latent:

$$z_t = f(o_t), \qquad \widehat{z}_{t+1} = g(z_{t-h+1:t}, a_{t-h+1:t}), \qquad s_{t+1} = \|\widehat{z}_{t+1} - z_{t+1}\|_2^2. \tag{1}$$

Here a_t contains the actions between observations t and $t+1$. Focused episodes have 21 observations, five simulator steps apart, with $h = 3$. Both scores use frames 0–2 as initial context and score $\mathcal{H} = \{3, \dots, 20\}$. *One-step* surprise uses the latest observed latents. *Open-loop* surprise feeds predictions back after frame 2, so s_t measures a $(t-2)$-step prediction error without further observed context (Appendix B).

We evaluate LeWorldModel (LeWM), Planning with Latent Dynamics Models (PLDM), and DINO-WM, which predicts pretrained visual features (Maes et al., 2026; Sobal et al., 2025; Zhou et al., 2025). For LeWM, *upstream* denotes the original recipe that regularizes latent embeddings toward a Gaussian distribution; *innovation* regularizes normalized residuals after predicting each latent coordinate from its previous value (Appendix C). The broad benchmark and TwoRoom reference study use open-loop surprise. The primary predictor and object studies use one-step surprise on innovation PushT to reassess cohort C's apparent head advantage and object–appearance effect (Sections 3.3 and 3.4); upstream PushT provides a secondary comparison.

2.2 EPISODE-LEVEL CALIBRATION AND INDEPENDENT AUDITING

Dependent frames are not separate calibration examples. Each reference episode contributes one maximum, $S_i = \max_{t \in \mathcal{H}} s_{i,t}$, under the same scoring protocol as the query. For n reference maxima and query maximum S, the standard split-conformal outlier rule (Shafer & Vovk, 2007; Bates et al., 2023) is

$$p(S) = \frac{1 + \sum_{i=1}^{n} \mathbf{1}\{S_i \geq S\}}{n+1}, \qquad q_\alpha = S_{(k)}, \qquad k = \lceil (n+1)(1-\alpha) \rceil, \tag{2}$$

where $\mathbf{1}$ is the indicator, $S_{(k)}$ the kth smallest reference maximum, and $q_\alpha = +\infty$ if $k > n$. An alarm occurs when $S > q_\alpha$, equivalently $p(S) \leq \alpha$. This p-value measures atypicality relative to the reference, not the probability of a physical violation. Conditional on the frozen scoring map, exchangeable reference and query scores give $\Pr(S > q_\alpha) \leq \alpha$ (Shafer & Vovk, 2007). The bound covers any crossing within one fixed-horizon episode, averaged over reference draws; it need not hold separately for each realized reference (Vovk, 2012; Bates et al., 2023). It also does not establish validity for an adapting planning loop.

Comparator fitting, reference calibration, nominal auditing, and intervention evaluation use disjoint parents. A threshold is *locked* after calibration. At $\alpha = 0.05$, it meets our audit tolerance when the one-sided 95% Clopper–Pearson upper bound (Clopper & Pearson, 1934) on its alarm rate over 1,000 nominal episodes is at most 0.07; otherwise the audit is *inconclusive* and the threshold is retained. Even independent draws from one continuous score distribution give inconclusive audits about 27% of the time at $n = 600$ (Appendix B). This outcome alone therefore does not establish distribution shift. TwoRoom uses fresh collector rollouts. Every PushT source episode contributed windows during world-model training: its disjoint audit roles describe that exposed finite pool, with binomial intervals as a working approximation, rather than generalization to unseen episodes.

2.3 VOE-BENCH AND MATCHED INTERVENTIONS

VoE-Bench edits simulator states or applied actions and preserves the commanded actions given to the monitor. Physical and no-physical twins share the appearance condition; onset labels are used only for evaluation. The broad audit covers TwoRoom navigation, PushT planar pushing, and Cube robotic manipulation. In PushT, a circular *pusher* moves a T-shaped *block*. Each focused parent receives an intervention start frame, or *onset*, $j \in \{8, \ldots, 14\}$, sampled uniformly and shared across its variants. Four physical conditions (none, pusher teleport, block teleport, 90° action mismatch) are crossed with clean rendering or 180° recolouring of either object, *abrupt* from j or *constant* from frame 0.

Both teleports move the object 75 pixels in a direction accepted by geometric checks without consulting scores. A teleport changes position at frame j; action mismatch can first change the observed position at $j+1$, leaving only two potentially affected frames in $[j, j+2]$. TwoRoom crosses no edit, teleport, and action mismatch with clean, abrupt, and constant agent recolouring. Observability checks record whether an edit survives to an observed frame (Appendix C).

2.4 ENDPOINTS AND CONFIRMATION

Endpoints and cohorts. For first crossing t^* and onset j, an alarm is *premature* if $t^* < j$, *detected* if $j \leq t^* \leq 20$, and *timely* if $j \leq t^* \leq j+2$; no crossing is a miss. For action mismatch, an onset-frame alarm cannot be attributed to the edit. Development cohorts A–C contain 128, 256, and 512 intervention parents per environment. Final cohort D uses 600 reference, 1,000 audit, and 512 intervention parents per environment, without pooling earlier cohorts.

Physical selectivity. A high detection rate can be uninformative when appearance alone triggers the alarm. We therefore compare each physical variant with its appearance-only twin before thresholding. Let $W_i(e, c)$ be parent i's maximum surprise over $[j_i, j_i + 2]$ for physical edit e and appearance c; $e = 0$ denotes its appearance-only twin. Paired ordering is the fraction of N parents whose physical variant scores higher, counting exact ties as one-half:

$$\theta(e, c) = \frac{1}{N} \sum_{i=1}^{N} \left[\mathbf{1}\{W_i(e,c) > W_i(0,c)\} + \tfrac{1}{2}\mathbf{1}\{W_i(e,c) = W_i(0,c)\} \right]. \tag{3}$$

Thus 0.5 denotes balanced ordering, not a detection rate. The object contrast $I_{\text{object}} = \theta_{\text{other}} - \theta_{\text{same}}$ averages over both teleported objects under constant recolouring of the other or same object. *State readouts* decode simulator coordinates from a frozen latent, testing what remains recoverable; per-frame differences and later windows test temporal masking (Appendix F).

Confirmation. After inspecting C, we fixed two primary D contrasts on innovation PushT: original predictor minus an action-aware multilayer perceptron (MLP) in timely detection of clean action mismatch at approximately 5% nominal false-positive rate (FPR), and I_{object}. Each uses 5,000 parent-bootstrap resamples and a 97.5% interval, an approximate Bonferroni adjustment over two contrasts. Common-FPR resamples re-estimate thresholds. A difference is *resolved* when its interval excludes zero; unresolved differences do not establish equivalence (Appendix G). Appendix B.5 lists every departure from the dated protocols and every post hoc correction.

2.5 Controlled comparisons

Table 1 identifies what each comparison holds fixed.

Table 1: Controlled comparisons; rows describe separate experiments. Temporal windows are onset-informed diagnostics. Appendix I gives each finding's scope.

Comparison	Held fixed	Varied
Reference	Model, score, and query episodes	Clean calibration reference
Predictor	Encoder; three latents and three action chunks	Head at a matched empirical false-alarm rate
Object–appearance	State path and commands within each physical condition	Which object is recoloured
Temporal	Paired physical/benign score sequences	Onset-containing versus later window
Decisions	Paired episode specifications; per-call search budget	Replanning rule; selection rule on fixed candidates

Reference sensitivity. An order-statistic forecast predicts the distribution of recolouring alarm rates across reference draws from estimated nominal and recoloured score distributions. It is frozen before scoring 9,600 new reference and 1,000 audit parents. References form 64 sets of 150, 16 of 600, and four of 2,400, disjoint within each size and nested across sizes (Appendix D).

Predictor contribution. Ridge regression and two-hidden-layer MLPs predict latent changes from the matched inputs in Table 1. Separately fitted action-free versions omit commands. Fitting uses 2,000 nominal parents and selection uses 400 others. MLP predictions are averaged across three initialization seeds before scoring. These are *equal-input* comparisons, not equal training exposure: the original PushT head also saw evaluation-source windows. We compare calibrated pipelines and detection at approximately equal empirical nominal FPR. The latter thresholds are estimated from audit records and have no independent conformal audit (Appendix E).

Decision utility. Replanning, subgoal screening, and the deployment-start audit use PushT with the upstream LeWM checkpoint; the reachability diagnostic uses TwoRoom. Surprise-triggered replanning reruns a cross-entropy-method (CEM) action-sequence optimizer when a frame-pooled per-step rule at $\alpha = 0.1$ alarms or the plan ends. Its control replans every five steps; both are warm-started from the previous plan wherever an unexecuted suffix remains. Subgoal screening rejects intermediate goals using distance to nearby reference latents or the goal error remaining after CEM search, calibrated against demonstrations or the proposer's outputs. Reachability labels require an executed plan or a geometric lower bound on travel time. We measure success and computation separately; these adaptive loops carry no episode-level guarantee (Appendix H).

2.6 Related work and the scope of this audit

Violation-of-expectation studies evaluate learned physical expectations using controlled possible and impossible events (Riochet et al., 2020; Piloto et al., 2022; Garrido et al., 2025; Bordes et al., 2025); LeWM already contrasts teleportation and recolouring (Maes et al., 2026). World-model failure monitors use predictive discrepancies or uncertainty (Domberg & Schildbach, 2025; Ward et al., 2026). CheckVLA verifies execution with a separately trained, frozen action-conditioned

world model and a calibrated intervention trigger (Liu et al., 2026); Foresight learns a failure detector from task outcomes and calibrates trajectory-level thresholds (Zhang et al., 2026). Our contribution is the controlled diagnosis of how the reference, prediction head, edited object, and temporal statistic affect a residual's physical interpretation.

We use standard split-conformal calibration (Shafer & Vovk, 2007; Angelopoulos & Bates, 2022); weighted and nonexchangeable variants require additional assumptions and are not implemented (Tibshirani et al., 2020; Barber et al., 2023). Plan-consistency diagnostics (Ruan et al., 2026) and generated-video benchmarks (Lin et al., 2026; Yang et al., 2026) address related questions through different outputs and labels. Appendix A details these distinctions; the contribution here is empirical evidence and an audit protocol, rather than a new conformal guarantee.

3 Results and Analysis

The broad benchmark establishes the variation to be explained; the focused controls in Table 1 then examine reference sensitivity, physical selectivity, and decision utility. Unless stated otherwise, focused PushT results use the innovation checkpoint and cohort D. Primary contrasts have 97.5% intervals; secondary intervals are 95% unless specified.

3.1 Nominal control and detection rates do not establish selectivity

At target $\alpha = 0.05$, historical nominal alarm rates span 0.025–0.060 across 29 checkpoints. Twenty-five meet the audit tolerance: a one-sided 95% Clopper–Pearson upper bound no greater than 0.07; four are inconclusive. Fresh references need not reproduce that verdict. In D, PushT's one-step residual and MLP comparator meet tolerance at 0.040 and 0.046 (upper bounds $0.052/0.058$), whereas the open-loop residual is inconclusive at 0.067 (upper 0.081). Both TwoRoom predictors are inconclusive at 0.057 (upper 0.071). Thresholds remain fixed after auditing (Appendix J).

Physical-edit detection also varies substantially (Table 2). Upstream TwoRoom LeWM and PLDM detect almost all action mismatches; PushT LeWM's rate is 0.03 upstream and 0.48 for the innovation checkpoint. Cube LeWM combines high physical-edit detection with almost universal colour alarms. Conversely, TwoRoom's low colour rate does not establish robust selectivity: the upstream family includes the reference draw underlying the near-zero recolouring rate examined next. The focused studies therefore hold queries fixed or pair physical edits with appearance-matched benign twins.

Table 2: Historical open-loop benchmark at $\alpha = 0.05$, showing every family–environment group. Entries are first-alarm rates at or after onset; premature alarms are excluded. Parentheses count checkpoints; $\pm$ is their sample SD. One collapsed TwoRoom innovation checkpoint is excluded from these detection means but retained in the nominal audits. Colour rates average benchmark severities and depend on the fitted reference. Full results and observability checks are in Appendix I.

Environment	Family (checkpoints)	Teleport	Action mismatch	Colour control
TwoRoom	LeWM upstream (3)	0.81±0.07	0.99±0.01	0.01±0.00
TwoRoom	LeWM innovation (2)	0.40±0.41	0.54±0.63	0.37±0.44
TwoRoom	PLDM (3)	0.76±0.01	0.99±0.01	0.17±0.26
TwoRoom	DINO-WM (3)	0.38±0.05	0.69±0.08	0.12±0.01
PushT	LeWM upstream (1)	0.09	0.03	0.00
PushT	LeWM innovation (1)	0.35	0.48	0.12
PushT	PLDM (3)	0.11±0.03	0.31±0.03	0.01±0.00
PushT	DINO-WM (3)	0.32±0.02	0.15±0.03	0.02±0.00
Cube	LeWM upstream (1)	0.98	0.87	0.99
Cube	LeWM innovation (1)	0.96	0.88	1.00
Cube	PLDM (3)	0.63±0.04	0.42±0.01	0.03±0.01
Cube	DINO-WM (4)	0.17±0.02	0.96±0.02	0.04±0.02

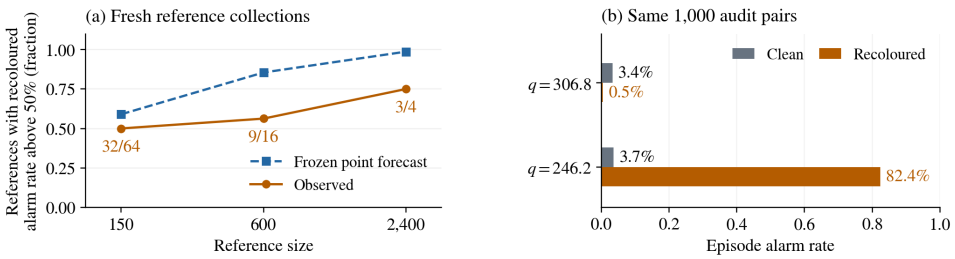


Figure 1: Reference sensitivity: TwoRoom open-loop residual. (a) Fresh-reference counts versus the frozen point forecast; sets are nested across sizes and share an audit. (b) Thresholds applied to the *same* 1,000 clean/recoloured audit pairs. Crossed nominal rates differ from each threshold's original audit. Appendix K gives query collections and the forecast's inferential scope.

3.2 Fresh references reveal unstable benign-alarm rates

Changing only the threshold changes the benign-alarm rate. For the same TwoRoom open-loop detector and 1,000 benign-recoloured queries, two 600-episode clean references yield alarm rates 0.005 and 0.824. Their thresholds are 306.8 and 246.2 in residual units. Re-scoring the stored episodes reproduces their cached scores exactly; applying both thresholds to each query collection preserves the contrast (Appendix K). Both original nominal audits meet tolerance, at 0.040 and 0.037. Thus a satisfactory nominal audit does not determine the response to this benign shift.

Why nominally similar thresholds can behave differently under recolouring. Let $F(x) = \Pr_P(S \le x)$ be the cumulative distribution function (CDF) of nominal scores and $G(x) = \Pr_Q(S \le x)$ that of recoloured scores. The population recolouring-alarm rate at a threshold q_α is $A_Q(q_\alpha)$. For independent, identically distributed reference scores with continuous F, the standard order-statistic law gives

$$U = F(q_\alpha) \sim \mathrm{Beta}(k, n+1-k), \qquad A_Q(q_\alpha) = 1 - G\big(F^{-1}(U)\big)\,, \tag{4}$$

where k is the rank in Eq. equation 2 and F^{-1} is the nominal quantile function. This separates variation in the nominal quantile rank U from its effect on a shifted population. A sparse nominal region makes F^{-1} sensitive to small rank changes; concentrated recoloured scores make G change rapidly over the resulting thresholds. Larger references concentrate U, but the shifted response also depends on these two distributions (Appendix D).

Fresh collections test both the sensitivity and its forecast. The two historical collections used different generator versions. We therefore hold the collector configuration fixed in the separate 9,600-parent study and apply every reference to 1,000 shared audit parents (Section 2.5). At $n = 600$, nine references alarm on more than half of the recoloured queries and seven on fewer than 5%. At $n = 2{,}400$, three exceed one-half and the fourth alarms on only 0.03. All thresholds act on the same query scores (Figure 1).

The measured distributions exhibit the configuration in Eq. equation 4. Only 1.5–2.3% of nominal scores lie between 100 and 200, while mass at or above 200 is close to the 5% target: 4.5% in the earlier reservoir and 4.6–6.0% across new blocks. Small tail-mass changes can move the threshold across the concentrated recoloured scores (roughly 234–286 between their 5th and 95th percentiles). A forecast using estimated CDFs, frozen before any new parent was scored, predicts probabilities $0.855/0.118$ of alarm rates above one-half/below 5% at $n = 600$, compared with observed fractions $9/16$ and $7/16$; its median threshold is 157 versus 238 observed. Sampling the rank U accounts for reference variability given F and G, whereas estimating those distributions adds uncertainty. The discrepancy concerns the frozen point forecast, not a test integrating that estimation uncertainty. Re-estimation improves agreement post hoc but neither separates estimation error from drift nor establishes a general reference-size requirement.

3.3 Predictor comparisons depend on the operating point

Does the action-mismatch signal require the original prediction head? Cohort C suggested a $+4.7$-point timely advantage for the original head, using a capped comparator and unequal nominal false-

alarm rates ($0.073/0.056$). Cohort D reassesses this result with final comparators sharing the frozen encoder and input history, fitted only on nominal transitions. Under separate calibration, the action-aware MLP has timely detection 0.414, versus 0.389 for the original head and 0.111 for the separately fitted action-free MLP (Table 3). The pixel-change control detects only 0.08 of these episodes (0.02 timely; nominal audit 0.036, upper 0.047). The action-conditioned predictors thus recover evidence missed by these controls.

Table 3: PushT action mismatch, innovation checkpoint, cohort D. Timely means first alarm in $[j, j+2]$; detected means in $[j, 20]$. Panel A uses separate clean conformal references; brackets give nominal one-sided 95% upper bounds. Panel B uses audit-estimated thresholds, both with nominal FPR bracket $[0.049, 0.050]$. Counts make the primary difference explicit.

A. Separately calibrated pipelines

Score	Nominal rate [upper]	Timely	Detected
Original one-step predictor	0.040 [0.052]	0.389	0.791
MLP with actions, ensemble	0.046 [0.058]	0.414	0.818
MLP without actions	0.044 [0.056]	0.111	0.275

B. Approximately 5% empirical nominal FPR

Score	Nominal threshold	Timely count	Timely rate
Original one-step predictor	audit-estimated	$204/512$	0.398
MLP with actions, ensemble	audit-estimated	$218/512$	0.426

Timely difference: $-14/512 = -2.7$ pp; 97.5% interval $[-7.8, +6.3]$.

At approximately 5% empirical nominal false-positive rate (FPR), the primary comparison gives $204/512$ timely detections for the original head and $218/512$ for the MLP. Their difference is -2.7 percentage points, 97.5% interval $[-7.8, +6.3]$, with thresholds re-estimated in each joint bootstrap replicate. The comparison resolves neither superiority nor equivalence; these audit-estimated thresholds carry no independent conformal audit. For the separate secondary endpoint of full-horizon detection at locked, separately calibrated thresholds, the MLP is favoured (original minus MLP: -2.7 points; 95% interval $[-5.1, -0.4]$; 12 versus 26 discordant pairs). This secondary interval conditions on the fitted thresholds.

The ranking changes with the operating point and checkpoint. PushT point estimates favour the original head at targets 1–3% (0.207 versus 0.154 at 1%) and the MLP at 5–10%; joint-bootstrap uncertainty was computed at 5%. At that same FPR, the TwoRoom difference is $+5.1$ points, 95% interval $[+2.9, +7.4]$, against an MLP still improving at its fitting cap. Upstream PushT instead favours the MLP (-6.3 points, 97.5% interval $[-10.4, -3.5]$), with low timely rates for both ($0.05/0.11$). The primary result therefore concerns head choice at a specified operating point, while retaining the shared pretrained representation (Appendix L).

3.4 Matched object controls expose appearance-dependent ordering

Crossing the edited and recoloured objects isolates appearance dependence. The head comparison leaves a second question: does appearance affect the evidence for the same physical edit? Paired ordering θ compares a physical twin's onset-window score with its appearance-only twin, counting exact ties as one-half (Eq. equation 3). We cross 75-pixel pusher and block teleports with recolouring of either object. For each edit, state trajectories are identical across colours, and D contains no null teleports. Constant recolouring begins at frame 0, so it changes the appearance throughout the context without introducing a colour transition at the teleport onset.

Ordering for pusher/block teleports is $0.395/0.625$ when the same object is recoloured and $0.797/0.959$ when the other object is recoloured (Figure 2a). The primary other-minus-same interaction is $I_{\text{object}} = 0.368$, 97.5% interval $[0.332, 0.405]$; 324 parents contribute positively and 22 negatively. Each side averages over both teleported objects and both recoloured objects, cancelling additive differences in their overall difficulty (Appendix F). This identifies a dependence on which object is recoloured; it does not require assuming perceptually equal colour changes or identify a unique encoder-level cause.

The interaction also appears upstream (0.167, 97.5% interval $[0.128, 0.205]$), at moderate 90° recolouring (0.398, 95% interval $[0.352, 0.445]$), and with the MLP (0.370, $[0.338, 0.402]$). Pixel

change has a near-zero interaction (0.001, $[-0.007, 0.011]$). Same-pusher ordering is below 0.5 (0.395, 95% interval $[0.35, 0.44]$): the physical twin more often receives the *smaller* residual. Action mismatch likewise orders at 0.820 with constant block recolouring but 0.382 with constant pusher recolouring. These reversals matter because residual ordering reflects predictability as well as inconsistency; it is not a detection probability (Appendix M).

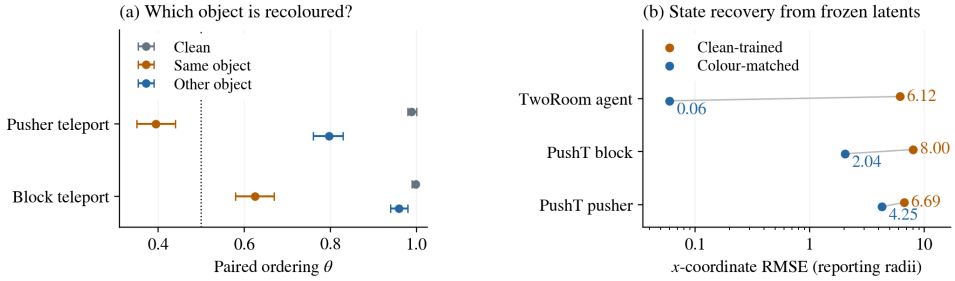


Figure 2: Object and readout controls, cohort D. (a) PushT teleport ordering under clean rendering and constant recolouring of the same or other object; bars are 95% parent-bootstrap intervals and the dashed line marks 0.5. (b) Coordinate-wise x root-mean-square error (RMSE) on no-physical-edit episodes with the named object recoloured; clean-trained and colour-matched MLP readouts share frozen latents. Reporting radii: TwoRoom 7 pixels, PushT 15. Appendix M gives full coordinates, jump ratios, and temporal controls.

Matched readouts test whether state remains accessible. Poor ordering could arise even when the latent retains the object's position. To test this possibility, a readout predicts simulator coordinates from one frozen latent; a colour-matched readout is fitted on nominal trajectories with the test recolouring. Its state supervision is unavailable to the monitor. Coordinate-wise root-mean-square error (RMSE) is reported in radii: seven pixels in TwoRoom and 15 in PushT. On the recoloured TwoRoom agent, matching reduces x error from 6.12 to 0.06 reporting radii and recovers a median decoded/true displacement-magnitude ratio of 1.00, although teleport ordering is 0.498. Accessible position therefore coexists with an unselective residual. In PushT, matched recoloured-block error falls from 8.00 to 2.04 radii, with jump ratio 0.88; matched pusher error remains 4.25 radii, with ratio 0.48 (Figure 2b). Clean-rendering errors are $1.17/0.46$ radii for block/pusher. Recovery is partial and object-dependent; unsuccessful probes do not establish information loss. Jump ratios measure magnitude, not direction.

3.5 Temporal masking and the costs of simple remedies

A shared onset transient can conceal later action evidence. Low window ordering can also arise from the way residuals are aggregated. Earlier cohort C (a separate 512-parent sample) illustrates alarm saturation: action mismatch plus abrupt *block* recolouring (from onset) has timely rate 0.95, versus 0.94 for its appearance-only twin, an increment of 0.01 (95% interval $[0.00, 0.02]$). Because appearance alone already alarms in roughly 94% of parents, the physical edit can add at most about six percentage points of timely alarms. This ceiling motivates inspecting the continuous scores on D. Both twins' maxima occur at the recolouring onset in 89% of parents: onset-window ordering is 0.547 (53 wins, five losses, 454 ties), yet the physical score is higher in 382–434 of 512 parents at each later offset $j+1, \ldots, j+5$. Shifting to the equal-length window $[j+1, j+3]$ raises ordering to 0.877; the five-frame window $[j+1, j+5]$ gives 0.881. The common maximum conceals later evidence already present in the residual sequence.

Constant pusher recolouring gives a different result: ordering is 0.382 on $[j, j+2]$ and 0.350 on $[j+1, j+3]$, despite no appearance transition at j. Removing the onset frame therefore does not remedy every appearance effect. Together, the two controls distinguish transient masking from weak ordering that persists beyond onset. They use the known intervention time for diagnosis; an online improvement would need a causal rule that locates useful evidence without that label.

Reference expansion and persistence have different costs. The preceding diagnoses clarify two simple changes evaluated in separate studies: broadening the reference to include expected appearance variation, and requiring residual evidence to persist. Adding known recolourings to the refer-

ence lowers PushT's seen-colour alarms from 0.910 to 0.105, while raising the one-step threshold from about 84 to 340 and reducing clean nominal alarms to zero. The mixture audit remains inconclusive; held-out 180° recolouring alarms at 0.170 (TwoRoom: 0.694). These benign rates alone do not establish preserved physical detection. Three-frame persistence, $s_t^{\text{pers}} = \min(s_{t-2}, s_{t-1}, s_t)$ with its own calibration, reduces D's abrupt-block alarms from 0.966 to 0.510 but increases constant-block alarms from 0.667 to 0.847. A lower recalibrated threshold can admit sustained recolouring residuals. Clean-action full-horizon detection falls from 0.79 to 0.68; the timely drop (0.39 to 0.04) is largely imposed by waiting three frames after an effect first visible at $j + 1$. These results show condition-specific reductions in benign alarms, with an unmeasured physical-detection cost for reference expansion and a measured cost for persistence (Table 26). Recalibration changes threshold crossings; improving physical-versus-benign raw-score ordering requires changing the score or representation.

3.6 Decision experiments test the value of acting on surprise

An informative score need not specify a useful intervention: replanning can consume computation without improving the action, and screening can discard useful goals. We test surprise-triggered replanning and subgoal screening on upstream PushT, controlling optimizer initialization and using matched per-call search budgets or fixed candidate sets. Against replanning every five steps, with both optimizers warm-started from any unexecuted plan suffix, the trigger changes success by 0.0 points (95% interval $[-2.9, +2.9]$; 384 episodes) and uses 7.5% more planner calls. This checkpoint has weak mismatch detection in the separate offline audit (Section 3.3); that audit uses a different thresholding rule and does not identify the cause of the replanning outcome. Screening against demonstration subgoals reduces success by 18.0 points; matching the reference to the proposer gives an unresolved $+1.6$ points $[-2.3, +5.6]$ on 768 episodes. On the same stream of four candidate subgoals, calibrated selection versus raw-residual ranking gives $+3.6$ points $[-1.0, +8.6]$ over 384 episodes, separating the selection rule from candidate count.

Diagnostics delimit the claim. A reference from deployment starts reduces first-proposal rejection from 0.911 to 0.080, but its upper bound 0.096 exceeds audit tolerance. In a separate TwoRoom diagnostic, the planning residual orders 39 certified reachable/unreachable target pairs correctly in 0.436, versus 0.590 for latent distance, with no resolved difference. Reference matching can reduce rejection without establishing a reachability certificate or a planning gain (Appendix N).

4 Discussion and Limitations

These studies distinguish physical information accessible in latents from evidence expressed by a score. Under separate calibration, the action-aware PushT MLP detects more mismatches than its action-free counterpart; the original head has no resolved timely advantage over the action-aware MLP at the primary common-FPR operating point. TwoRoom readouts recover position despite unselective residuals, while later PushT residuals reveal action evidence masked by an appearance transient.

Each diagnosis suggests a response: examine nominal tails and shifted-score concentration for reference sensitivity; change the score or representation for reversed raw ordering; evaluate causal aggregation with detection delay for temporal masking. Each response also requires nominal auditing, matched physical/appearance controls, and a separate decision-utility test (Table 8).

5 Conclusion

Calibrated surprise detects some simulator-relative inconsistencies, but its physical interpretation depends on the reference, appearance, and temporal statistic. VoE-Bench shows that changing the clean reference moves alarms on fixed benign TwoRoom queries from 0.5% to 82.4%, PushT teleport ordering depends on which object is recoloured, and temporal aggregation can conceal action evidence. Matched TwoRoom readouts further show that accessible position can coexist with an unselective residual. The tested planning rules provide no resolved success gain. These findings identify distinct targets for improvement and provide controlled tests for interpreting an alarm and evaluating whether acting on it improves decisions.

AI USE STATEMENT

We used generative AI tools to refine manuscript wording, assist with code debugging, and support selected aspects of methodological development, hypothesis refinement, proof preparation, and interpretation of results. The authors reviewed the AI-assisted material, checked the mathematical arguments and interpretations, and verified LLM-generated or modified code for correctness. All final research decisions and conclusions remained with the authors, who take responsibility for the manuscript and accompanying artifacts.

ETHICS STATEMENT

This work audits latent world-model monitors in simulated navigation and manipulation environments. A central concern is that calibrated alarms may be interpreted as evidence of physical inconsistency or safe operation without sufficient validation. Our benchmark addresses this concern through appearance-matched controls, independent nominal audits, and separate evaluations of planning utility. The reported calibration guarantees concern nominal episodes under the stated assumptions; they do not certify physical correctness or safe deployment. Applications to physical systems would require application-specific validation of detection performance, false alarms, and the consequences of acting on monitor outputs.

REPRODUCIBILITY STATEMENT

Section 2 specifies the monitoring protocol, data roles, evaluation endpoints, and controlled comparisons. Appendix B provides the statistical assumptions, proofs, and audit pseudocode, while Appendix C details intervention construction and observation controls. Comparator fitting, diagnostic analyses, uncertainty estimation, and planning protocols are documented in Appendices E, F, G, and H, respectively. We will provide an anonymized codebase as supplementary material for review. Model checkpoints are omitted from the submission because they exceed the supplementary upload size limit. Upon acceptance, we will publicly release the codebase and all model checkpoints used in this study.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL https://arxiv.org/abs/2107.07511.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics, 2025. URL https://arxiv.org/abs/2511.08544.

Rina Foygel Barber, Emmanuel J. Candes, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability, 2023. URL https://arxiv.org/abs/2202.13415.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1), February 2023. ISSN 0090-5364. doi: 10.1214/22-aos2244. URL http://dx.doi.org/10.1214/22-AOS2244.

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments, 2025. URL https://arxiv.org/abs/2506.09849.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

Fabian Domberg and Georg Schildbach. World models for anomaly detection during model-based reinforcement learning inference, 2025. URL https://arxiv.org/abs/2503.02552.

T. W. Epps and L. B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983. URL https://doi.org/10.1093/biomet/70.3.723.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos, 2025. URL https://arxiv.org/abs/2502.11831.

Michelle Ho, Muhammad Fadhil Ginting, Isaac R. Ward, Andrzej Reinke, Mykel J. Kochenderfer, Ali akbar Agha-Mohammadi, and Shayegan Omidshafiei. World model failure classification and anomaly detection for autonomous inspection, 2026. URL https://arxiv.org/abs/2602.16182.

Juyi Lin, Arash Akbari, Yumei He, Lin Zhao, Haichao Zhang, Arman Akbari, Xingchen Xu, Zoe Y. Lu, Enfu Nan, Hokin Deng, Edmund Yeh, Sarah Ostadabbas, Yun Fu, Jennifer Dy, Pu Zhao, and Yanzhi Wang. Phyground: Benchmarking physical reasoning in generative world models, 2026. URL https://arxiv.org/abs/2605.10806.

Yushan Liu, Peibo Sun, Xintao Chao, Zhenyang Yang, Yifan Xie, Lingfeng Zhang, Shoujie Li, Chenyu Tang, Fang Chen, Xiao-Ping Zhang, and Wenbo Ding. Checkvla: Execution-time verification with action-conditioned world model for long-horizon mobile manipulation, 2026. URL https://arxiv.org/abs/2607.26789.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels, 2026. URL https://arxiv.org/abs/2603.19312.

Sergi Masip, Jonathan Swinnen, Yutong Hu, Renaud Detry, and Tinne Tuytelaars. Ff-jepa: Long-horizon planning in world models with latent planners, 2026. URL https://arxiv.org/abs/2606.09311.

Luis S. Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6(9): 1257–1267, 2022. doi: 10.1038/s41562-022-01394-8. URL https://www.nature.com/articles/s41562-022-01394-8.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning, 2020. URL https://arxiv.org/abs/1803.07616.

Bo-Kai Ruan, Teng-Fang Hsiao, Ling Lo, and Hong-Han Shuai. Is the future compatible? diagnosing dynamic consistency in world action models, 2026. URL https://arxiv.org/abs/2605.07514.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction, 2007. URL https://arxiv.org/abs/0706.3188.

Vlad Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim G. J. Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models, 2025. URL https://arxiv.org/abs/2502.14819.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candes, and Aaditya Ramdas. Conformal prediction under covariate shift, 2020. URL https://arxiv.org/abs/1904.06019.

Vladimir Vovk. Conditional validity of inductive conformal predictors, 2012. URL https://arxiv.org/abs/1209.2673.

Isaac R. Ward, Michelle Ho, Houjun Liu, Aaron Feldman, Joseph Vincent, Liam Kruse, Sean Cheong, Duncan Eddy, Mykel J. Kochenderfer, and Mac Schwager. Foundational world models accurately detect bimanual manipulator failures, 2026. URL https://arxiv.org/abs/2603.06987.

Tianzhuo Yang, Zihan Shen, Zirui Mi, Zhaoyi Zhang, Jiayi Zhou, Jiaming Ji, Juntao Dai, Jiawei Chen, Boyuan Chen, and Yaodong Yang. Mirabench: Evaluating action-conditioned reliability in robotic world models, 2026. URL https://arxiv.org/abs/2605.29360.

Haoran Zhang, Yifu Lu, Boyang Wang, Xuhui Kang, Yen-Ling Kuo, Zezhou Cheng, Mengdi Wang, and Odest Chadwicke Jenkins. Foresight: Failure detection for long-horizon robotic manipulation with action-conditioned world model latents, 2026. URL https://arxiv.org/abs/2606.23085.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning, 2025. URL https://arxiv.org/abs/2411.04983.

A Extended Related Work and Positioning

A.1 Physical expectation in learned representations

Latent prediction provides a common basis for representation learning, planning, and violation-of-expectation evaluation, but these uses test different properties. PLDM studies planning from reward-free offline trajectories using a learned latent dynamics model (Sobal et al., 2025). DINO-WM predicts pretrained visual features for goal-directed action optimization (Zhou et al., 2025), while LeWM trains a joint-embedding world model from pixels (Maes et al., 2026). Their planning results motivate studying the information carried by latent predictions. They do not determine which deviations a thresholded residual will identify.

The violation-of-expectation approach offers a more direct test of physical expectation. Garrido et al. (2025) evaluate possible and impossible events using models pretrained on natural videos; IntPhys 2 organizes controlled video tests around physical principles such as permanence and continuity (Bordes et al., 2025). LeWM compares nominal, recoloured, and teleported trajectories in TwoRoom, PushT, and Cube (Maes et al., 2026). We retain this use of controlled perturbations and extend the analysis to reference-dependent decisions, equal-input prediction heads, and physical edits crossed with the identity and appearance of the affected object. Appearance is matched within each physical/no-physical pair, making score ordering interpretable even when both episodes trigger an alarm.

A.2 Failure monitoring and conformal calibration

Domberg & Schildbach (2025) use discrepancies between world-model predictions and observations to monitor departures from familiar operating conditions. Ward et al. (2026) calibrate uncertainty and prediction-error scores from a probabilistic world model for manipulation failure detection, while explicitly discussing benign-shift false alarms. Ho et al. (2026) combine supervised failure classification and anomaly detection for autonomous inspection. These objectives include task failure and unfamiliarity; a failed task can follow lawful dynamics, just as a benign appearance change can be unfamiliar.

Foresight trains a causal failure detector using task-level outcomes and predicted action-conditioned latents, then calibrates detection thresholds over trajectories (Zhang et al., 2026). CheckVLA combines an action-conditioned world predictor, a learned risk head, episode-level calibration, and execution repair; its ablations already address benign negatives, action inputs, refreshing predictions from observations, and temporal aggregation (Liu et al., 2026). Our principal scores are frozen latent residuals, and our small predictor comparators are fitted on nominal transitions. We examine what these scores distinguish through controlled interventions, without attributing novelty to episode-level calibration or benign-shift testing themselves. The present planning results do not evaluate or refute the gains reported by those different systems.

The statistical construction follows standard conformal rank calibration (Shafer & Vovk, 2007; Angelopoulos & Bates, 2022). Its marginal guarantee concerns the specified exchangeable population and averages over the calibration sample; auditing a locked threshold answers a separate empirical question. Covariate-shift weighting requires an appropriate weighting model (Tibshirani et al., 2020), and broader nonexchangeable analyses quantify validity under their stated conditions (Barber et al., 2023). These results motivate possible extensions, but they do not turn nominal calibration into a guarantee for arbitrary appearance subgroups. Our new evidence is the measured sensitivity of particular shifted-score distributions to reference resampling and the limits of the frozen forecast.

A.3 Plan evaluation and benchmark ground truth

Action-state consistency can help diagnose imagined futures and select candidate rollouts (Ruan et al., 2026). FF-JEPA instead introduces a latent subgoal planner to support longer-horizon execution (Masip et al., 2026). Our imagined-rollout and reachability analyses concern additional scores placed around such planning procedures. A high residual from a budgeted search may reflect model error or optimization failure; a simulator-respecting certificate is needed to label a target unreachable. Accordingly, the corrected paired analysis and the measured effect of screening are kept separate from any interpretation of score magnitude as physical infeasibility.

PhyGround evaluates generated videos with criteria grounded in physical laws, human annotations, and a learned judge (Lin et al., 2026). MiraBench separates physical adherence, action fidelity, and optimism about failure-inducing actions (Yang et al., 2026). These benchmarks address the reliability of generated futures. VoE-Bench instead starts from known simulator states and interventions, preserving the commanded actions and logging onset and observability. This control supports paired tests of observed-transition monitors, while limiting conclusions to the simulated interventions and observation schedules used here.

Table 4: Relation to selected prior work. Each row identifies an established contribution and the particular question examined here. The comparison concerns study objectives and controls; it makes no claim that a listed control is absent from every related method.

Setting	Established contribution	Question examined in this audit
Physical expectation (Garrido et al., 2025; Bordes et al., 2025; Maes et al., 2026)	Evaluate responses to controlled physical violations and visual changes.	How do the reference threshold, edited object, and matched appearance affect the alarm and score ordering?
Runtime monitoring (Domberg & Schildbach, 2025; Ward et al., 2026)	Use predictive discrepancies or uncertainty to detect anomalous execution.	What additional evidence supports a physical interpretation of a residual-based alarm?
Calibrated failure detectors (Zhang et al., 2026; Liu et al., 2026)	Combine predictive representations, learned failure or risk scores, and calibrated execution monitoring.	How much of the original residual's discrimination is recovered by a nominally fitted predictor with the same latent and action inputs?
Conformal inference (Shafer & Vovk, 2007; Tibshirani et al., 2020; Barber et al., 2023)	Establish error guarantees under explicit sampling or shift assumptions.	How sensitive is a shifted subgroup's observed alarm rate to the realized clean reference?
Planning (Ruan et al., 2026; Masip et al., 2026)	Diagnose future consistency or plan through latent subgoals.	Do the tested calibrated triggers and screens improve success against matched planning controls?
Generated-video evaluation (Lin et al., 2026; Yang et al., 2026)	Assess physical or action-conditioned reliability with structured annotations.	What does a monitor detect when intervention onset, nominal evolution, and appearance are controlled in a simulator?

B Statistical Definitions and Study Design

B.1 A frozen score and an episode as the calibration unit

An episode E contains observations and commanded actions over a fixed observation schedule. A scoring map Ψ specifies the encoder and predictor, input normalization, context length, rule for refreshing predictions from observations, scalar residual, and horizon $\mathcal{H}$. Its episode score is

$$S(E;\Psi) = \max_{t\in\mathcal{H}} s_t(E;\Psi). \tag{5}$$

Any fitted score transformation belongs to Ψ and is fixed before calibration. The rank argument requires exchangeable reference and query scores conditional on Ψ. Fitting on independent data is

one sufficient construction; symmetric role assignment within a fixed population is another, with a narrower interpretation given below. Calibration is performed separately for each checkpoint and score. No equality of residual scale between checkpoints is assumed. Changing the policy, rendering process, observation rate, or horizon can change the reference distribution even when the score formula is unchanged.

The focused studies sample 101 raw frames at a stride of five, giving 21 observations indexed 0 through 20. The initial context contains frames $0, 1, 2$, and $\mathcal{H} = \{3, \ldots, 20\}$ contains 18 target frames. In the implementation's length-20 score arrays, array index $t - 1$ corresponds to observed frame t; therefore the common evaluation slice is indices 2 through 19. Early one-step scores, when available, are excluded to match the open-loop horizon. Neither this horizon nor the threshold is aligned to the intervention onset.

For clarity, write $\bar{z}_t = z_t$ for $t = 0, 1, 2$ and recursively define the open-loop predictions by

$$\bar{z}_t = g(\bar{z}_{t-3:t-1}, a_{t-3:t-1}), \qquad s_t^{\mathrm{OL}} = \|\bar{z}_t - z_t\|_2^2, \qquad t \in \mathcal{H}. \tag{6}$$

In contrast, one-step scoring replaces the three inputs $\bar{z}_{t-3:t-1}$ with the observed $z_{t-3:t-1}$ at every prediction. Both use the same commanded-action history. The observed target z_t is used to compute the error in either case; only one-step scoring then incorporates it into future prediction contexts.

Proposition 1 (Episode-level split-conformal control)**.** *Condition on a scoring map Ψ. Suppose the finite scores $S_1, \ldots, S_n, S_{n+1}$ of the reference episodes and a nominal query episode are exchangeable. For the deterministic p-value in Eq. equation 2,*

$$\Pr\{p(S_{n+1}) \le \alpha \mid \Psi\} \le \alpha. \tag{7}$$

If the scores are almost surely distinct, this probability is $\lfloor \alpha(n+1) \rfloor / (n+1)$. The alarm is equivalently $S_{n+1} > q_\alpha$, with $q_\alpha = +\infty$ when $k > n$.

Proof. Without ties, exchangeability makes the query's descending rank among the $n + 1$ scores uniform on $\{1, \ldots, n + 1\}$. The numerator of p is that rank, so exactly $\lfloor \alpha(n + 1) \rfloor$ ranks cause an alarm. With ties, randomly ordering tied scores gives a uniform rank, while counting every reference tie as at least as large as the query can only increase the deterministic rank and p-value. This proves super-uniformity. Finally, writing $r_\alpha = \lfloor \alpha(n+1) \rfloor$ gives $k = n+1-r_\alpha$. The condition $p(S_{n+1}) \le \alpha$ is equivalent to at least k reference scores being strictly smaller than the query, hence to $S_{n+1} > S_{(k)}$ when $k \le n$. If $k = n + 1$, no alarm is attainable. □

This is the standard rank argument (Shafer & Vovk, 2007; Angelopoulos & Bates, 2022), applied to an episode scalar. The smallest possible p-value is $1/(n + 1)$, so level α is resolvable only when $n \ge \lceil 1/\alpha \rceil - 1$. At $n = 600$ and $\alpha = 0.05$, $k = 571$; the threshold is the 571st smallest reference maximum, with strict exceedance used for alarms. Quantile interpolation is not part of this rule.

Training exposure and finite-pool scope. The PushT training split was random over windows, so every source episode contributed to world-model training. Evaluation-role disjointness therefore does not mean an unseen training episode. Conditional on the trained model and an eligible finite pool, however, the scores are fixed values. If reference and query roles are assigned by a uniform random permutation of that same pool, the query's rank among any $n + 1$ selected scores is uniform (or conservative with ties). This gives the same marginal bound within that pool, without requiring the model to be independent of its training episodes. The focused partition procedure applies source-overlap exclusions before randomly assigning reference, audit, and intervention roles. This symmetry argument concerns nominal queries sampled by that procedure; it gives no guarantee for unseen episodes, differently selected queries, or recoloured variants. Training exposure can also make nominal residuals optimistic for deployment on new episodes.

Why frame pooling fails. Consider ten independent continuous episode values, each repeated over twenty frames. Pooling these 200 dependent values gives a query frame a p-value of $1/201$ whenever its episode value exceeds all ten reference values. That event has probability $1/11 > 0.05$. The apparent calibration sample size is therefore misleading. Reducing each reference episode to one maximum retains the actual exchangeable unit. Dependence within an episode is allowed; exchangeability across episode scores is the assumption.

What online threshold crossings inherit. Define $t^*(E) = \min\{t \in \mathcal{H} : s_t(E) > q_\alpha\}$, with $t^* = \infty$ if the set is empty. Then

$$\{t^*(E) < \infty\} = \{S(E) > q_\alpha\}. \tag{8}$$

The maximum-calibrated threshold thus controls the probability of any crossing over this fixed episode protocol. It provides neither an anytime-valid guarantee for an unbounded horizon nor simultaneous control over multiple checkpoints, episodes, or policy changes. In particular, replanning can alter the future episode distribution and is evaluated separately.

B.2 Auditing one realized threshold

For a fixed threshold q_α, let K of m nominal audit episodes cross it. We report K/m and the one-sided 95% Clopper–Pearson upper bound

$$U_{.95}(K, m) = \begin{cases} B_{.95}^{-1}(K+1, m-K), & K < m, \\ 1, & K = m, \end{cases} \tag{9}$$

where $B_u^{-1}(a, b)$ denotes the u quantile of a Beta distribution with shape parameters a, b. The audit criterion is $U_{.95} \leq 0.07$ at target level $\alpha = 0.05$. The 0.02 margin is an empirical tolerance, not an alteration of the conformal target. An upper bound above 0.07 means that the audit has not established the tolerance; the upper bound alone is not a lower-confidence claim that the true rate exceeds it. The threshold is retained regardless of the audit outcome.

The binomial interval is exact for independent, identically distributed Bernoulli audit outcomes conditional on the locked procedure. TwoRoom's fresh collector episodes approximate this sampling design. PushT roles are sampled without replacement from a finite dataset population that was exposed during world-model training. Its source-disjoint evaluation roles do not imply independence from model training, and the reported binomial bounds are a working approximation for that population. An audit neither proves exchangeability nor diagnoses its unique failure mode. Likewise, the marginal theorem does not require every realized reference to have conditional error at most α.

The audit's operating characteristic makes this distinction quantitative. Assume independent, identically distributed reference and audit scores from a continuous nominal cumulative distribution function (CDF) F. The probability integral transform gives

$$U = F(q_\alpha) \sim \mathrm{Beta}(k, n+1-k), \qquad A_P(q_\alpha) = 1 - U, \tag{10}$$

where $A_P(q_\alpha)$ is the population alarm probability conditional on that reference. At $n = 600$ and $\alpha = 0.05$, $A_P \sim \mathrm{Beta}(30, 571)$, with mean $30/601 \simeq 0.04992$ and standard deviation 0.00888. Conditional on A_P, the independent audit count is $K \sim \mathrm{Binomial}(1000, A_P)$; averaging over references gives a beta-binomial count. The upper bound is 0.06947 at $K = 56$ and 0.07056 at $K = 57$, so

$$\Pr\{K \geq 57\} = 0.2663, \qquad K \sim \mathrm{BetaBinomial}(1000, 30, 571). \tag{11}$$

This is an analytic property of the stated ideal sampling design, not a measured frequency across our checkpoints and not a consequence of exchangeability alone. Scores sharing a reference and audit sample can have correlated audit outcomes; several inconclusive scores need not provide independent evidence of drift. The calculation does not apply unchanged to PushT's sampling without replacement or to shifted appearance groups.

B.3 Distribution shift and reference mixtures

Let P denote the nominal episode-score law and Q a shifted law, with the score and observation protocol held fixed. The following bound isolates a change in the query distribution.

Table 5: Parent-episode roles in the focused studies. Appearance variants inherit their parent's role. The reference-sensitivity experiment uses a separate collection described in Appendix D.

Role	Parents per environment	Purpose
Fitting	2,000	Fit comparator heads and state readouts
Validation	400	Select nominal fitting settings
Reference (R2-D)	600	Set each operational threshold
Nominal audit (R3-D)	1,000	Audit that threshold; estimate separate FPR curves
Confirmation (D)	512	Evaluate frozen intervention comparisons
Moderate-colour subset	256 of D	Secondary appearance diagnostic

Proposition 2 (A total-variation bound)**.** *If the reference scores are independent draws from P, and the query is independently drawn from Q, then*

$$\Pr_{P^n \times Q}\{S > q_\alpha\} \le \min\{1, \alpha + d_{\mathrm{TV}}(P, Q)\}, \tag{12}$$

where $d_{\mathrm{TV}}(P, Q) = \sup_A |P(A) - Q(A)|$.

Proof. For each realized reference $\mathcal{C}$, the alarm set is $A_{\mathcal{C}} = (q_\alpha(\mathcal{C}), \infty)$. Thus $Q(A_{\mathcal{C}}) \le P(A_{\mathcal{C}}) + d_{\mathrm{TV}}(P, Q)$. Averaging over $\mathcal{C} \sim P^n$ and applying Proposition 1 proves the claim. □

This bound applies to the *score distributions*, not a measured distance between simulator states or images. A binned empirical distance is only a diagnostic and is not a certified upper bound on the underlying total variation. We do not implement importance-weighted or anytime-valid calibration. Related extensions require their own assumptions (Tibshirani et al., 2020; Barber et al., 2023).

For a reference mixture $P_{\mathrm{mix}} = \sum_c \pi_c P_c$, where c indexes appearance conditions and π_c are their sampling probabilities, a mixture alarm rate bounds the weighted average $\sum_c \pi_c P_c(S > q_\alpha)$. It does not bound each subgroup rate individually. The earlier fixed-composition appearance mixture and the later randomly sampled mixture are distinct studies; their calibration and audit generators must be identified separately. Re-renderings of one parent remain paired observations, not additional independent episodes.

B.4 Roles, confirmation stages, and pseudocode

Table 5 gives the final confirmation roles. Fitting and validation parents are reused across comparator families to hold the comparison data fixed; the 400 validation parents are separate from the 2,000 fitting parents. Reference, audit, and confirmation parents are source-disjoint from those fitting roles and earlier evaluation roles. Earlier stages A, B, and C contain 128, 256, and 512 intervention parents per environment, respectively. The two primary D contrasts were fixed after inspecting C, including the apparent PushT head advantage and the object/appearance diagnostics. A–C therefore provide development evidence for those hypotheses and are not pooled into D. Each confirmation analysis has a dated protocol and a hash record of frozen artifacts.

Algorithm 1 separates fitting a score, calibrating an alarm, and interpreting that alarm. Here $\tau = 0.07$ and $\alpha = 0.05$ for the episode audits. Its resampling unit is always the parent episode. The reference-sensitivity branch has its own nominal/recoloured audit pairs; it does not reuse the stage-D confirmation parents. Planning experiments are outside this fixed-horizon procedure.

B.5 Protocol deviations and post hoc corrections

Table 6 collects every known departure from the dated protocols and every correction made after results were first computed. The protocol-specified versions of the first two items were not re-run.

Algorithm 1 Reference calibration, nominal auditing, and physical-selectivity evaluation

Inputs: frozen encoder/predictor; nominal fitting and validation episodes; disjoint reference $\mathcal{C}$, nominal audit $\mathcal{A}$, and intervention parents $\mathcal{D}$; horizon $\mathcal{H}$; level α; audit tolerance τ.

1. Fit any comparator or score transformation using fitting/validation data only. Freeze its scoring map Ψ, input history, and rule for using observed context.
2. For each $E_i \in \mathcal{C}$, compute $S_i = \max_{t \in \mathcal{H}} s_t(E_i; \Psi)$. Set $k = \lceil (|\mathcal{C}| + 1)(1 - \alpha) \rceil$ and $q_\alpha = S_{(k)}$; use $+\infty$ if $k > |\mathcal{C}|$.
3. On $\mathcal{A}$, count $K = \sum_{E \in \mathcal{A}} \mathbf{1}\{S(E; \Psi) > q_\alpha\}$. Report $K/|\mathcal{A}|$, Eq. equation 9, and whether $U_{.95} \leq \tau$. Retain q_α.
4. For each parent in $\mathcal{D}$, score all frozen state/action and appearance variants. The detector receives observations and commanded actions, without onset labels.
5. Using the logged onset only during evaluation, compute first-alarm outcomes, paired physical increments, and tie-aware score ordering. Keep all variants of each parent together when resampling.
6. For the separate common-FPR comparison, estimate empirical thresholds on $\mathcal{A}$; evaluate on $\mathcal{D}$. Bootstrap $\mathcal{A}$ and $\mathcal{D}$ independently by parent, sharing indices across scores and refitting thresholds in every replicate.
7. For reference sensitivity, repeat step 2 on each prespecified new reference set and score the same fixed nominal/recoloured audit pairs. Preserve nesting across reference sizes in interpretation.

Outputs: operational thresholds and nominal audits; detection and appearance-selectivity estimates; uncertainty and reference-sensitivity summaries. Step 6 has no independent audit of its newly selected thresholds.

Table 6: Protocol deviations and post hoc corrections.

Item	Issue	Resolution
Common-FPR threshold (primary contrast)	The protocol specified labelled interpolation between attainable empirical FPR points	Implemented as a single strict threshold at $V_{(m-\lfloor \beta m \rfloor + 1)}$, attaining empirical FPR 0.049 at $\beta = 0.05$ (Appendix E)
MLP comparator checkpoint	The protocol specified minimum-validation-loss weights	The implementation retains the last update meeting the 0.1% improvement criterion, which can precede the absolute minimum (Appendix E)
PushT jump ratios	Missing from the initial stage-D readout analysis because of a variant-name filter	Added after the filter was corrected; existing outcomes and both primary contrasts unchanged (Appendix M)
Reference-forecast uncertainty	A 200-iteration ECDF bootstrap was frozen with the point forecast (Appendix D)	Not analysed in this submission; conclusions concern the frozen point forecast

C Intervention Construction and Observation Controls

C.1 Simulator-relative ground truth

Let x_u denote the simulator state at raw step u, a_u^{cmd} the commanded action, and a_u^{app} the applied action. Nominal evolution is $x_{u+1} = F_{\mathrm{sim}}(x_u, a_u^{\mathrm{cmd}})$. A state intervention replaces a state component at its logged onset and, for dynamically simulated categories, resumes the simulator from the edited state. An action intervention changes a_u^{app} while leaving a_u^{cmd} available to the detector. An appearance condition changes the renderer R_c in $o_u = R_c(x_u)$ without changing x_u or the action streams. The index u denotes raw simulator time; t denotes the strided observation index used by the monitor.

Table 7 separates the broad benchmark's intervention types. A velocity edit that is overwritten by the next simulator update, an overlap removed before observation, or a clipped action with no visible effect cannot support a claim about detecting a visible physical inconsistency. The implementation therefore includes no-edit re-simulation checks and observability diagnostics. We retain the simulator and observation-rate qualifications when reporting these categories. Occlusion/reappearance is a kinematic visibility and state-rendering construction and is not described as a dynamically simulated teleport trajectory.

Table 7: Intervention semantics. Appearance-only twins isolate rendering effects; a consistent hide/reappear control isolates disappearance from an inconsistent reappearance.

Condition	Change and interpretation
Teleport	Displace the selected object's state at onset; resume nominal dynamics.
Velocity discontinuity	Scale or reverse velocity; check whether the edit survives to an observed frame.
Interpenetration	Introduce geometric overlap; check its persistence under contact resolution.
Action mismatch	Rotate the applied action stream while retaining the original commanded stream as predictor input.
Occlusion/reappearance	Hide an object and render reappearance with a state displacement; compare with zero-displacement reappearance.
Colour or lighting	Change renderer parameters while preserving state and actions; report benign alarm rates separately.

C.2 The focused object and appearance design

For each stage-D PushT parent, a shared onset j is sampled uniformly from observed frames 8 through 14, corresponding to raw frame $5j$. The four physical conditions are no edit, pusher teleport, block teleport, and a 90° applied-action rotation from onset. They are crossed with five appearance conditions: clean, abruptly recoloured pusher, constantly recoloured pusher, abruptly recoloured block, and constantly recoloured block. Abrupt recolouring starts at j; constant recolouring starts at frame zero. For each fixed appearance condition, physical and no-physical twins share the rendering rule and are identical before the physical edit. They need not be identical to the clean-rendering trajectory before j.

A teleport changes the rendered position at raw step $5j$, so observed frame j can already differ. For action mismatch, the applied actions change from raw step $5j$; the onset position itself is unchanged, and the first potentially different observed position is at $j + 1$. The generator also rotates the target velocity at onset, which is not directly rendered as position. Consequently, the timely window $[j, j + 2]$ includes two potentially affected observations for action mismatch but three for teleportation. An alarm at j is retained under the prespecified endpoint even though it cannot be caused by the command–execution mismatch. First-effect metadata records realized observability; this timing statement describes the generator, not a guarantee that every edit is visible at its earliest possible frame.

Both PushT teleports request a displacement of 75 pixels. Directions are sampled uniformly and accepted only when the displaced center remains in the allowed box and satisfies the code's center-distance clearance test, at least the sum of the two simulator radii plus five pixels from the other object. Thus accepted directions are conditioned on geometry, not uniformly distributed over all angles. Up to 2,000 proposals are attempted; an unsuccessful search is logged as a null edit. This clearance rule uses simulator radius parameters and is not a polygonal separation certificate for the T-shaped block. Displacements, accepted directions, first state effects, and null edits are recorded without consulting score values. No null teleports were recorded in D.

The primary appearance magnitude is the renderer's $+180^\circ$ setting; a $+90^\circ$ constant-colour panel uses 256 hash-selected D parents. For the coloured pusher, this rotates its base hue. The block starts grey, so the renderer assigns a saturated colour to its achromatic base. Consequently, the block manipulation also changes chromatic saturation and is more accurately described as recolouring than as a pure hue rotation. TwoRoom crosses no edit, teleport, and action mismatch with clean, abrupt, and constant agent recolouring. Appearance severity is a renderer setting, not a perceptually matched distance across objects or environments.

C.3 Commands, temporal alignment, and model scope

Each observation interval contains five raw commanded actions. These are concatenated into a_t and passed through the checkpoint's training-time action normalization before the world model's action encoder. Applied actions are retained as ground-truth metadata but are not substituted for commands when testing action mismatch. Such a substitution would change the question from

whether execution follows a command to whether the predictor explains the altered command. The input-history comparison uses the same three action chunks and three observed latents for every equal-input predictor.

The broad suite contains LeWorldModel (LeWM), PLDM adaptations that plan with latent dynamics, and DINO-WM variants using pretrained visual features. LeWM's upstream recipe applies sketched isotropic Gaussian regularization (SIGReg) (Balestriero & LeCun, 2025) to latent embeddings (Maes et al., 2026). The innovation recipe instead applies it to standardized transition residuals $\eta_t = (\widetilde{z}_{t+1} - \rho \odot \widetilde{z}_t)/\sqrt{1-\rho^2}$, where $\widetilde{z}$ is coordinate-standardized and ρ is a coordinatewise lag-one correlation. The evaluated recipe also retains a marginal regularization term with relative weight 0.1. This is a training regularizer; η_t is neither the action-conditioned prediction residual in Eq. equation 1 nor a newly fitted monitoring score. The innovation checkpoints were trained in a companion study and remain frozen in this audit.

The primary D contrasts use innovation PushT to reassess the development findings: an apparent timely-detection advantage over a capped MLP and an appearance effect that motivated crossing the edited and recoloured objects. This choice was fixed before D was read. Upstream PushT provides a secondary comparison, and upstream TwoRoom is used for the reference study. These are checkpoint-specific contrasts; differences between upstream and innovation checkpoints do not isolate the training objective from training-seed or other checkpoint effects. The focused findings are not assumed to hold for all 29 checkpoints.

D Reference Sensitivity and Its Prospective Forecast

D.1 Separating score changes from threshold changes

To compare two reference collections, we first keep the checkpoint, preprocessing, observation schedule, and scalar score identical. The bridge analysis re-scores the earlier stored trajectories with the current exporter, compares nominal and recoloured score distributions, and cross-applies each collection's clean threshold to both query populations. The earlier and later collections use generator versions 0.1.0 and 0.2.0, respectively, so this contrast alone does not isolate repeated sampling from one unchanged population. Crossing fixed thresholds with the same queries isolates the threshold effect; the separate prospective collection then holds the generator configuration fixed while varying the reference. It does not declare either reference valid solely because a nominal audit meets tolerance.

D.2 The order-statistic calculation

Equation equation 4 in the main text combines the nominal order-statistic law with the shifted survival function. Here F is the nominal episode-maximum CDF, G is the CDF for one fixed appearance condition, and $F^{-1}(u) = \inf\{x : F(x) \geq u\}$ is the generalized quantile function. Under the independent, continuous-score assumptions of Eq. equation 10, $U = F(q_\alpha)$ has the stated Beta law and $A_Q(q_\alpha) = 1 - G(q_\alpha)$. The resulting map describes variation across reference draws; it does not assign an episode a probability of physical invalidity. Estimating F and G introduces a second source of uncertainty beyond sampling U.

When densities f_F and f_G exist and f_F is positive near the threshold,

$$\frac{\mathrm{d}}{\mathrm{d}u}\left[1 - G(F^{-1}(u))\right] = -\frac{f_G(F^{-1}(u))}{f_F(F^{-1}(u))}. \tag{13}$$

Thus shifted scores concentrated where nominal scores are sparse amplify reference variation. This local calculation does not establish that the CDFs are estimated precisely. The empirical forecast uses step CDFs, not fitted densities; Eq. equation 13 explains its sensitivity rather than supplying another fitted model.

D.3 Frozen estimates and new reference collections

We estimate F using a separate 4,300-parent nominal reservoir and G using 1,000 earlier hue-180° audit maxima. For each $n \in \{150, 600, 2400\}$, 20,000 Beta draws are transformed by the empirical

inverse CDF and shifted survival function. The calculation uses the inverted empirical CDF and strict exceedance, preserving its ties. It is a plug-in continuous-score approximation, not an exact finite-reservoir prediction. A further 200 bootstrap iterations independently resample the F and G records, using 2,000 Beta draws per iteration to describe CDF-estimation sensitivity. All forecasts are frozen before the new reference and audit scores are inspected.

The prospective collection contains 9,600 unique nominal reference parents partitioned into four disjoint blocks of 2,400. Each block splits into four sets of 600 and sixteen sets of 150, giving 84 reference sets overall. Sets are disjoint within each size and nested across sizes; there are four largest references, not 84 independent largest-sample replications. Every threshold is applied to the same 1,000 new audit parents rendered clean and with an abrupt $+180^\circ$ recolouring. This appearance law uses the earlier raw-onset range 25–75, rather than D's physical-intervention onset law. Sharing the audit pairs removes query-sample differences from threshold comparisons but induces shared audit uncertainty. We report discrepancies from the frozen point forecast without replacing it after inspection. CDF re-estimation using the new collection is a labeled post hoc diagnostic.

E Equal-Input Predictor Comparisons

E.1 Inputs, targets, and scores

For an observed transition from t to $t+1$, define

$$x_t^+ = [z_{t-2}; z_{t-1}; z_t; a_{t-2}; a_{t-1}; a_t], \quad x_t^- = [z_{t-2}; z_{t-1}; z_t], \quad y_t = z_{t+1} - z_t. \tag{14}$$

Semicolons denote concatenation. The superscripts $+$ and $-$ indicate whether commanded-action chunks are included. Each comparator b_ω predicts y_t and receives score

$$s_{t+1}^b = \|y_t - b_\omega(x_t^\pm)\|_2^2. \tag{15}$$

This is also the squared error of the next-latent prediction $z_t + b_\omega(x_t^\pm)$. Both variants are trained separately; the action-free model is not an action-aware model evaluated with zeroed actions. All share the frozen pretrained encoder. Equal-input therefore means the same latent history and commanded actions at evaluation, not matched architecture, optimization, or training exposure. In PushT, the original head was trained on windows from the evaluation-source episodes, whereas the comparators use only their separate fitting parents. This asymmetry may favour the original head; it is not a proof that the observed difference is conservative. The comparison assesses the fitted heads given those latents and does not test whether encoder pretraining is unnecessary.

E.2 Fitting and model selection

The final comparators use 2,000 nominal fitting parents and 400 validation parents. Each parent contributes 18 transitions. Input features and change targets are standardized using fitting-set means and standard deviations, with 10^{-6} added to each standard deviation. Outputs are transformed back to original latent units before computing Eq. equation 15. Ridge regression minimizes summed squared error in standardized target coordinates plus $\lambda\|W\|_F^2$, with $\lambda \in \{10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$ selected by nominal validation error in original units. Centering supplies the intercept after inverse standardization.

The MLP has two 256-unit ReLU hidden layers and a linear output. Each action condition uses three independent initialization seeds, AdamW with learning rate 10^{-3}, batch size 512, and weight decay selected from $\{0, 10^{-4}, 10^{-2}\}$ using standardized nominal validation loss. The frozen final rule allows at most 300 epochs and stops after 15 epochs without a relative validation improvement of at least 0.1%. In the implementation, checkpoint updates and patience resets both require this improvement, so the retained weights correspond to the last qualifying validation improvement. The written protocol requested the minimum-validation-loss weights; this implementation can retain an earlier checkpoint if later improvements are smaller than the criterion. We report the implemented selection rule, without claiming that the absolute validation minimum was retained. The primary comparator averages the three predicted latent changes *before* squaring the residual. All initialization seeds use the same parent splits.

The final PushT comparator satisfies its stopping criterion, while the upstream TwoRoom comparator reaches the 300-epoch cap with validation loss still improving. We retain this convergence

limitation in the interpretation of the TwoRoom comparison. Earlier one-action-chunk ridge fits and 50-epoch MLP fits remain historical comparisons and do not define the final equal-input baseline.

E.3 Operational pipelines and empirical common-FPR comparisons

Operational comparison assigns each score its own Eq. equation 2 threshold from the 600 R2-D parents. The fixed thresholds are audited on R3-D and evaluated on D. Their realized nominal alarm rates can differ, so differences in timely detection are differences between complete calibrated pipelines.

The common-FPR comparison instead uses the 1,000 clean R3-D maxima as an operating-point estimation sample. Let $V_{(1)} \le \cdots \le V_{(m)}$ denote those maxima and let β be a requested empirical FPR. The implemented rule sets $r = \lfloor \beta m \rfloor$ and

$$\widetilde{q}_\beta = V_{(m-r+1)} \quad (r \ge 1), \qquad \widehat{a}_- = \frac{1}{m}\sum_i \mathbf{1}\{V_i > \widetilde{q}_\beta\}, \quad \widehat{a}_+ = \frac{1}{m}\sum_i \mathbf{1}\{V_i \ge \widetilde{q}_\beta\}. \tag{16}$$

For $r = 0$, $\widetilde{q}_\beta = +\infty$. Detection uses the strict $>$ rule; $\widehat{a}_-$ and $\widehat{a}_+$ report the attainable bracket. With 1,000 distinct maxima and $\beta = 0.05$, they are 0.049 and 0.050. The written protocol described labelled interpolation between achievable FPR points. The implemented analysis instead uses the single strict threshold above, without interpolation or tie randomization; the reported primary estimate follows that implementation. We therefore describe it as approximately 5% empirical FPR. R3-D cannot independently audit a threshold chosen from those same records. This comparison estimates relative discrimination; it does not establish a new 5% deployment guarantee.

F Appearance Selectivity, State Readouts, and Temporal Aggregation

F.1 Threshold-free paired ordering

The paired ordering $\theta(e, c)$ in Eq. equation 3 uses the three-frame maximum

$$W_i(e, c) = \max_{t \in \{j_i, j_i+1, j_i+2\}} s_{i,t}(e, c). \tag{17}$$

Here 0 denotes no physical edit, N is the number of paired parents, and equality means exact equality of the stored scores. A value above 0.5 indicates a directional ordering advantage under this statistic. It is not a per-episode detection probability, an unpaired area under the receiver operating characteristic curve (AUROC), or evidence of equality of the two score distributions when it equals 0.5. We report strict wins, losses, and ties because a shared transient can make the two maxima identical even when later scores differ.

Changing the reference or alarm level while holding s_t fixed leaves every $W_i(e, c)$, and hence $\theta(e, c)$, exactly unchanged. Furthermore, for a fixed reference the conformal map is nonincreasing: $S_a > S_b$ implies $p(S_a) \le p(S_b)$. Discrete ranks can merge distinct scores into ties but cannot reverse their order. Recalibration therefore changes the operating threshold without supplying a new ordering of the raw evidence. A new temporal statistic, predictor, or representation changes the score and must be evaluated separately.

Let p, b denote pusher and block teleports, and $c_\mathrm{p}, c_\mathrm{b}$ constant recolouring of the respective object. The primary contrast is

$$I_{\text{object}} = \tfrac{1}{2}\big[\theta(\mathrm{p}, c_\mathrm{b}) + \theta(\mathrm{b}, c_\mathrm{p})\big] - \tfrac{1}{2}\big[\theta(\mathrm{p}, c_\mathrm{p}) + \theta(\mathrm{b}, c_\mathrm{b})\big]. \tag{18}$$

Positive I_{object} means higher teleport ordering when the other object is recoloured. To see what the crossing controls, suppose the four ordering means have an additive form $\theta(e, c) = \mu + a_e + b_c$, with an edit effect a_e and an appearance effect b_c. Both halves of Eq. equation 18 then equal $\mu + (a_\mathrm{p} + a_\mathrm{b} + b_{c_\mathrm{p}} + b_{c_\mathrm{b}})/2$, giving $I_{\text{object}} = 0$. A nonzero contrast therefore detects a departure from that additive description. It does not assert equal perceptual colour distances, invariance to recolouring the other object, or an encoder-level cause. Abrupt and moderate recolourings, the upstream checkpoint, the MLP comparator, open-loop surprise, and pixel change are secondary checks.

F.2 THRESHOLDED INCREMENTS AND SATURATION

Let $Y_i(e, c) \in \{0, 1\}$ be a timely-alarm outcome under one locked threshold. The paired physical increment and the difference from its clean-rendering value are

$$\Delta(e, c) = \frac{1}{N} \sum_i [Y_i(e, c) - Y_i(0, c)], \qquad J(e, c) = \Delta(e, c) - \Delta(e, \text{clean}). \tag{19}$$

Let $b(c) = N^{-1} \sum_i Y_i(0, c)$ be the appearance-only timely rate. Since $Y_i(e, c) \leq 1$, the increment satisfies

$$\Delta(e, c) \leq 1 - b(c). \tag{20}$$

This is an algebraic ceiling on the binary endpoint, independent of how informative the continuous residual may be. At the approximately 0.94 benign timely rate in cohort C, the largest possible increment is about 0.06. The observed 0.01 increment should therefore be interpreted alongside the unthresholded comparison, rather than as evidence that the score contains no physical signal. Conversely, a positive ordering advantage does not ensure useful alarm timing. We report the appearance-only rate and the four joint outcomes $(Y_i(0, c), Y_i(e, c))$, not only Δ. The same construction applies to detected and any-alarm endpoints, with the endpoint named explicitly.

F.3 STATE READOUTS AND WHAT THEY IDENTIFY

Readouts map a single current-frame latent to simulator coordinates. PushT targets are pusher and block x/y coordinates and block orientation represented by sine and cosine; TwoRoom targets are agent x/y. We fit ridge regression and an MLP with two 128-unit ReLU hidden layers, using the same 2,000 fitting and 400 validation parents. The two fitting regimes are clean rendering and appearance-matched rendering of those nominal parents. Encoders remain frozen and physical intervention labels are not used to fit the readouts.

Ridge penalties are $\{10^{-4}, 10^{-2}, 1, 100\}$. The readout MLP uses AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size 512, at most 200 epochs, and the same 15-epoch, 0.1% improvement rule. Inputs and targets are standardized on fitting parents. The primary neural readout averages predictions from seeds 0, 1, and 2. These are diagnostic models trained with simulator state supervision, not additional inputs to the monitor.

For target coordinate ℓ, let $y_{i,t,\ell}$ be its true value and $\widehat{y}_{i,t,\ell}$ its decoded value. Post-onset error is

$$\text{RMSE}_\ell = \left[\frac{1}{N} \sum_i \frac{1}{21 - j_i} \sum_{t=j_i}^{20} (\widehat{y}_{i,t,\ell} - y_{i,t,\ell})^2 \right]^{1/2}. \tag{21}$$

We report coordinate errors in pixels and normalized by the reporting radius, 15 pixels for PushT and seven for TwoRoom. The compact mechanism plot uses x-coordinate RMSE for each object; the detailed tables report coordinates separately. These coordinate errors are distinct from two-dimensional Euclidean position RMSE. Orientation coordinates are dimensionless and are not divided by a spatial radius.

For a teleported object's two-dimensional position $v_{i,t}$, jump recovery uses

$$R_{\text{jump}} = \text{median}_i \frac{\|\widehat{v}_{i,j_i} - \widehat{v}_{i,j_i-1}\|_2}{\max\{\|v_{i,j_i} - v_{i,j_i-1}\|_2, 10^{-6}\}}. \tag{22}$$

The denominator is the actual observed inter-frame displacement, which may include nominal motion in addition to the imposed teleport. This ratio measures jump magnitude and does not by itself test its direction. The prespecified reporting scales are post-onset coordinate-wise RMSE within two reporting radii and within three times the corresponding clean-trained clean-rendering error, and median recovered jump magnitude at least half of the true inter-frame displacement. They are interpretive scales, not exclusion or significance gates.

Successful appearance-matched decoding shows that the tested state variables remain accessible to that readout family. Failure of a clean-trained readout alone can reflect representation shift. Failure of both tested families limits the demonstrated recoverability but cannot prove that all state information is absent. Readouts also do not establish that every feature needed by a dynamics predictor is preserved. We combine their evidence with the intervention and temporal diagnostics rather than treating them as a unique causal decomposition.

F.4 Temporal masking and persistence

For offsets $\ell = 0, \ldots, 5$, we inspect the paired residual difference

$$d_{i,\ell}(e, c) = s_{i,j_i+\ell}(e, c) - s_{i,j_i+\ell}(0, c). \tag{23}$$

We report its median and quartiles, strict wins/losses/ties, and the fraction of pairs whose two three-frame maxima both occur at onset. The dedicated temporal analysis also recomputes the ordering in Eq. equation 3 using $\{j_i + 1, \ldots, j_i + 5\}$. Later positive differences together with onset-dominated maxima support aggregation masking. Persistently weak ordering under constant recolouring requires a different explanation. These onset-aligned windows are offline diagnostics and are never used to construct the operational threshold.

A secondary three-frame persistence score is $s_t^{\text{pers}} = \min\{s_{t-2}, s_{t-1}, s_t\}$ on complete scored windows. Its episode maximum is calibrated separately. Crossing a persistence threshold requires all three constituent residuals to exceed that threshold, trading sensitivity to isolated spikes for detection delay. It is a separate scoring procedure, not an uncalibrated alteration of the original alarm rule.

G Endpoints and Statistical Inference

Table 8 collects the measurements used along the audit. Their denominators, reference populations, and inferential targets differ; a strong result on one row cannot substitute for the measurement on another. The parent is the resampling unit for the focused offline comparisons, with all variants kept together. Planning comparisons instead pair outcomes under the same evaluation episode specification.

Table 8: What each reported measurement establishes. Operational detection uses a locked reference threshold; common-FPR comparisons re-estimate thresholds from nominal audit records and have no independent audit of those new thresholds.

Measurement	Population and comparison	Question answered
Nominal alarm rate	Fraction of nominal audit parents with any crossing under one locked threshold	Does this realized threshold meet the stated nominal tolerance?
Timely / detected rate	Fraction of intervention parents whose first crossing is in $[j, j+2]$ / $[j, 20]$	How often, and how soon, does the operational rule detect an edit?
Paired ordering θ	Within-parent physical versus appearance-only window maxima; ties count one-half	Does the physical edit raise the score above its matched benign twin?
Object contrast I_{object}	Other-minus-same recolouring contrast in θ, averaged over both teleports	Does the effect of appearance depend on which object is teleported?
Common-FPR difference	Paired timely outcomes at audit-estimated nominal operating points	How do predictors compare at approximately equal empirical false-alarm rate?
Planning success and cost	Paired episode outcomes, planner calls, and time under specified control policies	Does using the score improve decisions at the measured computational cost?

G.1 First-alarm outcomes and AUROC populations

For the first crossing t_i^* and onset j_i, premature, detected, timely, and missed outcomes are defined as in the main text. A timely alarm is a subset of detected alarms; premature episodes are not subsequently reclassified as detected when another crossing occurs. Any-alarm rate equals premature plus detected rate, and missed means no crossing. All parents remain in the denominator. A lower threshold can move a first crossing from timely to premature, so timely-detection curves need not be monotone in nominal FPR. Delay summaries condition on detected episodes and are not a replacement for the missed rate.

For episode scalar V^+ from a physical condition and V^- from its stated negative population, ordinary AUROC estimates $\Pr(V^+ > V^-) + \frac{1}{2}\Pr(V^+ = V^-)$. The stage-D common-FPR export compares full-horizon physical maxima with nominal R3-D maxima. Appearance diagnostics compare

full-horizon maxima with appearance-only twins using all cross-parent positive/negative comparisons, while resampling paired parents for uncertainty. Earlier onset-aligned benchmark summaries use their stated post-onset scalar. These different populations and reductions are labeled explicitly; none is interchangeable with within-parent ordering in Eq. equation 3.

G.2 Paired uncertainty and the two primary contrasts

For a difference between detectors or conditions evaluated on the same parents, each bootstrap draw resamples parents and carries all of their variants and detector outcomes together. The two stage-D primary endpoints are (i) one-step world-model minus action-aware MLP timely detection for clean PushT action mismatch at approximately 5% empirical nominal FPR, and (ii) Eq. equation 18 for one-step surprise under constant recolouring. Both use the innovation-trained PushT checkpoint and 5,000 resamples. The reported 97.5% percentile intervals use the 1.25th and 98.75th percentiles. Allocating 2.5% error to each of two intervals is an approximate Bonferroni familywise treatment because bootstrap coverage itself is approximate.

For the common-FPR contrast, each replicate independently resamples nominal R3-D parents and intervention D parents, while using common indices across the compared scores within each role. Both thresholds are re-estimated from that replicate's nominal maxima before computing the paired timely difference. This captures nominal operating-point estimation uncertainty as well as intervention-sample uncertainty. By contrast, bootstrap intervals for the fixed operational pipelines condition on the realized R2-D thresholds; they do not integrate over new reference draws. The separate reference experiment addresses that source of variation.

Secondary paired effects generally use 95% intervals. The dedicated temporal and state-readout analyses use 2,000 parent resamples; the ordinary appearance AUROC helper uses 1,000. Exact McNemar tests, when reported for binary comparisons, condition on discordant parent pairs. The primary intervals are percentile intervals, not bias-corrected intervals. Re-estimating discrete empirical thresholds can yield asymmetric bootstrap distributions; the reported interval is not assumed to be centred on the point estimate. A confidence interval containing zero is treated as an unresolved difference, not as an equivalence result. Parent-level intervals are conditional on the tested checkpoints and fitted heads; three ensemble initialization seeds do not supply three independent world-model replications.

H Plan-Time Scores and Decision Experiments

H.1 Imagined-rollout typicality

Execution-time surprise compares an imagined latent with an observation. Before execution, observations of the candidate future are unavailable. The plan-time score therefore measures a different property: how the imagined sequence compares with a reference of model rollouts driven by held-out demonstration actions under the same context and horizon.

Let $\widehat{z}_0, \dots, \widehat{z}_L$ be an imagined sequence, and let $\widehat{\mu}$, $\widehat{v}$, and $\widehat{\rho}$ be fitted per-dimension latent means, variances, and lag-one correlations. Componentwise standardization and first-order autoregressive, or AR(1), residual normalization give

$$\widehat{u}_t = \frac{\widehat{z}_t - \widehat{\mu}}{\sqrt{\widehat{v} + \epsilon}}, \qquad \widehat{\eta}_t = \frac{\widehat{u}_{t+1} - \widehat{\rho} \odot \widehat{u}_t}{\sqrt{1 - \widehat{\rho}^2}}, \qquad s_{\mathrm{dyn}} = \frac{1}{L} \sum_{t=0}^{L-1} \|\widehat{\eta}_t\|_2^2. \tag{24}$$

Here $\odot$ denotes componentwise multiplication, L is the number of imagined transitions, $\epsilon = 10^{-8}$, and correlations are clipped to $[-0.995, 0.995]$. Mean and variance normalization precede the AR residual. This is diagonal AR(1) normalization, not full spatial whitening or a proof of temporal independence or Gaussianity.

The sliced alternative projects $\widehat{\eta}_t$ onto 16 fixed-seed unit directions v_ℓ. For projection $x_t = v_\ell^\top \widehat{\eta}_t$, define its empirical characteristic function $\widehat{\varphi}_\ell(\omega) = L^{-1} \sum_t \exp(\mathrm{i}\omega x_t)$. The implemented Epps–Pulley-type discrepancy (Epps & Pulley, 1983) is

$$s_{\mathrm{EP}} = \frac{1}{16} \sum_{\ell=1}^{16} L \int_{-5}^{5} \left| \widehat{\varphi}_\ell(\omega) - e^{-\omega^2/2} \right|^2 \frac{e^{-\omega^2/2}}{\sqrt{2\pi}} \, \mathrm{d}\omega, \tag{25}$$

evaluated by trapezoidal integration on 65 equally spaced points. The frequency ω is an integration variable and i is the imaginary unit. The score compares projected innovations with a standard-normal characteristic function; it is used as a nonconformity statistic, not an independent-sample Gaussianity test.

This imagined-rollout diagnostic, denoted S2 in the result tables, estimates normalization statistics from a calibration split that includes its ranked reference trajectories. We therefore do not apply Proposition 1 to that experiment. Legal but unusual action sequences are controls for typicality sensitivity. Their elevated scores cannot establish physical impossibility of a candidate plan. A future validity claim would require a suitable independent fitting/reference design or another justified calibration construction.

H.2 Subgoal support and budgeted reachability

For subgoal latent z_{sg}, the support score $s_{\mathrm{knn}}(z_{\mathrm{sg}})$ is its Euclidean distance to the k_{NN}th nearest latent in a separate index. The index and k_{NN} are fixed before reference ranking. A reachability score uses the executor's cross-entropy-method (CEM) search,

$$\widehat{a}_{1:L} = \mathrm{CEM}_{\mathcal{B}}(z, z_{\mathrm{sg}}), \qquad s_{\mathrm{reach}}(z_{\mathrm{sg}}; z) = \|z_{\mathrm{sg}} - g^{(L)}(z, \widehat{a}_{1:L})\|_2. \tag{26}$$

Here z is the initial latent context, $g^{(L)}$ denotes the model's terminal imagined latent, and $\mathcal{B}$ fixes the search budget, action bounds, and sampling rule. The evaluated executor budget uses 300 candidates, 30 iterations, and 30 elites. The score is the residual of the returned action sequence, not a global minimum. A large residual may reflect model error or unsuccessful optimization and is not an infeasibility certificate.

Reference candidates are either true future demonstration latents or the learned proposer's own candidates at held-out contexts, under a matching horizon and search budget. The deployment-start audit instead draws one first proposal per random-start episode, using 600 calibration and 1,000 audit episodes. At a single query, the support/reachability rule rejects when $p_{\mathrm{knn}} \leq 0.025$ or $p_{\mathrm{reach}} \leq 0.025$. A union bound gives a 0.05 target only if both individual p-values have their required marginal validity under the query-generating process. Matching the proposer while changing its contexts does not establish this condition. No validity claim is extended to adaptively visited contexts or repeated resampling after rejections.

H.3 Reachability controls and valid labels

The geometric diagnostic proposes pairs of TwoRoom targets at Euclidean distances matched within 5% from the same context. Reachable labels require an executed witness within 0.8 of a 25-raw-step budget. Unreachable labels require a lower bound of at least 1.2 budgets for entering the target's two-radius tolerance ball. The lower bound respects coordinatewise speed limits, the complete permitted door bands, and all doors; it does not infer unreachability from a failed planner. Along a required door crossing, the implementation minimizes a Chebyshev path-length lower bound and subtracts a Lipschitz correction for its finite grid. The analysis retains 39 pairs with both labels certified. Comparators are initial latent distance, support distance, and initial-population random shooting; returned CEM plans are also executed.

H.4 Replanning and screening utility

Surprise-triggered replanning evaluates a loop-specific per-step calibrated rule at $\alpha = 0.1$, replanning when it alarms or the plan ends. The rule uses a reference scored with the same refresh of prediction context from observations, and its per-step p-values rank each residual among the pooled per-frame scores of that reference. Because these frames are dependent, the rule is an operational threshold with no calibrated per-step or episode-level false-alarm guarantee, unlike the fixed-horizon episode-max construction above. The primary comparison is against warm-started fixed-five-step replanning on matched episode specifications, using evaluation seeds 3–5 and 128 episodes per seed after development on seeds 0–2. Per-call CEM search budgets are identical across methods; warm starts retain the unexecuted suffix of the previous plan. Success, planner calls, and computation are reported together, with paired uncertainty for success. No adaptive-loop false-alarm guarantee is inferred from these measurements.

Subgoal screening rejects candidates using the support/reachability rule, subject to a resampling cap, and then falls back to a flat goal or selects the least-rejected candidate according to the specified variant. Demonstration-reference and proposer-reference screens are separate pipelines. A further comparison uses fixed sets of four proposals to keep the candidate budget matched across selection rules. These experiments ask whether the scores improve decisions under the tested proposer and compute budget; they do not convert support, typicality, or a CEM residual into a certificate of physical validity.

I Evidence Scope and the Broader Benchmark

I.1 Which study supports each conclusion

The historical benchmark establishes the breadth of the audit across checkpoints and environments. The focused studies then ask controlled questions within selected checkpoints. Development cohorts A, B, and C use 128, 256, and 512 intervention parents per environment; confirmation cohort D also uses 512. D follows inspection of C and has two new prespecified primary contrasts. These stages remain separate: larger numbers of rendered variants or repeated score evaluations do not increase the number of independent parent episodes. A parent and all of its descendants retain one fitting, validation, calibration, audit, or intervention role.

The final predictor and object contrasts use PushT's innovation-trained LeWM checkpoint. Upstream PushT, upstream TwoRoom, and PLDM comparisons delimit their scope; they are not independent replications of the same primary test. Each final environment has 600 calibration, 1,000 nominal-audit, and 512 intervention parents. The comparator and readout fits use 2,000 nominal parents and 400 validation parents. Three fitting seeds are averaged at the prediction level before forming a residual. Parent-bootstrap intervals condition on the checkpoint and fitted heads, and do not capture uncertainty across new world-model training runs.

Table 9: Evidence roadmap. Results are interpreted at the scope of their controlling comparison.

Finding	Controlling evidence	Inferential boundary
Nominal alarms	Historical audit and D audit tables	Locked thresholds; specified sampling populations
Reference sensitivity	Crossed-threshold bridge; fresh nested references	One TwoRoom checkpoint and open-loop score
Predictor contribution	D common-FPR primary and action-free controls	Original head versus fitted alternatives sharing its encoder
Object dependence	D within-parent teleport–colour interaction	Two PushT objects and specified recolourings
Recovery and masking	Matched readouts; temporal diagnostics	Tested readout families and onset-defined windows
Utility	Paired replanning and screening cohorts	Fixed planner/proposer implementations and budgets

I.2 Detection varies across models and environments

The broad open-loop results establish that the audit includes sensitive detectors. On TwoRoom, upstream LeWM and PLDM detect action mismatches at mean rates near 0.99. In PushT, upstream and innovation LeWM have action-mismatch rates 0.03 and 0.48, respectively. Cube LeWM detects many physical edits while alarming on almost every colour control. Sensitivity to an edit and selectivity for physical inconsistency are therefore distinct even before the paired appearance study.

Tables 10 and 11 separate physical and appearance outcomes to make that distinction visible. Both physical and appearance entries use the exported first-alarm-at-or-after-onset endpoint; premature alarms are counted separately in the underlying code. The appearance entries are therefore benign post-onset alarm rates, not full-episode any-alarm rates. Family means average the available checkpoint-level rates, with sample standard deviations where multiple checkpoints exist. They are descriptive aggregates over the benchmark severities and do not give a common-FPR comparison be-

tween families. The simulator-consistent occlusion control preserves lawful hidden motion, whereas the intervention changes the trajectory across the occlusion interval.

Table 10: Historical open-loop benchmark, target 0.05. Parentheses give checkpoint counts; ± denotes sample standard deviation across checkpoints, not a confidence interval. Entries are post-onset first-alarm rates; consistent occlusion is the lawful control. The collapsed TwoRoom innovation seed is excluded from these family means and described in the text.

Environment	Family (count)	Teleport	Interpen.	Action	Occlusion	Consistent
TwoRoom	LeWM up. (3)	0.81±0.07	–	0.99±0.01	0.75±0.11	0.02±0.01
TwoRoom	LeWM innov. (2)	0.40±0.41	–	0.54±0.63	0.46±0.58	0.44±0.59
TwoRoom	PLDM (3)	0.76±0.01	–	0.99±0.01	0.63±0.02	0.04±0.02
TwoRoom	DINO-WM (3)	0.38±0.05	–	0.69±0.08	0.67±0.08	0.65±0.07
PushT	LeWM up. (1)	0.09	0.07	0.03	0.07	0.04
PushT	LeWM innov. (1)	0.35	0.37	0.48	0.10	0.08
PushT	PLDM (3)	0.11±0.03	0.08±0.05	0.31±0.03	0.05±0.01	0.02±0.01
PushT	DINO-WM (3)	0.32±0.02	0.21±0.04	0.15±0.03	0.05±0.01	0.03±0.01
Cube	LeWM up. (1)	0.98	0.86	0.87	1.00	0.41
Cube	LeWM innov. (1)	0.96	0.80	0.88	0.99	0.24
Cube	PLDM (3)	0.63±0.04	0.21±0.01	0.42±0.01	0.33±0.04	0.05±0.03
Cube	DINO-WM (4)	0.17±0.02	0.07±0.03	0.96±0.02	0.07±0.01	0.04±0.01

Table 11: Historical open-loop benchmark, target 0.05. Parentheses give checkpoint counts; ± denotes sample standard deviation across checkpoints, not a confidence interval. Appearance entries count first alarms at or after the control onset; premature alarms are excluded. These cells do not use the later object-factorial design. The collapsed TwoRoom innovation seed is excluded from these family means and described in the text.

Environment	Family (count)	Colour	Lighting
TwoRoom	LeWM up. (3)	0.01±0.00	0.01±0.01
TwoRoom	LeWM innov. (2)	0.37±0.44	0.08±0.08
TwoRoom	PLDM (3)	0.17±0.26	0.02±0.00
TwoRoom	DINO-WM (3)	0.12±0.01	0.17±0.04
PushT	LeWM up. (1)	0.00	0.09
PushT	LeWM innov. (1)	0.12	0.10
PushT	PLDM (3)	0.01±0.00	0.04±0.01
PushT	DINO-WM (3)	0.02±0.00	1.00±0.00
Cube	LeWM up. (1)	0.99	0.84
Cube	LeWM innov. (1)	1.00	0.85
Cube	PLDM (3)	0.03±0.01	0.02±0.01
Cube	DINO-WM (4)	0.04±0.02	0.18±0.05

The upstream TwoRoom family includes checkpoint s3072, whose historical clean reference is the old reference in the crossed-threshold study (Table 17). Its nominal rate is 0.040 and its $180°$ recolouring rate is 0.002 on the original query collection. The benchmark colour cell (0.01 for this checkpoint) averages the three renderer severities 0.25, 0.5, and 1.0; the last is the $180°$ setting. That aggregate is not a subgroup-specific robustness result, and the bridge shows its sensitivity to the reference draw.

One TwoRoom innovation seed collapsed during training and was excluded from the family detection mean under the recorded exclusion rule, while remaining in the nominal audit. Its teleport/action detection rates are $0.06/0.08$, versus $0.69/0.99$ and $0.11/0.10$ for the two other reported seeds. Colour-alarm rates of those seeds are 0.01, 0.68, and 0.06, respectively. This heterogeneity cautions against treating the focused checkpoint as representative of all training runs.

I.3 Observable edits and interpretation of misses

An intervention in simulator state need not leave persistent visual evidence. Table 12 records whether its state effect is visible at the first observed frame, later, or never. In the historical PushT velocity cells, 35% never change the observed state; the remaining cells cannot be dismissed as unobservable. Action mismatches often first affect a later frame, whereas teleports are immediately observable almost everywhere. These strata describe the historical generator. The final PushT object generator instead records zero null teleports under the common 75-pixel displacement policy. Observability is diagnosed from state trajectories, not selected by detector scores.

Table 12: Historical benchmark observability. First/later/never describe when the intervention first changes the observed simulator state. Conditional detection is averaged across evaluated arms, conditional on ever-observable edits, with parent-clustered 95% intervals. Counts refer to edit records, not independent model replications.

Environment	Edit	n	First	Later	Never	Cond. detection
TwoRoom	Action	300	0.83	0.17	0.00	0.760 [0.745, 0.776]
TwoRoom	Occlusion	450	1.00	0.00	0.00	0.595 [0.571, 0.617]
TwoRoom	Consistent occl.	450	0.87	0.00	0.13	0.288 [0.276, 0.300]
TwoRoom	Teleport	750	1.00	0.00	0.00	0.560 [0.540, 0.581]
PushT	Action	300	0.78	0.22	0.00	0.223 [0.199, 0.249]
PushT	Interpenetration	450	1.00	0.00	0.00	0.164 [0.147, 0.182]
PushT	Occlusion	450	1.00	0.00	0.00	0.060 [0.050, 0.071]
PushT	Consistent occl.	450	0.86	0.00	0.14	0.035 [0.027, 0.043]
PushT	Teleport	750	1.00	0.00	0.00	0.215 [0.198, 0.233]
PushT	Velocity	450	0.49	0.15	0.35	0.035 [0.027, 0.044]
Cube	Action	300	0.78	0.22	0.00	0.749 [0.724, 0.774]
Cube	Interpenetration	450	1.00	0.00	0.00	0.287 [0.270, 0.304]
Cube	Occlusion	450	1.00	0.00	0.00	0.363 [0.349, 0.378]
Cube	Consistent occl.	450	0.89	0.00	0.11	0.114 [0.103, 0.126]
Cube	Teleport	600	1.00	0.00	0.00	0.500 [0.483, 0.517]
Cube	Velocity	450	0.68	0.20	0.11	0.053 [0.043, 0.064]

J Nominal Audits and Appearance Reference Mixtures

J.1 Historical and final-cohort intervals

All nominal-audit tables use the one-sided 95% Clopper–Pearson bound defined in Appendix B. Historical bounds are recomputed from the reported integer counts out of 1,000 episodes; thresholds and verdicts are unchanged. Twenty-five of 29 meet the 0.07 tolerance at target $\alpha = 0.05$. The largest passing count is $53/1{,}000$ (upper 0.06617), and the smallest inconclusive count is $57/1{,}000$ (upper 0.07056). These are separate pipeline audits, not simultaneous bounds across checkpoints. No threshold is replaced after auditing.

Table 13: Historical nominal audits, 600 reference and 1,000 audit parents per pipeline. Upper bounds are one-sided 95% Clopper–Pearson bounds recomputed from the reported counts. Verdicts use unrounded bounds and tolerance 0.07.

Environment	Checkpoint	Rate	Upper	Tolerance
TwoRoom	LeWM up. s0	0.031	0.042	Met
TwoRoom	LeWM up. s1	0.035	0.046	Met
TwoRoom	LeWM up. s3072	0.040	0.052	Met
TwoRoom	LeWM innov. s0	0.057	0.071	Inconclusive
TwoRoom	LeWM innov. s1	0.052	0.065	Met
TwoRoom	LeWM innov. s3072	0.032	0.043	Met
TwoRoom	PLDM s0	0.050	0.063	Met
TwoRoom	PLDM s1	0.042	0.054	Met

Continued on next page

Table 13 continued

Environment	Checkpoint	Rate	Upper	Tolerance
TwoRoom	PLDM s2	0.049	0.062	Met
TwoRoom	DINO-WM s0	0.048	0.061	Met
TwoRoom	DINO-WM s1	0.060	0.074	Inconclusive
TwoRoom	DINO-WM s2	0.041	0.053	Met
PushT	LeWM up.	0.038	0.050	Met
PushT	LeWM innov. s1	0.030	0.040	Met
PushT	PLDM s0	0.049	0.062	Met
PushT	PLDM s1	0.046	0.058	Met
PushT	PLDM s2	0.037	0.048	Met
PushT	DINO-WM s0	0.044	0.056	Met
PushT	DINO-WM s1	0.038	0.050	Met
PushT	DINO-WM s2	0.041	0.053	Met
Cube	LeWM up.	0.058	0.072	Inconclusive
Cube	LeWM innov.	0.059	0.073	Inconclusive
Cube	PLDM s0	0.034	0.045	Met
Cube	PLDM s1	0.049	0.062	Met
Cube	PLDM s2	0.025	0.035	Met
Cube	DINO-WM no-prop. s0	0.053	0.066	Met
Cube	DINO-WM s0	0.035	0.046	Met
Cube	DINO-WM s1	0.050	0.063	Met
Cube	DINO-WM s2	0.041	0.053	Met

In D, nine of the ten displayed PushT scores meet the clean-rendering tolerance; open-loop surprise does not. Seven of the ten TwoRoom scores are inconclusive; the two persistence scores and encoder change meet tolerance. These scores share one reference and audit draw, so their inconclusive outcomes are correlated and do not by themselves establish a common distribution shift (Appendix B). A calibrated-pipeline difference compares the entire operational procedures, including their realized false-positive rates. It does not isolate ranking quality at a shared rate.

The recoloured columns in Tables 14–15 are stress tests of thresholds calibrated on clean episodes. On PushT, one-step benign rates span 0.667–0.979 across the four object/timing conditions. On TwoRoom, one-step alarms occur on every recoloured audit episode, whereas the MLP's rates are 0.171 for abrupt recolouring and zero for constant recolouring. A low rate on a shifted subgroup alone does not establish useful detection there; the paired physical controls must be read alongside it.

Table 14: PushT innovation checkpoint, cohort D. Rates [one-sided 95% Clopper–Pearson upper bounds] on 1,000 no-physical-edit audit parents, using clean references of 600. Only the clean column matches the calibration population; recoloured columns are shifted-subgroup stress tests.

Score	Clean	Block abrupt	Block const.	Pusher abrupt	Pusher const.
Open-loop	0.067 [0.081]	0.336 [0.361]	0.305 [0.330]	0.514 [0.540]	0.770 [0.792]
WM	0.040 [0.052]	0.966 [0.975]	0.667 [0.692]	0.979 [0.986]	0.830 [0.849]
WM persist	0.028 [0.038]	0.510 [0.536]	0.847 [0.865]	0.513 [0.539]	0.828 [0.847]
Ridge+a	0.042 [0.054]	0.829 [0.848]	0.393 [0.419]	0.950 [0.961]	0.507 [0.533]
Ridge-a	0.045 [0.057]	0.743 [0.766]	0.333 [0.358]	0.896 [0.911]	0.378 [0.404]
MLP+a	0.046 [0.058]	0.981 [0.988]	0.770 [0.792]	0.988 [0.993]	0.839 [0.858]
MLP-a	0.044 [0.056]	0.843 [0.862]	0.418 [0.444]	0.943 [0.955]	0.417 [0.443]
MLP persist	0.045 [0.057]	0.716 [0.739]	0.907 [0.922]	0.532 [0.558]	0.841 [0.860]
Encoder	0.047 [0.060]	0.576 [0.602]	0.056 [0.069]	0.787 [0.808]	0.140 [0.159]
Pixel	0.036 [0.047]	0.015 [0.023]	0.000 [0.003]	0.033 [0.044]	0.029 [0.039]

Table 15: TwoRoom upstream checkpoint, cohort D. Rates [one-sided 95% Clopper–Pearson upper bounds] on 1,000 no-physical-edit audit parents, using clean references of 600. Only the clean column matches the calibration population; recoloured columns are shifted-subgroup stress tests.

Score	Clean	Agent abrupt	Agent constant
Open-loop	0.072 [0.087]	1.000 [1.000]	1.000 [1.000]
WM	0.057 [0.071]	1.000 [1.000]	1.000 [1.000]
WM persist	0.048 [0.061]	1.000 [1.000]	1.000 [1.000]
Ridge+a	0.060 [0.074]	0.041 [0.053]	0.000 [0.003]
Ridge-a	0.065 [0.079]	0.044 [0.056]	0.000 [0.003]
MLP+a	0.057 [0.071]	0.171 [0.192]	0.000 [0.003]
MLP-a	0.074 [0.089]	0.050 [0.063]	0.000 [0.003]
MLP persist	0.047 [0.060]	0.032 [0.043]	0.000 [0.003]
Encoder	0.055 [0.068]	0.028 [0.038]	0.000 [0.003]
Pixel	0.057 [0.071]	0.021 [0.030]	0.000 [0.003]

J.2 A REFERENCE MIXTURE DOES NOT GUARANTEE EVERY APPEARANCE SUBGROUP

An earlier study calibrates on a random mixture with probability 0.5 of clean rendering and otherwise a 45° or 90° recolouring onset, equally likely. Its direct-mixture audit follows that same law. Table 16 also reports its component and held-out appearance groups. For the PushT one-step score, the mixture reference raises the threshold from approximately 84 to 340 residual units, reducing the seen-recolouring rate from 0.910 to 0.105 and clean nominal alarms to zero. The higher threshold can also remove physical alarms; the reported one-step comparison does not quantify that detection cost, so the benign-alarm reduction alone does not establish improved physical selectivity. The direct-mixture audit is nevertheless inconclusive at 0.058 with upper bound 0.072, and the unseen 180° rate is 0.170. On TwoRoom the corresponding one-step mixture audit meets tolerance, but the unseen-recolouring rate is 0.694. These outcomes are consistent with a mixture-level target controlling a weighted average rather than every subgroup.

Table 16: Earlier A/B study, separate from cohort D. Rates [one-sided 95% Clopper–Pearson upper bounds] on 1,000 audit parents. A random mixture combines clean and seen-recolouring episodes; the direct-mixture column matches that reference law. Seen recolourings use 45 or 90 degrees; the 180-degree condition is held out from that reference.

Environment	Score	Reference	Clean	Mixture	Seen	180 deg.
PushT	Open-loop	clean	0.044 [0.056]	0.065 [0.079]	0.068 [0.083]	0.250 [0.274]
PushT	Open-loop	mixture	0.038 [0.050]	0.048 [0.061]	0.048 [0.061]	0.194 [0.216]
PushT	One-step	clean	0.044 [0.056]	0.485 [0.512]	0.910 [0.924]	0.971 [0.979]
PushT	One-step	mixture	0.000 [0.003]	0.058 [0.072]	0.105 [0.122]	0.170 [0.191]
TwoRoom	Open-loop	clean	0.037 [0.048]	0.096 [0.113]	0.144 [0.164]	0.824 [0.844]
TwoRoom	Open-loop	mixture	0.036 [0.047]	0.041 [0.053]	0.048 [0.061]	0.456 [0.482]
TwoRoom	One-step	clean	0.034 [0.045]	0.464 [0.490]	0.897 [0.912]	1.000 [1.000]
TwoRoom	One-step	mixture	0.015 [0.023]	0.044 [0.056]	0.090 [0.106]	0.694 [0.718]

The later C reference draw produces seven inconclusive clean-reference audits among ten PushT scores and three among ten TwoRoom scores, according to the recorded full audit. Its PushT one-step and MLP rates are 0.073 and 0.056, unlike D's 0.040 and 0.046. Those changes motivate reporting reference sensitivity. An earlier frame-pooled dataset calibrator is superseded because dependent frames are not exchangeable calibration episodes; Appendix B gives the counterexample and the episode-level guarantee. The surprise-triggered replanning rule, by design, pools per-frame reference scores and carries no calibration guarantee (Appendix H).

K Reference Sensitivity and the Frozen Forecast

K.1 Crossing references while holding scoring fixed

The bridge uses upstream TwoRoom LeWM (s3072), open-loop residuals, and the episode maximum over observed frames 3–20. Re-scoring the earlier stored episodes through the later exporter gives zero score discrepancy. Table 17 applies both clean-reference thresholds to both query collections. Within the newer collection, recolouring rates are 0.005 and 0.824; within the older collection they are 0.002 and 0.842. Each collection contains 1,000 clean/recoloured pairs. Original nominal audits are 0.040 for the old threshold on the old collection and 0.037 for the new threshold on the new collection. These are distinct from the crossed nominal comparisons. Holding queries and scoring fixed isolates the threshold's effect without requiring a change in physics or score computation.

Table 17: Crossed reference thresholds and query collections. All entries are episode alarm rates; both thresholds use 600 clean references. The headline and Figure 1b compare the two rows within the newer query collection.

Threshold	Old clean	Old recoloured	New clean	New recoloured
306.8 (old reference)	0.040	0.002	0.034	0.005
246.2 (new reference)	0.046	0.842	0.037	0.824

K.2 Why the nominal and shifted distributions both matter

Equations equation 10, equation 4, and equation 13 define the forecast: nominal quantile variation is mapped through the recoloured score distribution to $A_Q(q_\alpha)$. The density ratio in Eq. equation 13 identifies the relevant combination, a sparse nominal region occupied densely by shifted scores. Heavy tails alone do not imply this behaviour.

The nominal CDF estimate uses a separate 4,300-rollout reservoir and the shifted CDF uses an earlier recoloured audit. The forecast was frozen before any new reference or audit parent was scored; the parent trajectories had already been generated. The subsequent 9,600 reference parents form four disjoint blocks of 2,400; each contains four 600-parent sets and sixteen 150-parent sets. Thus the 84 sets are disjoint at a fixed size but nested across sizes, and all use the same 1,000 audit parents. Pooling them as 84 independent replications would exaggerate the evidence. The number of independent large blocks is four.

Table 18: Frozen forecast versus fresh references. A shifted rate is the fraction of the shared recoloured audit that alarms. Observed fractions use counts of disjoint references at each size; forecasts are point probabilities fixed before the fresh parents were scored.

n	Sets	Fraction with shifted rate > 0.5		Fraction with rate < 0.05	
		Observed	Forecast	Observed	Forecast
150	64	$32/64$	0.59	$30/64$	0.39
600	16	$9/16$	0.855	$7/16$	0.118
2,400	4	$3/4$	0.987	$1/4$	0.007

Table 19: Threshold summaries in latent residual units and mean shifted rates. The four observed thresholds at $n = 2{,}400$ are individual values, not percentiles. The forecast quantiles refer to the reference-sampling distribution under the frozen empirical CDFs.

n	Observed threshold 5/50/95%	Forecast 5/50/95%	Mean rate obs./forecast
150	90 / 254 / 383	79 / 208 / 376	0.51 / 0.58
600	112 / 238 / 343	105 / 157 / 316	0.55 / 0.85
2,400	152, 174, 243, 292 (all four)	127 / 153 / 228	0.72 / 0.98

K.3 What the point-forecast discrepancy identifies

The observed threshold median at $n = 600$ is 238 rather than the forecast 157. The old and new recoloured $5/25/50/75/95$ percentiles are approximately $234/251/262/272/284$ and $236/251/263/274/286$, respectively. Nominal mass at or above 200 is 4.5% in the reservoir and 5.1% in the pooled new references, ranging from 4.6% to 6.0% by block. Only about 1.5–2.3% lies between 100 and 200. In this region, modest probability differences have substantial quantile consequences.

Replacing the nominal CDF estimate with the pooled new records yields post hoc forecast probabilities 0.46, 0.65, and 0.81 for a shifted rate above one-half at sizes 150, 600, and 2,400. This improves agreement but is not independent confirmation, since the records helped diagnose the discrepancy. The frozen point forecast overestimates the fraction of high-alarm references at $n = 600$. As an exploratory count check, treating the 16 disjoint references as independent and the logged forecast probabilities as fixed gives

$$\Pr\{\mathrm{Bin}(16, 0.855) \le 9\} \simeq 0.0046, \qquad \Pr\{\mathrm{Bin}(16, 0.118) \ge 7\} \simeq 0.0014. \tag{27}$$

These one-sided tail probabilities use rounded logged inputs. They are two dependent descriptions of the same references, and ignore estimation error in both CDFs and uncertainty from the shared audit; they are not tests of the full predictive distribution. At $n = 150$, analogous checks give approximately 0.091 and 0.123. At $n = 2{,}400$, observing at most three high-alarm references gives approximately 0.051, based on only four blocks. The count discrepancy is therefore clearest at $n = 600$ under the fixed-input approximation. Assessing the forecast with CDF-estimation uncertainty would require the frozen ECDF bootstrap described in Appendix D, which this submission does not analyse (Appendix B.5); our conclusions concern the frozen point forecast only. The observed sensitivity itself does not establish drift or a general failure of larger references.

L Prediction Baselines and Operating-Point Ablations

L.1 Fitting, convergence, and the action control

The final comparator receives $[z_{t-2}, z_{t-1}, z_t, a_{t-2}, a_{t-1}, a_t, 1]$ and predicts $z_{t+1} - z_t$. The action-free comparator is fitted separately with its action features omitted. The original predictor and comparators therefore share observed history, commanded actions, and the frozen encoder; none receives the physical onset or applied-action intervention labels. Full feature standardization, ridge selection, and MLP stopping rules are in Appendix E.

The PushT action-aware MLP validation errors are 6.80, 6.91, and 6.80 for the three fits, stopping after 142, 142, and 157 epochs. The action-free errors are 12.66, 13.03, and 12.85, and the action-aware ridge error is 17.17. On TwoRoom, the action-aware MLP errors remain 79.8–81.7 at the 300-epoch cap, compared with approximately 169.7 without actions and 161.79 for ridge with actions. The TwoRoom MLP therefore had not converged when fitting stopped, and the reported TwoRoom head advantage is conditional on this fitting limit.

L.2 Operational comparisons retain their realized audit rates

Tables 20 and 21 report the independently calibrated pipelines. On PushT, the original predictor minus MLP has a timely difference of -2.5 points $[-5.1, 0.0]$, with 16 discordant pairs favouring the original and 29 favouring the MLP (exact McNemar $p = 0.072$). Its detected difference is -2.7 points $[-5.1, -0.4]$, with 12 versus 26 discordant pairs ($p = 0.034$). These secondary 95% intervals condition on the fitted calibration thresholds. They answer a different question from the primary comparison, which re-estimates audit-based operating points in each bootstrap replicate.

On TwoRoom, the corresponding timely difference favours the original predictor by 4.9 points $[2.5, 7.2]$, with 30 versus five discordant pairs. The separately reported teleport difference is 2.5 points $[0.8, 4.5]$. Both one-step and MLP nominal audits are inconclusive. At 5% empirical nominal FPR, TwoRoom has an original-minus-MLP timely difference of $+5.1$ points, 95% interval $[+2.9, +7.4]$. Upstream PushT instead gives -6.3 points, 97.5% interval $[-10.4, -3.5]$, with timely rates $0.049/0.111$ and any-alarm AUROCs $0.608/0.724$. Both heads are weak on this upstream

checkpoint. The upstream PushT predictor and constant-object contrasts were prespecified as secondary; we retain their more conservative reported 97.5% intervals rather than relabel them as 95%.

Table 20: PushT innovation checkpoint, clean action mismatch at 90 degrees, cohort D. Each score uses its own calibrated threshold. Nominal rates [CP upper] use 1,000 audit parents; intervention and unedited-twin rates use the separate 512-parent cohort. Intervention and twin rates are shown to two decimals; nominal rates and upper bounds use three.

Score	Nominal [upper]	Timely	Detected	Premature	Twin timely/any
WM	0.040 [0.052]	0.39	0.79	0.02	0.01 / 0.04
MLP+a	0.046 [0.058]	0.41	0.82	0.02	0.01 / 0.06
MLP-a	0.044 [0.056]	0.11	0.28	0.01	0.01 / 0.05
Ridge+a	0.042 [0.054]	0.10	0.32	0.01	0.01 / 0.04
Open-loop	0.067 [0.081]	0.06	0.47	0.00	0.01 / 0.05
Pixel	0.036 [0.047]	0.02	0.08	0.01	0.01 / 0.04
WM persist	0.028 [0.038]	0.04	0.68	0.02	0.01 / 0.04

Table 21: TwoRoom upstream checkpoint, clean action mismatch at 90 degrees, cohort D. Each score uses its own calibrated threshold. Nominal rates [CP upper] use 1,000 audit parents; intervention and unedited-twin rates use the separate 512-parent cohort. Intervention and twin rates are shown to two decimals; nominal rates and upper bounds use three.

Score	Nominal [upper]	Timely	Detected	Premature	Twin timely/any
WM	0.057 [0.071]	0.99	0.99	0.01	0.00 / 0.04
MLP+a	0.057 [0.071]	0.94	0.97	0.02	0.01 / 0.06
MLP-a	0.074 [0.089]	0.01	0.03	0.06	0.01 / 0.07
Ridge+a	0.060 [0.074]	0.02	0.05	0.05	0.01 / 0.07
Open-loop	0.072 [0.087]	0.98	0.99	0.01	0.01 / 0.05
Pixel	0.057 [0.071]	0.01	0.03	0.04	0.01 / 0.06
WM persist	0.048 [0.061]	0.06	0.97	0.03	0.01 / 0.05

L.3 COMMON-FPR CURVES AND UNCERTAINTY

For an empirical target β and m finite nominal maxima sorted as $v_{(1)} \leq \cdots \leq v_{(m)}$, the implementation selects $v_{(m-\lfloor \beta m \rfloor + 1)}$ when $\lfloor \beta m \rfloor > 0$ and uses strict exceedance. Thus at $m = 1{,}000$ with distinct maxima, a target 0.05 yields a strict rate 0.049 and weak-exceedance bracket endpoint 0.050. No interpolation is used. Thresholds are fitted on audit records specifically for this descriptive comparison; those same records cannot supply a new independent audit of the resulting threshold.

The primary uncertainty calculation resamples the 1,000 nominal-audit parents and 512 intervention parents independently, while preserving paired detector outputs within each population. Each replicate re-estimates both thresholds. The resulting 97.5% interval, $[-7.8, +6.3]$ percentage points, is wider than the fixed-calibration interval and covers differences in either direction. The original head therefore shows no resolved advantage over the MLP at this operating point: the data establish neither superiority nor equivalence, and the point estimate favours the MLP.

Table 22: PushT innovation checkpoint, cohort D, clean 90-degree action mismatch. Empirical operating points are estimated on nominal audit parents; brackets show strict and weak exceedance rates at the selected threshold. No interpolation is performed. Timely rates can decrease when a lower threshold moves the first alarm before onset.

Score	Target FPR	Bracket	Timely	Detected	Premature	Twin any
WM	0.01	[0.009, 0.010]	0.21	0.57	0.01	0.01
WM	0.02	[0.019, 0.020]	0.29	0.70	0.01	0.02

Continued on next page

Table 22 continued

Score	Target FPR	Bracket	Timely	Detected	Premature	Twin any
WM	0.03	[0.029, 0.030]	0.35	0.77	0.01	0.04
WM	0.05	[0.049, 0.050]	0.40	0.80	0.02	0.05
WM	0.075	[0.074, 0.075]	0.44	0.84	0.04	0.10
WM	0.1	[0.099, 0.100]	0.48	0.84	0.04	0.12
MLP+a	0.01	[0.009, 0.010]	0.15	0.42	0.00	0.01
MLP+a	0.02	[0.019, 0.020]	0.22	0.61	0.00	0.03
MLP+a	0.03	[0.029, 0.030]	0.30	0.73	0.01	0.04
MLP+a	0.05	[0.049, 0.050]	0.43	0.82	0.02	0.06
MLP+a	0.075	[0.074, 0.075]	0.49	0.86	0.03	0.09
MLP+a	0.1	[0.099, 0.100]	0.53	0.86	0.04	0.12
Ridge+a	0.01	[0.009, 0.010]	0.03	0.09	0.00	0.01
Ridge+a	0.02	[0.019, 0.020]	0.05	0.15	0.00	0.02
Ridge+a	0.03	[0.029, 0.030]	0.08	0.26	0.01	0.03
Ridge+a	0.05	[0.049, 0.050]	0.11	0.34	0.01	0.05
Ridge+a	0.075	[0.074, 0.075]	0.13	0.42	0.02	0.07
Ridge+a	0.1	[0.099, 0.100]	0.15	0.50	0.03	0.10
Open-loop	0.01	[0.009, 0.010]	0.01	0.18	0.00	0.01
Open-loop	0.02	[0.019, 0.020]	0.04	0.26	0.00	0.01
Open-loop	0.03	[0.029, 0.030]	0.04	0.31	0.00	0.02
Open-loop	0.05	[0.049, 0.050]	0.05	0.40	0.00	0.04
Open-loop	0.075	[0.074, 0.075]	0.07	0.51	0.01	0.05
Open-loop	0.1	[0.099, 0.100]	0.10	0.66	0.02	0.08

Table 23: Any-alarm AUROC for PushT innovation, using physical episode maxima versus the separate nominal audit population. This unpaired, threshold-free statistic ignores onset timing and differs from paired ordering and AUROC against intervention twins.

Score	AUROC
WM	0.962
MLP+a	0.968
Ridge+a	0.813
Open-loop	0.917

Table 24: PushT upstream checkpoint, cohort D, clean 90-degree action mismatch. Empirical operating points are estimated on nominal audit parents; brackets show strict and weak exceedance rates at the selected threshold. No interpolation is performed. Timely rates can decrease when a lower threshold moves the first alarm before onset.

Score	Target FPR	Bracket	Timely	Detected	Premature	Twin any
WM	0.01	[0.009, 0.010]	0.02	0.05	0.01	0.01
WM	0.02	[0.019, 0.020]	0.03	0.08	0.02	0.03
WM	0.03	[0.029, 0.030]	0.04	0.09	0.02	0.04
WM	0.05	[0.049, 0.050]	0.05	0.11	0.04	0.06
WM	0.075	[0.074, 0.075]	0.06	0.14	0.06	0.08
WM	0.1	[0.099, 0.100]	0.07	0.16	0.08	0.12
MLP+a	0.01	[0.009, 0.010]	0.06	0.14	0.01	0.01
MLP+a	0.02	[0.019, 0.020]	0.07	0.16	0.01	0.03
MLP+a	0.03	[0.029, 0.030]	0.09	0.20	0.01	0.05
MLP+a	0.05	[0.049, 0.050]	0.11	0.24	0.02	0.07
MLP+a	0.075	[0.074, 0.075]	0.15	0.29	0.04	0.11
MLP+a	0.1	[0.099, 0.100]	0.18	0.33	0.05	0.15
Ridge+a	0.01	[0.009, 0.010]	0.03	0.07	0.01	0.02
Ridge+a	0.02	[0.019, 0.020]	0.03	0.09	0.01	0.03
Ridge+a	0.03	[0.029, 0.030]	0.04	0.12	0.01	0.05

Continued on next page

Table 24 continued

Score	Target FPR	Bracket	Timely	Detected	Premature	Twin any
Ridge+a	0.05	[0.049, 0.050]	0.05	0.13	0.01	0.05
Ridge+a	0.075	[0.074, 0.075]	0.06	0.14	0.03	0.09
Ridge+a	0.1	[0.099, 0.100]	0.08	0.16	0.06	0.14
Open-loop	0.01	[0.009, 0.010]	0.00	0.02	0.01	0.01
Open-loop	0.02	[0.019, 0.020]	0.00	0.03	0.01	0.03
Open-loop	0.03	[0.029, 0.030]	0.01	0.06	0.01	0.04
Open-loop	0.05	[0.049, 0.050]	0.02	0.07	0.02	0.08
Open-loop	0.075	[0.074, 0.075]	0.04	0.11	0.02	0.12
Open-loop	0.1	[0.099, 0.100]	0.04	0.12	0.02	0.14

Table 25: Any-alarm AUROC for PushT upstream, using physical episode maxima versus the separate nominal audit population. This unpaired, threshold-free statistic ignores onset timing and differs from paired ordering and AUROC against intervention twins.

Score	AUROC
WM	0.608
MLP+a	0.724
Ridge+a	0.564
Open-loop	0.501

The original predictor has higher timely point estimates at PushT innovation targets 0.01, 0.02, and 0.03, while the MLP is higher at 0.05, 0.075, and 0.10. Only the 0.05 contrast has joint-bootstrap uncertainty; the other operating points do not establish resolved differences. Any-alarm AUROCs are $0.962/0.968$ for the original predictor/MLP. This crossing limits any uniform ranking of the heads. AUROC is a separate, onset-agnostic comparison of physical episode maxima against nominal-audit maxima. It neither summarizes alarm timing nor replaces the paired object-ordering statistic.

L.4 Head cost versus shared encoding cost

The architecture microbenchmark uses an RTX 3060, 20 warm-up iterations, 200 timed iterations, and CUDA synchronization. On PushT at batch size one, the shared encoder takes 2.25 ms per frame, the world-model predictor head 1.08 ms, a single MLP head 0.04 ms, and the ensemble 0.12 ms. At batch size 64 these are 60.61, 3.12, 0.04, and 0.12 ms per batch. The action encoder is timed separately ($0.04/0.05$ ms for batches 1/64). TwoRoom gives closely similar figures. These are component timings; the MLP microbenchmark instantiates the fitted architecture and does not time an end-to-end monitor with feature preparation. The shared encoding cost prevents interpreting the head ratio as an end-to-end speedup.

L.5 Persistence changes the latency–selectivity trade-off

The persistence score uses $S_{\text{persist}} = \max_{t=5}^{20} \min\{s_{t-2}, s_{t-1}, s_t\}$ and its own episode-level threshold. An alarm requires three consecutive residuals above that threshold. In D, PushT benign abrupt-block alarms fall from 0.966 to 0.510, but constant-block alarms rise from 0.667 to 0.847 (Table 26). Since each three-frame minimum is no larger than the original episode maximum, calibration on the same parents produces a threshold no larger than the original threshold. A sustained recolouring residual can therefore cross the persistence threshold even when it did not cross the original threshold.

The action rotation starts at raw step $5j$, leaving observed state j unchanged; its first possible effect is at observed frame $j + 1$. Three affected residuals are therefore available only at $j + 3$, outside the timely window $[j, j + 2]$. The timely drop from about 0.39 to 0.04 largely reflects this timing constraint. A timely persistence crossing can still occur when pre-effect residuals already exceed its threshold, but cannot be attributed to three affected frames. Full-horizon detection, which falls from about 0.79 to 0.68, is the more informative operational cost.

The separate C study gives clean-teleport detection 0.94 for one-step scoring and 0.09 for persistence, with both nominal audits inconclusive. A single-frame teleport spike cannot by itself satisfy the persistence rule. Its action-mismatch median delay rises from two to three observed frames. These results characterize the selected statistic and latency window; they do not establish that persistence generally preserves physical detection.

Table 26: Reference expansion and persistence, separate studies. Values are rates; brackets give one-sided 95% Clopper–Pearson upper bounds. A: the mixture combines clean $(1/2)$, 45° $(1/4)$, and 90° $(1/4)$ renderings; its direct audit matches only the mixture reference. B: PushT innovation checkpoint, clean reference and separately calibrated scores. Audit/appearance rates use 1,000 parents; timely detection uses 512. The mixture panel reports benign rates, not its physical-detection cost.

A. Reference mixture, earlier A/B study; one-step score				
Environment	Reference	Direct-mixture [upper]	Seen colour	Held-out 180°
PushT	Clean	0.485 $[0.512]$	0.910	0.971
PushT	Mixture	0.058 $[0.072]$	0.105	0.170
TwoRoom	Mixture	0.044 $[0.056]$	0.090	0.694
B. Three-frame persistence, PushT cohort D				
Score	Nominal [upper]	Abrupt block	Constant block	Timely action
One-step	0.040 $[0.052]$	0.966	0.667	0.39
Persistence	0.028 $[0.038]$	0.510	0.847	0.04

M Object, Readout, and Temporal Diagnostics

M.1 Matched appearance-only alarms and physical increments

The earlier cohort C uses 512 PushT parents and a one-step residual threshold calibrated before intervention evaluation. In this study the recoloured object is the block. For each appearance condition, physical and no-physical twins share their parent and rendering. A timely alarm means that the first threshold crossing occurs between the physical onset j and $j + 2$, using the same evaluation window for the appearance-only twin. The paired increment is the physical timely rate minus the twin timely rate; it equals the number of new timely alarms minus the number of lost timely alarms, divided by 512. All variants of a parent are carried together when bootstrapping. These are secondary 95% intervals, not the two final-cohort primary tests.

Table 27: Appearance-only controls in cohort C, PushT innovation checkpoint, 90° action mismatch and block recolouring. The same one-step threshold is used throughout. Rates and intervals retain the exported two-decimal precision. New/lost counts refer to timely-alarm status after adding the physical edit. The clean nominal audit is inconclusive at 0.073, with Clopper–Pearson upper bound 0.088.

Rendering	Twin timely	Physical timely	Increment [95%]	New / lost
Clean	0.03	0.44	0.41 $[0.37, 0.46]$	215 / 4
Abrupt block 180°	0.94	0.95	0.01 $[0.00, 0.02]$	6 / 1
Constant block 180°	0.09	0.21	0.12 $[0.09, 0.14]$	64 / 5

The abrupt-recolouring cell illustrates saturation of the binary endpoint: most twins already alarm in the onset window, leaving little room for a further increase. Its combined full-horizon detected rate is 0.96, which alone does not attribute the alarms to the physical edit. In the constant-colour cell, whole-episode benign alarms occur at rate 0.81 even though only 0.09 are timely; many first crossings precede the physical onset. Thus timely, detected, and any-alarm outcomes answer different questions. The later D object study and its threshold-free paired ordering provide separate confirmation evidence and are not pooled with these historical rates.

M.2 THE FACTORIAL OBJECT TEST

The core PushT design crosses four physical conditions (none, pusher teleport, block teleport, action mismatch) with five renderings (clean, abrupt pusher recolouring, constant pusher recolouring, abrupt block recolouring, constant block recolouring). The moderate-colour subset uses 256 prespecified parents. Abrupt recolouring starts at the physical onset; constant recolouring starts at frame 0 and changes the predictor's observed context as well as its targets. Teleports use a common 75-pixel policy with geometric feasibility checks; appearance variants retain identical underlying states and commanded actions.

The other-minus-same contrast is defined in Eq. equation 18. All four outcomes of a parent stay together in each bootstrap replicate. Of 512 parents, 324 contribute a positive difference, 22 a negative difference, and 166 zero. Average other-object ordering is 0.878, compared with 0.510 for same-object recolouring.

Table 28: PushT innovation checkpoint, cohort D. P/P and B/B mean pusher/block teleports with the same object recoloured; P/B and B/P recolour the other object. I is the average other-minus-same ordering. Constant and abrupt rows use 512 parents; 90-degree rows use 256. Intervals are 97.5% for constant WM contrasts and 95% otherwise; only the innovation WM contrast is a study primary.

Score	Colour	P/P	B/B	P/B	B/P	I [interval]
WM	Constant	0.395	0.625	0.797	0.959	0.368 [0.332, 0.405] (97.5%)
WM	Abrupt	0.518	0.393	0.631	0.869	0.295 [0.255, 0.334] (95%)
WM	90 deg.	0.520	0.441	0.836	0.922	0.398 [0.352, 0.445] (95%)
Open-loop	Constant	0.391	0.539	0.676	0.562	0.154 [0.115, 0.192] (95%)
Open-loop	Abrupt	0.590	0.447	0.658	0.844	0.232 [0.194, 0.269] (95%)
Open-loop	90 deg.	0.348	0.426	0.645	0.465	0.168 [0.115, 0.223] (95%)
MLP+a	Constant	0.395	0.633	0.809	0.959	0.370 [0.338, 0.402] (95%)
MLP+a	Abrupt	0.523	0.400	0.633	0.861	0.285 [0.247, 0.323] (95%)
MLP+a	90 deg.	0.500	0.465	0.812	0.914	0.381 [0.336, 0.426] (95%)
Pixel	Constant	0.631	0.986	0.632	0.988	0.001 [-0.007, 0.011] (95%)
Pixel	Abrupt	0.643	0.986	0.762	0.988	0.061 [0.035, 0.085] (95%)
Pixel	90 deg.	0.631	0.992	0.648	0.992	0.009 [-0.005, 0.023] (95%)

Table 29: PushT upstream checkpoint, cohort D. P/P and B/B mean pusher/block teleports with the same object recoloured; P/B and B/P recolour the other object. I is the average other-minus-same ordering. Constant and abrupt rows use 512 parents; 90-degree rows use 256. Intervals are 97.5% for constant WM contrasts and 95% otherwise; only the innovation WM contrast is a study primary.

Score	Colour	P/P	B/B	P/B	B/P	I [interval]
WM	Constant	0.408	0.676	0.498	0.920	0.167 [0.128, 0.205] (97.5%)
WM	Abrupt	0.445	0.467	0.434	0.932	0.227 [0.191, 0.262] (95%)
WM	90 deg.	0.406	0.605	0.645	0.914	0.273 [0.225, 0.318] (95%)
Open-loop	Constant	0.473	0.463	0.393	0.615	0.036 [-0.003, 0.074] (95%)
Open-loop	Abrupt	0.510	0.398	0.428	0.689	0.104 [0.067, 0.142] (95%)
Open-loop	90 deg.	0.422	0.438	0.516	0.582	0.119 [0.062, 0.174] (95%)
MLP+a	Constant	0.424	0.559	0.467	0.932	0.208 [0.171, 0.244] (95%)
MLP+a	Abrupt	0.484	0.461	0.424	0.955	0.217 [0.182, 0.252] (95%)
MLP+a	90 deg.	0.445	0.602	0.625	0.922	0.250 [0.203, 0.297] (95%)
Pixel	Constant	0.631	0.986	0.632	0.988	0.001 [-0.007, 0.011] (95%)
Pixel	Abrupt	0.643	0.986	0.762	0.988	0.061 [0.035, 0.085] (95%)
Pixel	90 deg.	0.631	0.992	0.648	0.992	0.009 [-0.005, 0.023] (95%)

The moderate-colour and alternative-predictor rows show that the interaction is not limited to a single hue magnitude or the original prediction head. The pixel control constrains a simpler explanation in which the physical edit disappears from the rendering. It does not establish equal

perceptual salience: the pusher and block differ in size and base appearance, and the originally grey block is assigned a saturated colour. Nor does the interaction claim that recolouring the other object leaves every score unchanged. Upstream PushT pusher-teleport ordering under block recolouring, for example, is only 0.498.

M.3 Full paired-ordering grids

The following tables retain strict wins, losses, and ties as well as ordering. A value of 0.5 can arise from many ties or from balanced wins and losses; it is not evidence that the underlying score distributions are equal. A value below 0.5 means the physical twin more often receives the lower window maximum.

The later-window column uses the three observed frames $j+1$ through $j+3$. The dedicated temporal diagnostic below instead uses five frames, $j+1$ through $j+5$. Per-cell intervals are descriptive and are not adjusted for the full grid of secondary comparisons.

Table 30: PushT innovation checkpoint, cohort D. Ordering compares physical and no-physical twins; $>$ / $<$ / $=$ are strict wins/losses/exact ties. Onset ordering uses j through $j+2$ and has 95% parent-bootstrap intervals. The later statistic uses $j+1$ through $j+3$, distinct from the dedicated five-frame diagnostic. AUROC uses full episode maxima against no-physical twins, not the nominal audit.

Score	Edit	Appearance	Ordering [95%]	W/L/T	Later (3)	AUROC
Open-loop	Action	B abr.	0.646 [0.61, 0.69]	331/181/0	0.689	0.678
Open-loop	Action	B const.	0.690 [0.65, 0.73]	353/158/1	0.746	0.688
Open-loop	Action	clean	0.938 [0.92, 0.96]	480/32/0	0.951	0.931
Open-loop	Action	P abr.	0.561 [0.52, 0.60]	286/224/2	0.561	0.592
Open-loop	Action	P const.	0.395 [0.35, 0.44]	201/309/2	0.389	0.412
Open-loop	Block	B abr.	0.447 [0.40, 0.49]	229/283/0	0.420	0.520
Open-loop	Block	B const.	0.539 [0.50, 0.58]	276/236/0	0.514	0.526
Open-loop	Block	B 90 deg.	0.426 [0.37, 0.49]	109/147/0	0.410	0.465
Open-loop	Block	clean	0.949 [0.93, 0.97]	486/26/0	0.928	0.910
Open-loop	Block	P abr.	0.844 [0.81, 0.88]	432/80/0	0.812	0.704
Open-loop	Block	P const.	0.562 [0.52, 0.61]	288/224/0	0.543	0.505
Open-loop	Block	P 90 deg.	0.465 [0.40, 0.53]	119/137/0	0.445	0.479
Open-loop	Pusher	B abr.	0.658 [0.62, 0.70]	337/175/0	0.650	0.592
Open-loop	Pusher	B const.	0.676 [0.63, 0.71]	346/166/0	0.658	0.575
Open-loop	Pusher	B 90 deg.	0.645 [0.59, 0.70]	165/91/0	0.656	0.526
Open-loop	Pusher	clean	0.957 [0.94, 0.97]	490/22/0	0.945	0.873
Open-loop	Pusher	P abr.	0.590 [0.55, 0.63]	302/210/0	0.615	0.625
Open-loop	Pusher	P const.	0.391 [0.35, 0.43]	200/312/0	0.396	0.431
Open-loop	Pusher	P 90 deg.	0.348 [0.29, 0.41]	89/167/0	0.352	0.465
WM	Action	B abr.	0.547 [0.53, 0.56]	53/5/454	0.877	0.596
WM	Action	B const.	0.820 [0.79, 0.85]	390/62/60	0.867	0.770
WM	Action	clean	0.949 [0.93, 0.97]	477/17/18	0.961	0.957
WM	Action	P abr.	0.485 [0.47, 0.50]	11/26/475	0.354	0.487
WM	Action	P const.	0.382 [0.35, 0.42]	113/234/165	0.350	0.438
WM	Block	B abr.	0.393 [0.35, 0.44]	201/311/0	0.533	0.494
WM	Block	B const.	0.625 [0.58, 0.67]	320/192/0	0.521	0.567
WM	Block	B 90 deg.	0.441 [0.38, 0.50]	113/143/0	0.355	0.505
WM	Block	clean	0.998 [0.99, 1.00]	511/1/0	0.492	0.992
WM	Block	P abr.	0.869 [0.84, 0.90]	445/67/0	0.354	0.722
WM	Block	P const.	0.959 [0.94, 0.98]	491/21/0	0.369	0.814
WM	Block	P 90 deg.	0.922 [0.89, 0.95]	236/20/0	0.352	0.743
WM	Pusher	B abr.	0.631 [0.59, 0.67]	323/189/0	0.562	0.580
WM	Pusher	B const.	0.797 [0.76, 0.83]	408/104/0	0.562	0.634
WM	Pusher	B 90 deg.	0.836 [0.79, 0.88]	214/42/0	0.426	0.549
WM	Pusher	clean	0.988 [0.98, 1.00]	506/6/0	0.531	0.928
WM	Pusher	P abr.	0.518 [0.47, 0.56]	265/247/0	0.377	0.491
WM	Pusher	P const.	0.395 [0.35, 0.44]	202/310/0	0.377	0.451
WM	Pusher	P 90 deg.	0.520 [0.46, 0.58]	133/123/0	0.414	0.491

Continued on next page

Table 30 continued

Score	Edit	Appearance	Ordering [95%]	W/L/T	Later (3)	AUROC
Pixel	Action	B abr.	0.502 [0.50, 0.51]	3/1/508	0.279	0.504
Pixel	Action	B const.	0.378 [0.35, 0.41]	102/227/183	0.279	0.492
Pixel	Action	clean	0.396 [0.36, 0.43]	102/209/201	0.287	0.495
Pixel	Action	P abr.	0.412 [0.38, 0.44]	94/184/234	0.297	0.495
Pixel	Action	P const.	0.399 [0.37, 0.43]	105/208/199	0.297	0.492
Pixel	Block	B abr.	0.986 [0.97, 1.00]	505/7/0	0.341	0.942
Pixel	Block	B const.	0.986 [0.97, 1.00]	505/7/0	0.341	0.972
Pixel	Block	B 90 deg.	0.992 [0.98, 1.00]	254/2/0	0.350	0.974
Pixel	Block	clean	0.988 [0.98, 1.00]	506/6/0	0.362	0.963
Pixel	Block	P abr.	0.988 [0.98, 1.00]	506/6/0	0.354	0.967
Pixel	Block	P const.	0.988 [0.98, 1.00]	506/6/0	0.354	0.966
Pixel	Block	P 90 deg.	0.992 [0.98, 1.00]	254/2/0	0.354	0.973
Pixel	Pusher	B abr.	0.762 [0.72, 0.80]	389/121/2	0.334	0.572
Pixel	Pusher	B const.	0.632 [0.59, 0.67]	323/188/1	0.334	0.533
Pixel	Pusher	B 90 deg.	0.648 [0.59, 0.71]	166/90/0	0.316	0.547
Pixel	Pusher	clean	0.638 [0.60, 0.68]	326/185/1	0.334	0.547
Pixel	Pusher	P abr.	0.643 [0.60, 0.68]	328/182/2	0.328	0.543
Pixel	Pusher	P const.	0.631 [0.59, 0.67]	322/188/2	0.328	0.542
Pixel	Pusher	P 90 deg.	0.631 [0.57, 0.69]	161/94/1	0.301	0.554
MLP+a	Action	B abr.	0.530 [0.52, 0.54]	39/8/465	0.820	0.580
MLP+a	Action	B const.	0.777 [0.74, 0.81]	368/84/60	0.824	0.749
MLP+a	Action	clean	0.960 [0.94, 0.98]	487/16/9	0.963	0.961
MLP+a	Action	P abr.	0.497 [0.49, 0.51]	11/14/487	0.328	0.494
MLP+a	Action	P const.	0.397 [0.36, 0.43]	126/231/155	0.314	0.435
MLP+a	Block	B abr.	0.400 [0.36, 0.44]	205/307/0	0.512	0.495
MLP+a	Block	B const.	0.633 [0.59, 0.67]	324/188/0	0.529	0.567
MLP+a	Block	B 90 deg.	0.465 [0.41, 0.53]	119/137/0	0.363	0.504
MLP+a	Block	clean	1.000 [1.00, 1.00]	512/0/0	0.775	0.987
MLP+a	Block	P abr.	0.861 [0.83, 0.89]	441/71/0	0.346	0.714
MLP+a	Block	P const.	0.959 [0.94, 0.97]	491/21/0	0.342	0.848
MLP+a	Block	P 90 deg.	0.914 [0.88, 0.95]	234/22/0	0.363	0.796
MLP+a	Pusher	B abr.	0.633 [0.59, 0.67]	324/188/0	0.559	0.588
MLP+a	Pusher	B const.	0.809 [0.78, 0.84]	414/98/0	0.555	0.642
MLP+a	Pusher	B 90 deg.	0.812 [0.76, 0.86]	208/48/0	0.434	0.580
MLP+a	Pusher	clean	0.990 [0.98, 1.00]	507/5/0	0.748	0.955
MLP+a	Pusher	P abr.	0.523 [0.48, 0.57]	268/244/0	0.367	0.494
MLP+a	Pusher	P const.	0.395 [0.35, 0.44]	202/310/0	0.365	0.448
MLP+a	Pusher	P 90 deg.	0.500 [0.44, 0.56]	128/128/0	0.430	0.506

Table 31: PushT upstream checkpoint, cohort D. Ordering compares physical and no-physical twins; $> / < / =$ are strict wins/losses/exact ties. Onset ordering uses j through $j + 2$ and has 95% parent-bootstrap intervals. The later statistic uses $j + 1$ through $j + 3$, distinct from the dedicated five-frame diagnostic. AUROC uses full episode maxima against no-physical twins, not the nominal audit.

Score	Edit	Appearance	Ordering [95%]	W/L/T	Later (3)	AUROC
Open-loop	Action	B abr.	0.449 [0.41, 0.49]	230/282/0	0.416	0.441
Open-loop	Action	B const.	0.438 [0.39, 0.48]	224/288/0	0.402	0.417
Open-loop	Action	clean	0.487 [0.44, 0.53]	249/262/1	0.520	0.486
Open-loop	Action	P abr.	0.474 [0.43, 0.52]	241/268/3	0.480	0.493
Open-loop	Action	P const.	0.399 [0.36, 0.44]	203/306/3	0.412	0.426
Open-loop	Block	B abr.	0.398 [0.36, 0.44]	204/308/0	0.393	0.449
Open-loop	Block	B const.	0.463 [0.42, 0.51]	237/275/0	0.434	0.462
Open-loop	Block	B 90 deg.	0.438 [0.38, 0.50]	112/144/0	0.465	0.502
Open-loop	Block	clean	0.654 [0.61, 0.69]	335/177/0	0.645	0.575
Open-loop	Block	P abr.	0.689 [0.65, 0.73]	353/159/0	0.674	0.606
Open-loop	Block	P const.	0.615 [0.57, 0.66]	315/197/0	0.588	0.550
Open-loop	Block	P 90 deg.	0.582 [0.52, 0.64]	149/107/0	0.578	0.549

Continued on next page

Table 31 continued

Score	Edit	Appearance	Ordering [95%]	W/L/T	Later (3)	AUROC
Open-loop	Pusher	B abr.	0.428 [0.38, 0.47]	219/293/0	0.426	0.451
Open-loop	Pusher	B const.	0.393 [0.35, 0.43]	201/311/0	0.418	0.448
Open-loop	Pusher	B 90 deg.	0.516 [0.45, 0.57]	132/124/0	0.559	0.508
Open-loop	Pusher	clean	0.518 [0.47, 0.56]	265/247/0	0.535	0.507
Open-loop	Pusher	P abr.	0.510 [0.47, 0.55]	261/251/0	0.521	0.515
Open-loop	Pusher	P const.	0.473 [0.43, 0.52]	242/270/0	0.451	0.461
Open-loop	Pusher	P 90 deg.	0.422 [0.36, 0.48]	108/148/0	0.367	0.446
WM	Action	B abr.	0.494 [0.49, 0.50]	2/8/502	0.434	0.499
WM	Action	B const.	0.496 [0.46, 0.54]	194/198/120	0.445	0.474
WM	Action	clean	0.704 [0.67, 0.74]	327/118/67	0.652	0.581
WM	Action	P abr.	0.454 [0.42, 0.48]	109/156/247	0.377	0.415
WM	Action	P const.	0.394 [0.36, 0.43]	130/239/143	0.365	0.432
WM	Block	B abr.	0.467 [0.42, 0.51]	239/273/0	0.391	0.481
WM	Block	B const.	0.676 [0.63, 0.71]	346/166/0	0.387	0.526
WM	Block	B 90 deg.	0.605 [0.55, 0.66]	155/101/0	0.414	0.478
WM	Block	clean	0.973 [0.96, 0.99]	498/14/0	0.379	0.905
WM	Block	P abr.	0.932 [0.91, 0.95]	477/35/0	0.279	0.823
WM	Block	P const.	0.920 [0.90, 0.94]	471/41/0	0.275	0.741
WM	Block	P 90 deg.	0.914 [0.88, 0.95]	234/22/0	0.297	0.723
WM	Pusher	B abr.	0.434 [0.39, 0.48]	222/290/0	0.359	0.475
WM	Pusher	B const.	0.498 [0.46, 0.54]	255/257/0	0.369	0.479
WM	Pusher	B 90 deg.	0.645 [0.59, 0.70]	165/91/0	0.422	0.478
WM	Pusher	clean	0.734 [0.70, 0.77]	376/136/0	0.414	0.557
WM	Pusher	P abr.	0.445 [0.40, 0.49]	228/284/0	0.369	0.442
WM	Pusher	P const.	0.408 [0.37, 0.45]	209/303/0	0.367	0.437
WM	Pusher	P 90 deg.	0.406 [0.35, 0.47]	104/152/0	0.328	0.434
Pixel	Action	B abr.	0.502 [0.50, 0.51]	3/1/508	0.279	0.504
Pixel	Action	B const.	0.378 [0.34, 0.41]	102/227/183	0.279	0.492
Pixel	Action	clean	0.396 [0.36, 0.43]	102/209/201	0.287	0.495
Pixel	Action	P abr.	0.412 [0.38, 0.44]	94/184/234	0.297	0.495
Pixel	Action	P const.	0.399 [0.37, 0.43]	105/208/199	0.297	0.492
Pixel	Block	B abr.	0.986 [0.97, 1.00]	505/7/0	0.341	0.942
Pixel	Block	B const.	0.986 [0.97, 1.00]	505/7/0	0.341	0.972
Pixel	Block	B 90 deg.	0.992 [0.98, 1.00]	254/2/0	0.350	0.974
Pixel	Block	clean	0.988 [0.98, 1.00]	506/6/0	0.362	0.963
Pixel	Block	P abr.	0.988 [0.98, 1.00]	506/6/0	0.354	0.967
Pixel	Block	P const.	0.988 [0.98, 1.00]	506/6/0	0.354	0.966
Pixel	Block	P 90 deg.	0.992 [0.98, 1.00]	254/2/0	0.354	0.973
Pixel	Pusher	B abr.	0.762 [0.72, 0.80]	389/121/2	0.334	0.572
Pixel	Pusher	B const.	0.632 [0.59, 0.67]	323/188/1	0.334	0.533
Pixel	Pusher	B 90 deg.	0.648 [0.59, 0.71]	166/90/0	0.316	0.547
Pixel	Pusher	clean	0.638 [0.60, 0.68]	326/185/1	0.334	0.547
Pixel	Pusher	P abr.	0.643 [0.60, 0.68]	328/182/2	0.328	0.543
Pixel	Pusher	P const.	0.631 [0.59, 0.67]	322/188/2	0.328	0.542
Pixel	Pusher	P 90 deg.	0.631 [0.57, 0.69]	161/94/1	0.301	0.554
MLP+a	Action	B abr.	0.486 [0.47, 0.50]	12/26/474	0.441	0.496
MLP+a	Action	B const.	0.455 [0.42, 0.49]	171/217/124	0.406	0.434
MLP+a	Action	clean	0.761 [0.73, 0.79]	365/98/49	0.752	0.689
MLP+a	Action	P abr.	0.478 [0.45, 0.51]	115/138/259	0.395	0.389
MLP+a	Action	P const.	0.410 [0.37, 0.45]	138/230/144	0.379	0.426
MLP+a	Block	B abr.	0.461 [0.42, 0.50]	236/276/0	0.361	0.480
MLP+a	Block	B const.	0.559 [0.52, 0.60]	286/226/0	0.396	0.480
MLP+a	Block	B 90 deg.	0.602 [0.54, 0.66]	154/102/0	0.414	0.470
MLP+a	Block	clean	0.986 [0.97, 1.00]	505/7/0	0.600	0.975
MLP+a	Block	P abr.	0.955 [0.94, 0.97]	489/23/0	0.326	0.865
MLP+a	Block	P const.	0.932 [0.91, 0.95]	477/35/0	0.316	0.783
MLP+a	Block	P 90 deg.	0.922 [0.89, 0.95]	236/20/0	0.312	0.765
MLP+a	Pusher	B abr.	0.424 [0.38, 0.47]	217/295/0	0.385	0.474
MLP+a	Pusher	B const.	0.467 [0.42, 0.51]	239/273/0	0.377	0.457
MLP+a	Pusher	B 90 deg.	0.625 [0.56, 0.68]	160/96/0	0.379	0.460
MLP+a	Pusher	clean	0.867 [0.84, 0.89]	444/68/0	0.639	0.746

Continued on next page

Table 31 continued

Score	Edit	Appearance	Ordering [95%]	W/L/T	Later (3)	AUROC
MLP+a	Pusher	P abr.	0.484 [0.44, 0.53]	248/264/0	0.395	0.456
MLP+a	Pusher	P const.	0.424 [0.38, 0.47]	217/295/0	0.375	0.457
MLP+a	Pusher	P 90 deg.	0.445 [0.38, 0.51]	114/142/0	0.355	0.444

Table 32: TwoRoom upstream checkpoint, cohort D. Ordering compares physical and no-physical twins; $> / < / =$ are strict wins/losses/exact ties. Onset ordering uses j through $j + 2$ and has 95% parent-bootstrap intervals. The later statistic uses $j + 1$ through $j + 3$, distinct from the dedicated five-frame diagnostic. AUROC uses full episode maxima against no-physical twins, not the nominal audit.

Score	Edit	Appearance	Ordering [95%]	W/L/T	Later (3)	AUROC
Open-loop	Action	clean	0.996 [0.99, 1.00]	510/2/0	0.998	0.995
Open-loop	Action	Agent abr.	0.515 [0.48, 0.55]	176/161/175	0.449	0.513
Open-loop	Action	Agent const.	0.471 [0.44, 0.51]	165/195/152	0.471	0.445
Open-loop	Teleport	clean	0.998 [0.99, 1.00]	511/1/0	0.998	0.994
Open-loop	Teleport	Agent abr.	0.432 [0.39, 0.47]	221/291/0	0.406	0.536
Open-loop	Teleport	Agent const.	0.451 [0.41, 0.49]	231/281/0	0.447	0.463
WM	Action	clean	1.000 [1.00, 1.00]	512/0/0	1.000	1.000
WM	Action	Agent abr.	0.501 [0.50, 0.50]	1/0/511	0.525	0.500
WM	Action	Agent const.	0.524 [0.49, 0.56]	196/171/145	0.520	0.490
WM	Teleport	clean	1.000 [1.00, 1.00]	512/0/0	0.623	0.999
WM	Teleport	Agent abr.	0.395 [0.35, 0.44]	202/310/0	0.467	0.481
WM	Teleport	Agent const.	0.498 [0.46, 0.54]	255/257/0	0.479	0.501
Pixel	Action	clean	0.519 [0.48, 0.56]	197/178/137	0.533	0.490
Pixel	Action	Agent abr.	0.500 [0.50, 0.50]	0/0/512	0.467	0.500
Pixel	Action	Agent const.	0.494 [0.46, 0.53]	188/194/130	0.467	0.446
Pixel	Teleport	clean	0.900 [0.87, 0.93]	461/51/0	0.473	0.811
Pixel	Teleport	Agent abr.	0.717 [0.68, 0.76]	367/145/0	0.486	0.512
Pixel	Teleport	Agent const.	0.943 [0.92, 0.96]	483/29/0	0.486	0.829
MLP+a	Action	clean	1.000 [1.00, 1.00]	512/0/0	1.000	0.998
MLP+a	Action	Agent abr.	0.501 [0.50, 0.50]	1/0/511	0.404	0.500
MLP+a	Action	Agent const.	0.477 [0.44, 0.51]	180/204/128	0.426	0.327
MLP+a	Teleport	clean	1.000 [1.00, 1.00]	512/0/0	0.576	0.996
MLP+a	Teleport	Agent abr.	0.367 [0.33, 0.41]	188/324/0	0.414	0.490
MLP+a	Teleport	Agent const.	0.430 [0.39, 0.47]	220/292/0	0.424	0.400

Table 33: TwoRoom PLDM checkpoint, cohort D. Ordering compares physical and no-physical twins; $> / < / =$ are strict wins/losses/exact ties. Onset ordering uses j through $j + 2$ and has 95% parent-bootstrap intervals. The later statistic uses $j + 1$ through $j + 3$, distinct from the dedicated five-frame diagnostic. AUROC uses full episode maxima against no-physical twins, not the nominal audit.

Score	Edit	Appearance	Ordering [95%]	W/L/T	Later (3)	AUROC
Open-loop	Action	clean	0.991 [0.98, 1.00]	507/4/1	0.992	0.988
Open-loop	Action	Agent abr.	0.499 [0.46, 0.54]	190/191/131	0.484	0.497
Open-loop	Action	Agent const.	0.510 [0.47, 0.55]	201/191/120	0.541	0.519
Open-loop	Teleport	clean	0.994 [0.99, 1.00]	509/3/0	0.994	0.986
Open-loop	Teleport	Agent abr.	0.529 [0.48, 0.57]	271/241/0	0.516	0.509
Open-loop	Teleport	Agent const.	0.562 [0.52, 0.61]	288/224/0	0.549	0.508
WM	Action	clean	1.000 [1.00, 1.00]	512/0/0	1.000	0.999
WM	Action	Agent abr.	0.503 [0.49, 0.51]	19/16/477	0.549	0.507
WM	Action	Agent const.	0.525 [0.49, 0.56]	207/181/124	0.545	0.518
WM	Teleport	clean	1.000 [1.00, 1.00]	512/0/0	0.650	0.994
WM	Teleport	Agent abr.	0.549 [0.51, 0.59]	281/231/0	0.529	0.520

Continued on next page

Table 33 continued

Score	Edit	Appearance	Ordering [95%]	W/L/T	Later (3)	AUROC
WM	Teleport	Agent const.	0.535 [0.49, 0.58]	274/238/0	0.529	0.507
Pixel	Action	clean	0.519 [0.48, 0.56]	197/178/137	0.533	0.490
Pixel	Action	Agent abr.	0.500 [0.50, 0.50]	0/0/512	0.467	0.500
Pixel	Action	Agent const.	0.494 [0.46, 0.53]	188/194/130	0.467	0.446
Pixel	Teleport	clean	0.900 [0.88, 0.92]	461/51/0	0.473	0.811
Pixel	Teleport	Agent abr.	0.717 [0.68, 0.76]	367/145/0	0.486	0.512
Pixel	Teleport	Agent const.	0.943 [0.92, 0.96]	483/29/0	0.486	0.829
MLP+a	Action	clean	0.997 [0.99, 1.00]	510/1/1	1.000	1.000
MLP+a	Action	Agent abr.	0.497 [0.49, 0.50]	0/3/509	0.520	0.499
MLP+a	Action	Agent const.	0.521 [0.48, 0.56]	213/192/107	0.525	0.464
MLP+a	Teleport	clean	0.998 [0.99, 1.00]	511/1/0	0.436	0.992
MLP+a	Teleport	Agent abr.	0.543 [0.50, 0.59]	278/234/0	0.539	0.515
MLP+a	Teleport	Agent const.	0.584 [0.54, 0.63]	299/213/0	0.521	0.527

Reversed ordering and prediction difficulty. The one-step residual reverses direction for pusher teleport under constant pusher recolouring (0.395, 95% interval $[0.35, 0.44]$), action mismatch under constant pusher recolouring (0.382, $[0.35, 0.42]$), block teleport under abrupt block recolouring (0.393, $[0.35, 0.44]$), and TwoRoom teleport under abrupt agent recolouring (0.395, $[0.35, 0.44]$). These cells indicate that the physical twin more often has a smaller maximum, not merely an unresolved positive effect. The orientation of the deployed alarm remains fixed in advance.

Residual size also depends on how predictable the edited trajectory is. The teleport generator conditions accepted directions on a center-distance clearance test, which does not certify polygonal separation of the T-shaped block or absence of later contact. Moreover, pixel-change ordering for clean action mismatch is 0.396 (102 wins, 209 losses, 201 ties): the edit often reduces this image-difference statistic. These observations motivate a prediction-difficulty explanation but do not identify it causally. Contact or displacement stratification would be required to test that explanation. In contrast, each teleport has identical underlying trajectories across recolourings, so those physical changes cannot account for its within-edit appearance contrast. The crossed interaction removes additive object and colour main effects on the ordering scale, while retaining the possibility of object-specific perceptual effects.

M.4 Coordinate recovery and displacement magnitude

State readouts use one current-frame frozen latent. They are trained on nominal states rendered clean or in the matched colour condition, without intervention labels. The coordinate RMSE averages squared errors after onset within each parent, then across parents, before taking the square root. Coordinates are reported separately, including x and y; orientation sine/cosine errors are dimensionless. A matched readout is diagnostic access to supervised simulator state, not an appearance-adapted monitor proposed for deployment.

TwoRoom provides the strongest positive recovery control. Under constant agent recolouring, the matched MLP gives x/y errors of $0.06/0.07$ reporting radii, while the clean-trained readout gives $6.12/5.70$. The matched median teleport displacement-magnitude ratio is 1.00, although the original residual's teleport ordering is 0.498. The tested position variables remain accessible despite a poor response from the original residual. This does not establish preservation of every feature relevant to dynamics.

On PushT innovation, matched block errors are $2.04/2.09$ radii and pusher errors $4.25/4.12$, compared with clean-trained clean-rendering errors $1.17/1.19$ for the block and $0.46/0.42$ for the pusher. Block recovery is therefore partial, and pusher recovery is more degraded. The protocol scales (two radii and three times the corresponding clean error) apply to coordinate-wise RMSE. Both matched block coordinates remain above two radii. Their combined Euclidean position RMSE is approximately $\sqrt{2.04^2 + 2.09^2} = 2.92$ radii, a different quantity. These scales aid interpretation rather than defining significance or exclusions.

Table 34: TwoRoom upstream MLP readout ensemble. Coordinates are evaluated after onset on no-physical-edit parents (512); brackets give 95% parent-bootstrap intervals. Reporting radii are seven pixels for TwoRoom and 15 for PushT. Orientation coordinates are dimensionless. Cross-regime errors show transfer, not new fit selection.

Fit rendering	Test rendering	Coordinate	RMSE px [95%]	Radii
clean	clean	agent x	0.2 [0.2, 0.2]	0.03
clean	clean	agent y	0.2 [0.2, 0.2]	0.03
clean	hue colour	agent x	42.9 [41.3, 44.5]	6.12
clean	hue colour	agent y	39.9 [37.6, 42.3]	5.70
agent-matched	clean	agent x	154.4 [150.2, 158.4]	22.06
agent-matched	clean	agent y	241.1 [235.7, 246.1]	34.44
agent-matched	hue colour	agent x	0.4 [0.4, 0.4]	0.06
agent-matched	hue colour	agent y	0.5 [0.4, 0.5]	0.07

Table 35: TwoRoom upstream displacement-magnitude recovery. Entries are median per-parent ratios of predicted to true inter-frame displacement at teleport onset. They do not measure direction and are not ratios to the nominal edit distance alone.

Fit rendering	Teleported object	Test rendering	Median ratio
clean	Agent	Clean	1.00
clean	Agent	Agent colour	0.16
agent-matched	Agent	Agent colour	1.00

Table 36: PushT innovation MLP readout ensemble. Coordinates are evaluated after onset on no-physical-edit parents (512); brackets give 95% parent-bootstrap intervals. Reporting radii are seven pixels for TwoRoom and 15 for PushT. Orientation coordinates are dimensionless. Cross-regime errors show transfer, not new fit selection.

Fit rendering	Test rendering	Coordinate	RMSE px [95%]	Radii
clean	block colour	pusher x	39.1 [36.2, 42.5]	2.61
clean	block colour	pusher y	25.7 [23.8, 28.0]	1.71
clean	block colour	block x	120.0 [114.9, 124.8]	8.00
clean	block colour	block y	93.9 [90.4, 97.7]	6.26
clean	block colour	block sin	1.3 [1.3, 1.4]	–
clean	block colour	block cos	1.1 [1.1, 1.2]	–
clean	clean	pusher x	7.0 [5.2, 8.9]	0.46
clean	clean	pusher y	6.3 [5.0, 7.7]	0.42
clean	clean	block x	17.6 [16.1, 18.9]	1.17
clean	clean	block y	17.9 [16.3, 19.6]	1.19
clean	clean	block sin	0.4 [0.4, 0.4]	–
clean	clean	block cos	0.4 [0.4, 0.4]	–
clean	pusher colour	pusher x	100.3 [96.6, 104.0]	6.69
clean	pusher colour	pusher y	88.5 [84.8, 92.0]	5.90
clean	pusher colour	block x	39.1 [37.4, 40.7]	2.61
clean	pusher colour	block y	49.4 [47.7, 51.0]	3.29
clean	pusher colour	block sin	0.9 [0.8, 0.9]	–
clean	pusher colour	block cos	1.0 [1.0, 1.1]	–
block-matched	block colour	pusher x	11.5 [9.5, 13.5]	0.77
block-matched	block colour	pusher y	9.1 [8.1, 10.1]	0.61
block-matched	block colour	block x	30.7 [28.0, 33.7]	2.04
block-matched	block colour	block y	31.4 [28.2, 34.7]	2.09
block-matched	block colour	block sin	0.5 [0.5, 0.5]	–
block-matched	block colour	block cos	0.4 [0.4, 0.5]	–
block-matched	clean	pusher x	75.0 [71.8, 78.3]	5.00

Continued on next page

Table 36 continued

Fit rendering	Test rendering	Coordinate	RMSE px [95%]	Radii
block-matched	clean	pusher y	68.5 [66.1, 70.9]	4.57
block-matched	clean	block x	104.4 [99.4, 109.7]	6.96
block-matched	clean	block y	95.1 [90.2, 100.3]	6.34
block-matched	clean	block sin	0.8 [0.8, 0.8]	–
block-matched	clean	block cos	0.7 [0.7, 0.7]	–
block-matched	pusher colour	pusher x	84.0 [81.1, 86.7]	5.60
block-matched	pusher colour	pusher y	95.0 [91.4, 98.3]	6.33
block-matched	pusher colour	block x	189.6 [181.7, 197.7]	12.64
block-matched	pusher colour	block y	99.6 [92.3, 106.8]	6.64
block-matched	pusher colour	block sin	1.1 [1.1, 1.2]	–
block-matched	pusher colour	block cos	1.3 [1.2, 1.3]	–
pusher-matched	block colour	pusher x	299.7 [287.7, 311.6]	19.98
pusher-matched	block colour	pusher y	197.0 [188.2, 206.0]	13.13
pusher-matched	block colour	block x	69.3 [65.6, 72.8]	4.62
pusher-matched	block colour	block y	63.2 [59.6, 66.8]	4.22
pusher-matched	block colour	block sin	1.4 [1.3, 1.5]	–
pusher-matched	block colour	block cos	1.1 [1.1, 1.2]	–
pusher-matched	clean	pusher x	174.4 [166.8, 181.8]	11.63
pusher-matched	clean	pusher y	156.3 [150.9, 162.2]	10.42
pusher-matched	clean	block x	51.1 [48.9, 53.4]	3.41
pusher-matched	clean	block y	56.8 [53.1, 60.4]	3.78
pusher-matched	clean	block sin	1.0 [1.0, 1.1]	–
pusher-matched	clean	block cos	0.9 [0.8, 0.9]	–
pusher-matched	pusher colour	pusher x	63.8 [60.0, 67.8]	4.25
pusher-matched	pusher colour	pusher y	61.8 [59.0, 64.5]	4.12
pusher-matched	pusher colour	block x	19.7 [18.5, 20.9]	1.32
pusher-matched	pusher colour	block y	19.0 [17.6, 20.3]	1.27
pusher-matched	pusher colour	block sin	0.4 [0.4, 0.5]	–
pusher-matched	pusher colour	block cos	0.4 [0.4, 0.4]	–

Table 37: PushT innovation displacement-magnitude recovery. Entries are median per-parent ratios of predicted to true inter-frame displacement at teleport onset. They do not measure direction and are not ratios to the nominal edit distance alone.

Fit rendering	Teleported object	Test rendering	Median ratio
clean	Block	Block colour	0.61
clean	Block	Clean	1.01
clean	Pusher	Clean	1.00
clean	Pusher	Pusher colour	0.28
block-matched	Block	Block colour	0.88
pusher-matched	Pusher	Pusher colour	0.48

Table 38: PushT upstream MLP readout ensemble. Coordinates are evaluated after onset on no-physical-edit parents (512); brackets give 95% parent-bootstrap intervals. Reporting radii are seven pixels for TwoRoom and 15 for PushT. Orientation coordinates are dimensionless. Cross-regime errors show transfer, not new fit selection.

Fit rendering	Test rendering	Coordinate	RMSE px [95%]	Radii
clean	block colour	pusher x	62.3 [59.9, 64.9]	4.16
clean	block colour	pusher y	60.1 [58.4, 62.0]	4.01
clean	block colour	block x	52.7 [50.2, 55.2]	3.51
clean	block colour	block y	61.7 [59.4, 64.3]	4.11
clean	block colour	block sin	0.6 [0.6, 0.7]	–
clean	block colour	block cos	1.1 [1.0, 1.1]	–

Continued on next page

Table 38 continued

Fit rendering	Test rendering	Coordinate	RMSE px [95%]	Radii
clean	clean	pusher x	15.9 [13.1, 18.9]	1.06
clean	clean	pusher y	15.8 [13.5, 18.3]	1.06
clean	clean	block x	9.0 [8.0, 10.0]	0.60
clean	clean	block y	8.8 [7.8, 9.8]	0.59
clean	clean	block sin	0.2 [0.1, 0.2]	–
clean	clean	block cos	0.2 [0.1, 0.2]	–
clean	pusher colour	pusher x	72.9 [70.1, 75.8]	4.86
clean	pusher colour	pusher y	65.5 [62.9, 68.0]	4.37
clean	pusher colour	block x	12.0 [11.1, 12.9]	0.80
clean	pusher colour	block y	11.9 [11.0, 12.9]	0.79
clean	pusher colour	block sin	0.2 [0.2, 0.2]	–
clean	pusher colour	block cos	0.2 [0.2, 0.2]	–
block-matched	block colour	pusher x	25.5 [23.4, 28.0]	1.70
block-matched	block colour	pusher y	24.5 [22.6, 26.7]	1.63
block-matched	block colour	block x	18.4 [17.2, 19.6]	1.23
block-matched	block colour	block y	19.5 [18.2, 20.9]	1.30
block-matched	block colour	block sin	0.4 [0.4, 0.4]	–
block-matched	block colour	block cos	0.4 [0.4, 0.4]	–
block-matched	clean	pusher x	313.4 [297.1, 329.8]	20.89
block-matched	clean	pusher y	435.4 [409.0, 463.1]	29.03
block-matched	clean	block x	217.4 [206.5, 228.3]	14.50
block-matched	clean	block y	221.3 [208.1, 235.1]	14.75
block-matched	clean	block sin	4.0 [3.8, 4.2]	–
block-matched	clean	block cos	4.5 [4.3, 4.8]	–
block-matched	pusher colour	pusher x	281.4 [266.5, 296.1]	18.76
block-matched	pusher colour	pusher y	400.4 [377.0, 426.5]	26.69
block-matched	pusher colour	block x	209.9 [198.3, 222.4]	13.99
block-matched	pusher colour	block y	224.9 [212.7, 237.3]	15.00
block-matched	pusher colour	block sin	3.6 [3.4, 3.8]	–
block-matched	pusher colour	block cos	4.6 [4.4, 4.8]	–
pusher-matched	block colour	pusher x	60.9 [58.5, 63.5]	4.06
pusher-matched	block colour	pusher y	61.7 [59.3, 64.2]	4.11
pusher-matched	block colour	block x	53.5 [51.0, 55.9]	3.57
pusher-matched	block colour	block y	62.8 [60.3, 65.2]	4.18
pusher-matched	block colour	block sin	0.6 [0.6, 0.7]	–
pusher-matched	block colour	block cos	1.1 [1.0, 1.1]	–
pusher-matched	clean	pusher x	61.0 [57.7, 64.5]	4.07
pusher-matched	clean	pusher y	63.3 [60.2, 66.4]	4.22
pusher-matched	clean	block x	15.2 [14.2, 16.3]	1.01
pusher-matched	clean	block y	15.4 [14.5, 16.2]	1.03
pusher-matched	clean	block sin	0.3 [0.3, 0.3]	–
pusher-matched	clean	block cos	0.3 [0.3, 0.3]	–
pusher-matched	pusher colour	pusher x	53.9 [49.7, 58.2]	3.60
pusher-matched	pusher colour	pusher y	51.6 [48.5, 54.7]	3.44
pusher-matched	pusher colour	block x	9.8 [8.8, 10.8]	0.65
pusher-matched	pusher colour	block y	9.3 [8.5, 10.2]	0.62
pusher-matched	pusher colour	block sin	0.2 [0.1, 0.2]	–
pusher-matched	pusher colour	block cos	0.2 [0.1, 0.2]	–

Table 39: PushT upstream displacement-magnitude recovery. Entries are median per-parent ratios of predicted to true inter-frame displacement at teleport onset. They do not measure direction and are not ratios to the nominal edit distance alone.

Fit rendering	Teleported object	Test rendering	Median ratio
clean	Block	Block colour	0.39
clean	Block	Clean	0.99
clean	Pusher	Clean	0.95
clean	Pusher	Pusher colour	0.69

Continued on next page

Table 39 continued

Fit rendering	Teleported object	Test rendering	Median ratio
block-matched	Block	Block colour	0.98
pusher-matched	Pusher	Pusher colour	0.67

The jump ratio in Eq. equation 22 measures decoded displacement magnitude, not direction; the true inter-frame displacement also contains nominal motion. For the recoloured innovation pusher, the MLP ratio improves from 0.28 with a clean-trained readout to 0.48 with a matched readout. The corresponding block ratios are 0.61 and 0.88. The clean-trained ridge pusher ratio is 0.20; the MLP ratio is 0.28. PushT jump rows were added after a filtering correction that left all other outcomes unchanged (Appendix B.5).

Coordinate errors above concern no-physical-edit trajectories, while jump ratios concern teleport trajectories. Poor recovery by the tested probes cannot establish that the encoder has discarded all information about the object.

M.5 Temporal masking and what a later window can show

The paired per-frame difference $d_{i,\ell}(e, c)$ in Eq. equation 23 compares physical and appearance-only twins at offset ℓ from onset. This separates framewise evidence from the difference between two maxima: a common large onset residual can make both maxima identical even when later differences are positive.

For D action mismatch with abrupt block recolouring, onset-window ordering is 0.547 with 53 strict wins, five losses, and 454 ties. Both maxima occur at onset in the reported 89% of parents. At each of the five later offsets the physical score is higher in 382–434 parents, and the dedicated five-frame diagnostic gives ordering 0.881.

Table 40: Temporal statistics must name their windows. PushT innovation, D, action mismatch and abrupt block recolouring. The two later estimates use different window lengths and are not interchangeable.

Statistic	Observed frames	Reported ordering
Primary onset window	j through $j+2$	0.547
Shifted three-frame window	$j+1$ through $j+3$	0.877
Dedicated temporal diagnostic	$j+1$ through $j+5$	0.881

This temporal explanation is conditional on the object and appearance condition. Under constant pusher recolouring, action ordering is 0.382 on $[j, j{+}2]$ and 0.350 on the shifted window $[j{+}1, j{+}3]$. These are both three-frame windows, distinct from the dedicated five-frame diagnostic above. Shifting this window does not restore the pattern observed under block recolouring. Constant recolouring also removes the abrupt appearance transition as an explanation for the object interaction. These controls distinguish several failure modes without identifying a unique causal decomposition into encoder, predictor, and aggregation components. Because the windows use ground-truth onset, their improvement cannot be claimed as online monitoring performance.

N Deployment References, Reachability, and Planning Utility

N.1 A deployment-start reference reduces mismatch but remains inconclusive

The support/reachability screen rejects a proposal if either channel has $p \leq 0.025$. A 0.05 combined target requires validity of both channel ranks under the proposal-generating population; it is not guaranteed merely by adding their nominal levels. A proposer-matched reference collected at demonstration contexts rejects $911/1{,}000$ first proposals at random deployment starts. A new reference consisting of 600 deployment-start first proposals reduces this to $80/1{,}000$, with upper bound 0.096. It therefore does not establish the specified 0.07 combined tolerance.

Table 41: First-proposal audit on the same 1,000 PushT deployment starts, upstream LeWM checkpoint. Bounds are one-sided 95% Clopper–Pearson. The 0.07 tolerance belongs to the combined rule, not to the channels' individual 0.025 targets.

Reference	Channel	Count	Rate	Upper
Demonstration context	Combined	911	0.911	0.925
	Support	14	0.014	0.022
	Reachability	908	0.908	0.923
Deployment start	Combined	80	0.080	0.096
	Support	43	0.043	0.055
	Reachability	38	0.038	0.050

Both deployment-start channel rates exceed their 0.025 targets. Resampling the same 600 reference records gives a combined-rate 5th–95th percentile range of 0.056–0.096 with the audit records fixed. This diagnoses sensitivity conditional on those records. It is not an independent reference replication and cannot rule out an unrepresentative initial draw. No guarantee for repeated adaptive screening or later visited contexts follows from this first-proposal audit.

N.2 Reachability evidence requires a valid certificate

The geometric study proposes target pairs at matched Euclidean distance from 192 TwoRoom contexts. It finds 154 candidate pairs with executed witnesses reaching the nominally reachable target. The simulator-respecting lower bound certifies 39 paired targets as unreachable within the stated budget and tolerance; only these 39 doubly certified pairs enter the reachability comparison. The final lower bound uses every permitted door band, the actual crossing rule, and the goal tolerance set as described in Appendix H.

On the 39 doubly certified pairs, the CEM residual orders the unreachable target above the reachable target in 0.436 of cases, with interval $[0.28, 0.59]$. Latent distance gives 0.590, kNN support 0.538, and initial-population random shooting 0.282. The CEM-minus-distance difference is -0.154 $[-0.36, +0.08]$. The small certified subset limits what can be concluded about the score; a fixed-budget optimization residual is not an infeasibility certificate.

A separate short-target control reaches 22 of 24 known-reachable targets, and obtains a residual at most that of the witness in 20 of 24 cases. The executor therefore reaches most easy targets; this control does not establish optimizer convergence on the harder paired targets.

N.3 Imagined-rollout alarms can measure typicality

The S2 sliced Epps–Pulley diagnostic tests the distribution of normalized imagined transition residuals. With demonstration plans as reference, its flag rate is 0.05 on demonstration actions but 0.19 on lawful half-speed actions, 0.13 on lawful wait-then-demonstration plans, and 0.19 when executing another context's demonstration. Zero actions are flagged at 1.00. This pattern is evidence of sensitivity to reference-relative plan typicality, not a rule for declaring impossible physics. S2 normalization was fitted using episodes that overlap its rank reference. Because the normalization uses reference trajectories without treating the query symmetrically, conditional exchangeability of the ranked scores is not established; no conformal validity claim is attached to these numbers.

N.4 Replanning with matched optimizer initialization

The loop runs on PushT with the upstream LeWM checkpoint and uses per-step surprise at level 0.1, thresholded against pooled per-frame reference scores (Appendix H); this operational rule is separate from the fixed-horizon episode-max monitor and carries no false-alarm guarantee. The paired comparison uses three fresh evaluation seeds, 128 episodes per seed, after development on separate seeds. Each CEM call has the same candidate/iteration/elite budget. Warm-started methods initialize the optimizer from the unexecuted suffix of the previous plan. Fixed-five warm and cold starts have identical success outcomes on all 384 episodes (zero discordant pairs). Warm starts improve fixed-two and fixed-one success by 8.9 and 5.5 points, respectively. The trigger's development-seed

advantage ($+4.4$ points, 95% interval $[+0.8, +8.1]$; cold starts) is unresolved on fresh seeds ($+1.8$, $[-1.3, +4.9]$). Warm-starting the trigger itself changes success by -1.8 points. Thus the primary fixed-five result cannot be attributed to a cold-start penalty in its baseline.

Table 42: Fresh replanning comparison on PushT (upstream LeWM checkpoint), $n = 384$. Success intervals are pooled 95% Wilson intervals; the paired primary difference is 0.0 pp $[-2.9, +2.9]$. Nominal and perturbed success columns are descriptive strata.

Warm-started rule	Calls/episode	Planner time (s)	Success % [95%]	Nominal	Perturbed
Fixed every five	4.41	7.6	25.8 [21.7,30.4]	28.6	22.9
Surprise, $\alpha = 0.1$	4.74	8.1	25.8 [21.7,30.4]	32.3	19.3
Fixed every two	9.59	16.2	26.0 [21.9,30.7]	31.8	20.3
Fixed every one	18.47	31.4	24.2 [20.2,28.7]	34.4	14.1

The primary comparison has 17 discordant pairs in each direction (McNemar $p = 1$). Its nominal and perturbed differences are $+3.6$ points $[+1.0, +6.8]$ and -3.6 $[-8.9, +1.6]$, respectively; these secondary strata provide no evidence that the trigger preferentially helps perturbed episodes. The trigger uses 4.74 rather than 4.41 planner calls per episode, about 7.5% more. It therefore establishes neither a success advantage nor an efficiency gain against the fixed-five baseline, while leaving modest effects in either direction unresolved. Because surprise from this checkpoint detects clean action mismatch weakly (Table 2; Section 3.3), the null result may partly reflect a weak trigger signal.

N.5 Subgoal screening and fixed candidate budgets

The initial screening comparison uses PushT, the upstream LeWM checkpoint, one trained proposer, and three evaluation seeds of 256 episodes each ($n = 768$). The FF-JEPA-style proposer is our reimplementation of the latent subgoal planner of Masip et al. (2026); its success is below that reported for the original system, and we have not identified the source of this gap, so the screening results concern our implementation. Screening against true demonstration subgoals and falling back to the flat goal produces a recorded paired loss of 18.0 points $[-22.1, -13.7]$. A proposer-matched demonstration-context reference with screen-select reduces the mismatch, but its paired improvement over the unscreened proposer is unresolved at $+1.6$ points $[-2.3, +5.6]$.

The subsequent fixed-candidate experiment uses the first 128 episode specifications from each evaluation seed ($n = 384$). At each decision, all selection rules see the same stream of four proposals. It is a distinct cohort and comparison, so its estimates are not interchangeable with the earlier $+1.6$-point estimate. Both scores and costs are retained in Table 43.

Table 43: Fixed-candidate screening comparison on PushT (upstream LeWM checkpoint), $n = 384$. Each selection rule uses the same four-candidate stream. Success carries pooled 95% intervals; planner time and search counts are per episode.

Rule	CEM solves	Proposals	Time (s)	Success % [95%]
First of four	10.7	42.8	18	32.0 [27.6,36.9]
Lowest raw residual	42.9	42.9	72	31.2 [26.8,36.1]
Highest calibrated minimum p	42.4	42.4	71	34.9 [30.3,39.8]

Calibrated minus raw ranking gives $+3.6$ points $[-1.0, +8.6]$, with 54 versus 40 discordant pairs ($p = 0.180$). Relative to executing the first candidate, raw and calibrated selection give -0.8 points $[-6.8, +5.5]$ and $+2.9$ $[-2.6, +8.6]$, at roughly four times the CEM solves. Across 8,185 logged candidate sets the two ranking rules choose the same candidate in 73% of decisions; 93% of their disagreements involve p-value ties. The support channel binds for some candidate in only 4% of sets. This helps explain why calibration can change the decision boundary without supplying a resolved planning benefit.

O Extended Interpretation and Future Directions

O.1 What reference sensitivity does and does not imply

The bridge is compatible with the split-conformal guarantee, which concerns a nominal population under its sampling assumptions. Two thresholds can have similar nominal audit rates and lie on opposite sides of most appearance-shifted scores. The shifted response depends jointly on nominal mass near the selected quantile and the concentration of shifted scores there. Heavy tails alone do not explain the contrast.

Under suitable i.i.d. sampling and a uniquely identified population quantile, increasing reference size can reduce threshold uncertainty. The tested references retain widely varying subgroup rates, and the frozen point forecast differs from the observed frequencies. Uncertainty in the estimated nominal tail is one possible contributor. Four largest blocks cannot determine a general sample-complexity requirement. A next experiment could freeze a tail estimator and test its predictions on additional disjoint collections, retaining failed forecasts and shared-audit dependence.

O.2 Calibration and discrimination require different remedies

The ordering argument in Appendix F explains why a threshold adjustment and a score change address different questions. Recalibration can change which scores cross an operating threshold, but leaves the raw paired ordering fixed. This matters when a benign transient dominates the maximum or an appearance change reduces physical/no-physical ordering. Adding shifted examples to the reference may lower benign alarms while also changing physical detection; both must be measured. Likewise, a satisfactory audit on a reference mixture controls an aggregate rate under the relevant assumptions, and leaves subgroup behaviour to separate evaluation.

The analyses establish specific information that the alarm can fail to expose. Action-conditioned fitting provides a stronger mismatch signal than action-free fitting in the focused PushT comparison; matched TwoRoom readouts recover position; later PushT windows expose action evidence masked by the onset maximum. These results motivate targeted follow-up tests. A state-supervised readout is not itself an unsupervised monitor, and a window chosen from the logged onset is not an online policy. The matched readouts refine the diagnosis without identifying a unique internal cause. When a colour-matched readout recovers position while the residual orders physical and benign twins poorly, accessible state coexists with weak selectivity of the tested score. When recovery remains poor, the evidence is limited to the readout families, data, and training procedures tried. Richer probes, direct interventions on predictor inputs, and comparisons across independently trained encoders could distinguish additional explanations. The current results support targeted diagnosis rather than a complete causal decomposition of the encoder and predictor.

O.3 Temporal evidence and adaptive decisions

The later-window analysis uses the known intervention onset to reveal evidence suppressed by a shared appearance transient. A deployed monitor does not receive that onset. Turning this finding into an improved detector would require specifying a causal temporal statistic, fixing its fitting and calibration procedure, and evaluating delayed and premature alarms on untouched episodes. Persistence can suppress brief transients while delaying short physical events, so success should be assessed using both selectivity and detection timing.

Decision evaluation introduces another change of population. A planner chooses contexts and proposals in response to previous scores, and rejection can itself change later queries. Matching a reference to a proposer at demonstration contexts therefore does not establish validity at deployment starts or under repeated adaptive screening. Weighted or sequential calibration would require a justified model of that process; neither is supplied by the current fixed-horizon guarantee. Any subsequent intervention study should also separate changes in ranking, candidate count, optimizer initialization, and computational budget.