BOUNDARY-FIDELITYBENCH: PROBING FIRST-CROSSING FIDELITY IN ACTION-CONDITIONED WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Action-conditioned world models support planning by predicting the consequences of candidate actions, yet accurate predictions do not guarantee that a model preserves the action conditions under which physical behavior changes. Existing evaluations assess physical consistency and action response, but do not explicitly characterize where the first physical-mode change occurs along controlled action perturbations. We therefore formulate first-crossing boundary fidelity as an evaluation target. We operationalize it through Boundary-FidelityBench, which combines controlled action probing with independent physical references, together with FCRG, a learned readout that uses scale-aware residual evidence for fine-grained estimates of boundary presence, first-crossing location, and supported transition types. Across seven mode-switching systems, 5,629 fresh evaluation rays, and six world-model outputs, boundary-resolved evaluation reveals differences not captured by predictive-quality and behavioral-outcome measures. FCRG achieves 80.76% exact-bin accuracy on fresh simulator responses and separates missed boundaries from incorrect first-interval assignments. First-crossing estimates can guide action-query allocation under a fixed world-model query budget, yielding gains in tested settings while showing that better boundary coverage does not always translate into better decisions. These results establish boundary fidelity as a fine-grained complement to world-model evaluation.

1 INTRODUCTION

Action-conditioned world models support planning by predicting the consequences of candidate actions (Zhou et al., 2025; Assran et al., 2025; Hafner et al., 2025). Yet accurate predictions do not ensure preservation of the action conditions under which physical behavior changes. A model may remain numerically close to the reference while misplacing a mode transition and assigning incorrect behavior near the true boundary. The question is therefore not only whether an outcome is accurate, but whether the model preserves the critical action conditions at which physical behavior changes.

Recent benchmarks evaluate world models beyond visual resemblance. Physics-IQ and WorldModelBench examine physical consistency and plausibility (Li et al., 2025; Motamed et al., 2026; Rädsch et al., 2026), while MiraBench evaluates action following and failure-related optimism (Yang et al., 2026). Most closely related, WorldSimProbe studies simulator-relative action responses and interaction fidelity, including the onset of persistent task failure under increasing action perturbations (Co et al., 2026). These evaluations characterize predicted action consequences. However, physical plausibility at sampled actions does not directly specify where a physical mode first changes, while task-failure thresholds remain tied to a chosen success criterion. The first physical-mode boundary along an action perturbation therefore remains a distinct evaluation target.

Measuring such a boundary is nontrivial because smooth within-mode variation, observation noise, and genuine mode changes can produce overlapping local response patterns (Pfrommer et al., 2021; Pang et al., 2023; Baranowski et al., 2019; Gavioli-Akilagun & Fryzlewicz, 2025; Li et al., 2024). We formalize this problem through first-crossing boundary fidelity. From a fixed initial state and reference action, actions are perturbed along a feasible one-sided path, and the first crossing is

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

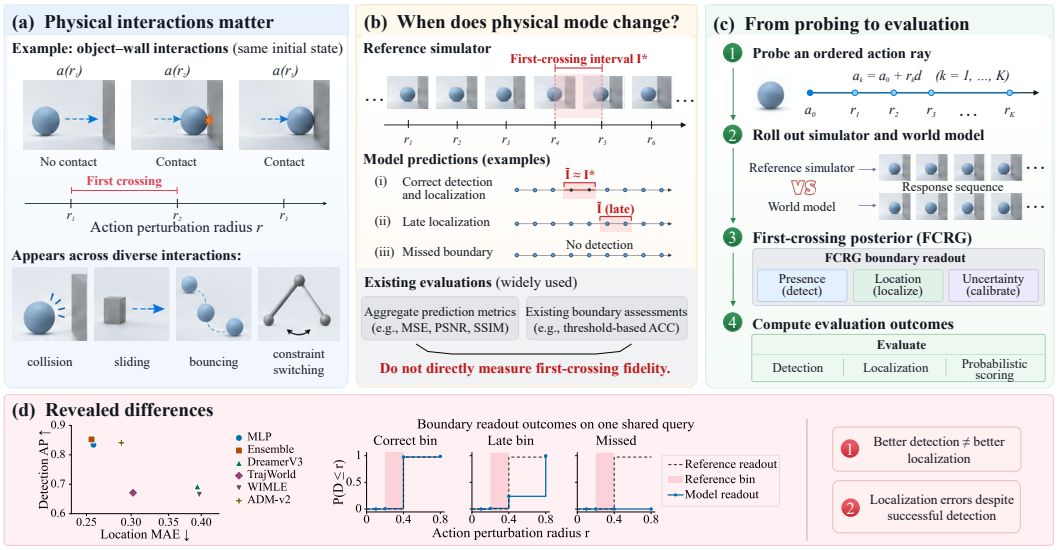

Figure 1: Overview of first-crossing boundary fidelity. Physical interactions, reference and predicted first crossings, the evaluation workflow, and complementary detection and localization outcomes.

the smallest perturbation at which the discrete physical signature departs from that of the reference action. This formulation makes boundary presence, first-crossing location, supported transition type, and directional displacement explicit evaluation targets.

We introduce Boundary-FidelityBench to operationalize this evaluation through controlled action probing and independent physical references. Simulator events and modes define reference boundaries, while world models are queried under matched initial states and action paths. Together with the benchmark, FCRG uses scale-aware residual evidence for fine-grained estimates of boundary presence, first-crossing location, and supported transition types from finite responses. The framework extends evaluation from predicted outcomes to the physical boundaries separating qualitatively different behaviors. Figure 1 summarizes the motivation and evaluation workflow.

Our contributions are threefold.

1. We introduce first-crossing boundary fidelity as a distinct evaluation target for action-conditioned world models. Building on existing evaluation criteria, it further assesses whether a model preserves the action conditions under which physical behavior first changes, and makes boundary existence, location, supported transition type, and directional displacement explicit and measurable model properties.
2. We develop Boundary-FidelityBench and FCRG as a first-boundary measurement framework grounded in independent physical references, enabling fine-grained boundary diagnosis through controlled action probing, independent physical references, and scale-aware residual evidence.
3. We show that boundary-resolved evaluation provides boundary information not explicitly captured by existing measures of predictive quality and behavioral outcomes, while distinguishing missed boundaries from mislocalized first crossings. Our downstream experiments further show that boundary estimates can improve action-query allocation under a fixed world-model query budget, demonstrating both diagnostic and downstream utility.

2 Related Work

2.1 Action-Conditioned World Models

Action-conditioned world models support planning through visual, latent, or state prediction. DINO-WM and the action-conditioned extension of V-JEPA 2 predict visual features for planning (Zhou

et al., 2025; Assran et al., 2025), while Dreamer and TD-MPC2 learn latent dynamics for control (Hafner et al., 2025; Hansen et al., 2024). TrajWorld extends prediction across heterogeneous state–action spaces (Yin et al., 2025), and Ctrl-World supports policy-in-the-loop robotic simulation (Guo et al., 2026). Contact-aware dynamics, manipulation planning, and neural event models address discontinuities and interaction-dependent transitions (Pfrommer et al., 2021; Pang et al., 2023; Chen et al., 2021; Suh et al., 2026). Our focus is not how such transitions are modeled, but whether learned responses preserve the action conditions at which they occur.

2.2 Evaluating Physical and Action-Conditioned Predictions

World-model evaluation has expanded from visual quality to physical consistency, physical commonsense, and functional utility (Li et al., 2025; Motamed et al., 2026; Rädsch et al., 2026; Meng et al., 2025; Huang et al., 2024; Zheng et al., 2025; Bansal et al., 2025; 2026; Shang et al., 2026; Bear et al., 2021). Controlled studies examine video prediction across physical regimes and distribution shifts (Kang et al., 2025; Xue et al., 2026), while WorldArena connects perceptual evaluation with downstream functionality (Shang et al., 2026).

More directly related diagnostics test whether predictions respond faithfully to supplied actions. MiraBench evaluates physics adherence, action following, and optimism under failure-inducing actions (Yang et al., 2026), while WorldEcho diagnoses off-expert action following and WorldSync improves intervention alignment (Chen et al., 2026). WorldSimProbe further studies simulator-relative action responses, local calibration, and interaction dynamics, including the onset of persistent task failure under increasing action perturbations (Co et al., 2026). This setting is closest to ours. However, task-failure thresholds depend on a chosen success criterion, whereas a physical-mode boundary marks where interaction behavior itself changes. Boundary-FidelityBench therefore targets the first physical-mode transition along an action perturbation, with reference transitions defined independently through simulator events and modes.

2.3 Change-Point Detection and Boundary Measurement

Change-point methods detect structural variation in ordered observations, including local interval searches, kernel segmentation, classification-based segmentation, differencing, and learned detectors (Baranowski et al., 2019; Gavioli-Akilagun & Fryzlewicz, 2025; Li et al., 2024; Arlot et al., 2019; Ermshaus et al., 2023). These methods provide tools for detection and localization, but a structural change in an observed response is not itself a physical-mode transition.

Physical evaluation additionally requires an independent definition of what constitutes a transition. Physion connects learned representations with physical contact labels through task-specific readouts (Bear et al., 2021). In our setting, simulator events and modes define the reference first crossing, while ordered action-conditioned responses provide only the evidence used to recover it. Separating the physical evaluation target from its readout allows Boundary-FidelityBench to compare prediction-quality metrics and alternative boundary measurements without conflating response variation with the physical transition itself.

3 Boundary-FidelityBench

3.1 First-Crossing Boundary Fidelity

Boundary-FidelityBench separates the physical reference from the responses used to measure it. For state x, feasible action set $\mathcal{A}$, horizon H, reference action a_0, and unit direction v, define the ray $a_\rho(r) = a_0 + rv$, where $\rho = (x, a_0, v)$ and $0 \le r \le R_\rho$ is the valid probing range. The simulator signature $\sigma_H(x, a) = (e_H(x, a), \ell_H(x, a), m_H(x, a))$ records event occurrence, first event type, and terminal mode; continuous within-mode variation does not change this signature.

The first-crossing distance is

$$d^*_\rho = \inf \{r \in (0, R_\rho] \mid \sigma_H(x, a_\rho(r)) \neq \sigma_H(x, a_0)\}. \tag{1}$$

The ordering is over action perturbations, not rollout time; later returns to the reference signature do not alter the first crossing. A resolved crossing has bracket $\mathcal{I}_\rho = [L_\rho, U_\rho]$ and supported transition label c_ρ; no crossing within a valid search is right-censored at R_ρ.

3.2 Controlled Probing and Physical References

At realized radii $0 = r_0 < \cdots < r_{K_\rho} = R_\rho$, simulator and model queries share the initial state, actions, horizon, and observation convention. Model responses are $Y_\rho^{\mathcal{M}} = (Q_H^{\mathcal{M}}(x, a_\rho(r_j)))_{j=0}^{K_\rho}$, with $Y_\rho^{\mathcal{E}}$ defined analogously. Realized distances and validity masks preserve query geometry (Figure 2). A separate, denser simulator search locates and refines the earliest signature change; sparse readout responses never determine reference labels. Unresolved cases remain distinct from valid no-crossing observations.

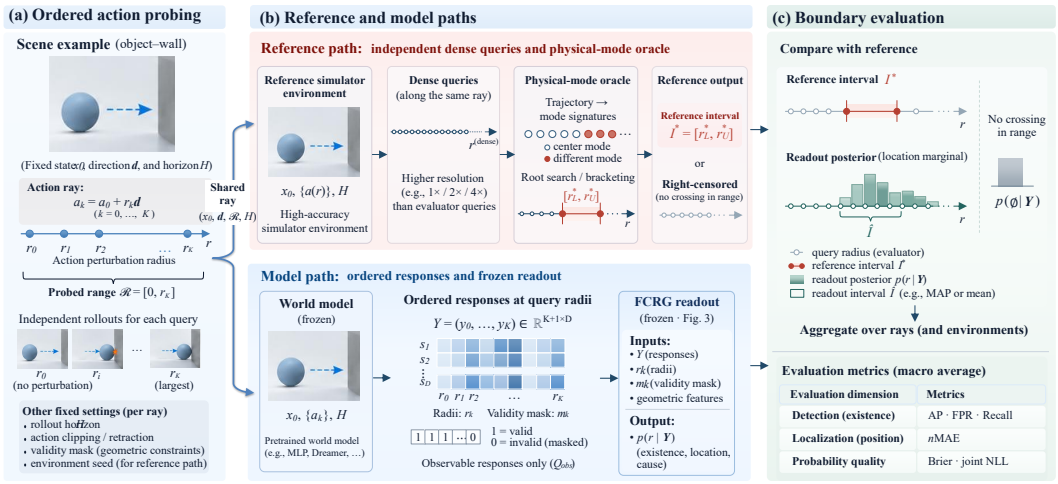

Figure 2: Boundary-FidelityBench evaluation protocol. Ordered action probing supports independent simulator-based reference construction and frozen world-model/readout evaluation under matched action rays.

Query intervals are $B_{\rho k} = (r_{k-1}, r_k]$. A reference bracket within one interval yields a unique label; otherwise all compatible labels are retained. Physical-reference precision and readout resolution therefore remain distinct (Appendix A.4.1). At inference, the readout uses observable responses, permitted state–action context, radii, and validity information; simulator signatures are reserved for supervision and scoring. Applying the same readout to simulator responses quantifies measurement error. Prediction and boundary scores use the same declared model output, with stochastic and ensemble conventions detailed in Appendix A.4.2.

3.3 Benchmark Construction and Evaluation Protocol

The suite contains seven mode-switching systems including PiecewiseAffine, SaturationKink, HardStop1D, BouncingBall, PuckWall2D, TwoBodyCollision, and StickSlipBlock, spanning piecewise, saturation, contact, impact, and frictional transitions, together with the smooth control SmoothDuffing. Each system contains eight parameter environments. The main protocol uses eight-step rollouts and four nominal nonzero radii $0.1, 0.2, 0.4, 0.8$, restricted by the feasible prefix.

After fitting and calibration on designated reference responses, readout parameters, normalization statistics, and thresholds are frozen within each system and shared across world models, which receive identical queries and references. The output satisfies $p_\rho(\varnothing) + \sum_{k,c} p_\rho(k, c) = 1$, over valid interval–type pairs and no crossing within range. Presence is $\pi_\rho = 1 - p_\rho(\varnothing)$, with detection at the calibrated threshold τ.

Table 1 summarizes the fine-grained dimensions. AP evaluates presence ranking; recall and FPR evaluate frozen-threshold detection. Exact-bin requires detection and the correct uniquely resolved interval. nMAE measures normalized distance to the reference bracket, including misses scored beyond the probing range. Miss, WrongBin, and Exact partition uniquely resolved reference-positive rays; early/late errors indicate displacement relative to the reference. Full definitions appear in Appendix A.4.3 and Table B2.

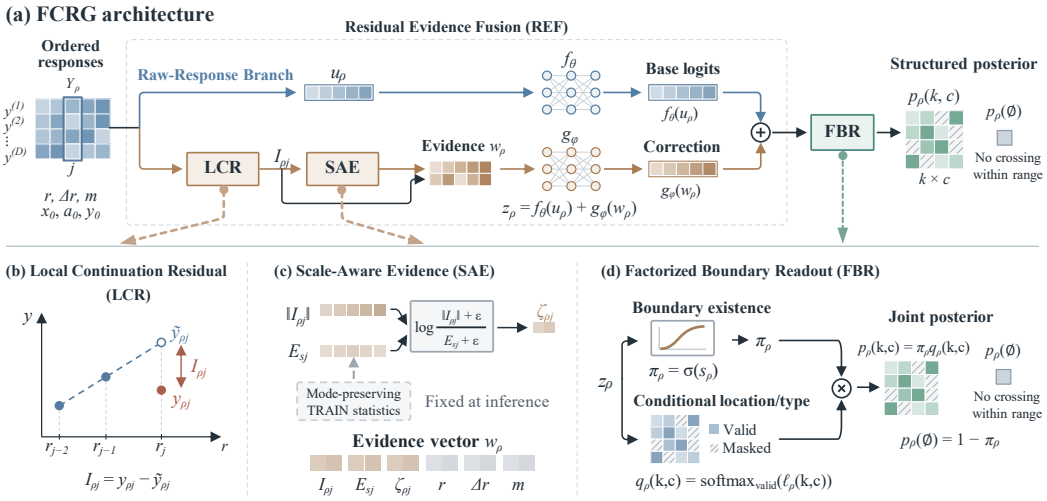

Figure 3: FCRG architecture. Local Continuation Residual (LCR), Scale-Aware Evidence (SAE), Residual Evidence Fusion (REF), and Factorized Boundary Readout (FBR) map ordered responses to a structured first-crossing posterior.

Table 1: Fine-grained evaluation dimensions and metrics.

Dimension	Information retained	Metrics
Boundary presence	Crossing presence and false alarms	AP, recall, FPR
First-crossing location	First interval and distance to reference	Exact-bin, interval-distance MAE, nMAE
Error attribution	Missed, incorrect, early or late crossings	Miss, WrongBin, Exact; early/late rates
Probabilistic agreement	Presence and compatible interval–type outcomes	Binary Brier score and joint NLL

Let b_ρ indicate a reference crossing and $\mathcal{S}_\rho$ its compatible interval–type labels. Joint NLL retains label ambiguity, supported types, and right-censoring:

$$\ell_\rho = \begin{cases} -\log\left(\sum_{(k,c)\in\mathcal{S}_\rho} p_\rho(k,c)\right), & b_\rho = 1, \\ -\log p_\rho(\varnothing), & b_\rho = 0. \end{cases} \tag{2}$$

Results use equal environment and system weights, with SmoothDuffing reported separately. Seed averaging and statistical procedures follow Section 5.1; undefined metrics retain their support counts (Appendix A.4.4).

4 FCRG for Boundary Measurement

FCRG combines ordered responses with scale-aware local residual evidence to produce the structured boundary posterior of Section 3. Figure 3 shows its four components.

4.1 Local Continuation Residual and Scale-Aware Evidence

Let $y_{\rho j} \in \mathbb{R}^D$ be the response at r_j, standardized using reference training data. With $\Delta r_j = r_j - r_{j-1}$, LCR extrapolates the preceding local trend:

$$\widetilde{y}_{\rho j} = y_{\rho,j-1} + \frac{\Delta r_j}{\Delta r_{j-1}}\left(y_{\rho,j-1} - y_{\rho,j-2}\right), \qquad I_{\rho j} = y_{\rho j} - \widetilde{y}_{\rho j}, \tag{3}$$

for eligible $j \geq 2$. The spacing ratio handles nonuniform probes. Signed residuals retain radius-indexed deviations, but smooth curvature and noise can also produce them, so they are evidence rather than boundary decisions.

SAE compares residual magnitudes with a coordinate-wise envelope E_{sj} estimated only from valid, reference-confirmed mode-preserving training prefixes:

$$\zeta_{\rho j} = \log \frac{\|I_{\rho j}\|_2 + \epsilon}{\|E_{sj}\|_2 + \epsilon}, \tag{4}$$

Here $\epsilon > 0$ ensures numerical stability. LCR preserves the signed response pattern, while SAE measures its magnitude relative to within-mode variation. Envelope estimation and validity rules are given in Appendix B.4.1.

4.2 Residual Evidence Fusion and Factorized Boundary Readout

REF combines raw-response inputs $\mathbf{u}_\rho$ and residual/scale evidence $\mathbf{w}_\rho$ through two multilayer perceptrons, $\mathbf{z}_\rho = f_\theta(\mathbf{u}_\rho) + g_\phi(\mathbf{w}_\rho)$. The raw branch retains the first probe interval, where LCR lacks a complete stencil; the evidence branch learns corrections. Inputs include permitted context, query geometry, and validity masks (Appendix B.4.2).

For fused scores $\mathbf{z}_\rho = (\eta_\rho, \{A_{\rho kc}\}_{k,c})$, FBR sets $\pi_\rho = \mathrm{sigmoid}(\eta_\rho)$ and $q_\rho = \mathrm{softmax}_{\Omega_\rho}(A_\rho)$, normalized over admissible interval–type pairs Ω_ρ:

$$p_\rho(k, c) = \pi_\rho q_\rho(k, c), \qquad p_\rho(\varnothing) = 1 - \pi_\rho. \tag{5}$$

Mass outside Ω_ρ is zero. This separates existence from conditional localization while retaining location–type dependence; the location marginal is $q_\rho(k) = \sum_c q_\rho(k, c)$. Appendix B.4.2 gives the complete masked normalization.

4.3 Training and Inference

FCRG minimizes the joint NLL in Eq. (2), supplemented by an auxiliary pairwise presence-ranking loss. We first train f_θ, then freeze it and learn g_ϕ through the combined logits. Noise augmentation mixes clean and perturbed reference responses while preserving their physical labels. The fitted readout remains frozen within each system and is shared across evaluated world models under Section 3.3. Loss construction, ranking weight, initialization, and data separation are detailed in Appendix B.4.3.

5 Experiments

Our experiments examine the necessity, validity, and downstream value of first-boundary measurement. We first identify boundary information not explicitly captured by existing predictive-quality and behavioral-outcome measures, then validate the readout and apply it to world-model diagnosis. We further test whether estimated first intervals can guide action-query allocation under a fixed world-model query budget. Robustness, ablations, and computational analysis characterize the conditions and design choices supporting these results.

5.1 Experimental Setup

For fresh-input evaluation (Sections 3–4), we evaluate 2,048 new centers sampled after model and calibration freezing, yielding 5,629 eligible rays comprising 1,577 reference positives and 4,052 negatives. We evaluate MLP, Ensemble, DreamerV3 (Hafner et al., 2025), TrajWorld (Yin et al., 2025), WIMLE (Aghabozorgi et al., 2026), and ADM-v2 (Lin et al., 2026) outputs, comparing FCRG with its clean-trained counterpart FCRG-clean and the noise-trained second-difference readout DD2-noise.

Results average readout and world-model seeds, then use equal environment and event-system weights; SmoothDuffing is reported separately. Confidence intervals use grouped bootstrap resampling, and prespecified readout contrasts use two-sided sign-flip tests with Holm correction. Data roles, training, output definitions, and statistical details appear in Appendices A and B.

5.2 Complementary Information from First-Boundary Evaluation

We compare boundary deletion with MSE-matched noise and shift a physical boundary from 0.3 to 0.6 to test whether aggregate and local-response scores resolve boundary changes. Table 2 shows identical MSEs for responses with different boundary presence and a near-unit local response ratio despite the location shift. These controlled contrasts establish the additional information supplied by explicit presence and first-location evaluation (Appendix C.1).

Table 2: Boundary changes under aggregate error and local-response scores. Both MSEs have zero parameter-cell SD; the reference ratio is unity. The shifted ratio reports mean ± SD across eight parameter cells.

Contrast	**Response condition**	**Aggregate / local score**	**Generator-defined target**
Error-matched	Complete boundary deletion	MSE = 0.0058	No boundary within the scan
Error-matched	Matched observation noise	MSE = 0.0058	Original generator boundary retained
Physical shift	Reference distance 0.3	Ratio = 1.0000	First boundary at 0.3
Physical shift	Shifted distance 0.6	Ratio = 0.9946 ± 0.0019	First boundary at 0.6

We extend this comparison to learned predictions and behavioral criteria. On SaturationKink, WIMLE combines small MSE and high AP with lower Exact-bin than simulator responses under the same readout (Appendix C.4). Matched PuckWall2D and StickSlipBlock scans reveal frequent disagreement between physical-mode changes and persistent task failure (Appendix C.2), while non-penetration can hold with incorrect first intervals (Appendix C.3). Boundary evaluation thus identifies transition errors alongside outcome accuracy, task success, and physical constraints.

5.3 Measurement Accuracy and Fine-Grained World-Model Diagnosis

We test boundary-change measurement using deletion, equal-error noise, and a physical shift. It raises no alarms after deletion, retains 94.79 ± 3.04% detection under noise, and recovers the destination bin for the $0.3 \rightarrow 0.6$ shift in all 32 paired blocks across three seeds. These interventions distinguish disappearance, preservation, and the tested displacement (Appendix D.1).

We next evaluate fresh simulator responses to isolate measurement accuracy from world-model prediction error. FCRG reaches 80.76% Exact-bin, versus 77.01% for FCRG-clean and 63.59% for DD2-noise, establishing the readout's accuracy on unseen reference responses (Appendix D.2).

Table 3: World-model boundary evaluation (95% CIs: Table D4a). Exact: end-to-end exact-bin accuracy. Bold in Tables 3–5 marks best estimates, not statistical significance.

Model	**AP↑**	**NLL↓**	**FPR (%)↓**	**Recall (%)↑**	**Exact (%)↑**	**nMAE↓**
MLP	0.8343	0.6084	9.93	84.00	59.40	0.2572
Ensemble	**0.8524**	0.5888	9.17	84.99	**59.45**	**0.2552**
DreamerV3	0.6927	0.9557	15.58	67.82	32.39	0.3938
TrajWorld	0.6712	1.3837	26.74	**85.90**	43.98	0.3025
WIMLE	0.6644	0.5742	9.31	60.24	44.09	0.3968
ADM-v2	0.8411	**0.4264**	**6.71**	80.62	55.36	0.2884

We then apply the shared readout to six world models to identify their boundary errors. Ensemble leads in AP, while ADM-v2 has the lowest NLL and FPR (Table 3). Figure 4 separates misses from incorrect intervals. On BouncingBall, MLP and TrajWorld have recalls of 85.88% and 82.00% but a 27.25-point Exact-bin gap, while WIMLE primarily misses boundaries. This decomposition identifies whether errors arise in presence detection or first-location assignment.

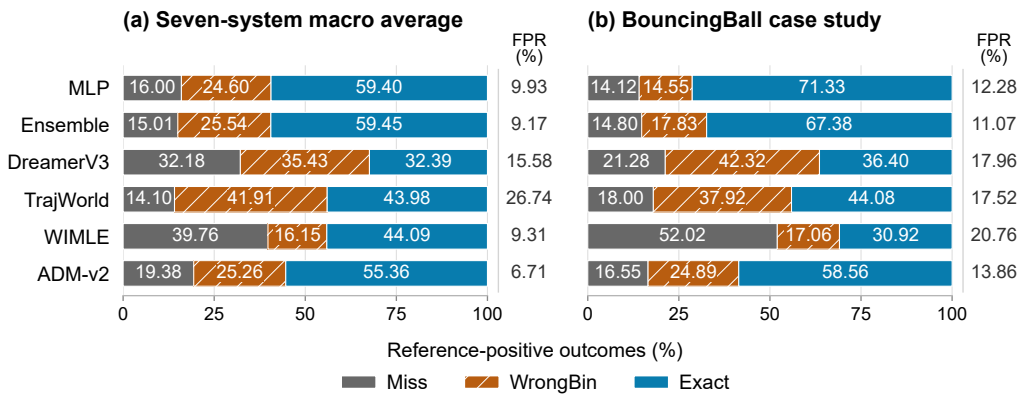

Figure 4: Boundary outcome decomposition: (a) seven-system macro average; (b) BouncingBall case study. Both panels include all six models. Miss, WrongBin, and Exact partition reference-positive rays; FPR uses reference negatives. Appendix Table D5 gives all per-system results.

Readout comparisons assess the consistency of these diagnoses. FCRG-clean retains the qualitative gaps, and Holm-corrected tests support FCRG's AP gains over DD2-noise for five of six outputs and NLL reductions over FCRG-clean for five of six, excluding TrajWorld. These results support improved presence ranking and joint probability quality across multiple outputs (Appendix D.4).

5.4 Boundary-Guided Action Search under a Fixed Query Budget

We test terminal-position success on StickSlipBlock and PuckWall2D using frozen MLP and TrajWorld to assess whether first intervals improve query allocation. FCRG retains original candidates and adds queries around predicted first intervals. Uniform and Local use equal query budgets; reference guidance uses privileged boundaries, and Base retains only original candidates (Appendix E.1).

Table 4: Fresh-input action success. Bold: best of Uniform, Local, and FCRG; ties at shown precision.

Search rule	**StickSlip / MLP (%)**	**StickSlip / Traj World (%)**	**PuckWall / MLP (%)**	**PuckWall / Traj World (%)**
Base	46.48 ± 50.54	30.08 ± 33.51	32.55 ± 9.89	28.52 ± 9.20
Uniform	67.06 ± 43.66	36.85 ± 28.09	33.98 ± 8.94	**31.25** ± 6.77
Local	82.81 ± 18.98	40.76 ± 14.62	**34.11** ± 9.91	29.69 ± 10.02
Reference boundary	95.44 ± 10.52	53.13 ± 18.93	34.11 ± 8.85	30.86 ± 7.06
FCRG guidance	**91.93** ± 11.93	**52.43** ± 19.89	**34.11** ± 8.85	29.90 ± 6.23

FCRG improves StickSlip success over Local by 9.11 percentage points with MLP and 11.68 with TrajWorld, with environment CIs [1.04, 20.31] and [4.12, 18.79]. Higher successful-candidate coverage links these gains to finding useful actions (Appendix E.2). PuckWall2D shows no comparable success gain. Coverage and selection analyses distinguish candidate discovery from the model's ability to select successful actions, explaining how boundary guidance contributes to search (Appendix E). FCRG reuses existing probe responses, making this information available without additional world-model probe trajectories (Appendix H).

5.5 Robustness to Observation Perturbations

We test noisy interval discrimination under matched augmentation on shared development responses, separate from the 5,629 fresh evaluation rays. FCRG attains 70.76% Exact-bin and NLL 0.3386, compared with 65.98% and 0.4263 for Raw and 65.87% and 0.3511 for MLP32. Its 90.09% recall

and 9.56% FPR compare with 97.53% and 51.95% for DD2-noise. FCRG improves localization and joint probability quality over Raw and MLP32, with fewer false alarms than DD2-noise.

We also halve response amplitudes while preserving boundaries to characterize measurement resolution. FCRG retains 100.00% correct-bin detection at distances 0.3 and 0.6, while Exact-bin at 0.05 and 0.15 falls from 18.75% and 40.63% to 6.25% and 5.21%. This contrast localizes sensitivity to weak nearby boundaries. Testing smooth negative responses yields no alarms on 512 fresh SmoothDuffing rays across six model outputs, supporting specificity on this control (Appendix F).

5.6 Module Ablations

On historical development-validation responses, we ablate direct LCR input or SAE to isolate signed residual and scale-aware evidence contributions, retaining the raw-response path, REF, and FBR (Table 5). The first variant retains residual magnitudes through SAE; the second removes the envelope and excess representation together. Appendix G reports all variants and per-system results.

Table 5: Development-validation ablations of direct LCR input and SAE. Exact denotes end-to-end exact-bin accuracy.

Configuration	**Clean**		**Noisy**	
	NLL↓	**Exact (%)↑**	**NLL↓**	**Exact (%)↑**
FCRG	**0.1773**	**81.65**	0.3386	**70.76**
w/o direct LCR input	0.1799	81.01	0.3481	70.31
w/o SAE	0.3238	69.55	**0.3308**	68.90

Removing SAE reduces clean Exact-bin by 12.10 percentage points and increases NLL from 0.1773 to 0.3238; removing the direct LCR vector reduces Exact-bin by 0.65 points. Full FCRG also achieves the highest noisy Exact-bin, while the SAE ablation has the lowest noisy NLL. These results identify scale-aware organization of LCR-derived evidence as the larger contributor to interval accuracy in these ablations and distinguish its localization benefit from joint probability scoring.

6 Conclusion

We study whether action-conditioned world models preserve the critical action conditions at which physical behavior first changes. To make this property explicit, we introduce first-crossing boundary fidelity and develop Boundary-FidelityBench together with FCRG to measure boundary presence, first-crossing location, and supported transition types from controlled action-conditioned responses. This extends world-model evaluation from agreement at predicted outcomes to agreement on the physical-mode boundaries separating them in action space.

Our experiments show that boundary-resolved evaluation provides information not explicitly captured by existing measures of predictive quality and behavioral outcomes. Responses with similar aggregate errors can preserve different boundary structures, while strong boundary-presence scores do not substitute for first-location accuracy. Fine-grained diagnosis further separates missed boundaries from incorrect interval assignments and probability-quality errors. First-crossing estimates can also guide action-query allocation under a fixed world-model query budget, improving useful candidate coverage and task success in some tested settings.

The results also delineate the current scope of the approach. Boundary information does not guarantee downstream gains across tasks, as its utility depends on the action space, candidate-allocation strategy, and the world model's ability to rank predicted consequences. FCRG also remains sensitive to some weak nearby boundaries and observation perturbations. Future work can extend boundary-fidelity evaluation to higher-dimensional action spaces, richer physical interactions, and visual world models, and investigate tighter integration of boundary uncertainty with planning and control. Overall, Boundary-FidelityBench offers a complementary fine-grained view of whether a world model preserves not only predicted outcomes, but also where physical behavior first switches as actions change.

AI USE STATEMENT

AI assistance was used to assist with research brainstorming, software implementation and debugging, literature search and summarization, and manuscript editing and formatting. All methodological decisions, experiments, results, and scientific claims were reviewed and verified by the authors. AI-generated suggestions were treated as auxiliary input only, and the authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

An anonymized implementation of the core components is available at `https://anonymous.4open.science/r/Boundary-FidelityBench-2F5F/`. The complete codebase will be made publicly available upon publication.

REFERENCES

Mehran Aghabozorgi, Alireza Moazeni, Yanshu Zhang, and Ke Li. WIMLE: Uncertainty-aware world models with IMLE for sample-efficient continuous control. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.14351.

Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of Machine Learning Research*, 20(162):1–56, 2019. URL https://jmlr.org/beta/papers/v20/16-155.html.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985, 2025. URL https://arxiv.org/abs/2506.09985.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xMlYKYFd03.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HA8KSQW7SO.

Rafal Baranowski, Yining Chen, and Piotr Fryzlewicz. Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3):649–672, 2019. doi: 10.1111/rssb.12322.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Fei-Fei Li, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith Fan. Physion: Evaluating physical prediction from vision in humans and machines. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/d09bf41544a3365a46c9077ebb5e35c3-Abstract-round1.html.

Ricky T. Q. Chen, Brandon Amos, and Maximilian Nickel. Learning neural event functions for ordinary differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=kW_zpEmMLdP.

Sixiang Chen, Jiaming Liu, Jixian Wu, Yichen Guo, Tinghao Wang, Siyuan Qian, Hao Chen, Jiajun Cao, Jian Tang, and Shanghang Zhang. Do robotic world models really follow actions? diagnosing and aligning action-conditioned generation for policy learning. arXiv preprint arXiv:2608.24885, 2026. URL https://arxiv.org/abs/2608.24885.

Peterson Co, Sicheng Hu, Chunxuan Jiao, Hongyang Cheng, Yulin Luo, Yijie Xu, Sixiang Chen, Zhongxia Zhao, Zihao Wang, DaFeng Chi, Peidong Liu, YuTong Chen, Henghua Liu, Zhihao Yuan, Huizhu Jia, Yuzheng Zhuang, Tianle Zhang, Liang Lin, Huajie Tan, and Shanghang Zhang. WorldSimProbe: Diagnosing simulator faithfulness in action-conditioned world models for embodied manipulation. arXiv preprint arXiv:2608.09298, 2026. URL https://arxiv.org/abs/2608.09298.

Arik Ermshaus, Patrick Schäfer, and Ulf Leser. ClaSP: Parameter-free time series segmentation. *Data Mining and Knowledge Discovery*, 37:1262–1300, 2023. doi: 10.1007/s10618-023-00923-x.

Shakeel Gavioli-Akilagun and Piotr Fryzlewicz. Fast and optimal inference for change points in piecewise polynomials via differencing. *Electronic Journal of Statistics*, 19(1):593–655, 2025. doi: 10.1214/25-EJS2345.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-World: A controllable generative world model for robot manipulation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.10125.

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

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/pdf/d7962b8e764953c9c4db96aa85307e9e9b689881.pdf.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28991–29017, 2025. URL https://proceedings.mlr.press/v267/kang25g.html.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1834. Datasets and Benchmarks Track.

Jie Li, Paul Fearnhead, Piotr Fryzlewicz, and Tengyao Wang. Automatic change-point detection in time series via deep learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):273–285, 2024. doi: 10.1093/jrsssb/qkae004.

Haoxin Lin, Siyuan Xiao, Yi-Chen Li, Zhilong Zhang, Yihao Sun, Chengxing Jia, and Yang Yu. ADM-v2: Pursuing full-horizon roll-out in dynamics models for offline policy learning and evaluation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ICbXEwqpga.

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

Learning, volume 267 of *Proceedings of Machine Learning Research*, pp. 43781–43806, 2025. URL https://proceedings.mlr.press/v267/meng25c.html.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 948–958, 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Motamed_Do_Generative_Video_Models_Understand_Physical_Principles_WACV_2026_paper.html.

Tao Pang, H. J. Terry Suh, Lujie Yang, and Russ Tedrake. Global planning for contact-rich manipulation via local smoothing of quasi-dynamic contact models. *IEEE Transactions on Robotics*, 39 (6):4691–4711, 2023. doi: 10.1109/TRO.2023.3300230.

Samuel Pfrommer, Mathew Halm, and Michael Posa. ContactNets: Learning discontinuous contact dynamics with smooth, implicit representations. In *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pp. 2279–2291, 2021. URL https://proceedings.mlr.press/v155/pfrommer21a.html.

Tim Rädsch, Yuki M. Asano, Hilde Kuehne, Stefan Bauer, Priyank Jaini, Robert Geirhos, and Carsten T. Lüth. Physics-IQ verified. arXiv preprint arXiv:2606.18943, 2026. URL https://arxiv.org/abs/2606.18943.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, Chen Gao, Wei Wu, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Wenwu Zhu, and Yong Li. WorldArena: A unified benchmark for evaluating perception and functional utility of embodied world models. arXiv preprint arXiv:2602.08971, 2026. URL https://arxiv.org/abs/2602.08971.

H. J. Terry Suh, Tao Pang, Tong Zhao, and Russ Tedrake. Dexterous contact-rich manipulation via the contact trust region. *The International Journal of Robotics Research*, 45(9):1418–1454, 2026. doi: 10.1177/02783649251398875.

Haotian Xue, Yipu Chen, Liqian Ma, Zelin Zhao, Lama Moukheiber, Yuchen Zhu, and Yongxin Chen. ACWM-Phys: Investigating generalized physical interaction in action-conditioned video world models. arXiv preprint arXiv:2605.08567, 2026. URL https://arxiv.org/abs/2605.08567.

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. arXiv preprint arXiv:2605.29360, 2026. URL https://arxiv.org/abs/2605.29360.

Shaofeng Yin, Jialong Wu, Siqiao Huang, Xingjian Su, Xu He, Jianye Hao, and Mingsheng Long. Trajectory world models for heterogeneous environments. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72462–72484, 2025. URL https://proceedings.mlr.press/v267/yin25f.html.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755, 2025. URL https://arxiv.org/abs/2503.21755v2.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135, 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.

APPENDIX: EVIDENCE, MECHANISMS, AND IMPLEMENTATION OF FIRST-BOUNDARY MEASUREMENT

The appendix follows the argument of the main paper: measurement targets and comparable settings, information beyond existing scores, measurement validity and model diagnosis, action-search utility, and finally robustness, module contributions, and cost. Each section motivates its comparison before presenting the corresponding evidence. Development and fresh-input evaluations remain separate.

- A Measurement Target and Data Construction
- B Models, Readouts, and Statistical Definitions
- C First-Boundary Information Beyond Existing Criteria
- D Measurement Validation and Complete World-Model Diagnosis
- E From Boundary Information to Fixed-Budget Action Search
- F Robustness and Measurement Resolution
- G Module Contributions and Internal Representations
- H Computation and Numerical Materials

Additional construction and evaluation details appear in Appendix A.4; additional FCRG implementation details appear in Appendix B.4.

A MEASUREMENT TARGET AND DATA CONSTRUCTION

A.1 ACTION RAYS AND FIRST MODE BOUNDARIES

Boundary-FidelityBench uses programmatically generated state–action queries and reference dynamical responses rather than field-recorded trajectories as its primary evaluation data. Each unit comprises an initial state, central action, and perturbation direction. After sampling the state and central action within a fixed parameter environment, we construct ordered perturbations along positive and negative coordinate directions. The ray is a query design: the simulator generates reference trajectories by executing each action, while world models independently predict the matched inputs through free-running rollouts. These responses have distinct generation sources.

The main scan requests radii 0.1, 0.2, 0.4, and 0.8, retaining the admissible prefix under action constraints. A constant-action rollout supplies eight observations. The reference boundary is the first departure of the finite-horizon discrete mode signature from that of the central action. "First" is defined along action space, not by the time of first contact within one trajectory.

This construction places numerical error and mode structure on comparable responses: preservation and first-interval assignment can be assessed at matched inputs without relying on a task-specific success threshold. Reference labels use simulator signatures and the established numerical-stability and eligibility rules; exact event time is not part of the primary signature-equality test. No change within the admissible scan is treated as range censoring, while ineligible requests remain in the manifest rather than becoming negative labels.

A.2 SYSTEMS, SHARED DATA, AND EVALUATION SPLITS

Table A1 summarizes the systems and reference support.

Table A1: Systems and reference support for the main evaluation. Eight fixed parameter environments per system. SmoothDuffing is a separate negative control and is excluded from event-system macro averages.

System	New centers	Requested rays	Eligible rays	Positive	Scan-censored
PiecewiseAffine	256	1024	1024	500	524
HardStop1D	256	512	512	95	417
SaturationKink	256	512	512	363	149
BouncingBall	256	512	510	120	390
StickSlipBlock	256	512	512	353	159
PuckWall2D	256	1024	1023	75	948
TwoBody Collision	256	1024	1024	71	953
SmoothDuffing	256	512	512	0	512
Total	2,048	5,632	5,629	1,577	4,052

The six world models share 512 training and 128 validation trajectories per parameter environment. The 2,048 fresh centers are generated once after models, training-derived scales, and calibration are frozen, and are shared by all models and readouts. Readout training, validation, and calibration partitions are separate from this evaluation. Fixed environments preserve paired comparisons; fresh centers assess previously unused state–action inputs rather than retraining or generalization to new parameter environments.

Table A2: Roles of the experimental datasets.

Dataset	Support	Role
Main evaluation	2,048 centers; 5,629 eligible rays	Frozen-model, fresh-input measurement
Controlled deletion	80 rays; 64 original positives; 8 parameter cells	Existence under error-matched intervention
Physical displacement	32 paired blocks; 5 distances	First-interval measurement
Geometry–boundary joint analysis	256 centers; 768 rays; 16 environments	Complementary diagnostics on the same output
Task-conditioned comparison	80 PuckWall rays; 155 Stick Slip rays	Mode boundary versus task failure
Robustness / ablations	Shared development responses	Matched-recipe design analysis
Fresh-input action search	2 tasks $\times$ 8 environments $\times$ 32 inputs	Fixed-rule decision utility

Table A2 distinguishes generation sources by experimental role. Main, geometric, and task analyses use simulator references and matched learned-model responses. Controlled deletion uses an explicit branch generator; physical displacement uses the original simulator to create pairs with known boundary locations; observation noise and amplitude halving modify specified responses. These controlled targets validate measurement of known changes and are not pooled with unmodified model predictions as one performance population. Fresh evaluation centers are fixed after model and calibration freezing but before simulation and scoring; censoring and eligibility follow reference rules rather than resampling on model success.

This construction supports matched queries, independent reference labels, and targeted controls for presence, location, and observation perturbations. Its value is experimental control over the measurement target, not treating sample count as a substitute for physical-mechanism coverage. Each dataset retains its stated role and analysis denominator.

A.3 Resolution and Observability

Physical reference brackets and sparse query bins provide different resolutions. Every eligible main-evaluation positive maps to a unique reference bin, without converting its continuous bracket into an exact point. Geometric checks instead operate directly on recorded states. Reference boundaries, observable output constraints, and response readouts thus complement one another while retaining distinct meanings.

Table A3: Direct contact witnesses in eight-observation reference responses. From the two-system observability audit at the base tolerance. Entries are query records, not independent centers.

System	Event-positive query records	Directly witnessed contact
BouncingBall	952	37
PuckWall2D	703	0

Table A3 distinguishes absence of a contact witness from absence of an event. Penetration, in contrast, can be checked directly from observed positions and supplies a geometric criterion independent of FCRG. This separation allows the analysis to use both observable evidence and boundary readouts without imputing unobserved events as model labels.

A.4 Additional Details of Boundary Construction and Evaluation

A.4.1 Query Geometry and Physical Reference Construction

Action coordinates and scales are fixed within each system, and opposite perturbation directions are treated as separate rays. In the state-based suite, each response contains a predicted state trajectory. Realized action distances and valid-prefix masks preserve the actual query geometry, while repeated or infeasible action locations terminate the valid prefix.

Reference generation performs an outward simulator scan followed by local bracket refinement around the earliest resolved signature change. Grid-refinement and solver-consistency checks are used to assess reference stability. Numerically unresolved cases retain explicit validity flags and remain distinct from valid no-crossing observations. The sparse responses supplied to the boundary readout are not used to generate these reference labels.

The physical-reference resolution and readout resolution are intentionally separated. A reference bracket fully contained within one sparse query interval produces a unique interval label, whereas a bracket spanning several query intervals retains all compatible labels. The query-interval midpoint is used only as a point summary of location. Two crossings that are distinguishable by the denser physical reference but fall within the same sparse interval therefore share the same readout label. This prevents physical-reference precision from being conflated with the resolution available to the readout.

A.4.2 Output Semantics for Stochastic and Ensemble Models

For stochastic and ensemble world models, prediction error and boundary metrics are computed from the same declared output. When the declared output is the complete mean response $\bar{Y}_\rho$, both evaluations use $\bar{Y}_\rho$, including the boundary readout $\mathcal{G}(\bar{Y}_\rho)$.

Averaging boundary distributions obtained from individual samples or ensemble members is reported as a separate output because

$$\mathcal{G}(\mathbb{E}[Y]) \neq \mathbb{E}[\mathcal{G}(Y)]$$

in general. The first quantity measures the boundary structure of the aggregated response, whereas the second aggregates boundary distributions obtained from individual responses. These procedures therefore represent different measurement objects and are not interchanged when relating prediction error to boundary fidelity.

Applying the same boundary readout to environment responses provides a reference-response measurement baseline without world-model prediction error. This baseline reflects the capability of the readout under the same sparse observation interface and should not be interpreted as a theoretical oracle score.

A.4.3 Additional Interpretation of Evaluation Metrics

Boundary-presence metrics describe different aspects of the same detection problem. AP evaluates the ranking of boundary-existence evidence across rays, while recall and FPR characterize behavior at the frozen operating threshold. Binary Brier score evaluates the probabilistic quality of the existence probability π_ρ. Strong ranking performance therefore does not necessarily imply accurate thresholded decisions or well-calibrated probabilities.

Joint NLL evaluates the probability assigned to the complete reference observation. For a reference-positive ray, compatible interval–type outcomes are summed before taking the logarithm, while a reference-negative ray is scored through the no-crossing probability. Supported transition labels therefore contribute to the joint target, allowing probability assigned to an incompatible transition type to be distinguished from probability assigned to a compatible reference outcome. A fixed probability floor is used only for numerical stability.

For detected boundaries, the predicted bin maximizes $\sum_c p_\rho(k, c)$, and its midpoint defines $\hat{d}_\rho$. Missed boundaries use $R_\rho + \frac{1}{2}(r_{K_\rho} - r_{K_\rho - 1})$ as the scoring coordinate. The interval-relative error and normalized MAE are

$$e_\rho = \hat{d}_\rho - \mathrm{proj}_{[L_\rho, U_\rho]}(\hat{d}_\rho), \qquad \mathrm{nMAE} = \frac{1}{N_+} \sum_{\rho \,|\, b_\rho = 1} \frac{|e_\rho|}{R_\rho}, \tag{A.1}$$

where N_+ is the number of reference-positive rays. This end-to-end definition retains missed detections, and exact-bin accuracy likewise requires detection and the correct uniquely resolved interval.

Localization metrics are evaluated end to end. Exact-bin accuracy requires both a boundary detection and selection of the uniquely resolved reference interval. nMAE retains missed detections through the scoring coordinate defined in Eq. (A.1), rather than evaluating distance only on successfully detected rays. Interval-distance MAE retains the corresponding unnormalized distance from the compatible reference bracket.

For additional error attribution, reference-positive rays are separated into missed, detected-in-wrong-bin, and detected-in-correct-bin outcomes. Detected localization errors can further be divided into detected-early and detected-late outcomes according to the sign of e_ρ. The late-or-missed rate combines missed detections with detected-late predictions while keeping both components available separately.

These rates use reference-positive rays as their denominator, whereas FPR is defined over reference-negative rays. Early and late describe the estimated boundary position relative to the physical reference and do not by themselves indicate beneficial, harmful, safe, or unsafe downstream behavior. Task consequences are evaluated separately.

A.4.4 Aggregation and Undefined Metrics

Metrics are first computed within parameter environments after averaging the designated readout and world-model seeds. Environment-level results are then aggregated with equal weights, followed by equal weighting across the seven mode-switching systems for the main benchmark summary. SmoothDuffing is reported separately as the smooth no-transition control.

Metrics that are undefined because the required class or event support is absent remain undefined rather than being replaced with zero. Their effective support is retained in reporting. This convention is particularly important for event-conditioned localization quantities and smooth negative controls.

The benchmark therefore preserves detection, localization, probabilistic agreement, and error attribution as separate evaluation dimensions rather than collapsing them into a single composite score.

B Models, Readouts, and Statistical Definitions

B.1 Prediction Outputs and Readout Conditions

Table B1 defines the six prediction outputs and their training budgets. The adapted external predictors are DreamerV3 (Hafner et al., 2025), TrajWorld (Yin et al., 2025), WIMLE (Aghabozorgi et al., 2026), and ADM-v2 (Lin et al., 2026).

Table B1: Definitions of the six evaluated outputs.

Model	Update cap	Main output
MLP	80,000	Member 0; one complete response
Ensemble	80,000 per member	Mean of three complete MLP responses
DreamerV3	20,000	Mean of eight sampled complete responses
TrajWorld	20,000	Native mean decoding
WIMLE	20,000	Mean of eight sampled complete responses
ADM-v2	20,000	Mean of eight sampled complete responses

Predictors are fitted per parameter environment, and the main MLP is nested within the three-member ensemble. MLP checkpoints follow one-step validation error, whereas adapted external models retain their prescribed multistep validation rules; boundary scores do not select checkpoints. All predictions free-run from the same true initial state without subsequent teacher forcing. Main tables compare the specified adapted outputs rather than compute-matched architecture training.

Response MSE is paired with the boundary readout of that same complete response. For stochastic or ensemble models, reading the response mean differs from averaging per-response posteriors; the latter is examined separately in D.5. FCRG-clean is the full clean-trained FCRG, not the Raw baseline. DD2-noise uses second-order differences and a fitted linear head; Raw directly reads response features, and MLP32 is a raw-response readout with 32 hidden units. These readouts are distinct from the MLP predictor.

FCRG retains a frozen raw-response branch and learns logit corrections from LCR and SAE evidence; REF combines the branches and FBR provides presence and conditional location–cause probabilities. The final training views comprise 50.00% clean and 50.00% noisy responses. FCRG, FCRG-clean, and DD2-noise each retain temperature and threshold values from their separate calibration data. World-model seeds are 2026091600, 2026091700, and 2026091800; readout seeds are 20260910, 20260911, and 20260912.

MLP performs feedforward state-transition regression from the current state and action. TrajWorld uses Transformer sequence modeling with discretized state representations and native-mean decoding. They share the state–action query interface while using different internal predictive representations, providing two predictor conditions for the downstream comparison in Appendix E.1 without retraining for boundary guidance.

B.2 From Fine-Grained Targets to Metrics

FBR represents the boundary distribution as $p_\rho(k, c) = \pi_\rho q_\rho(k, c)$, with $p_\rho(\varnothing) = 1 - \pi_\rho$. Presence is coupled to a conditional joint location–cause distribution, rather than three independent classifiers. Ranking, fixed-threshold alarms, first intervals, and probability quality can consequently be examined separately. Fine granularity resides both in the ray-level evaluation unit and in this attribute and error decomposition.

Table B2: Metrics and diagnostic targets.

Field	Short label	Target
`average_precision`	AP	Presence ranking

Continued on next page

Table B2 (continued)

Field	Short label	Target
`normalized_ap`	nAP	Prevalence-adjusted AP
`binary_brier`	Binary Brier	Squared error of presence probabilities
`adaptive_ece`	ECE	Adaptive-bin calibration error
`calibration_bias`	Cal. bias	Signed presence-probability bias
`joint_interval_nll`	Joint NLL	Joint NLL for interval–cause and censoring outcomes
`joint_categorical_brier_singleton_only`	Joint Brier	Joint squared probability loss on singleton-bin labels
`interval_nll`	Interval NLL	Probability assigned to event intervals
`survival_nll`	Survival NLL	Probability assigned to absence/ censoring
`false_positive_rate`	FPR	Alarms on reference negatives
`operating_recall`	Recall	Detections on reference positives
`operating_precision`	Precision	Reference positives among alarms
`exact_bin_accuracy`	Exact-bin	Alarm and correct reference bin
`within_one_bin_accuracy`	Within-one	Alarm within the permitted adjacent-bin range
`interval_distance_mae`	Dist. MAE	Prescribed interval distance and missed-detection handling
`normalized_mae`	nMAE	Normalized interval distance, including missed-detection penalty
`signed_bias`	Loc. bias	Signed location bias
`dangerous_late_rate`	Late	Historical composite of late readout and missed detection
`conservative_early_rate`	Early	Early-location rate
`deletion_rate`	Miss	Missed reference-boundary rate
`hallucination_rate`	Hallucination	Equal to reference-negative FPR in this interface

For a reference-positive ray with one reference bin and a frozen-threshold alarm $\widehat{y}$,

$$\Pr(\widehat{y}=0 \mid y=1)+\Pr(\widehat{y}=1,\widehat{k}\neq k^{*} \mid y=1)+\Pr(\widehat{y}=1,\widehat{k}=k^{*} \mid y=1)=1.$$

These terms are Miss, WrongBin, and Exact. They separate absent detections from incompatible locations after detection. Closed-interval overlap in the geometry analysis, Exact-bin in the main evaluation, and the conditional median bin used for search retain their respective definitions.

Late, Miss, and Hallucination in Table B2 describe response readouts, not observed accidents, deleted physical events, or internal event hallucinations. Cause probabilities are interpreted within labeled support; joint NLL does not replace a separate cause-accuracy evaluation. AP and event localization are NA without positives, and unsupported class metrics or insufficient-environment SDs remain NA.

B.3 Paired Statistics, Aggregation, and Missingness

Main scores average readout seeds, then world-model seeds, then equally weight environments and the seven event systems. Reference responses are not replicated by world-model seeds. Directions from one center, repeated readouts, and training seeds are paired repetitions, not additional independent centers. Environment SD describes variation across the fixed environments. Main-score 95% intervals resample 32 centers within each environment 10,000 times with shared weights across methods, directions, and seeds, recomputing AP, probability, and distance metrics.

The readout family contains 36 contrasts: six outputs, FCRG versus FCRG-clean or DD2-noise, and AP, joint NLL, or nMAE. Mean-response versus posterior-mixture comparisons form a separate 12-test family. Paired environment intervals resample the eight environments within each system 10,000

times. Two-sided tests use 10,000 whole-environment sign flips with the Monte Carlo plus-one correction and separate Holm adjustment per family. All primary contrasts retain 56 environments. Sign exchangeability of paired differences is the working null assumption. Environment intervals and adjusted tests are not inversions of the same inferential construction, so both are retained when their conclusions differ.

Planning has 24 prespecified contrasts across four task–model combinations, FCRG versus Local or Uniform, and coverage, success, or terminal cost. Environment-block and within-environment input resampling each use 10,000 draws to form individual 95% intervals, without additional p-values or simultaneous-coverage claims. Undefined class-dependent bootstrap draws remain undefined with counts reported. Unknown planning costs remain NA, invalid selected actions count as success failures, and complete-case means do not replace full-denominator estimates.

B.4 Additional Implementation Details of FCRG

B.4.1 Residual Construction and Validity Handling

LCR is evaluated only when the current and two preceding query locations are valid and the corresponding spacings are positive. The ratio $\Delta r_j/\Delta r_{j-1}$ in Eq. (3) accommodates nonuniform probe spacing. The reference response $y_{\rho 0}$ provides the initial history. The first nonzero probe remains available to the raw-response branch even though a complete three-point residual stencil is unavailable at that position.

Retaining the full signed residual vector rather than only its norm preserves coordinate-level and radius-indexed structure. The sign of an LCR residual describes deviation in the response coordinates and should not be interpreted as an earlier or later boundary. Early and late localization are defined only after the predicted crossing is compared with the independent physical reference.

SAE interprets this evidence relative to mode-preserving training variation. For system s, radius index j, and coordinate d,

$$E_{sjd} = \mathrm{Quantile}_{\gamma}\left(\left\{|I_{\rho j,d}| \;\middle|\; \rho \in \mathcal{T}^0_{s,j}\right\}\right), \tag{B.1}$$

where $\mathcal{T}^0_{s,j}$ contains rays with valid residuals and reference-confirmed mode-preserving prefixes through r_j. We use $\gamma = 0.99$. SAE retains both positive and negative values of $\zeta_{\rho j}$. Positive values indicate residual magnitudes above the empirical mode-preserving envelope, whereas negative values indicate magnitudes below it. Neither case is independently thresholded into a boundary label.

B.4.2 REF Sequence Processing and Output Masking

The raw-response input $\mathbf{u}_\rho$ collects ordered responses, permitted state–action context, query geometry, and validity information. The evidence input $\mathbf{w}_\rho$ collects LCR residuals, SAE outputs, radii, spacings, and validity masks. Both branches are multilayer perceptrons whose output logits are added as $\mathbf{z}_\rho = f_\theta(\mathbf{u}_\rho) + g_\phi(\mathbf{w}_\rho)$.

Both REF branches process the complete valid sequence of a ray. This allows observations at larger perturbation radii to contribute evidence when interpreting an earlier candidate crossing. Invalid response locations and residual features are masked, and different action rays are processed independently.

The geometric branch acts additively on the logits produced by the raw-response branch rather than replacing them. Consequently, removing the direct LCR residual input does not necessarily remove all information derived from local residual magnitude when SAE features remain available.

The complete masked FBR normalization is

$$\begin{aligned} \pi_\rho &= \mathrm{sigmoid}(\eta_\rho), \\ q_\rho(k,c) &= \frac{\exp(A_{\rho kc})}{\sum_{(k',c')\in\Omega_\rho} \exp(A_{\rho k'c'})}, \qquad (k,c) \in \Omega_\rho, \\ p_\rho(k,c) &= \pi_\rho q_\rho(k,c), \qquad p_\rho(\varnothing) = 1-\pi_\rho. \end{aligned} \tag{B.2}$$

For FBR, probability outside the admissible set Ω_ρ is fixed to zero. The smooth control retains a generic boundary category for detection even though it contains no simulator-defined physical mode transition under the reference construction.

The factorization in Eq. (5) is a probability-chain parameterization. It separates the marginal probability of boundary existence from the conditional distribution over location–type outcomes while retaining the dependence between location and supported transition type.

B.4.3 Training Procedure and Data Separation

FCRG is trained using the physical references defined in Section 3. Substituting Eq. (5) into the interval likelihood gives

$$\ell_\rho = \begin{cases} -\log \pi_\rho - \log\left(\sum_{(k,c)\in\mathcal{S}_\rho} q_\rho(k,c)\right), & b_\rho = 1, \\ -\log(1-\pi_\rho), & b_\rho = 0. \end{cases} \tag{B.3}$$

Positive rays supervise existence and compatible location–type outcomes, while right-censored rays supervise no crossing within the observed range. Ambiguous references contribute through the total probability assigned to compatible labels.

An auxiliary ranking objective acts on boundary-presence scores. For positive–negative pairs $\mathcal{P}$ drawn within the same system and matched probe-coverage strata,

$$\begin{aligned} \mathcal{L}_{\text{rank}} &= \frac{1}{|\mathcal{P}|} \sum_{(\rho^+,\rho^-)\in\mathcal{P}} \log\big(1+\exp\big[\eta_{\rho^-} - \eta_{\rho^+}\big]\big), \\ \mathcal{L} &= \frac{1}{N}\sum_{\rho=1}^{N} \ell_\rho + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}}. \end{aligned} \tag{B.4}$$

We use $\lambda_{\text{rank}} = 0.1$. Batches without eligible positive–negative pairs contribute zero ranking loss. The joint likelihood remains responsible for the structured boundary outcome, whereas the auxiliary ranking term specifically improves ordering of boundary-presence scores.

Training follows the additive REF construction. The raw-response branch f_θ is optimized first. Its parameters are then frozen while g_ϕ is trained through the combined logits. The output layer of g_ϕ is initialized to zero, ensuring that the initial combined predictor is identical to the already fitted raw branch. Subsequent updates therefore learn corrections rather than replacing the raw-response prediction.

The noise-augmented recipe mixes clean and perturbed reference responses while preserving their physical labels. LCR and SAE features are recomputed from the perturbed inputs, whereas the SAE envelope remains the estimate obtained from the training data.

Normalization statistics and SAE envelopes are estimated only from the designated training split. Model checkpoints are selected using the validation split, and operating thresholds are determined using a separate calibration split. Training budgets and augmentation settings are reported with the experimental protocol.

B.4.4 Inference and Computational Cost

At inference, FCRG parameters, normalization statistics, SAE envelopes, and calibrated operating rules remain fixed within each physical system and are shared across evaluated world models. No simulator event or mode signature is supplied to FCRG as an inference feature.

FBR provides the existence probability and conditional interval distribution used to derive the detection, localization, probabilistic, and error-decomposition metrics defined in Section 3. Reported crossing locations remain query intervals, with their midpoints used only when a point summary is required.

Constructing LCR and SAE requires $O(K_\rho D)$ arithmetic for a ray with K_ρ probe locations and response dimension D, followed by the two REF branches and FBR. FCRG operates on responses

already produced for the benchmark queries and therefore requires neither additional world-model queries nor gradients through the evaluated world model.

C First-Boundary Information Beyond Existing Criteria

This section moves beyond observing different scores to identifying different evaluation targets. Controlled responses isolate aggregate error, while matched-ray task and geometry analyses hold the evaluated outputs fixed to examine what boundary information remains unspecified.

C.1 Error Matching and Local Response Ratios

The analytic set contains 80 rays in eight parameter cells: 64 original positives and 16 scan-censored negatives. Deletion fractions of 25.00%, 50.00%, 75.00%, and 100.00% replace 16, 32, 48, and 64 positive rays with an explicit single-branch generator. The fraction controls the number of fully modified rays, not attenuation of each boundary. Observation-noise controls retain the original generator labels and match deletion MSE on every ray. Table C1 reports all four fractions. At complete deletion, both conditions have parameter-cell mean MSE 0.0058 and zero between-cell SD; the reference response ratio is unity by construction. Thus the nonzero SD in Table 2 describes only the shifted-ratio variation, not a confidence interval.

Table C1: Error matching for all four deletion fractions. MSE is aggregated over all 80 paired rays. Generator labels determine structural preservation; readout accuracy is analyzed in D.1.

Deletion (%)	**Modified positives**	**Deletion MSE**	**Noise MSE**
25.00	16	0.0013	0.0013
50.00	32	0.0026	0.0026
75.00	48	0.0041	0.0041
100.00	64	0.0058	0.0058

These pairs give direct examples in which aggregate error does not identify boundary preservation. Mode structure differs despite identical error magnitude. First-boundary evaluation consequently adds a structural attribute missing from the scalar error rather than measuring the same numerical discrepancy again.

Physical shifts use 32 paired blocks from the original piecewise-affine simulator at distances 0.05, 0.15, 0.3, 0.6, and 0.95. The numerical ratio core uses the center and two queries, 0.1/0.8 in the main example; FCRG uses the center and four admissible queries. When the boundary moves from 0.3 to 0.6, the ratio remains 0.9946 ± 0.0019 (mean $\pm$ sample SD across eight parameter cells). It describes similarity in relative response magnitude, while the first interval additionally specifies where the change begins in action space. Table D2 retains all four shifts and readout outcomes; distance 0.95 is censored within the scan ending at 0.8.

C.2 Task Outcomes and Physical Mode Changes

Task-conditioned comparisons evaluate persistent task failure and first physical-mode changes on the same action scans. PuckWall2D retains 80 reference-eligible rays, while StickSlipBlock contains 155 eligible rays from distinct centers. Eligibility is determined from reference conditions rather than model performance. Table C2 gives the joint mode-change and task-failure counts.

Table C2: Joint counts of physical-mode changes and persistent task failure.

Task	Reference mode	Persistent failure	No persistent suffix	Final point only	Total
PuckWall2D	Mode change	0	14	1	15
PuckWall2D	No mode change	28	21	16	65
StickSlipBlock	Mode change	51	0	0	51
StickSlipBlock	No mode change	104	0	0	104

Mode transitions and task failures are not interchangeable labels. Contact or sliding can remain compatible with a task objective, while task failure can also arise without a physical-mode change. A first boundary is therefore better interpreted as a structural region worth examining than as an automatically prohibited action.

Table C3a: PuckWall2D task-conditioned six-model comparison. Mean ± environment SD; the local ratio is a score on a 0–100 scale, not a success rate. Recall uses 6 positive-support environments; the other displayed metrics use 8.

Model	Local ratio score	Local MSE	Joint NLL	Recall (%)	FPR (%)
MLP	90.5252 ± 6.2743	0.0278 ± 0.0223	0.3976 ± 0.1985	98.77 ± 3.02	32.59 ± 19.50
Ensemble	91.1720 ± 5.5892	0.0208 ± 0.0188	0.2940 ± 0.1633	98.15 ± 4.54	34.41 ± 17.96
DreamerV3	68.7714 ± 6.6064	0.2051 ± 0.0771	0.8228 ± 0.5897	53.58 ± 12.89	33.13 ± 23.60
TrajWorld	30.5731 ± 11.7113	0.0568 ± 0.0347	1.0105 ± 0.7551	95.31 ± 6.05	61.90 ± 26.32
WIMLE	97.8658 ± 2.4494	0.2484 ± 0.1681	1.0524 ± 0.7304	20.74 ± 25.38	29.92 ± 17.32
ADM-v2	90.7651 ± 4.7182	0.1860 ± 0.1509	0.6662 ± 0.5022	72.65 ± 30.24	22.50 ± 13.18

Table C3b: StickSlipBlock task-conditioned six-model comparison. Mean ± environment SD; the local ratio is a score on a 0–100 scale, not a success rate. Recall uses 5 positive-support environments; the other displayed metrics use 8.

Model	Local ratio score	Local MSE	Joint NLL	Recall (%)	FPR (%)
MLP	92.8928 ± 4.3618	0.0037 ± 0.0029	2.2846 ± 2.0004	78.49 ± 16.49	0.32 ± 0.92
Ensemble	94.0601 ± 3.3234	0.0035 ± 0.0030	3.0525 ± 2.3353	80.00 ± 18.81	0.65 ± 1.34
DreamerV3	79.8417 ± 8.2328	0.1038 ± 0.0542	0.2704 ± 0.2983	2.47 ± 5.52	0.00 ± 0.00
TrajWorld	46.6059 ± 15.6698	0.0937 ± 0.0865	3.1703 ± 2.2274	61.99 ± 28.60	13.22 ± 11.67
WIMLE	78.1448 ± 13.9290	0.0027 ± 0.0032	2.0899 ± 2.0376	59.35 ± 37.50	4.53 ± 6.69
ADM-v2	82.9522 ± 13.6795	0.0033 ± 0.0045	2.3568 ± 2.6516	50.12 ± 42.42	2.81 ± 5.41

Tables C3a and C3b show that high local-ratio scores can accompany low response error or substantially different boundary readouts. On PuckWall, WIMLE combines a high ratio with larger

response error, whereas MLP and Ensemble combine high ratios with lower error. Retaining the full model set demonstrates complementarity rather than using one model to dismiss all existing criteria. DreamerV3's low StickSlip recall additionally distinguishes ranking from fixed-threshold alarms; high AP alone does not determine realized detection.

Table C3c: Controlled checks of the local-ratio formula. These constructions evaluate the ratio formula only, not FCRG performance.

Task	**Transformation**	**Local ratio score**	**Local MSE**
PuckWall2D	action_ignored	0.0000 ± 0.0000	0.0029 ± 0.0036
PuckWall2D	half_response	$100.0000 \pm 1.07 \times 10^{-13}$	0.0007 ± 0.0009
PuckWall2D	reversed_query_order	0.0000 ± 0.0000	0.0039 ± 0.0065
PuckWall2D	unchanged_reference	100.0000 ± 0.0000	0.0000 ± 0.0000
StickSlipBlock	action_ignored	0.0000 ± 0.0000	$0.0070 \pm 1.19 \times 10^{-17}$
StickSlipBlock	half_response	$100.0000 \pm 5.71 \times 10^{-14}$	$0.0018 \pm 2.17 \times 10^{-19}$
StickSlipBlock	reversed_query_order	0.0000 ± 0.0000	$0.0028 \pm 1.64 \times 10^{-19}$
StickSlipBlock	unchanged_reference	100.0000 ± 0.0000	0.0000 ± 0.0000

The controls in Table C3c show that an unchanged ratio after amplitude halving reflects the diagnostic's scaling property, not a failure. Together with boundary-deletion and location targets in D.1, these controls distinguish invariance to observation transforms from sensitivity to structural changes, rather than treating every response difference as a boundary event.

C.3 Additional Diagnosis Among Geometrically Admissible Outputs

Geometric admissibility and first-boundary fidelity test different properties. BouncingBall and PuckWall2D permit direct checks of ground and wall non-penetration on saved states. These checks are independent of FCRG and can therefore be compared with first-boundary agreement on the same rays. Figure C1 summarizes the joint diagnosis.

DreamerV3 and TrajWorld show no observed penetration in PuckWall2D, yet their FCRG alarm-and-overlap rates are $20.19 \pm 4.39\%$ and $47.11 \pm 16.98\%$, with FPRs of $16.62 \pm 4.23\%$ and $26.14 \pm 5.43\%$. A geometrically valid state sequence can therefore still disagree with the reference on whether and where a physical-mode change occurs.

The complete three-readout decomposition, negative-ray false positives, penetration-conditioned subsets, and geometry-tolerance analysis are retained in Tables C4, C5, C6, and C7, respectively.

The geometric rule independently checks recorded positions at the center or any legal query for ground or wall penetration. Its base tolerance is 32 times float32 machine precision, approximately 3.81×10^{-6} in state units. Geometry and readout records are joined by system, environment, center, ray, model, and world-model seed, retaining 16 environments, 256 centers, and 768 rays. BouncingBall has 59 reference positives and 197 negatives; PuckWall has 41 positives and 471 negatives.

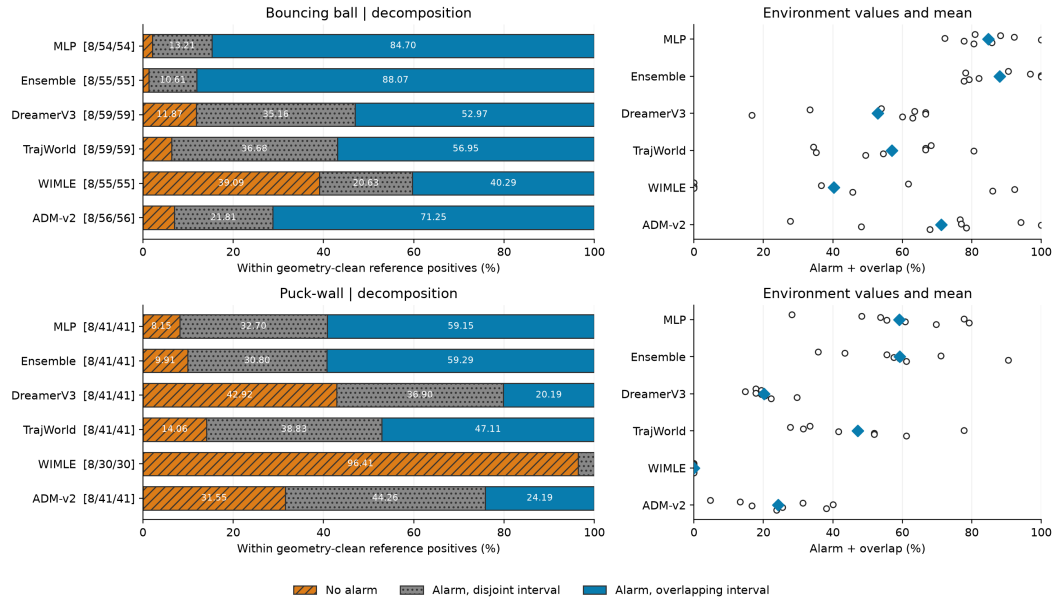

Figure C1: Boundary diagnosis of geometrically valid responses. Reference-positive rays for six model outputs are partitioned into no alarm, alarm with a disjoint interval, and alarm with an overlapping interval. Environment-level values and equal-weight means are shown together with their effective support.

Table C4: Three-readout decomposition of geometry-valid reference positives. Mean ± environment SD. Conditional support can differ by model; the three outcomes share each row's reference-positive denominator.

System	Model	Readout	Env/ center/ ray	No alarm (%)	Disjoint alarm (%)	Overlapping alarm (%)
Bouncing Ball	MLP	FCRG	8/54/54	2.08 ± 4.87	13.21 ± 7.95	84.70 ± 8.77
Bouncing Ball	MLP	FCRG-clean	8/54/54	2.08 ± 4.87	16.83 ± 11.91	81.09 ± 11.47
Bouncing Ball	MLP	DD2-noise	8/54/54	2.68 ± 3.86	17.06 ± 8.98	80.26 ± 9.48
Bouncing Ball	Ensemble	FCRG	8/55/55	1.32 ± 2.77	10.61 ± 9.17	88.07 ± 9.89
Bouncing Ball	Ensemble	FCRG-clean	8/55/55	1.32 ± 2.77	13.48 ± 13.36	85.20 ± 14.42
Bouncing Ball	Ensemble	DD2-noise	8/55/55	2.57 ± 4.06	22.09 ± 13.62	75.34 ± 14.10
Bouncing Ball	Dreamer V3	FCRG	8/59/59	11.87 ± 12.46	35.16 ± 12.29	52.97 ± 18.28
Bouncing Ball	Dreamer V3	FCRG-clean	8/59/59	12.16 ± 12.48	36.92 ± 9.37	50.92 ± 16.99
Bouncing Ball	Dreamer V3	DD2-noise	8/59/59	17.47 ± 15.10	37.51 ± 10.09	45.03 ± 14.93
Bouncing Ball	TrajWorld	FCRG	8/59/59	6.37 ± 5.43	36.68 ± 13.44	56.95 ± 16.55

Continued on next page

Table C4 (continued)

System	Model	Readout	Env/ center/ ray	No alarm (%)	Disjoint alarm (%)	Overlapping alarm (%)
Bouncing Ball	TrajWorld	FCRG-clean	8/59/59	7.15 ± 6.11	37.91 ± 11.48	54.94 ± 16.05
Bouncing Ball	TrajWorld	DD2-noise	8/59/59	0.23 ± 0.65	42.38 ± 11.12	57.38 ± 11.49
Bouncing Ball	WIMLE	FCRG	8/55/55	39.09 ± 50.51	20.63 ± 26.37	40.29 ± 38.10
Bouncing Ball	WIMLE	FCRG-clean	8/55/55	39.09 ± 50.51	24.22 ± 27.10	36.69 ± 34.76
Bouncing Ball	WIMLE	DD2-noise	8/55/55	44.92 ± 45.89	21.16 ± 19.31	33.91 ± 29.37
Bouncing Ball	ADM-v2	FCRG	8/56/56	6.94 ± 8.83	21.81 ± 16.78	71.25 ± 23.60
Bouncing Ball	ADM-v2	FCRG-clean	8/56/56	6.94 ± 8.83	23.98 ± 16.88	69.08 ± 23.50
Bouncing Ball	ADM-v2	DD2-noise	8/56/56	8.72 ± 9.83	38.02 ± 11.93	53.26 ± 14.49
Puck Wall2D	MLP	FCRG	8/41/41	8.15 ± 9.44	32.70 ± 18.96	59.15 ± 16.81
Puck Wall2D	MLP	FCRG-clean	8/41/41	9.70 ± 10.35	30.70 ± 17.49	59.60 ± 14.10
Puck Wall2D	MLP	DD2-noise	8/41/41	3.99 ± 6.73	42.24 ± 17.56	53.78 ± 18.66
Puck Wall2D	Ensemble	FCRG	8/41/41	9.91 ± 13.13	30.80 ± 18.62	59.29 ± 16.64
Puck Wall2D	Ensemble	FCRG-clean	8/41/41	11.02 ± 11.59	30.65 ± 18.92	58.33 ± 14.85
Puck Wall2D	Ensemble	DD2-noise	8/41/41	7.79 ± 11.33	41.13 ± 18.43	51.08 ± 24.21
Puck Wall2D	Dreamer V3	FCRG	8/41/41	42.92 ± 21.87	36.90 ± 21.65	20.19 ± 4.39
Puck Wall2D	Dreamer V3	FCRG-clean	8/41/41	54.50 ± 22.73	29.39 ± 22.37	16.11 ± 9.16
Puck Wall2D	Dreamer V3	DD2-noise	8/41/41	21.04 ± 25.66	46.92 ± 19.93	32.04 ± 20.76
Puck Wall2D	TrajWorld	FCRG	8/41/41	14.06 ± 13.14	38.83 ± 16.32	47.11 ± 16.98
Puck Wall2D	TrajWorld	FCRG-clean	8/41/41	17.95 ± 15.16	39.66 ± 23.17	42.38 ± 16.75
Puck Wall2D	TrajWorld	DD2-noise	8/41/41	1.39 ± 3.93	61.19 ± 14.69	37.42 ± 13.34
Puck Wall2D	WIMLE	FCRG	8/30/30	96.41 ± 5.19	3.59 ± 5.19	0.00 ± 0.00
Puck Wall2D	WIMLE	FCRG-clean	8/30/30	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Puck Wall2D	WIMLE	DD2-noise	8/30/30	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Puck Wall2D	ADM-v2	FCRG	8/41/41	31.55 ± 26.29	44.26 ± 23.07	24.19 ± 12.24
Puck Wall2D	ADM-v2	FCRG-clean	8/41/41	33.08 ± 24.90	47.21 ± 20.02	19.71 ± 9.56
Puck Wall2D	ADM-v2	DD2-noise	8/41/41	11.07 ± 15.39	57.34 ± 19.44	31.58 ± 22.14

Table C4 retains both stronger and weaker outputs. In geometry-valid BouncingBall cases, FCRG alarm-and-overlap rates reach 84.70% for MLP and 88.07% for Ensemble, while DreamerV3 and TrajWorld exhibit more interval disagreement despite no penetration. The diagnosis therefore does more than add failure labels: it distinguishes detection and first-location compatibility among outputs receiving the same geometric assessment. Alternative readouts retain this information distinction, although its magnitude varies.

Table C5: Full-support geometry and negative-ray false alarms. Mean ± environment SD. Penetration, positive overlap, and negative FPR have different denominators and are not subtracted.

System	Model	Penetration (%)	Overlap on all positives (%)	FPR on all negatives (%)	FPR on valid negatives (%)
Bouncing Ball	MLP	20.83 ± 11.19	77.30 ± 15.14	11.78 ± 10.90	12.29 ± 13.48
Bouncing Ball	Ensemble	10.29 ± 9.12	81.38 ± 17.60	11.56 ± 10.81	12.25 ± 12.78
Bouncing Ball	DreamerV3	0.00 ± 0.00	52.97 ± 18.28	19.61 ± 20.77	19.61 ± 20.77
Bouncing Ball	TrajWorld	0.00 ± 0.00	56.95 ± 16.55	18.65 ± 13.90	18.65 ± 13.90
Bouncing Ball	WIMLE	12.24 ± 15.23	32.51 ± 28.51	21.48 ± 18.30	21.80 ± 19.34
Bouncing Ball	ADM-v2	23.70 ± 13.31	66.50 ± 25.53	12.13 ± 11.86	11.13 ± 10.12
PuckWall2D	MLP	3.84 ± 2.46	56.23 ± 14.56	11.88 ± 4.80	10.52 ± 4.70
PuckWall2D	Ensemble	0.85 ± 1.39	58.52 ± 16.23	11.01 ± 5.22	10.84 ± 5.29
PuckWall2D	DreamerV3	0.00 ± 0.00	20.19 ± 4.39	16.62 ± 4.23	16.62 ± 4.23
PuckWall2D	TrajWorld	0.00 ± 0.00	47.11 ± 16.98	26.14 ± 5.43	26.14 ± 5.43
PuckWall2D	WIMLE	20.77 ± 6.28	0.56 ± 1.57	12.13 ± 4.60	2.37 ± 2.13
PuckWall2D	ADM-v2	6.84 ± 6.36	28.44 ± 14.29	9.34 ± 3.44	8.19 ± 3.59

Table C6: Reference-positive rays with penetration. NA denotes no conditional support; zero-penetration models are not assigned perfect scores in this stratum.

System	Model	Env/ center/ ray	No alarm (%)	Disjoint alarm (%)	Overlapping alarm (%)
BouncingBall	MLP	6/28/28	22.67 ± 27.83	23.12 ± 17.24	54.20 ± 36.90
BouncingBall	Ensemble	4/9/9	53.49 ± 41.44	17.94 ± 19.04	28.57 ± 48.09
BouncingBall	DreamerV3	0/0/0	NA	NA	NA
BouncingBall	TrajWorld	0/0/0	NA	NA	NA
BouncingBall	WIMLE	4/15/15	42.80 ± 40.42	30.15 ± 24.62	27.05 ± 26.90
BouncingBall	ADM-v2	7/33/33	15.42 ± 18.47	27.49 ± 25.41	57.09 ± 30.87
PuckWall2D	MLP	7/12/12	28.97 ± 21.02	38.10 ± 31.50	32.94 ± 26.37

Continued on next page

Table C6 (continued)

System	Model	Env/ center/ ray	No alarm (%)	Disjoint alarm (%)	Overlapping alarm (%)
PuckWall2D	Ensemble	4/4/4	8.33 ± 16.67	54.17 ± 36.96	37.50 ± 43.83
PuckWall2D	DreamerV3	0/0/0	NA	NA	NA
PuckWall2D	TrajWorld	0/0/0	NA	NA	NA
PuckWall2D	WIMLE	7/23/23	65.73 ± 17.65	32.37 ± 19.25	1.90 ± 5.04
PuckWall2D	ADM-v2	7/20/20	9.95 ± 17.25	46.72 ± 41.22	43.33 ± 35.75

Table C7: Tolerance sensitivity of the joint diagnosis.

System	Model	Readout	1× (%)	10× (%)	100× (%)	Max. change (pp)
BouncingBall	ADM-v2	FCRG-clean	69.08	69.08	68.63	0.45
BouncingBall	ADM-v2	FCRG	71.25	71.25	71.19	0.06
BouncingBall	ADM-v2	DD2-noise	53.26	53.26	52.99	0.27
BouncingBall	DreamerV3	FCRG-clean	50.92	50.92	50.92	0.00
BouncingBall	DreamerV3	FCRG	52.97	52.97	52.97	0.00
BouncingBall	DreamerV3	DD2-noise	45.03	45.03	45.03	0.00
BouncingBall	Ensemble	FCRG-clean	85.20	85.20	83.19	2.01
BouncingBall	Ensemble	FCRG	88.07	88.07	86.05	2.02
BouncingBall	Ensemble	DD2-noise	75.34	75.34	74.83	0.51
BouncingBall	MLP	FCRG-clean	81.09	81.09	81.09	0.00
BouncingBall	MLP	FCRG	84.70	84.70	84.70	0.00
BouncingBall	MLP	DD2-noise	80.26	80.26	80.26	0.00
BouncingBall	TrajWorld	FCRG-clean	54.94	54.94	54.94	0.00
BouncingBall	TrajWorld	FCRG	56.95	56.95	56.95	0.00
BouncingBall	TrajWorld	DD2-noise	57.38	57.38	57.38	0.00
BouncingBall	WIMLE	FCRG-clean	36.69	36.69	36.69	0.00
BouncingBall	WIMLE	FCRG	40.29	40.29	40.29	0.00
BouncingBall	WIMLE	DD2-noise	33.91	33.91	33.91	0.00
PuckWall2D	ADM-v2	FCRG-clean	19.71	19.71	19.71	0.00
PuckWall2D	ADM-v2	FCRG	24.19	24.19	24.19	0.00
PuckWall2D	ADM-v2	DD2-noise	31.58	31.58	31.58	0.00
PuckWall2D	DreamerV3	FCRG-clean	16.11	16.11	16.11	0.00
PuckWall2D	DreamerV3	FCRG	20.19	20.19	20.19	0.00
PuckWall2D	DreamerV3	DD2-noise	32.04	32.04	32.04	0.00
PuckWall2D	Ensemble	FCRG-clean	58.33	58.33	58.26	0.07
PuckWall2D	Ensemble	FCRG	59.29	59.29	58.49	0.80
PuckWall2D	Ensemble	DD2-noise	51.08	51.08	50.77	0.32
PuckWall2D	MLP	FCRG-clean	59.60	59.60	59.60	0.00
PuckWall2D	MLP	FCRG	59.15	59.15	59.15	0.00
PuckWall2D	MLP	DD2-noise	53.78	53.78	53.78	0.00
PuckWall2D	TrajWorld	FCRG-clean	42.38	42.38	42.38	0.00
PuckWall2D	TrajWorld	FCRG	47.11	47.11	47.11	0.00
PuckWall2D	TrajWorld	DD2-noise	37.42	37.42	37.42	0.00
PuckWall2D	WIMLE	FCRG-clean	0.00	0.00	0.00	0.00
PuckWall2D	WIMLE	FCRG	0.00	0.00	0.00	0.00
PuckWall2D	WIMLE	DD2-noise	0.00	0.00	0.00	0.00

Across the 36 model–readout–system combinations, the largest change in alarm-and-overlap rate from the base tolerance is 2.02 percentage points. Only geometric grouping changes, not readout calibration. The finding supports stability of this joint diagnosis under the tested tolerance choices.

Table C8: Compatible and incompatible interval cases in the same environment. Narrow reference brackets are displayed through midpoint and width. Midpoints do not replace brackets as point truth. Metric panel 1 of 2; identifying columns are repeated.

System	Center	Relation	Reference midpoint ≈	Bracket width	Readout bin
BouncingBall	episode10	disjoint	0.3171	1.52×10^{-6}	[0.4, 0.8]
BouncingBall	episode14	overlap	0.5739	3.05×10^{-6}	[0.4, 0.8]
PuckWall2D	episode16	disjoint	0.2684	1.53×10^{-6}	[0, 0.1]
PuckWall2D	episode14	overlap	0.4055	3.05×10^{-6}	[0.4, 0.8]

Table C8: Compatible and incompatible interval cases in the same environment. Narrow reference brackets are displayed through midpoint and width. Midpoints do not replace brackets as point truth. Metric panel 2 of 2; identifying columns are repeated.

System	Center	Presence probability	Threshold
BouncingBall	episode10	0.9621	0.1464
BouncingBall	episode14	0.8884	0.1464
PuckWall2D	episode16	0.9196	0.2469
PuckWall2D	episode14	0.5712	0.2469

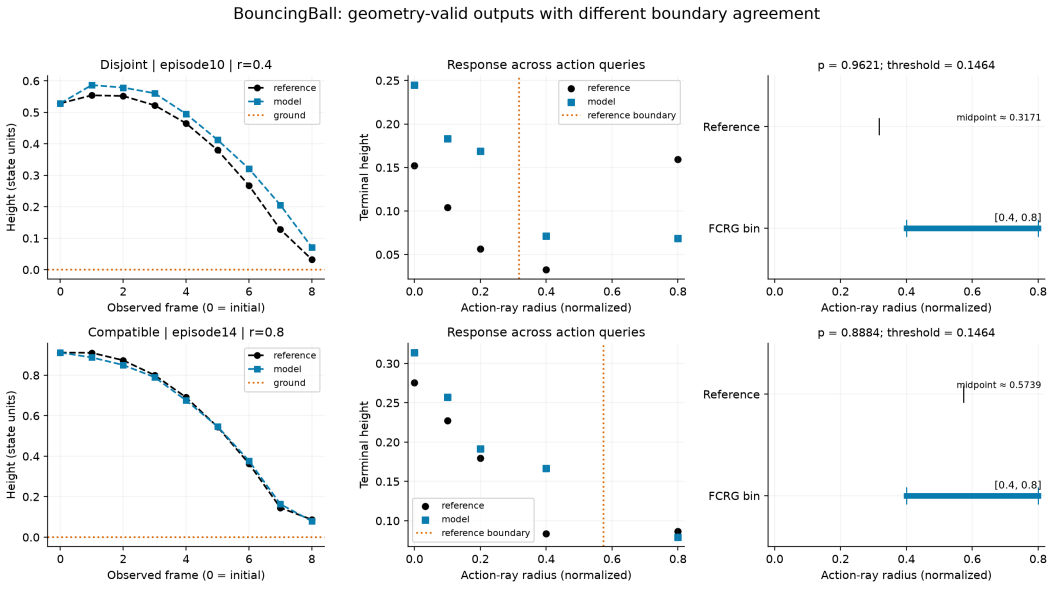

Figure C2: Two geometry-valid BouncingBall rays with incompatible and compatible first intervals. Cases follow the predefined lexicographic selection; connecting lines only indicate observation order.

The cases in Table C8 and Figures C2 and C3 turn the statistical distinction into inspectable action scans: recorded states remain admissible while reference changes and readout intervals may disagree. Compatible examples provide a contrast, making the measured attribute visible without interpreting every discrepancy as a fully verified internal physical-event error.

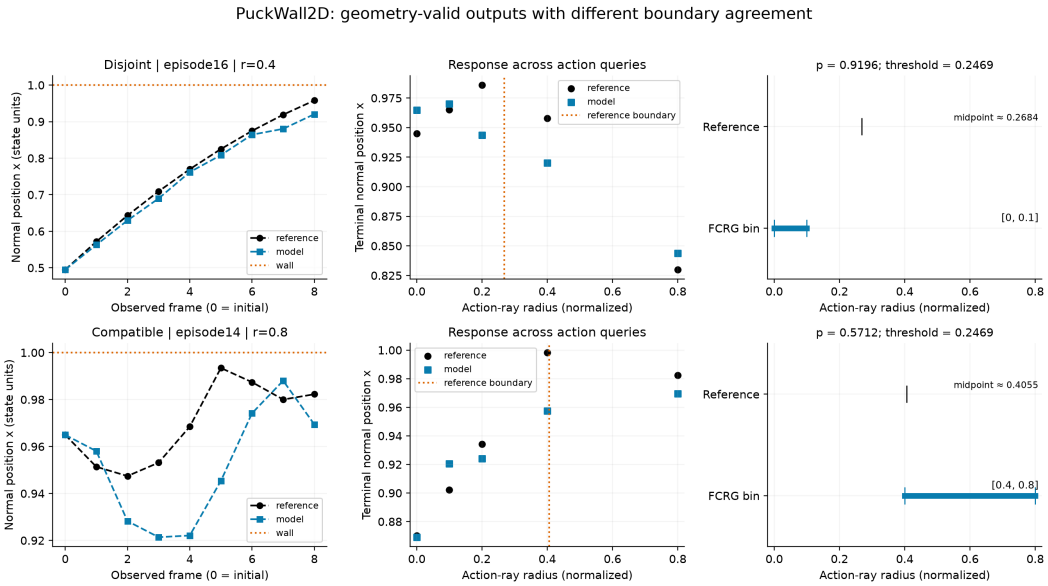

Figure C3: Corresponding geometry-valid PuckWall rays. Compatibility concerns the coarse interval, not the correctness of every predicted state.

C.4 Boundary Differences at High Aggregate Accuracy

Table C9: Complete new-input SaturationKink comparison.

Model	**Scaled MSE**	**AP**	**Recall (%)**	**Exact-bin (%)**	**FPR (%)**
MLP	2.48×10^{-5}	0.9870	85.30	76.14	4.84
Ensemble	1.63×10^{-5}	0.9874	85.23	76.60	4.38
DreamerV3	0.1670	0.9411	89.49	34.34	21.74
TrajWorld	0.0553	0.8987	94.77	42.03	33.94
WIMLE	7.08×10^{-8}	0.9989	84.66	82.56	0.00
ADM-v2	6.50×10^{-6}	0.9977	82.87	77.52	0.00

WIMLE's small MSE, high AP, and incomplete first-bin agreement show that local mode structure can yield different readouts under very small numerical deviations. Table C9 retains the other models, while Figure C4 includes all seven systems, showing both aligned and non-aligned metric behavior. The value lies in identifying these specific differences rather than requiring every system to exhibit the same counterexample.

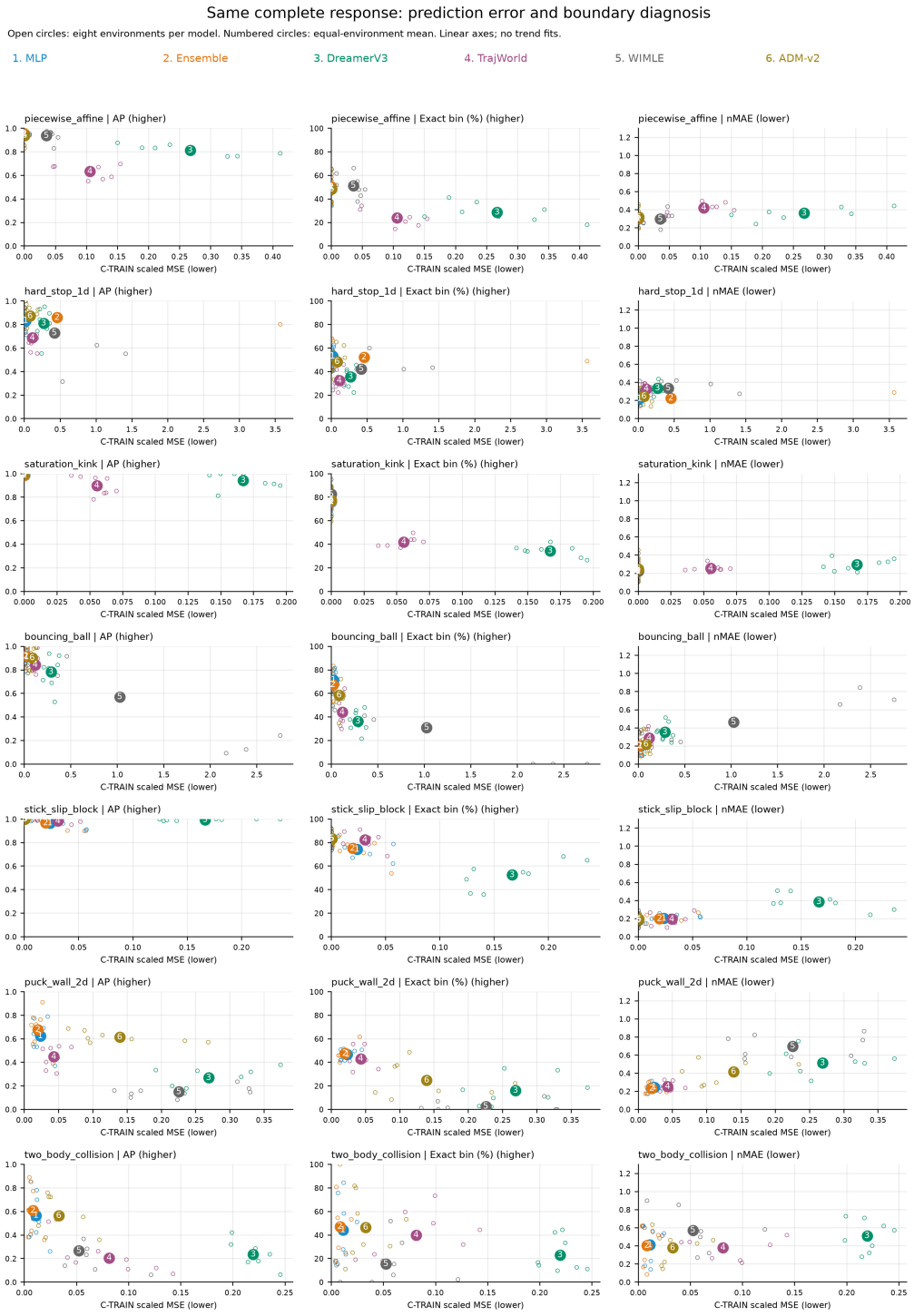

Figure C4: Same-response error and boundary-score relationships for all seven event systems and six outputs. Points retain the original environment and seed aggregation; raw-unit errors across systems are not treated as one scale.

D Measurement Validation and Complete World-Model Diagnosis

D.1 Post-Intervention Truth: Presence and First Location

A modified response can lose fidelity to the original system while being correctly described by the readout. We therefore retain NLL against both the original reference and the post-intervention generator target. The former measures fidelity; the latter validates measurement. Comparing with the altered target distinguishes correct recognition of boundary deletion from generic sensitivity to anomalous responses.

Table D1: Presence readouts under deletion and equal-error noise. Mean $\pm$ parameter-cell SD. Alarm rates use the same operated originally positive rays; NLL uses each full response set.

Deletion (%)	**Readout**	**Deletion alarm (%)**	**Matched-noise alarm (%)**	**Original-target NLL**	**Altered-target NLL**
25.00	FCRG	0.00 ± 0.00	95.14 ± 7.53	2.5929 ± 0.5482	1.0343 ± 0.2146
25.00	FCRG-clean	0.00 ± 0.00	70.14 ± 20.13	2.1095 ± 0.2775	1.2278 ± 0.1929
25.00	DD2-noise	0.00 ± 0.00	100.00 ± 0.00	1.0808 ± 0.1029	0.5454 ± 0.0209
50.00	FCRG	0.00 ± 0.00	96.53 ± 4.39	3.9610 ± 0.5315	0.8017 ± 0.2180
50.00	FCRG-clean	0.00 ± 0.00	69.79 ± 9.79	2.6161 ± 0.3093	0.9655 ± 0.2020
50.00	DD2-noise	0.00 ± 0.00	100.00 ± 0.00	1.5689 ± 0.1249	0.4489 ± 0.0304
75.00	FCRG	0.00 ± 0.00	96.30 ± 3.70	5.0909 ± 0.6381	0.4584 ± 0.1794
75.00	FCRG-clean	0.00 ± 0.00	68.29 ± 9.57	3.0482 ± 0.2521	0.6037 ± 0.1721
75.00	DD2-noise	0.00 ± 0.00	100.00 ± 0.00	2.0882 ± 0.0896	0.3412 ± 0.0340
100.00	FCRG	0.00 ± 0.00	94.79 ± 3.04	6.0308 ± 0.1748	0.0417 ± 0.0071
100.00	FCRG-clean	0.00 ± 0.00	64.93 ± 8.55	3.2755 ± 0.0645	0.1631 ± 0.0151
100.00	DD2-noise	0.00 ± 0.00	100.00 ± 0.00	$2.6842 \pm 1.14 \times 10^{-15}$	$0.2463 \pm 2.58 \times 10^{-16}$

Table D1 shows that under complete deletion, FCRG no longer alarms on modified rays, while most equal-error noisy responses retain alarms and the altered-target NLL remains low. This directly supports presence measurement. FCRG-clean and DD2-noise also recognize deletion, so the construction primarily establishes that the boundary attribute is measurable rather than uniquely favoring one readout. Subsequent comparisons distinguish probability and location behavior.

Table D2: Direction and interval readouts for all physical shifts. 32 paired blocks and three readout seeds. An unchanged predicted bin is not a correct directional shift. Correct destination bins and correct tracking of both endpoints are distinct requirements.

True shift	Readout	Ratio	Correct bin direction (%)	New-target Exact (%)	Original NLL	Altered NLL
0.15 → 0.05	FCRG-clean	0.8208	23.96	23.96	1.4688	1.6927
0.15 → 0.05	FCRG	0.8208	41.67	18.75	2.3562	2.0074
0.15 → 0.05	DD2-noise	0.8208	30.21	6.25	2.0648	2.4341
0.15 → 0.3	FCRG-clean	0.9974	96.88	100.00	2.5948	0.3899
0.15 → 0.3	FCRG	0.9974	43.75	100.00	5.5600	0.0901
0.15 → 0.3	DD2-noise	0.9974	56.25	18.75	4.0443	1.4576
0.3 → 0.15	FCRG-clean	0.7103	96.88	88.54	2.2046	0.6058
0.3 → 0.15	FCRG	0.7103	43.75	40.63	1.2539	1.3892
0.3 → 0.15	DD2-noise	0.7103	56.25	3.13	0.8739	2.7105
0.3 → 0.6	FCRG-clean	0.9946	100.00	100.00	2.4137	0.2703
0.3 → 0.6	FCRG	0.9946	100.00	100.00	4.2988	0.0527
0.3 → 0.6	DD2-noise	0.9946	18.75	37.50	2.6301	1.4914

For 0.3→0.6, Table D2 shows that FCRG and FCRG-clean recover both bins, providing an explicit location change despite a near-ideal ratio. Nearer shifts separate measurement stages: for 0.15→0.3, FCRG always recovers the destination bin but changes its predicted bin in the required direction on 43.75% of evaluations. The complete table thus characterizes both capability and resolution, distinguishing correct new locations from complete shift tracking.

D.2 Readout Performance on Simulator Responses

Table D3: Readout reference on fresh simulator responses.

Readout	AP	Joint NLL	Exact-bin (%) [95% CI]	nMAE	FPR (%)
FCRG	0.9306	0.1979	80.76 [77.90, 83.78]	0.1542	4.63
FCRG-clean	0.9327	0.2682	77.01 [73.99, 80.24]	0.1933	3.85
DD2-noise	0.7640	0.4600	63.59 [60.35, 66.88]	0.2513	12.29

Reference responses remove learned prediction error, so Table D3 directly characterizes the readout pipeline. FCRG and FCRG-clean have similar AP, while FCRG improves joint NLL and interval hits, showing that the measurement gain extends beyond presence ranking. This provides a basis for interpreting model scores without treating the reference error as a constant that can be subtracted from every model discrepancy.

D.3 Per-System Model Profiles and Error Decomposition

Table D4: Complete six-model results on seven event systems. Frozen FCRG on matched complete outputs; mean ± environment SD.

System	Model	AP	Joint NLL	Exact-bin (%)	nMAE	FPR (%)
Piecewise Affine	MLP	0.9524 ± 0.0425	0.7094 ± 0.1727	50.22 ± 10.28	0.3006 ± 0.0713	4.50 ± 3.42
Piecewise Affine	Ensemble	0.9530 ± 0.0421	0.7057 ± 0.1699	50.12 ± 10.24	0.3014 ± 0.0718	4.44 ± 3.43
Piecewise Affine	DreamerV3	0.8135 ± 0.0437	1.6240 ± 0.3604	28.59 ± 7.68	0.3626 ± 0.0648	28.74 ± 8.11
Piecewise Affine	TrajWorld	0.6337 ± 0.0563	4.0512 ± 0.6592	23.90 ± 6.50	0.4220 ± 0.0546	49.71 ± 4.23
Piecewise Affine	WIMLE	0.9374 ± 0.0464	0.7502 ± 0.1235	51.34 ± 9.40	0.2990 ± 0.0751	6.91 ± 5.47
Piecewise Affine	ADM-v2	0.9392 ± 0.0469	0.7751 ± 0.1574	48.34 ± 10.67	0.3189 ± 0.0841	5.02 ± 3.57
HardStop1D	MLP	0.8331 ± 0.0999	0.4224 ± 0.1960	53.03 ± 8.95	0.2212 ± 0.0555	10.76 ± 2.36
HardStop1D	Ensemble	0.8591 ± 0.0809	0.4082 ± 0.2017	52.19 ± 10.54	0.2267 ± 0.0606	10.53 ± 2.95
HardStop1D	DreamerV3	0.8115 ± 0.1281	0.5252 ± 0.2078	35.71 ± 8.36	0.3364 ± 0.0564	11.92 ± 5.28
HardStop1D	TrajWorld	0.6871 ± 0.1132	0.5953 ± 0.2044	32.39 ± 7.84	0.3299 ± 0.0529	23.98 ± 6.30
HardStop1D	WIMLE	0.7282 ± 0.2254	0.4605 ± 0.1815	42.29 ± 9.95	0.3374 ± 0.0863	12.06 ± 7.49
HardStop1D	ADM-v2	0.8722 ± 0.1033	0.4037 ± 0.2096	48.36 ± 12.29	0.2467 ± 0.0871	6.90 ± 2.94
Saturation Kink	MLP	0.9870 ± 0.0146	0.4121 ± 0.0708	76.14 ± 7.73	0.2205 ± 0.1003	4.84 ± 5.22
Saturation Kink	Ensemble	0.9874 ± 0.0141	0.3930 ± 0.0861	76.60 ± 8.53	0.2216 ± 0.1011	4.38 ± 2.86
Saturation Kink	DreamerV3	0.9411 ± 0.0674	2.1671 ± 0.1828	34.34 ± 4.88	0.2947 ± 0.0652	21.74 ± 21.70
Saturation Kink	TrajWorld	0.8987 ± 0.0801	2.6071 ± 0.3124	42.03 ± 3.97	0.2532 ± 0.0367	33.94 ± 16.95
Saturation Kink	WIMLE	0.9989 ± 0.0025	0.1897 ± 0.0712	82.56 ± 9.48	0.2265 ± 0.0983	0.00 ± 0.00
Saturation Kink	ADM-v2	0.9977 ± 0.0035	0.3279 ± 0.1360	77.52 ± 10.86	0.2479 ± 0.1181	0.00 ± 0.00
BouncingBall	MLP	0.9110 ± 0.0724	0.4743 ± 0.3914	71.33 ± 12.42	0.1985 ± 0.1051	12.28 ± 5.80
BouncingBall	Ensemble	0.9165 ± 0.0686	0.4751 ± 0.3926	67.38 ± 12.57	0.2023 ± 0.1010	11.07 ± 5.05
BouncingBall	DreamerV3	0.7835 ± 0.1351	0.8702 ± 0.5325	36.40 ± 9.00	0.3526 ± 0.0962	17.96 ± 12.29
BouncingBall	TrajWorld	0.8435 ± 0.0723	0.7128 ± 0.3649	44.08 ± 12.51	0.2865 ± 0.0790	17.52 ± 9.19
BouncingBall	WIMLE	0.5707 ± 0.3519	1.4973 ± 1.0706	30.92 ± 26.97	0.4619 ± 0.2377	20.76 ± 18.90
BouncingBall	ADM-v2	0.9015 ± 0.0920	0.5775 ± 0.4649	58.56 ± 11.76	0.2185 ± 0.1152	13.86 ± 9.91
StickSlip Block	MLP	0.9664 ± 0.0388	1.6403 ± 1.0674	74.15 ± 7.35	0.2028 ± 0.0425	13.08 ± 13.61
StickSlip Block	Ensemble	0.9668 ± 0.0426	1.5868 ± 1.2175	75.09 ± 9.78	0.1986 ± 0.0450	10.64 ± 13.53

Continued on next page

Table D4 (continued)

System	Model	AP	Joint NLL	Exact-bin (%)	nMAE	FPR (%)
StickSlip Block	DreamerV3	0.9940 ± 0.0069	0.7139 ± 0.2199	52.58 ± 11.76	0.3854 ± 0.0915	0.00 ± 0.00
StickSlip Block	TrajWorld	0.9834 ± 0.0154	0.6157 ± 0.3697	82.38 ± 7.09	0.1931 ± 0.0667	3.25 ± 3.33
StickSlip Block	WIMLE	0.9995 ± 0.0008	0.1568 ± 0.1499	83.45 ± 7.17	0.1828 ± 0.0684	0.00 ± 0.00
StickSlip Block	ADM-v2	0.9992 ± 0.0012	0.2168 ± 0.1997	83.49 ± 5.22	0.1934 ± 0.0568	0.00 ± 0.00
PuckWall2D	MLP	0.6241 ± 0.1398	0.3101 ± 0.0583	46.78 ± 4.36	0.2466 ± 0.0433	11.35 ± 3.90
PuckWall2D	Ensemble	0.6765 ± 0.1623	0.2728 ± 0.0658	47.55 ± 6.24	0.2315 ± 0.0523	11.02 ± 3.97
PuckWall2D	DreamerV3	0.2688 ± 0.0868	0.4430 ± 0.1401	16.05 ± 12.55	0.5128 ± 0.1372	15.09 ± 2.63
PuckWall2D	TrajWorld	0.4469 ± 0.0931	0.5494 ± 0.1038	43.10 ± 8.66	0.2525 ± 0.0471	26.48 ± 4.47
PuckWall2D	WIMLE	0.1493 ± 0.0472	0.6112 ± 0.1949	2.57 ± 4.18	0.6968 ± 0.1236	13.50 ± 4.34
PuckWall2D	ADM-v2	0.6140 ± 0.0452	0.3981 ± 0.1252	24.74 ± 14.33	0.4158 ± 0.1252	9.28 ± 2.51
TwoBody Collision	MLP	0.5659 ± 0.2067	0.2905 ± 0.1115	44.17 ± 27.08	0.4104 ± 0.2010	12.71 ± 4.55
TwoBody Collision	Ensemble	0.6075 ± 0.2110	0.2797 ± 0.1144	47.21 ± 30.73	0.4046 ± 0.2121	12.12 ± 5.05
TwoBody Collision	DreamerV3	0.2364 ± 0.1076	0.3467 ± 0.0943	23.04 ± 14.54	0.5117 ± 0.1711	13.63 ± 7.33
TwoBody Collision	TrajWorld	0.2051 ± 0.1390	0.5545 ± 0.1090	40.00 ± 20.24	0.3803 ± 0.1234	32.27 ± 11.86
TwoBody Collision	WIMLE	0.2665 ± 0.1635	0.3537 ± 0.1289	15.50 ± 18.46	0.5731 ± 0.2223	11.95 ± 6.51
TwoBody Collision	ADM-v2	0.5638 ± 0.1574	0.2860 ± 0.1163	46.51 ± 26.01	0.3777 ± 0.1737	11.91 ± 4.30

Table D4 shows that model strengths vary across systems and evaluation attributes rather than following a universal ordering. Fine-grained evaluation localizes this variation: high AP may accompany different fixed-threshold outcomes, and low joint NLL need not coincide with the highest interval accuracy. Interpreting these trade-offs across the complete system set is more informative than using a global rank alone for applications requiring boundary locations. Table D4a reports presence ranking, probability, and false alarms; its second panel reports detection and first location, retaining the confidence intervals for Table 3.

Table D4a: (a) Presence ranking, joint probability, and false alarms. Complete estimates and grouped-bootstrap 95% confidence intervals for Table 3.

World-model output	AP ↑	Joint NLL ↓	FPR (%) ↓
MLP	0.8343 [0.8240, 0.8620]	0.6084 [0.5655, 0.6523]	9.93 [8.99, 10.95]
Ensemble	0.8524 [0.8405, 0.8801]	0.5888 [0.5455, 0.6331]	9.17 [8.26, 10.20]
DreamerV3	0.6927 [0.6871, 0.7226]	0.9557 [0.9084, 1.0058]	15.58 [14.39, 16.88]
TrajWorld	0.6712 [0.6683, 0.7047]	1.3837 [1.3256, 1.4437]	26.74 [25.35, 28.26]
WIMLE	0.6644 [0.6591, 0.6910]	0.5742 [0.5267, 0.6225]	9.31 [8.39, 10.31]
ADM-v2	0.8411 [0.8317, 0.8677]	0.4264 [0.3933, 0.4614]	6.71 [5.97, 7.54]

Table D4a: (b) Detection and first location. Complete estimates and grouped-bootstrap 95% confidence intervals for Table 3.

World-model output	Recall (%) ↑	Exact-bin (%) ↑	nMAE ↓
MLP	84.00 [81.93, 86.13]	59.40 [56.70, 62.27]	0.2572 [0.2375, 0.2752]
Ensemble	84.99 [82.87, 87.19]	59.45 [56.58, 62.62]	0.2552 [0.2340, 0.2734]
DreamerV3	67.82 [65.35, 70.07]	32.39 [29.89, 34.95]	0.3938 [0.3717, 0.4127]
TrajWorld	85.90 [84.33, 87.57]	43.98 [41.52, 46.64]	0.3025 [0.2872, 0.3164]
WIMLE	60.24 [58.04, 62.22]	44.09 [42.02, 46.24]	0.3968 [0.3737, 0.4170]
ADM-v2	80.62 [78.33, 82.78]	55.36 [52.57, 58.42]	0.2884 [0.2685, 0.3055]

Table D5: Error composition on all reference positives. The three outcomes share the positive denominator and sum to 100.00%, up to displayed rounding.

System	Model	Miss (%)	WrongBin (%)	Exact (%)
PiecewiseAffine	MLP	25.55	24.24	50.22
PiecewiseAffine	Ensemble	25.59	24.29	50.12
PiecewiseAffine	DreamerV3	28.86	42.55	28.59
PiecewiseAffine	TrajWorld	27.32	48.78	23.90
PiecewiseAffine	WIMLE	24.43	24.23	51.34
PiecewiseAffine	ADM-v2	27.01	24.66	48.34
HardStop1D	MLP	6.97	40.00	53.03
HardStop1D	Ensemble	5.97	41.83	52.19
HardStop1D	DreamerV3	15.03	49.26	35.71
HardStop1D	TrajWorld	12.63	54.98	32.39
HardStop1D	WIMLE	16.55	41.15	42.29
HardStop1D	ADM-v2	12.95	38.69	48.36
SaturationKink	MLP	14.70	9.16	76.14
SaturationKink	Ensemble	14.77	8.63	76.60
SaturationKink	DreamerV3	10.51	55.14	34.34
SaturationKink	TrajWorld	5.23	52.74	42.03
SaturationKink	WIMLE	15.34	2.10	82.56
SaturationKink	ADM-v2	17.13	5.35	77.52
BouncingBall	MLP	14.12	14.55	71.33
BouncingBall	Ensemble	14.80	17.83	67.38
BouncingBall	DreamerV3	21.28	42.32	36.40
BouncingBall	TrajWorld	18.00	37.92	44.08
BouncingBall	WIMLE	52.02	17.06	30.92
BouncingBall	ADM-v2	16.55	24.89	58.56
StickSlipBlock	MLP	11.65	14.21	74.15
StickSlipBlock	Ensemble	11.05	13.86	75.09
StickSlipBlock	DreamerV3	40.39	7.03	52.58
StickSlipBlock	TrajWorld	12.54	5.08	82.38
StickSlipBlock	WIMLE	14.46	2.09	83.45
StickSlipBlock	ADM-v2	13.11	3.40	83.49
PuckWall2D	MLP	14.43	38.80	46.78
PuckWall2D	Ensemble	9.41	43.04	47.55
PuckWall2D	DreamerV3	51.66	32.29	16.05
PuckWall2D	TrajWorld	10.50	46.40	43.10
PuckWall2D	WIMLE	85.62	11.81	2.57
PuckWall2D	ADM-v2	26.36	48.90	24.74
TwoBodyCollision	MLP	24.59	31.24	44.17
TwoBodyCollision	Ensemble	23.46	29.33	47.21

Continued on next page

Table D5 (continued)

System	Model	Miss (%)	WrongBin (%)	Exact (%)
TwoBodyCollision	DreamerV3	57.56	19.41	23.04
TwoBodyCollision	TrajWorld	12.51	47.49	40.00
TwoBodyCollision	WIMLE	69.88	14.61	15.50
TwoBodyCollision	ADM-v2	22.53	30.95	46.51

Table D5 shows that on BouncingBall, MLP and TrajWorld have relatively close recalls but markedly different interval assignments after alarms; WIMLE loses more cases through missed detection. These are not alternative names for the same error: they identify different stages for further investigation—insufficient presence evidence versus incompatible localization despite an alarm. The decomposition therefore directs diagnosis rather than merely ranking models.

D.4 Alternative Readouts and Paired Effects

The main text summarizes whether readout improvements survive multiplicity correction. The principal paired effects are reported in Table D6a, while the complete family of 36 prespecified AP, joint-NLL, and nMAE contrasts is retained in Table D6.

Table D6a: Main paired readout effects. Differences are FCRG minus the comparator. Positive AP and negative NLL favor FCRG.

Model output	ΔAP vs. DD2-noise [95% CI]	Holm p	ΔNLL vs. FCRG-clean [95% CI]	Holm p
MLP	0.1702 [0.1412, 0.1976]	**0.0036**	-0.2991 [-0.3492, -0.2448]	**0.0036**
Ensemble	0.1753 [0.1452, 0.2041]	**0.0036**	-0.2971 [-0.3581, -0.2237]	**0.0036**
DreamerV3	0.1500 [0.1285, 0.1719]	**0.0036**	-0.1539 [-0.2011, -0.1059]	**0.0378**
TrajWorld	0.1988 [0.1729, 0.2260]	**0.0036**	0.1809 [0.1368, 0.2253]	0.7849
WIMLE	0.0340 [0.0044, 0.0632]	0.6225	-0.3639 [-0.4017, -0.3274]	**0.0036**
ADM-v2	0.1124 [0.0838, 0.1400]	**0.0036**	-0.3563 [-0.3816, -0.3294]	**0.0036**

Five of six AP improvements over DD2-noise survive Holm correction. Five of six joint-NLL reductions relative to FCRG-clean also survive correction, with TrajWorld showing the opposite mean direction. The complete test family prevents these selected summaries from being interpreted as the only performed comparisons.

Table D6: All 36 prespecified readout contrasts. Effects follow the contrast direction. Positive AP and negative NLL/nMAE favor FCRG. Bold denotes family-adjusted p<0.05.

Model	Contrast	Metric	Effect	Environment 95% CI	Raw p	Holm p
MLP	FCRG − FCRG-clean	AP	0.0040	[-0.0020, 0.0097]	0.4487	1.0000
MLP	FCRG − FCRG-clean	Joint NLL	-0.2991	[-0.3492, -0.2448]	1.00×10^{-4}	**0.0036**

Continued on next page

Table D6 (continued)

Model	Contrast	Metric	Effect	Environment 95% CI	Raw p	Holm p
MLP	FCRG − FCRG-clean	nMAE	-0.0212	[-0.0337, -0.0086]	0.0077	0.1309
MLP	FCRG − DD2-noise	AP	0.1702	[0.1412, 0.1976]	1.00×10^{-4}	**0.0036**
MLP	FCRG − DD2-noise	Joint NLL	-0.3422	[-0.3915, -0.2886]	1.00×10^{-4}	**0.0036**
MLP	FCRG − DD2-noise	nMAE	-0.0347	[-0.0521, -0.0163]	0.0198	0.2772
Ensemble	FCRG − FCRG-clean	AP	0.0051	[-0.0012, 0.0109]	0.3248	1.0000
Ensemble	FCRG − FCRG-clean	Joint NLL	-0.2971	[-0.3581, -0.2237]	1.00×10^{-4}	**0.0036**
Ensemble	FCRG − FCRG-clean	nMAE	-0.0169	[-0.0307, -0.0030]	0.0395	0.4740
Ensemble	FCRG − DD2-noise	AP	0.1753	[0.1452, 0.2041]	1.00×10^{-4}	**0.0036**
Ensemble	FCRG − DD2-noise	Joint NLL	-0.3671	[-0.4466, -0.2854]	1.00×10^{-4}	**0.0036**
Ensemble	FCRG − DD2-noise	nMAE	-0.0464	[-0.0660, -0.0267]	0.0054	0.1026
DreamerV3	FCRG − FCRG-clean	AP	0.0263	[0.0185, 0.0340]	1.00×10^{-4}	**0.0036**
DreamerV3	FCRG − FCRG-clean	Joint NLL	-0.1539	[-0.2011, -0.1059]	0.0018	**0.0378**
DreamerV3	FCRG − FCRG-clean	nMAE	-0.0026	[-0.0110, 0.0057]	0.7891	1.0000
DreamerV3	FCRG − DD2-noise	AP	0.1500	[0.1285, 0.1719]	1.00×10^{-4}	**0.0036**
DreamerV3	FCRG − DD2-noise	Joint NLL	-0.3502	[-0.4201, -0.2845]	0.0002	**0.0044**
DreamerV3	FCRG − DD2-noise	nMAE	0.0371	[0.0177, 0.0557]	0.0785	0.7849
TrajWorld	FCRG − FCRG-clean	AP	0.0063	[-0.0016, 0.0139]	0.4159	1.0000
TrajWorld	FCRG − FCRG-clean	Joint NLL	0.1809	[0.1368, 0.2253]	0.0827	0.7849
TrajWorld	FCRG − FCRG-clean	nMAE	-0.0228	[-0.0309, -0.0146]	0.0050	0.1000
TrajWorld	FCRG − DD2-noise	AP	0.1988	[0.1729, 0.2260]	1.00×10^{-4}	**0.0036**
TrajWorld	FCRG − DD2-noise	Joint NLL	-0.8412	[-0.9901, -0.6992]	1.00×10^{-4}	**0.0036**
TrajWorld	FCRG − DD2-noise	nMAE	0.0045	[-0.0088, 0.0182]	0.6995	1.0000
WIMLE	FCRG − FCRG-clean	AP	0.0113	[0.0054, 0.0177]	0.0222	0.2886
WIMLE	FCRG − FCRG-clean	Joint NLL	-0.3639	[-0.4017, -0.3274]	1.00×10^{-4}	**0.0036**
WIMLE	FCRG − FCRG-clean	nMAE	0.0136	[-0.0001, 0.0277]	0.1411	1.0000
WIMLE	FCRG − DD2-noise	AP	0.0340	[0.0044, 0.0632]	0.0566	0.6225
WIMLE	FCRG − DD2-noise	Joint NLL	-0.0832	[-0.1593, 0.0061]	0.1477	1.0000

Continued on next page

Table D6 (continued)

Model	Contrast	Metric	Effect	Environment 95% CI	Raw p	Holm p
WIMLE	FCRG − DD2-noise	nMAE	-0.0417	[-0.0663, -0.0178]	0.0177	0.2655
ADM-v2	FCRG − FCRG-clean	AP	0.0113	[0.0061, 0.0161]	0.0055	0.1026
ADM-v2	FCRG − FCRG-clean	Joint NLL	-0.3563	[-0.3816, -0.3294]	1.00×10^{-4}	**0.0036**
ADM-v2	FCRG − FCRG-clean	nMAE	0.0223	[0.0096, 0.0360]	0.0078	0.1309
ADM-v2	FCRG − DD2-noise	AP	0.1124	[0.0838, 0.1400]	1.00×10^{-4}	**0.0036**
ADM-v2	FCRG − DD2-noise	Joint NLL	-0.1530	[-0.1920, -0.1128]	1.00×10^{-4}	**0.0036**
ADM-v2	FCRG − DD2-noise	nMAE	-0.0121	[-0.0309, 0.0072]	0.4674	1.0000

The main AP and joint-NLL gains each survive correction for five outputs, supporting FCRG as a detection and probability-diagnosis tool; nMAE does not show a uniform corrected advantage. Alongside the decomposition, the complete tests distinguish better detection and probability assignment from uniformly lower distance errors. The contribution rests on the supported measurement capabilities rather than winning every metric.

Table D7: Model-order agreement between readouts. Descriptive ordering agreement, not independent physical-label validation.

Metric	Readouts	Spearman	Concordant pairs	Ties in either
AP	FCRG / FCRG-clean	0.9429	14/15	0
AP	FCRG / DD2-noise	0.7714	12/15	0
Joint NLL	FCRG / FCRG-clean	0.8286	13/15	0
Joint NLL	FCRG / DD2-noise	0.9429	14/15	0
nMAE	FCRG / FCRG-clean	0.7714	12/15	0
nMAE	FCRG / DD2-noise	0.6000	11/15	0

Table D7 reports agreement on 14 of 15 AP orderings between FCRG and FCRG-clean, which, together with the retained BouncingBall interval gap, indicates that the main differences are not isolated to one frozen readout. nMAE agreement is 12 of 15, and DD2 may change gap magnitudes. Alternative readouts thus assess robustness of the findings and provide a cross-check on score interpretation, rather than turning agreement into physical ground truth.

D.5 The Separate Effect of Response Aggregation

For a nonlinear readout C, generally $C(\mathbb{E}[Y]) \neq \mathbb{E}[C(Y)]$. The main paper pairs mean-response MSE with $C(\mathbb{E}[Y])$, while posterior mixtures remain auxiliary outputs. Specifying what is predicted before comparing scores prevents aggregation differences from being mistaken for changes in the quality of the same response.

Table D8: Twelve contrasts of mean-response readout minus posterior mixture. A separate multiplicity family from Table D6.

Model	Metric	Effect	Environment 95% CI	Raw p	Holm p
Ensemble	AP	-0.0049	[-0.0085, -0.0013]	0.0165	0.1155
Ensemble	Joint NLL	0.0873	[0.0493, 0.1355]	1.00×10^{-4}	**0.0012**
Ensemble	nMAE	-6.84×10^{-5}	[-0.0044, 0.0044]	0.9838	1.0000
DreamerV3	AP	-0.0014	[-0.0057, 0.0026]	0.5353	1.0000
DreamerV3	Joint NLL	0.1421	[0.1222, 0.1622]	1.00×10^{-4}	**0.0012**
DreamerV3	nMAE	0.0138	[0.0076, 0.0202]	0.0581	0.3486
WIMLE	AP	-0.0016	[-0.0043, 0.0014]	0.2970	1.0000
WIMLE	Joint NLL	0.0269	[0.0123, 0.0462]	0.0012	**0.0110**
WIMLE	nMAE	0.0023	[-0.0036, 0.0079]	0.5748	1.0000
ADM-v2	AP	-0.0075	[-0.0151, -0.0008]	0.0769	0.3845
ADM-v2	Joint NLL	0.0205	[0.0110, 0.0315]	0.0011	**0.0110**
ADM-v2	nMAE	0.0128	[0.0065, 0.0194]	0.0050	**0.0400**

Table D8 shows that joint NLL differs by aggregation order for all four outputs, whereas AP changes do not survive correction in this family. Probability-output definitions can therefore materially affect evaluation even when rankings change little. Both outputs are retained, while the main text fixes the response object shared with MSE to keep complementarity comparisons aligned.

D.6 Per-System 21-Metric Matrices

The following matrices retain all six models, three readouts, and simulator references, grouped into presence, probability, operating-point, and location metrics. Detection, hit, and directional error rates in this section are percentages; AP, nAP, and calibration scores retain their defined scales, and distances retain their defined units. Event metrics on Smooth are NA. Table D8 reports the response-aggregation contrasts. Tables D9, D10, D11, D12, D13, D14, D15, and D16 give the matrices for PiecewiseAffine, HardStop1D, SaturationKink, BouncingBall, StickSlipBlock, PuckWall2D, TwoBodyCollision, and SmoothDuffing, respectively; panels (a)–(d) cover the four metric groups.

Table D9: (a) PiecewiseAffine.

Model	Readout	AP	nAP	Binary Brier	ECE	Cal. bias
MLP	FCRG	0.9524	0.9071	0.0898	0.0646	-0.0283
MLP	FCRG-clean	0.8770	0.7608	0.2322	0.2627	0.2420
MLP	DD2-noise	0.9384	0.8781	0.1198	0.1613	-0.0605
Ensemble	FCRG	0.9530	0.9084	0.0895	0.0680	-0.0306
Ensemble	FCRG-clean	0.8792	0.7651	0.2292	0.2599	0.2392
Ensemble	DD2-noise	0.9389	0.8792	0.1194	0.1619	-0.0633
DreamerV3	FCRG	0.8135	0.6350	0.2141	0.1725	0.0751
DreamerV3	FCRG-clean	0.6866	0.3899	0.2577	0.1938	0.1511
DreamerV3	DD2-noise	0.7705	0.5492	0.2381	0.2092	0.1978
TrajWorld	FCRG	0.6337	0.2827	0.3362	0.3093	0.1661
TrajWorld	FCRG-clean	0.6795	0.3744	0.2542	0.1827	0.0593
TrajWorld	DD2-noise	0.5235	0.0697	0.4495	0.4494	0.4476
WIMLE	FCRG	0.9374	0.8780	0.1015	0.0692	0.0048

Continued on next page

Table D9 (continued)

Model	Readout	AP	nAP	Binary Brier	ECE	Cal. bias
WIMLE	FCRG-clean	0.8667	0.7419	0.2258	0.2458	0.2234
WIMLE	DD2-noise	0.8432	0.6923	0.1684	0.1259	0.0343
ADM-v2	FCRG	0.9392	0.8820	0.1012	0.0651	-0.0097
ADM-v2	FCRG-clean	0.8788	0.7648	0.2302	0.2594	0.2380
ADM-v2	DD2-noise	0.8981	0.8017	0.1374	0.1159	-0.0171
Reference	FCRG	0.9761	0.9538	0.0488	0.0443	-0.0122
Reference	FCRG-clean	0.9722	0.9468	0.0514	0.0780	-0.0349
Reference	DD2-noise	0.9435	0.8881	0.1235	0.1951	-0.0895

Table D9: (b) PiecewiseAffine.

Model	Readout	Joint NLL	Joint Brier	Interval NLL	Survival NLL
MLP	FCRG	0.7094	0.3528	1.2391	0.2138
MLP	FCRG-clean	1.2723	0.6558	1.4655	1.0982
MLP	DD2-noise	0.8821	0.4393	1.5281	0.2671
Ensemble	FCRG	0.7057	0.3526	1.2369	0.2089
Ensemble	FCRG-clean	1.2646	0.6516	1.4633	1.0853
Ensemble	DD2-noise	0.8801	0.4391	1.5298	0.2617
DreamerV3	FCRG	1.6240	0.6887	2.3666	0.9353
DreamerV3	FCRG-clean	1.5130	0.7268	1.9780	1.0806
DreamerV3	DD2-noise	1.3485	0.6557	1.5351	1.1777
TrajWorld	FCRG	4.0512	1.0520	5.1121	3.0669
TrajWorld	FCRG-clean	2.1012	0.8071	3.3446	0.9322
TrajWorld	DD2-noise	3.8062	1.2751	2.7878	4.7814
WIMLE	FCRG	0.7502	0.3801	1.2202	0.3134
WIMLE	FCRG-clean	1.2524	0.6538	1.4923	1.0357
WIMLE	DD2-noise	1.0299	0.5258	1.5123	0.5709
ADM-v2	FCRG	0.7751	0.3811	1.3114	0.2750
ADM-v2	FCRG-clean	1.2572	0.6476	1.4520	1.0811
ADM-v2	DD2-noise	0.9494	0.4696	1.5520	0.3763
Reference	FCRG	0.3883	0.1945	0.6259	0.1668
Reference	FCRG-clean	0.4161	0.2066	0.6877	0.1598
Reference	DD2-noise	0.8410	0.4352	1.4863	0.2249

Table D9: (c) PiecewiseAffine.

Model	Readout	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one (%)
MLP	FCRG	4.50	74.45	94.08	50.22	70.30
MLP	FCRG-clean	37.09	87.97	69.57	37.23	67.10
MLP	DD2-noise	7.04	82.90	92.15	48.21	72.07
Ensemble	FCRG	4.44	74.41	94.18	50.12	70.24
Ensemble	FCRG-clean	36.35	87.93	70.00	36.96	67.04
Ensemble	DD2-noise	6.76	82.81	92.40	48.16	72.34
DreamerV3	FCRG	28.74	71.14	70.67	28.59	59.87
DreamerV3	FCRG-clean	35.22	60.93	62.30	19.76	45.06
DreamerV3	DD2-noise	69.69	94.73	56.50	36.62	79.33
TrajWorld	FCRG	49.71	72.68	58.21	23.90	50.10
TrajWorld	FCRG-clean	24.59	53.70	67.69	18.58	36.45

Continued on next page

Table D9 (continued)

Model	Readout	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one (%)
TrajWorld	DD2-noise	96.66	98.66	49.35	19.04	61.10
WIMLE	FCRG	6.91	75.57	91.41	51.34	69.20
WIMLE	FCRG-clean	28.47	80.28	73.16	37.58	63.93
WIMLE	DD2-noise	36.24	85.27	69.35	43.15	74.85
ADM-v2	FCRG	5.02	72.99	93.23	48.34	68.04
ADM-v2	FCRG-clean	33.79	87.08	71.36	37.53	68.31
ADM-v2	DD2-noise	22.66	83.78	78.20	46.79	72.56
Reference	FCRG	6.12	90.19	93.36	76.29	86.29
Reference	FCRG-clean	3.83	83.34	95.17	71.64	80.75
Reference	DD2-noise	5.34	78.83	93.59	48.45	71.38

Table D9: (d) PiecewiseAffine. Metric panel 1 of 2; identifying columns are repeated.

Model	Readout	Dist. MAE	nMAE	Loc. bias	Late (%)
MLP	FCRG	0.2149	0.3006	0.1433	62.67
MLP	FCRG-clean	0.2146	0.2993	-0.0130	44.39
MLP	DD2-noise	0.2018	0.2787	0.1298	65.34
Ensemble	FCRG	0.2155	0.3014	0.1437	62.60
Ensemble	FCRG-clean	0.2151	0.2998	-0.0123	44.44
Ensemble	DD2-noise	0.2012	0.2782	0.1294	65.29
DreamerV3	FCRG	0.2658	0.3626	0.1581	67.50
DreamerV3	FCRG-clean	0.3485	0.4680	0.2371	72.98
DreamerV3	DD2-noise	0.1827	0.2539	-0.0168	49.12
TrajWorld	FCRG	0.3135	0.4220	0.1388	62.41
TrajWorld	FCRG-clean	0.3901	0.5242	0.2895	75.77
TrajWorld	DD2-noise	0.2406	0.3183	-0.1335	33.57
WIMLE	FCRG	0.2143	0.2990	0.1325	61.98
WIMLE	FCRG-clean	0.2420	0.3337	0.0639	52.29
WIMLE	DD2-noise	0.1995	0.2762	0.0890	61.69
ADM-v2	FCRG	0.2294	0.3189	0.1554	63.30
ADM-v2	FCRG-clean	0.2107	0.2944	0.0033	45.47
ADM-v2	DD2-noise	0.2036	0.2824	0.1234	64.67
Reference	FCRG	0.1315	0.1852	0.0519	55.36
Reference	FCRG-clean	0.1668	0.2388	0.0933	57.99
Reference	DD2-noise	0.2063	0.2873	0.1343	67.91

Table D9: (d) PiecewiseAffine. Metric panel 2 of 2; identifying columns are repeated.

Model	Readout	Early (%)	Miss (%)	Hallucination (%)
MLP	FCRG	37.33	25.55	4.50
MLP	FCRG-clean	55.61	12.03	37.09
MLP	DD2-noise	34.66	17.10	7.04
Ensemble	FCRG	37.40	25.59	4.44
Ensemble	FCRG-clean	55.56	12.07	36.35
Ensemble	DD2-noise	34.71	17.19	6.76
DreamerV3	FCRG	32.50	28.86	28.74
DreamerV3	FCRG-clean	27.02	39.07	35.22
DreamerV3	DD2-noise	50.88	5.27	69.69
TrajWorld	FCRG	37.59	27.32	49.71
TrajWorld	FCRG-clean	24.23	46.30	24.59

Continued on next page

Table D9 (continued)

Model	Readout	Early (%)	Miss (%)	Hallucination (%)
TrajWorld	DD2-noise	66.43	1.34	96.66
WIMLE	FCRG	38.02	24.43	6.91
WIMLE	FCRG-clean	47.71	19.72	28.47
WIMLE	DD2-noise	38.31	14.73	36.24
ADM-v2	FCRG	36.70	27.01	5.02
ADM-v2	FCRG-clean	54.53	12.92	33.79
ADM-v2	DD2-noise	35.33	16.22	22.66
Reference	FCRG	44.64	9.81	6.12
Reference	FCRG-clean	42.01	16.66	3.83
Reference	DD2-noise	32.09	21.17	5.34

Table D10: (a) HardStop1D.

Model	Readout	AP	nAP	Binary Brier	ECE	Cal. bias
MLP	FCRG	0.8331	0.7969	0.0660	0.0817	0.0138
MLP	FCRG-clean	0.8569	0.8234	0.0656	0.0861	0.0197
MLP	DD2-noise	0.4905	0.3864	0.1357	0.1427	-0.0374
Ensemble	FCRG	0.8591	0.8261	0.0617	0.0800	0.0089
Ensemble	FCRG-clean	0.8751	0.8453	0.0628	0.0869	0.0159
Ensemble	DD2-noise	0.5155	0.4146	0.1333	0.1373	-0.0457
DreamerV3	FCRG	0.8115	0.7664	0.0894	0.1269	0.0025
DreamerV3	FCRG-clean	0.8142	0.7680	0.0910	0.1355	0.0225
DreamerV3	DD2-noise	0.3736	0.2241	0.1503	0.1420	-0.0732
TrajWorld	FCRG	0.6871	0.6164	0.1031	0.1176	0.0745
TrajWorld	FCRG-clean	0.6958	0.6258	0.1011	0.1092	0.0634
TrajWorld	DD2-noise	0.2884	0.1197	0.2325	0.2765	0.2254
WIMLE	FCRG	0.7282	0.6732	0.0834	0.0990	-0.0028
WIMLE	FCRG-clean	0.7002	0.6390	0.0876	0.1066	0.0229
WIMLE	DD2-noise	0.6156	0.5331	0.1448	0.1563	-0.1211
ADM-v2	FCRG	0.8722	0.8427	0.0639	0.0889	-0.0090
ADM-v2	FCRG-clean	0.8661	0.8348	0.0657	0.0977	0.0023
ADM-v2	DD2-noise	0.5377	0.4367	0.1304	0.1330	-0.0689
Reference	FCRG	0.8960	0.8751	0.0531	0.0652	-0.0237
Reference	FCRG-clean	0.9095	0.8897	0.0500	0.0669	-0.0171
Reference	DD2-noise	0.6202	0.5405	0.1210	0.1270	-0.0813

Table D10: (b) HardStop1D.

Model	Readout	Joint NLL	Joint Brier	Interval NLL	Survival NLL
MLP	FCRG	0.4224	0.1982	1.5625	0.1416
MLP	FCRG-clean	0.4461	0.2084	1.6924	0.1450
MLP	DD2-noise	0.6012	0.3032	2.4753	0.1525
Ensemble	FCRG	0.4082	0.1905	1.5438	0.1288
Ensemble	FCRG-clean	0.4351	0.2021	1.6606	0.1368
Ensemble	DD2-noise	0.5943	0.3002	2.4640	0.1405
DreamerV3	FCRG	0.5252	0.2435	2.1182	0.1541
DreamerV3	FCRG-clean	0.5439	0.2473	2.0816	0.1808
DreamerV3	DD2-noise	0.7287	0.3230	3.2472	0.1143
TrajWorld	FCRG	0.5953	0.2735	2.0150	0.2645

Continued on next page

Table D10 (continued)

Model	Readout	Joint NLL	Joint Brier	Interval NLL	Survival NLL
TrajWorld	FCRG-clean	0.5766	0.2663	1.9873	0.2427
TrajWorld	DD2-noise	0.9260	0.4479	2.1706	0.6222
WIMLE	FCRG	0.4605	0.2246	1.8690	0.1389
WIMLE	FCRG-clean	0.5054	0.2387	1.9388	0.1755
WIMLE	DD2-noise	0.6655	0.3149	3.3213	0.0509
ADM-v2	FCRG	0.4037	0.1915	1.6458	0.1083
ADM-v2	FCRG-clean	0.4350	0.2015	1.7249	0.1260
ADM-v2	DD2-noise	0.5910	0.2940	2.6435	0.1045
Reference	FCRG	0.3106	0.1514	1.2189	0.0747
Reference	FCRG-clean	0.3114	0.1454	1.1874	0.0801
Reference	DD2-noise	0.5159	0.2784	2.2349	0.0810

Table D10: (c) HardStop1D.

Model	Readout	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one (%)
MLP	FCRG	10.76	93.03	63.79	53.03	85.40
MLP	FCRG-clean	12.58	93.94	60.69	38.27	85.80
MLP	DD2-noise	25.06	89.98	43.71	49.14	75.75
Ensemble	FCRG	10.53	94.03	64.34	52.19	85.70
Ensemble	FCRG-clean	12.00	95.09	62.26	44.37	86.84
Ensemble	DD2-noise	24.82	92.01	44.28	51.10	78.72
DreamerV3	FCRG	11.92	84.97	60.85	35.71	70.23
DreamerV3	FCRG-clean	17.63	93.29	54.47	28.52	80.28
DreamerV3	DD2-noise	48.41	73.56	24.05	37.94	59.75
TrajWorld	FCRG	23.98	87.37	43.78	32.39	67.73
TrajWorld	FCRG-clean	23.90	87.38	44.00	34.75	71.32
TrajWorld	DD2-noise	94.68	97.93	19.02	38.62	69.45
WIMLE	FCRG	12.06	83.45	61.98	42.29	70.62
WIMLE	FCRG-clean	17.32	91.86	54.99	34.45	81.37
WIMLE	DD2-noise	11.13	51.81	55.53	27.04	44.44
ADM-v2	FCRG	6.90	87.05	72.69	48.36	78.77
ADM-v2	FCRG-clean	8.85	89.20	68.77	43.64	80.39
ADM-v2	DD2-noise	23.39	82.44	43.57	46.46	71.87
Reference	FCRG	4.94	90.31	78.77	67.93	84.71
Reference	FCRG-clean	5.67	91.91	77.55	67.50	85.90
Reference	DD2-noise	16.70	87.53	52.17	60.71	77.53

Table D10: (d) HardStop1D. Metric panel 1 of 2; identifying columns are repeated.

Model	Readout	Dist. MAE	nMAE	Loc. bias	Late (%)
MLP	FCRG	0.1631	0.2212	0.0385	58.29
MLP	FCRG-clean	0.1932	0.2583	0.0204	55.34
MLP	DD2-noise	0.2143	0.2895	0.1082	62.20
Ensemble	FCRG	0.1681	0.2267	0.0505	60.77
Ensemble	FCRG-clean	0.1801	0.2413	0.0123	56.55
Ensemble	DD2-noise	0.2096	0.2833	0.1172	62.69
DreamerV3	FCRG	0.2478	0.3364	0.1286	66.46
DreamerV3	FCRG-clean	0.2413	0.3248	0.0178	52.16
DreamerV3	DD2-noise	0.2987	0.4049	0.2415	77.99

Continued on next page

Table D10 (continued)

Model	Readout	Dist. MAE	nMAE	Loc. bias	Late (%)
TrajWorld	FCRG	0.2474	0.3299	0.0449	53.76
TrajWorld	FCRG-clean	0.2398	0.3199	0.0560	55.91
TrajWorld	DD2-noise	0.2526	0.3367	0.0757	57.14
WIMLE	FCRG	0.2458	0.3374	0.1399	67.28
WIMLE	FCRG-clean	0.2109	0.2834	0.0160	52.45
WIMLE	DD2-noise	0.3622	0.4822	0.3387	85.55
ADM-v2	FCRG	0.1810	0.2467	0.0837	64.53
ADM-v2	FCRG-clean	0.1854	0.2503	0.0461	56.96
ADM-v2	DD2-noise	0.2543	0.3443	0.1910	72.20
Reference	FCRG	0.1458	0.1983	0.0524	60.06
Reference	FCRG-clean	0.1456	0.1977	0.0336	59.34
Reference	DD2-noise	0.1993	0.2682	0.0870	63.85

Table D10: (d) HardStop1D. Metric panel 2 of 2; identifying columns are repeated.

Model	Readout	Early (%)	Miss (%)	Hallucination (%)
MLP	FCRG	41.71	6.97	10.76
MLP	FCRG-clean	44.66	6.06	12.58
MLP	DD2-noise	37.80	10.02	25.06
Ensemble	FCRG	39.23	5.97	10.53
Ensemble	FCRG-clean	43.45	4.91	12.00
Ensemble	DD2-noise	37.31	7.99	24.82
DreamerV3	FCRG	33.54	15.03	11.92
DreamerV3	FCRG-clean	47.84	6.71	17.63
DreamerV3	DD2-noise	22.01	26.44	48.41
TrajWorld	FCRG	46.24	12.63	23.98
TrajWorld	FCRG-clean	44.09	12.62	23.90
TrajWorld	DD2-noise	42.86	2.07	94.68
WIMLE	FCRG	32.72	16.55	12.06
WIMLE	FCRG-clean	47.55	8.14	17.32
WIMLE	DD2-noise	14.45	48.19	11.13
ADM-v2	FCRG	35.47	12.95	6.90
ADM-v2	FCRG-clean	43.04	10.80	8.85
ADM-v2	DD2-noise	27.80	17.56	23.39
Reference	FCRG	39.94	9.69	4.94
Reference	FCRG-clean	40.66	8.09	5.67
Reference	DD2-noise	36.15	12.47	16.70

Table D11: (a) SaturationKink.

Model	Readout	AP	nAP	Binary Brier	ECE	Cal. bias
MLP	FCRG	0.9870	0.9583	0.0593	0.0881	-0.0256
MLP	FCRG-clean	0.9948	0.9826	0.2194	0.2502	0.2438
MLP	DD2-noise	0.9946	0.9825	0.0947	0.1167	0.1061
Ensemble	FCRG	0.9874	0.9601	0.0578	0.0828	-0.0268
Ensemble	FCRG-clean	0.9942	0.9810	0.2173	0.2487	0.2421
Ensemble	DD2-noise	0.9955	0.9855	0.0774	0.1002	0.0905
DreamerV3	FCRG	0.9411	0.8126	0.0904	0.1152	0.0139
DreamerV3	FCRG-clean	0.9334	0.7873	0.2220	0.2504	0.2462
DreamerV3	DD2-noise	0.9102	0.7119	0.1006	0.1066	0.0691

Continued on next page

Table D11 (continued)

Model	Readout	AP	nAP	Binary Brier	ECE	Cal. bias
TrajWorld	FCRG	0.8987	0.6710	0.1188	0.1171	0.0932
TrajWorld	FCRG-clean	0.8960	0.6572	0.2324	0.2545	0.2528
TrajWorld	DD2-noise	0.8395	0.4672	0.1636	0.1641	0.1600
WIMLE	FCRG	0.9989	0.9968	0.0335	0.0893	-0.0652
WIMLE	FCRG-clean	0.9997	0.9992	0.1848	0.2280	0.2150
WIMLE	DD2-noise	1.0000	1.0000	0.0224	0.0635	0.0070
ADM-v2	FCRG	0.9977	0.9928	0.0397	0.0920	-0.0616
ADM-v2	FCRG-clean	0.9981	0.9940	0.1801	0.2211	0.2110
ADM-v2	DD2-noise	0.9999	0.9997	0.0248	0.0634	0.0148
Reference	FCRG	1.0000	1.0000	0.0005	0.0052	-0.0047
Reference	FCRG-clean	1.0000	1.0000	0.0111	0.0856	-0.0076
Reference	DD2-noise	1.0000	1.0000	0.0217	0.0635	0.0008

Table D11: (b) SaturationKink.

Model	Readout	Joint NLL	Joint Brier	Interval NLL	Survival NLL
MLP	FCRG	0.4121	0.2237	0.4475	0.3212
MLP	FCRG-clean	1.9081	0.8144	1.2767	2.9667
MLP	DD2-noise	0.7380	0.4111	0.2961	0.9264
Ensemble	FCRG	0.3930	0.2159	0.4261	0.3087
Ensemble	FCRG-clean	1.8668	0.8018	1.2475	2.8960
Ensemble	DD2-noise	0.6353	0.3728	0.2455	0.6815
DreamerV3	FCRG	2.1671	0.7523	2.1586	0.7150
DreamerV3	FCRG-clean	2.4636	0.9288	1.6460	3.1268
DreamerV3	DD2-noise	3.6027	1.0271	3.9795	1.6656
TrajWorld	FCRG	2.6071	0.8022	2.4643	1.9220
TrajWorld	FCRG-clean	2.5748	0.9705	1.7330	3.6054
TrajWorld	DD2-noise	6.8131	1.2735	6.5728	5.6042
WIMLE	FCRG	0.1897	0.0937	0.2339	0.0829
WIMLE	FCRG-clean	1.4268	0.6427	0.9661	2.0759
WIMLE	DD2-noise	0.5127	0.2913	0.2845	0.1731
ADM-v2	FCRG	0.3279	0.1617	0.4126	0.1277
ADM-v2	FCRG-clean	1.5058	0.6787	1.0621	2.1142
ADM-v2	DD2-noise	0.4893	0.2854	0.2396	0.1992
Reference	FCRG	0.0081	0.0014	0.0080	0.0014
Reference	FCRG-clean	0.3486	0.1138	0.2814	0.1590
Reference	DD2-noise	0.5177	0.2912	0.2981	0.1572

Table D11: (c) SaturationKink.

Model	Readout	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one (%)
MLP	FCRG	4.84	85.30	97.63	76.14	84.54
MLP	FCRG-clean	58.85	100.00	80.81	28.69	78.06
MLP	DD2-noise	21.73	96.47	92.23	86.35	95.49
Ensemble	FCRG	4.38	85.23	97.82	76.60	84.60
Ensemble	FCRG-clean	58.85	100.00	80.81	30.84	81.04
Ensemble	DD2-noise	8.81	95.28	96.39	87.05	94.36
DreamerV3	FCRG	21.74	89.49	90.26	34.34	70.39

Continued on next page

Table D11 (continued)

Model	Readout	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one (%)
DreamerV3	FCRG-clean	58.85	100.00	80.81	22.60	69.38
DreamerV3	DD2-noise	23.26	88.21	89.75	8.75	49.12
TrajWorld	FCRG	33.94	94.77	87.04	42.03	75.46
TrajWorld	FCRG-clean	58.25	99.78	80.85	23.61	67.56
TrajWorld	DD2-noise	51.48	97.76	82.32	19.03	66.40
WIMLE	FCRG	0.00	84.66	100.00	82.56	84.40
WIMLE	FCRG-clean	55.99	100.00	81.51	51.61	89.58
WIMLE	DD2-noise	0.00	90.21	100.00	81.61	89.00
ADM-v2	FCRG	0.00	82.87	100.00	77.52	82.66
ADM-v2	FCRG-clean	53.34	100.00	82.25	44.52	88.57
ADM-v2	DD2-noise	0.00	90.64	100.00	82.49	89.43
Reference	FCRG	0.00	99.70	100.00	99.70	99.70
Reference	FCRG-clean	0.00	84.99	100.00	84.99	84.99
Reference	DD2-noise	0.00	89.42	100.00	80.53	88.21

Table D11: (d) SaturationKink. Metric panel 1 of 2; identifying columns are repeated.

Model	Readout	Dist. MAE	nMAE	Loc. bias	Late (%)
MLP	FCRG	0.1595	0.2205	0.1138	62.96
MLP	FCRG-clean	0.1811	0.2300	-0.1432	27.82
MLP	DD2-noise	0.0687	0.0890	0.0307	59.49
Ensemble	FCRG	0.1601	0.2216	0.1117	62.38
Ensemble	FCRG-clean	0.1716	0.2181	-0.1337	28.01
Ensemble	DD2-noise	0.0703	0.0928	0.0347	59.44
DreamerV3	FCRG	0.2158	0.2947	0.0536	49.00
DreamerV3	FCRG-clean	0.2033	0.2579	-0.1595	28.26
DreamerV3	DD2-noise	0.3367	0.4468	0.0806	55.18
TrajWorld	FCRG	0.1903	0.2532	0.0070	48.78
TrajWorld	FCRG-clean	0.2059	0.2614	-0.1615	27.83
TrajWorld	DD2-noise	0.2326	0.3010	-0.0690	42.46
WIMLE	FCRG	0.1716	0.2265	0.1294	61.35
WIMLE	FCRG-clean	0.1281	0.1634	-0.0814	38.70
WIMLE	DD2-noise	0.1020	0.1396	0.0746	61.81
ADM-v2	FCRG	0.1817	0.2479	0.1381	63.41
ADM-v2	FCRG-clean	0.1375	0.1754	-0.0944	33.95
ADM-v2	DD2-noise	0.0969	0.1319	0.0697	61.36
Reference	FCRG	0.0577	0.0733	0.0101	53.36
Reference	FCRG-clean	0.1436	0.1933	0.1106	61.27
Reference	DD2-noise	0.1038	0.1420	0.0779	62.49

Table D11: (d) SaturationKink. Metric panel 2 of 2; identifying columns are repeated.

Model	Readout	Early (%)	Miss (%)	Hallucination (%)
MLP	FCRG	37.04	14.70	4.84
MLP	FCRG-clean	72.18	0.00	58.85
MLP	DD2-noise	40.51	3.53	21.73
Ensemble	FCRG	37.62	14.77	4.38
Ensemble	FCRG-clean	71.99	0.00	58.85
Ensemble	DD2-noise	40.56	4.72	8.81
DreamerV3	FCRG	51.00	10.51	21.74

Continued on next page

Table D11 (continued)

Model	Readout	Early (%)	Miss (%)	Hallucination (%)
DreamerV3	FCRG-clean	71.74	0.00	58.85
DreamerV3	DD2-noise	44.82	11.79	23.26
TrajWorld	FCRG	51.22	5.23	33.94
TrajWorld	FCRG-clean	72.17	0.22	58.25
TrajWorld	DD2-noise	57.54	2.24	51.48
WIMLE	FCRG	38.65	15.34	0.00
WIMLE	FCRG-clean	61.30	0.00	55.99
WIMLE	DD2-noise	38.19	9.79	0.00
ADM-v2	FCRG	36.59	17.13	0.00
ADM-v2	FCRG-clean	66.05	0.00	53.34
ADM-v2	DD2-noise	38.64	9.36	0.00
Reference	FCRG	46.64	0.30	0.00
Reference	FCRG-clean	38.73	15.01	0.00
Reference	DD2-noise	37.51	10.58	0.00

Table D12: (a) BouncingBall.

Model	Readout	AP	nAP	Binary Brier	ECE	Cal. bias
MLP	FCRG	0.9110	0.8761	0.0637	0.0597	-0.0248
MLP	FCRG-clean	0.9117	0.8770	0.0630	0.0588	-0.0265
MLP	DD2-noise	0.4402	0.2772	0.1692	0.1503	-0.0148
Ensemble	FCRG	0.9165	0.8830	0.0642	0.0575	-0.0266
Ensemble	FCRG-clean	0.9165	0.8830	0.0636	0.0569	-0.0282
Ensemble	DD2-noise	0.4706	0.3150	0.1636	0.1415	-0.0286
DreamerV3	FCRG	0.7835	0.7027	0.1073	0.1044	-0.0544
DreamerV3	FCRG-clean	0.7848	0.7042	0.1071	0.1037	-0.0548
DreamerV3	DD2-noise	0.4884	0.3321	0.1801	0.1673	-0.1097
TrajWorld	FCRG	0.8435	0.7888	0.0935	0.0853	-0.0127
TrajWorld	FCRG-clean	0.8471	0.7933	0.0929	0.0839	-0.0154
TrajWorld	DD2-noise	0.4523	0.2843	0.1761	0.1746	0.0878
WIMLE	FCRG	0.5707	0.4532	0.1508	0.1685	-0.0562
WIMLE	FCRG-clean	0.5706	0.4532	0.1503	0.1673	-0.0566
WIMLE	DD2-noise	0.4789	0.3228	0.1859	0.1811	-0.1482
ADM-v2	FCRG	0.9015	0.8602	0.0720	0.0713	-0.0288
ADM-v2	FCRG-clean	0.9023	0.8612	0.0716	0.0723	-0.0300
ADM-v2	DD2-noise	0.5961	0.4702	0.1601	0.1572	-0.0602
Reference	FCRG	0.9377	0.9105	0.0556	0.0555	-0.0302
Reference	FCRG-clean	0.9391	0.9121	0.0548	0.0551	-0.0291
Reference	DD2-noise	0.5800	0.4576	0.1529	0.1296	-0.0715

Table D12: (b) BouncingBall.

Model	Readout	Joint NLL	Joint Brier	Interval NLL	Survival NLL
MLP	FCRG	0.4743	0.1671	1.5983	0.0691
MLP	FCRG-clean	0.4828	0.1703	1.6517	0.0659
MLP	DD2-noise	0.7472	0.3784	2.2674	0.2373
Ensemble	FCRG	0.4751	0.1733	1.6087	0.0665
Ensemble	FCRG-clean	0.4878	0.1784	1.6811	0.0636
Ensemble	DD2-noise	0.7480	0.3742	2.3478	0.2138

Continued on next page

Table D12 (continued)

Model	Readout	Joint NLL	Joint Brier	Interval NLL	Survival NLL
DreamerV3	FCRG	0.8702	0.3043	3.1408	0.1180
DreamerV3	FCRG-clean	0.8718	0.3046	3.1595	0.1176
DreamerV3	DD2-noise	0.9806	0.4021	3.4584	0.1243
TrajWorld	FCRG	0.7128	0.2844	2.4106	0.1641
TrajWorld	FCRG-clean	0.7222	0.2856	2.4778	0.1582
TrajWorld	DD2-noise	0.9024	0.4066	2.4519	0.3799
WIMLE	FCRG	1.4973	0.3336	6.5408	0.1473
WIMLE	FCRG-clean	1.4985	0.3343	6.5530	0.1467
WIMLE	DD2-noise	1.0511	0.4070	3.8971	0.0811
ADM-v2	FCRG	0.5775	0.2043	1.9689	0.0797
ADM-v2	FCRG-clean	0.5867	0.2086	2.0293	0.0779
ADM-v2	DD2-noise	0.8390	0.3769	2.8258	0.1709
Reference	FCRG	0.3901	0.1359	1.2276	0.0500
Reference	FCRG-clean	0.3900	0.1343	1.2244	0.0492
Reference	DD2-noise	0.6518	0.3504	2.1806	0.1495

Table D12: (c) BouncingBall.

Model	Readout	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one (%)
MLP	FCRG	12.28	85.88	68.98	71.33	84.10
MLP	FCRG-clean	10.56	85.13	71.78	68.22	83.03
MLP	DD2-noise	59.82	90.84	32.13	52.17	73.85
Ensemble	FCRG	11.07	85.20	71.22	67.38	82.89
Ensemble	FCRG-clean	9.30	84.83	74.32	63.13	81.40
Ensemble	DD2-noise	58.83	88.70	31.88	47.14	69.98
DreamerV3	FCRG	17.96	78.72	60.68	36.40	65.00
DreamerV3	FCRG-clean	16.73	77.63	62.63	34.80	64.40
DreamerV3	DD2-noise	41.21	71.20	34.98	28.45	53.30
TrajWorld	FCRG	17.52	82.00	60.48	44.08	70.74
TrajWorld	FCRG-clean	15.76	81.30	62.94	43.52	69.37
TrajWorld	DD2-noise	96.01	98.63	24.02	39.09	73.09
WIMLE	FCRG	20.76	47.98	35.99	30.92	44.82
WIMLE	FCRG-clean	18.97	47.72	37.91	30.13	44.30
WIMLE	DD2-noise	17.99	40.66	53.01	15.30	30.43
ADM-v2	FCRG	13.86	83.45	68.22	58.56	79.10
ADM-v2	FCRG-clean	12.41	82.69	70.33	55.18	77.25
ADM-v2	DD2-noise	52.72	86.40	33.51	34.70	63.62
Reference	FCRG	9.66	87.11	74.01	78.96	85.88
Reference	FCRG-clean	8.49	86.96	76.23	78.55	85.72
Reference	DD2-noise	47.69	82.60	34.62	54.27	68.18

Table D12: (d) BouncingBall. Metric panel 1 of 2; identifying columns are repeated.

Model	Readout	Dist. MAE	nMAE	Loc. bias	Late (%)
MLP	FCRG	0.1420	0.1985	0.0767	60.25
MLP	FCRG-clean	0.1500	0.2099	0.0850	61.29
MLP	DD2-noise	0.2223	0.3051	0.1637	69.82
Ensemble	FCRG	0.1457	0.2023	0.0769	58.43
Ensemble	FCRG-clean	0.1538	0.2133	0.0832	59.49

Continued on next page

Table D12 (continued)

Model	Readout	Dist. MAE	nMAE	Loc. bias	Late (%)
Ensemble	DD2-noise	0.2432	0.3345	0.1830	69.70
DreamerV3	FCRG	0.2533	0.3526	0.1849	74.98
DreamerV3	FCRG-clean	0.2538	0.3529	0.1845	74.54
DreamerV3	DD2-noise	0.3246	0.4552	0.2957	87.95
TrajWorld	FCRG	0.2048	0.2865	0.1021	59.64
TrajWorld	FCRG-clean	0.2089	0.2920	0.1052	60.87
TrajWorld	DD2-noise	0.2495	0.3391	0.1970	78.41
WIMLE	FCRG	0.3425	0.4619	0.3030	80.80
WIMLE	FCRG-clean	0.3446	0.4651	0.3055	81.90
WIMLE	DD2-noise	0.4325	0.5855	0.4194	93.64
ADM-v2	FCRG	0.1562	0.2185	0.0784	59.27
ADM-v2	FCRG-clean	0.1621	0.2282	0.0832	60.04
ADM-v2	DD2-noise	0.2799	0.3897	0.2496	85.76
Reference	FCRG	0.1280	0.1732	0.0713	60.95
Reference	FCRG-clean	0.1301	0.1752	0.0730	60.69
Reference	DD2-noise	0.2286	0.3182	0.1488	67.38

Table D12: (d) BouncingBall. Metric panel 2 of 2; identifying columns are repeated.

Model	Readout	Early (%)	Miss (%)	Hallucination (%)
MLP	FCRG	39.75	14.12	12.28
MLP	FCRG-clean	38.71	14.87	10.56
MLP	DD2-noise	30.18	9.16	59.82
Ensemble	FCRG	41.57	14.80	11.07
Ensemble	FCRG-clean	40.51	15.17	9.30
Ensemble	DD2-noise	30.30	11.30	58.83
DreamerV3	FCRG	25.02	21.28	17.96
DreamerV3	FCRG-clean	25.46	22.37	16.73
DreamerV3	DD2-noise	12.05	28.80	41.21
TrajWorld	FCRG	40.36	18.00	17.52
TrajWorld	FCRG-clean	39.13	18.70	15.76
TrajWorld	DD2-noise	21.59	1.37	96.01
WIMLE	FCRG	19.20	52.02	20.76
WIMLE	FCRG-clean	18.10	52.28	18.97
WIMLE	DD2-noise	6.36	59.34	17.99
ADM-v2	FCRG	40.73	16.55	13.86
ADM-v2	FCRG-clean	39.96	17.31	12.41
ADM-v2	DD2-noise	14.24	13.60	52.72
Reference	FCRG	39.05	12.89	9.66
Reference	FCRG-clean	39.31	13.04	8.49
Reference	DD2-noise	32.62	17.40	47.69

Table D13: (a) StickSlipBlock.

Model	Readout	AP	nAP	Binary Brier	ECE	Cal. bias
MLP	FCRG	0.9664	0.8968	0.0979	0.0911	0.0794
MLP	FCRG-clean	0.9781	0.9373	0.1940	0.2187	0.2173
MLP	DD2-noise	0.8882	0.6514	0.1928	0.1988	0.1926
Ensemble	FCRG	0.9668	0.8982	0.1057	0.1062	0.0905

Continued on next page

Table D13 (continued)

Model	Readout	AP	nAP	Binary Brier	ECE	Cal. bias
Ensemble	FCRG-clean	0.9853	0.9572	0.1927	0.2171	0.2159
Ensemble	DD2-noise	0.8769	0.6130	0.1819	0.1884	0.1825
DreamerV3	FCRG	0.9940	0.9823	0.0417	0.0426	-0.0357
DreamerV3	FCRG-clean	0.9909	0.9745	0.2127	0.2448	0.2405
DreamerV3	DD2-noise	0.9848	0.9542	0.2126	0.2124	0.2058
TrajWorld	FCRG	0.9834	0.9453	0.0346	0.0316	0.0202
TrajWorld	FCRG-clean	0.9902	0.9684	0.1768	0.2148	0.2120
TrajWorld	DD2-noise	0.8545	0.5057	0.1460	0.1677	0.1637
WIMLE	FCRG	0.9995	0.9986	0.0186	0.0182	0.0025
WIMLE	FCRG-clean	0.9959	0.9891	0.1762	0.2066	0.2035
WIMLE	DD2-noise	0.9927	0.9798	0.0658	0.0984	0.0726
ADM-v2	FCRG	0.9992	0.9977	0.0189	0.0180	0.0080
ADM-v2	FCRG-clean	0.9956	0.9879	0.1807	0.2130	0.2098
ADM-v2	DD2-noise	0.9829	0.9424	0.0828	0.1011	0.0812
Reference	FCRG	1.0000	1.0000	2.43×10^{-5}	0.0002	0.0002
Reference	FCRG-clean	1.0000	1.0000	0.0037	0.0289	0.0071
Reference	DD2-noise	0.9949	0.9860	0.0210	0.0508	-0.0183

Table D13: (b) StickSlipBlock.

Model	Readout	Joint NLL	Joint Brier	Interval NLL	Survival NLL
MLP	FCRG	1.6403	0.3815	1.0394	2.8364
MLP	FCRG-clean	1.6179	0.6610	0.9827	2.8024
MLP	DD2-noise	3.0731	0.8715	2.0406	4.6911
Ensemble	FCRG	1.5868	0.3865	0.9511	2.9262
Ensemble	FCRG-clean	1.5650	0.6528	0.9516	2.7422
Ensemble	DD2-noise	3.2342	0.8810	2.0947	5.0566
DreamerV3	FCRG	0.7139	0.2450	1.0147	0.0673
DreamerV3	FCRG-clean	1.5669	0.7193	1.0208	2.4387
DreamerV3	DD2-noise	1.6560	0.5992	0.7735	3.1599
TrajWorld	FCRG	0.6157	0.1398	0.5385	0.8039
TrajWorld	FCRG-clean	1.1052	0.5079	0.6461	1.9829
TrajWorld	DD2-noise	1.7123	0.4438	0.6985	3.5947
WIMLE	FCRG	0.1568	0.0707	0.1416	0.1951
WIMLE	FCRG-clean	1.0699	0.4929	0.6311	1.8926
WIMLE	DD2-noise	0.5032	0.2459	0.2868	0.5426
ADM-v2	FCRG	0.2168	0.0790	0.1797	0.2999
ADM-v2	FCRG-clean	1.0514	0.4969	0.6149	1.8969
ADM-v2	DD2-noise	0.5939	0.2721	0.2975	0.8828
Reference	FCRG	0.0024	0.0016	0.0032	0.0008
Reference	FCRG-clean	0.1233	0.0273	0.1089	0.0691
Reference	DD2-noise	0.3022	0.1459	0.2268	0.0772

Table D13: (c) StickSlipBlock.

Model	Readout	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one (%)
MLP	FCRG	13.08	88.35	94.19	74.15	83.56
MLP	FCRG-clean	45.79	98.76	82.99	53.54	78.87
MLP	DD2-noise	49.46	98.35	81.78	40.88	84.75
Ensemble	FCRG	10.64	88.95	94.90	75.09	84.02
Ensemble	FCRG-clean	46.18	98.64	82.81	51.87	76.28
Ensemble	DD2-noise	47.48	98.22	82.65	37.43	84.65
DreamerV3	FCRG	0.00	59.61	100.00	52.58	57.95
DreamerV3	FCRG-clean	34.41	97.51	87.02	39.80	69.86
DreamerV3	DD2-noise	62.91	98.44	77.90	68.11	95.24
TrajWorld	FCRG	3.25	87.46	98.46	82.38	86.11
TrajWorld	FCRG-clean	21.84	97.94	91.16	76.87	91.80
TrajWorld	DD2-noise	34.67	95.75	87.36	78.26	91.17
WIMLE	FCRG	0.00	85.54	100.00	83.45	84.85
WIMLE	FCRG-clean	19.80	98.47	91.73	78.53	90.71
WIMLE	DD2-noise	4.72	95.50	97.52	86.20	94.87
ADM-v2	FCRG	0.00	86.89	100.00	83.49	86.02
ADM-v2	FCRG-clean	13.28	98.04	94.79	81.82	97.19
ADM-v2	DD2-noise	13.17	91.44	94.91	81.97	90.68
Reference	FCRG	0.00	97.53	100.00	97.25	97.25
Reference	FCRG-clean	0.00	86.50	100.00	86.50	86.50
Reference	DD2-noise	0.00	83.55	100.00	79.12	83.07

Table D13: (d) StickSlipBlock. Metric panel 1 of 2; identifying columns are repeated.

Model	Readout	Dist. MAE	nMAE	Loc. bias	Late (%)
MLP	FCRG	0.1443	0.2028	0.0708	62.23
MLP	FCRG-clean	0.1634	0.2158	-0.1005	44.63
MLP	DD2-noise	0.1706	0.2288	-0.0844	43.61
Ensemble	FCRG	0.1403	0.1986	0.0697	61.82
Ensemble	FCRG-clean	0.1724	0.2274	-0.1086	45.30
Ensemble	DD2-noise	0.1800	0.2407	-0.0941	42.36
DreamerV3	FCRG	0.2716	0.3854	0.2452	77.07
DreamerV3	FCRG-clean	0.1992	0.2662	-0.1379	35.77
DreamerV3	DD2-noise	0.1097	0.1537	0.0071	59.13
TrajWorld	FCRG	0.1346	0.1931	0.0848	63.11
TrajWorld	FCRG-clean	0.1042	0.1425	-0.0204	52.91
TrajWorld	DD2-noise	0.0971	0.1396	0.0239	61.13
WIMLE	FCRG	0.1283	0.1828	0.0911	66.41
WIMLE	FCRG-clean	0.1034	0.1409	-0.0306	54.99
WIMLE	DD2-noise	0.0992	0.1399	0.0364	63.07
ADM-v2	FCRG	0.1370	0.1934	0.0912	63.56
ADM-v2	FCRG-clean	0.0775	0.1097	-0.0017	55.23
ADM-v2	DD2-noise	0.0895	0.1254	0.0549	68.09
Reference	FCRG	0.0838	0.1085	0.0247	55.94
Reference	FCRG-clean	0.1230	0.1693	0.0849	64.25
Reference	DD2-noise	0.1279	0.1886	0.0966	71.03

Table D13: (d) StickSlipBlock. Metric panel 2 of 2; identifying columns are repeated.

Model	Readout	Early (%)	Miss (%)	Hallucination (%)
MLP	FCRG	37.77	11.65	13.08
MLP	FCRG-clean	55.37	1.24	45.79
MLP	DD2-noise	56.39	1.65	49.46
Ensemble	FCRG	38.18	11.05	10.64
Ensemble	FCRG-clean	54.70	1.36	46.18
Ensemble	DD2-noise	57.64	1.78	47.48
DreamerV3	FCRG	22.93	40.39	0.00
DreamerV3	FCRG-clean	64.23	2.49	34.41
DreamerV3	DD2-noise	40.87	1.56	62.91
TrajWorld	FCRG	36.89	12.54	3.25
TrajWorld	FCRG-clean	47.09	2.06	21.84
TrajWorld	DD2-noise	38.87	4.25	34.67
WIMLE	FCRG	33.59	14.46	0.00
WIMLE	FCRG-clean	45.01	1.53	19.80
WIMLE	DD2-noise	36.93	4.50	4.72
ADM-v2	FCRG	36.44	13.11	0.00
ADM-v2	FCRG-clean	44.77	1.96	13.28
ADM-v2	DD2-noise	31.91	8.56	13.17
Reference	FCRG	44.06	2.47	0.00
Reference	FCRG-clean	35.75	13.50	0.00
Reference	DD2-noise	28.97	16.45	0.00

Table D14: (a) PuckWall2D.

Model	Readout	AP	nAP	Binary Brier	ECE	Cal. bias
MLP	FCRG	0.6241	0.5956	0.0517	0.0468	0.0319
MLP	FCRG-clean	0.6067	0.5769	0.0578	0.0643	0.0615
MLP	DD2-noise	0.4898	0.4510	0.0588	0.0607	0.0502
Ensemble	FCRG	0.6765	0.6526	0.0421	0.0340	0.0213
Ensemble	FCRG-clean	0.6601	0.6348	0.0486	0.0552	0.0533
Ensemble	DD2-noise	0.5278	0.4922	0.0530	0.0498	0.0357
DreamerV3	FCRG	0.2688	0.2109	0.0735	0.0606	0.0038
DreamerV3	FCRG-clean	0.2082	0.1455	0.0984	0.1068	0.0900
DreamerV3	DD2-noise	0.1433	0.0759	0.0909	0.1221	0.0980
TrajWorld	FCRG	0.4469	0.4042	0.1075	0.1251	0.1177
TrajWorld	FCRG-clean	0.3281	0.2755	0.1577	0.2037	0.2004
TrajWorld	DD2-noise	0.2368	0.1763	0.2243	0.3629	0.3624
WIMLE	FCRG	0.1493	0.0821	0.0840	0.0888	-0.0219
WIMLE	FCRG-clean	0.1699	0.1043	0.0717	0.0686	-0.0084
WIMLE	DD2-noise	0.2393	0.1799	0.0697	0.0695	-0.0490
ADM-v2	FCRG	0.6140	0.5832	0.0458	0.0343	0.0004
ADM-v2	FCRG-clean	0.6282	0.5988	0.0461	0.0472	0.0314
ADM-v2	DD2-noise	0.5654	0.5323	0.0461	0.0439	0.0098
Reference	FCRG	0.9304	0.9248	0.0184	0.0131	0.0020
Reference	FCRG-clean	0.9260	0.9196	0.0193	0.0159	0.0130
Reference	DD2-noise	0.6216	0.5926	0.0425	0.0411	0.0089

Table D14: (b) PuckWall2D.

Model	Readout	Joint NLL	Joint Brier	Interval NLL	Survival NLL
MLP	FCRG	0.3101	0.1282	2.4017	0.1436
MLP	FCRG-clean	0.3178	0.1460	2.1801	0.1690
MLP	DD2-noise	0.2980	0.1305	2.2897	0.1358
Ensemble	FCRG	0.2728	0.1126	2.3309	0.1078
Ensemble	FCRG-clean	0.2854	0.1310	2.1483	0.1356
Ensemble	DD2-noise	0.2860	0.1245	2.4478	0.1092
DreamerV3	FCRG	0.4430	0.1519	4.7362	0.0967
DreamerV3	FCRG-clean	0.4722	0.1904	3.3044	0.2396
DreamerV3	DD2-noise	0.4540	0.1762	3.4258	0.2020
TrajWorld	FCRG	0.5494	0.2251	2.5173	0.3920
TrajWorld	FCRG-clean	0.7279	0.3057	2.3185	0.5990
TrajWorld	DD2-noise	0.7869	0.3591	2.0656	0.6813
WIMLE	FCRG	0.6112	0.1676	7.2863	0.0714
WIMLE	FCRG-clean	0.4243	0.1492	4.8062	0.0708
WIMLE	DD2-noise	0.4554	0.1423	5.1510	0.0242
ADM-v2	FCRG	0.3981	0.1259	4.5817	0.0627
ADM-v2	FCRG-clean	0.3295	0.1241	3.2182	0.0966
ADM-v2	DD2-noise	0.2945	0.1152	2.9484	0.0718
Reference	FCRG	0.1100	0.0509	1.1261	0.0298
Reference	FCRG-clean	0.1000	0.0491	0.7997	0.0452
Reference	DD2-noise	0.1973	0.1011	1.7965	0.0709

Table D14: (c) PuckWall2D.

Model	Readout	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one (%)
MLP	FCRG	11.35	85.57	39.24	46.78	75.53
MLP	FCRG-clean	9.48	81.72	42.22	30.28	72.31
MLP	DD2-noise	18.92	90.88	28.31	38.10	70.01
Ensemble	FCRG	11.02	90.59	41.25	47.55	78.95
Ensemble	FCRG-clean	9.05	84.91	44.00	31.66	74.18
Ensemble	DD2-noise	18.72	92.11	28.48	35.27	68.42
DreamerV3	FCRG	15.09	48.34	20.20	16.05	32.95
DreamerV3	FCRG-clean	15.33	44.60	18.80	12.80	27.93
DreamerV3	DD2-noise	66.50	76.18	8.45	25.04	52.22
TrajWorld	FCRG	26.48	89.50	21.46	43.10	77.52
TrajWorld	FCRG-clean	28.61	80.79	18.46	29.52	66.56
TrajWorld	DD2-noise	90.14	98.53	7.93	32.29	69.59
WIMLE	FCRG	13.50	14.38	6.28	2.57	6.96
WIMLE	FCRG-clean	3.54	2.20	4.12	1.56	1.81
WIMLE	DD2-noise	0.00	0.00	NA	0.00	0.00
ADM-v2	FCRG	9.28	73.64	39.81	24.74	51.09
ADM-v2	FCRG-clean	6.33	72.13	48.55	21.76	48.59
ADM-v2	DD2-noise	13.86	87.53	33.83	29.29	61.84
Reference	FCRG	4.68	97.04	63.61	75.17	92.75
Reference	FCRG-clean	3.42	93.35	69.45	80.25	90.63
Reference	DD2-noise	7.07	91.15	50.86	58.05	73.80

Table D14: (d) PuckWall2D. Metric panel 1 of 2; identifying columns are repeated.

Model	Readout	Dist. MAE	nMAE	Loc. bias	Late (%)
MLP	FCRG	0.1798	0.2466	0.0909	58.60
MLP	FCRG-clean	0.2439	0.3315	0.1298	63.15
MLP	DD2-noise	0.2319	0.3278	0.1541	70.80
Ensemble	FCRG	0.1696	0.2315	0.0595	53.93
Ensemble	FCRG-clean	0.2220	0.3059	0.1042	60.96
Ensemble	DD2-noise	0.2473	0.3499	0.1671	71.91
DreamerV3	FCRG	0.3716	0.5128	0.2593	73.32
DreamerV3	FCRG-clean	0.3940	0.5427	0.2728	73.58
DreamerV3	DD2-noise	0.3033	0.4265	0.2630	85.17
TrajWorld	FCRG	0.1825	0.2525	0.0615	53.13
TrajWorld	FCRG-clean	0.2347	0.3251	0.1244	63.36
TrajWorld	DD2-noise	0.2477	0.3466	0.1379	68.89
WIMLE	FCRG	0.5236	0.6968	0.5053	96.98
WIMLE	FCRG-clean	0.5571	0.7404	0.5567	98.61
WIMLE	DD2-noise	0.5621	0.7532	0.5621	100.00
ADM-v2	FCRG	0.3052	0.4158	0.1934	70.18
ADM-v2	FCRG-clean	0.3093	0.4227	0.1849	71.80
ADM-v2	DD2-noise	0.2748	0.3787	0.2055	77.56
Reference	FCRG	0.0955	0.1333	0.0152	49.48
Reference	FCRG-clean	0.1092	0.1516	0.0333	47.35
Reference	DD2-noise	0.2010	0.2954	0.0700	56.42

Table D14: (d) PuckWall2D. Metric panel 2 of 2; identifying columns are repeated.

Model	Readout	Early (%)	Miss (%)	Hallucination (%)
MLP	FCRG	41.40	14.43	11.35
MLP	FCRG-clean	36.85	18.28	9.48
MLP	DD2-noise	29.20	9.12	18.92
Ensemble	FCRG	46.07	9.41	11.02
Ensemble	FCRG-clean	39.04	15.09	9.05
Ensemble	DD2-noise	28.09	7.89	18.72
DreamerV3	FCRG	26.68	51.66	15.09
DreamerV3	FCRG-clean	26.42	55.40	15.33
DreamerV3	DD2-noise	14.83	23.82	66.50
TrajWorld	FCRG	46.87	10.50	26.48
TrajWorld	FCRG-clean	36.64	19.21	28.61
TrajWorld	DD2-noise	31.11	1.47	90.14
WIMLE	FCRG	3.02	85.62	13.50
WIMLE	FCRG-clean	1.39	97.80	3.54
WIMLE	DD2-noise	0.00	100.00	0.00
ADM-v2	FCRG	29.82	26.36	9.28
ADM-v2	FCRG-clean	28.20	27.87	6.33
ADM-v2	DD2-noise	22.44	12.47	13.86
Reference	FCRG	50.52	2.96	4.68
Reference	FCRG-clean	52.65	6.65	3.42
Reference	DD2-noise	43.58	8.85	7.07

Table D15: (a) TwoBodyCollision.

Model	Readout	AP	nAP	Binary Brier	ECE	Cal. bias
MLP	FCRG	0.5659	0.5355	0.0495	0.0596	0.0344
MLP	FCRG-clean	0.5866	0.5571	0.0481	0.0502	0.0207
MLP	DD2-noise	0.4065	0.3649	0.0594	0.0625	-0.0222
Ensemble	FCRG	0.6075	0.5804	0.0466	0.0582	0.0296
Ensemble	FCRG-clean	0.6206	0.5936	0.0454	0.0469	0.0170
Ensemble	DD2-noise	0.4143	0.3739	0.0590	0.0626	-0.0289
DreamerV3	FCRG	0.2364	0.1809	0.0664	0.0814	0.0515
DreamerV3	FCRG-clean	0.2468	0.1918	0.0618	0.0539	0.0017
DreamerV3	DD2-noise	0.1278	0.0637	0.0731	0.0815	0.0566
TrajWorld	FCRG	0.2051	0.1484	0.1329	0.1691	0.1620
TrajWorld	FCRG-clean	0.2177	0.1623	0.1418	0.1544	0.1381
TrajWorld	DD2-noise	0.1121	0.0464	0.1668	0.2640	0.2635
WIMLE	FCRG	0.2665	0.2147	0.0656	0.0707	0.0122
WIMLE	FCRG-clean	0.2681	0.2166	0.0734	0.0757	0.0103
WIMLE	DD2-noise	0.2428	0.1882	0.0661	0.0682	-0.0470
ADM-v2	FCRG	0.5638	0.5319	0.0484	0.0560	0.0310
ADM-v2	FCRG-clean	0.5392	0.5055	0.0506	0.0481	0.0191
ADM-v2	DD2-noise	0.5206	0.4833	0.0562	0.0624	-0.0260
Reference	FCRG	0.7742	0.7603	0.0308	0.0401	0.0186
Reference	FCRG-clean	0.7824	0.7687	0.0299	0.0302	0.0082
Reference	DD2-noise	0.5878	0.5598	0.0430	0.0473	-0.0121

Table D15: (b) TwoBodyCollision.

Model	Readout	Joint NLL	Joint Brier	Interval NLL	Survival NLL
MLP	FCRG	0.2905	0.1132	2.8180	0.0999
MLP	FCRG-clean	0.3073	0.1117	3.2733	0.0837
MLP	DD2-noise	0.3149	0.1253	3.7588	0.0451
Ensemble	FCRG	0.2797	0.1086	2.7864	0.0904
Ensemble	FCRG-clean	0.2962	0.1067	3.2052	0.0759
Ensemble	DD2-noise	0.3135	0.1247	3.8614	0.0375
DreamerV3	FCRG	0.3467	0.1354	3.2254	0.1301
DreamerV3	FCRG-clean	0.3362	0.1293	3.8530	0.0730
DreamerV3	DD2-noise	0.3710	0.1454	3.3979	0.1402
TrajWorld	FCRG	0.5545	0.2382	2.6206	0.3991
TrajWorld	FCRG-clean	0.6118	0.2557	3.1007	0.4241
TrajWorld	DD2-noise	0.6278	0.2766	2.6938	0.4751
WIMLE	FCRG	0.3537	0.1366	3.8902	0.0866
WIMLE	FCRG-clean	0.3897	0.1500	4.3250	0.0936
WIMLE	DD2-noise	0.3841	0.1343	4.8926	0.0222
ADM-v2	FCRG	0.2860	0.1118	2.8463	0.0944
ADM-v2	FCRG-clean	0.3135	0.1152	3.3693	0.0849
ADM-v2	DD2-noise	0.2987	0.1207	3.6590	0.0389
Reference	FCRG	0.1757	0.0742	1.5435	0.0675
Reference	FCRG-clean	0.1882	0.0725	1.7724	0.0601
Reference	DD2-noise	0.1940	0.0996	2.1949	0.0438

Table D15: (c) TwoBodyCollision.

Model	Readout	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one (%)
MLP	FCRG	12.71	75.41	31.94	44.17	58.97
MLP	FCRG-clean	11.86	75.33	33.79	46.51	59.67
MLP	DD2-noise	8.31	47.16	34.46	23.07	34.10
Ensemble	FCRG	12.12	76.54	33.50	47.21	60.02
Ensemble	FCRG-clean	11.22	75.28	34.94	48.48	60.06
Ensemble	DD2-noise	7.96	44.26	36.35	25.17	33.91
DreamerV3	FCRG	13.63	42.44	20.74	23.04	30.40
DreamerV3	FCRG-clean	8.88	31.44	24.85	18.12	22.58
DreamerV3	DD2-noise	65.58	81.91	8.57	45.89	60.21
TrajWorld	FCRG	32.27	87.49	18.55	40.00	58.09
TrajWorld	FCRG-clean	26.74	81.46	20.66	37.89	55.45
TrajWorld	DD2-noise	98.08	98.96	7.00	45.55	71.47
WIMLE	FCRG	11.95	30.12	13.78	15.50	22.01
WIMLE	FCRG-clean	13.49	31.76	12.17	19.05	25.07
WIMLE	DD2-noise	0.00	0.87	NA	0.00	0.35
ADM-v2	FCRG	11.91	77.47	33.74	46.51	59.12
ADM-v2	FCRG-clean	11.67	74.35	33.89	45.78	57.48
ADM-v2	DD2-noise	6.36	58.91	44.85	29.38	42.99
Reference	FCRG	6.99	95.37	51.03	70.01	77.24
Reference	FCRG-clean	5.53	94.55	56.32	69.67	74.73
Reference	DD2-noise	9.22	89.99	42.40	63.97	75.10

Table D15: (d) TwoBodyCollision. Metric panel 1 of 2; identifying columns are repeated.

Model	Readout	Dist. MAE	nMAE	Loc. bias	Late (%)
MLP	FCRG	0.2979	0.4104	0.2340	71.67
MLP	FCRG-clean	0.2921	0.4042	0.2286	70.44
MLP	DD2-noise	0.3822	0.5249	0.3387	82.47
Ensemble	FCRG	0.2940	0.4046	0.2420	72.40
Ensemble	FCRG-clean	0.2908	0.3990	0.2387	72.17
Ensemble	DD2-noise	0.3889	0.5317	0.3541	83.67
DreamerV3	FCRG	0.3756	0.5117	0.3470	86.98
DreamerV3	FCRG-clean	0.4130	0.5619	0.3936	90.49
DreamerV3	DD2-noise	0.2485	0.3557	0.1976	74.18
TrajWorld	FCRG	0.2705	0.3803	0.1133	58.56
TrajWorld	FCRG-clean	0.2934	0.4118	0.1513	62.33
TrajWorld	DD2-noise	0.2189	0.3050	0.0982	62.38
WIMLE	FCRG	0.4249	0.5731	0.4114	90.27
WIMLE	FCRG-clean	0.4089	0.5555	0.3960	89.57
WIMLE	DD2-noise	0.5158	0.6926	0.5158	100.00
ADM-v2	FCRG	0.2731	0.3777	0.2208	73.20
ADM-v2	FCRG-clean	0.2775	0.3822	0.2283	74.54
ADM-v2	DD2-noise	0.3276	0.4511	0.2969	82.23
Reference	FCRG	0.1467	0.2075	0.0748	64.08
Reference	FCRG-clean	0.1607	0.2274	0.1082	67.97
Reference	DD2-noise	0.1885	0.2591	0.0906	66.40

Table D15: (d) TwoBodyCollision. Metric panel 2 of 2; identifying columns are repeated.

Model	Readout	Early (%)	Miss (%)	Hallucination (%)
MLP	FCRG	28.33	24.59	12.71
MLP	FCRG-clean	29.56	24.67	11.86
MLP	DD2-noise	17.53	52.84	8.31
Ensemble	FCRG	27.60	23.46	12.12
Ensemble	FCRG-clean	27.83	24.72	11.22
Ensemble	DD2-noise	16.33	55.74	7.96
DreamerV3	FCRG	13.02	57.56	13.63
DreamerV3	FCRG-clean	9.51	68.56	8.88
DreamerV3	DD2-noise	25.82	18.09	65.58
TrajWorld	FCRG	41.44	12.51	32.27
TrajWorld	FCRG-clean	37.67	18.54	26.74
TrajWorld	DD2-noise	37.62	1.04	98.08
WIMLE	FCRG	9.73	69.88	11.95
WIMLE	FCRG-clean	10.43	68.24	13.49
WIMLE	DD2-noise	0.00	99.13	0.00
ADM-v2	FCRG	26.80	22.53	11.91
ADM-v2	FCRG-clean	25.46	25.65	11.67
ADM-v2	DD2-noise	17.77	41.09	6.36
Reference	FCRG	35.92	4.63	6.99
Reference	FCRG-clean	32.03	5.45	5.53
Reference	DD2-noise	33.60	10.01	9.22

Table D16: (a) SmoothDuffing.

Model	Readout	AP	nAP	Binary Brier	ECE	Cal. bias
MLP	FCRG	NA	NA	7.10×10^{-14}	NA	1.75×10^{-8}
MLP	FCRG-clean	NA	NA	2.18×10^{-9}	NA	2.26×10^{-6}
MLP	DD2-noise	NA	NA	0.0002	NA	0.0127
Ensemble	FCRG	NA	NA	6.57×10^{-14}	NA	1.66×10^{-8}
Ensemble	FCRG-clean	NA	NA	6.65×10^{-11}	NA	5.97×10^{-7}
Ensemble	DD2-noise	NA	NA	0.0003	NA	0.0153
DreamerV3	FCRG	NA	NA	1.16×10^{-15}	NA	1.35×10^{-9}
DreamerV3	FCRG-clean	NA	NA	0.0115	NA	0.0247
DreamerV3	DD2-noise	NA	NA	0.0007	NA	0.0007
TrajWorld	FCRG	NA	NA	4.70×10^{-20}	NA	1.09×10^{-11}
TrajWorld	FCRG-clean	NA	NA	0.0176	NA	0.0318
TrajWorld	DD2-noise	NA	NA	0.0029	NA	0.0030
WIMLE	FCRG	NA	NA	7.23×10^{-14}	NA	1.71×10^{-8}
WIMLE	FCRG-clean	NA	NA	1.17×10^{-6}	NA	0.0001
WIMLE	DD2-noise	NA	NA	6.82×10^{-5}	NA	0.0047
ADM-v2	FCRG	NA	NA	2.39×10^{-15}	NA	3.03×10^{-9}
ADM-v2	FCRG-clean	NA	NA	5.98×10^{-5}	NA	0.0008
ADM-v2	DD2-noise	NA	NA	0.0298	NA	0.0342
Reference	FCRG	NA	NA	6.99×10^{-16}	NA	1.44×10^{-9}
Reference	FCRG-clean	NA	NA	4.15×10^{-18}	NA	2.11×10^{-10}
Reference	DD2-noise	NA	NA	0.0009	NA	0.0299

Table D16: (b) SmoothDuffing.

Model	Readout	Joint NLL	Joint Brier	Interval NLL	Survival NLL
MLP	FCRG	1.75×10^{-8}	9.75×10^{-14}	NA	1.75×10^{-8}
MLP	FCRG-clean	2.26×10^{-6}	2.98×10^{-9}	NA	2.26×10^{-6}
MLP	DD2-noise	0.0128	0.0003	NA	0.0128
Ensemble	FCRG	1.66×10^{-8}	9.02×10^{-14}	NA	1.66×10^{-8}
Ensemble	FCRG-clean	5.97×10^{-7}	9.14×10^{-11}	NA	5.97×10^{-7}
Ensemble	DD2-noise	0.0154	0.0004	NA	0.0154
DreamerV3	FCRG	1.35×10^{-9}	1.61×10^{-15}	NA	1.35×10^{-9}
DreamerV3	FCRG-clean	0.0380	0.0151	NA	0.0380
DreamerV3	DD2-noise	0.0081	0.0013	NA	0.0081
TrajWorld	FCRG	1.09×10^{-11}	6.86×10^{-20}	NA	1.09×10^{-11}
TrajWorld	FCRG-clean	0.0533	0.0237	NA	0.0533
TrajWorld	DD2-noise	0.0379	0.0058	NA	0.0379
WIMLE	FCRG	1.71×10^{-8}	9.91×10^{-14}	NA	1.71×10^{-8}
WIMLE	FCRG-clean	0.0001	1.50×10^{-6}	NA	0.0001
WIMLE	DD2-noise	0.0047	0.0001	NA	0.0047
ADM-v2	FCRG	3.03×10^{-9}	3.27×10^{-15}	NA	3.03×10^{-9}
ADM-v2	FCRG-clean	0.0008	8.27×10^{-5}	NA	0.0008
ADM-v2	DD2-noise	0.2208	0.0589	NA	0.2208
Reference	FCRG	1.44×10^{-9}	9.57×10^{-16}	NA	1.44×10^{-9}
Reference	FCRG-clean	2.11×10^{-10}	5.83×10^{-18}	NA	2.11×10^{-10}
Reference	DD2-noise	0.0304	0.0012	NA	0.0304

Table D16: (c) SmoothDuffing.

Model	Readout	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one (%)
MLP	FCRG	0.00	NA	NA	NA	NA
MLP	FCRG-clean	0.00	NA	NA	NA	NA
MLP	DD2-noise	0.00	NA	NA	NA	NA
Ensemble	FCRG	0.00	NA	NA	NA	NA
Ensemble	FCRG-clean	0.00	NA	NA	NA	NA
Ensemble	DD2-noise	0.00	NA	NA	NA	NA
DreamerV3	FCRG	0.00	NA	NA	NA	NA
DreamerV3	FCRG-clean	1.35	NA	NA	NA	NA
DreamerV3	DD2-noise	0.07	NA	NA	NA	NA
TrajWorld	FCRG	0.00	NA	NA	NA	NA
TrajWorld	FCRG-clean	2.60	NA	NA	NA	NA
TrajWorld	DD2-noise	0.30	NA	NA	NA	NA
WIMLE	FCRG	0.00	NA	NA	NA	NA
WIMLE	FCRG-clean	0.00	NA	NA	NA	NA
WIMLE	DD2-noise	0.00	NA	NA	NA	NA
ADM-v2	FCRG	0.00	NA	NA	NA	NA
ADM-v2	FCRG-clean	0.00	NA	NA	NA	NA

Continued on next page

Table D16 (continued)

Model	Readout	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one (%)
ADM-v2	DD2-noise	3.34	NA	0.00	NA	NA
Reference	FCRG	0.00	NA	NA	NA	NA
Reference	FCRG-clean	0.00	NA	NA	NA	NA
Reference	DD2-noise	0.00	NA	NA	NA	NA

Table D16: (d) SmoothDuffing. Metric panel 1 of 2; identifying columns are repeated.

Model	Readout	Dist. MAE	nMAE	Loc. bias	Late (%)
MLP	FCRG	NA	NA	NA	NA
MLP	FCRG-clean	NA	NA	NA	NA
MLP	DD2-noise	NA	NA	NA	NA
Ensemble	FCRG	NA	NA	NA	NA
Ensemble	FCRG-clean	NA	NA	NA	NA
Ensemble	DD2-noise	NA	NA	NA	NA
DreamerV3	FCRG	NA	NA	NA	NA
DreamerV3	FCRG-clean	NA	NA	NA	NA
DreamerV3	DD2-noise	NA	NA	NA	NA
TrajWorld	FCRG	NA	NA	NA	NA
TrajWorld	FCRG-clean	NA	NA	NA	NA
TrajWorld	DD2-noise	NA	NA	NA	NA
WIMLE	FCRG	NA	NA	NA	NA
WIMLE	FCRG-clean	NA	NA	NA	NA
WIMLE	DD2-noise	NA	NA	NA	NA
ADM-v2	FCRG	NA	NA	NA	NA
ADM-v2	FCRG-clean	NA	NA	NA	NA
ADM-v2	DD2-noise	NA	NA	NA	NA
Reference	FCRG	NA	NA	NA	NA
Reference	FCRG-clean	NA	NA	NA	NA
Reference	DD2-noise	NA	NA	NA	NA

Table D16: (d) SmoothDuffing. Metric panel 2 of 2; identifying columns are repeated.

Model	Readout	Early (%)	Miss (%)	Hallucination (%)
MLP	FCRG	NA	NA	0.00
MLP	FCRG-clean	NA	NA	0.00
MLP	DD2-noise	NA	NA	0.00
Ensemble	FCRG	NA	NA	0.00
Ensemble	FCRG-clean	NA	NA	0.00
Ensemble	DD2-noise	NA	NA	0.00
DreamerV3	FCRG	NA	NA	0.00
DreamerV3	FCRG-clean	NA	NA	1.35
DreamerV3	DD2-noise	NA	NA	0.07
TrajWorld	FCRG	NA	NA	0.00
TrajWorld	FCRG-clean	NA	NA	2.60
TrajWorld	DD2-noise	NA	NA	0.30
WIMLE	FCRG	NA	NA	0.00
WIMLE	FCRG-clean	NA	NA	0.00
WIMLE	DD2-noise	NA	NA	0.00
ADM-v2	FCRG	NA	NA	0.00
ADM-v2	FCRG-clean	NA	NA	0.00

Continued on next page

Table D16 (continued)

Model	Readout	Early (%)	Miss (%)	Hallucination (%)
ADM-v2	DD2-noise	NA	NA	3.34
Reference	FCRG	NA	NA	0.00
Reference	FCRG-clean	NA	NA	0.00
Reference	DD2-noise	NA	NA	0.00

The matrices retain models and metrics not discussed individually in the main text, allowing any diagnosis to be inspected within the same system and response. Together with the error decomposition, they support attribute-specific interpretation rather than selecting the most favorable score as an overall conclusion.

E From Boundary Information to Fixed-Budget Action Search

E.1 Tasks, Candidate Generation, and Comparators

The downstream experiment tests whether first intervals guide query allocation rather than only describing prediction quality. StickSlipBlock and PuckWall2D both define terminal-position objectives independently of boundary labels. They cover one-dimensional sticking–sliding transitions and contact during two-dimensional motion, allowing the same allocation principle to be examined under different action geometries and mode mechanisms. MLP and TrajWorld extend this comparison to feedforward transition regression and Transformer sequence prediction. The target is the use of boundary information with fixed predictors, not the training of a new controller or an overall architecture ranking.

All densified methods retain the original candidates and select by the fixed world model's predicted terminal cost. Simulator evaluation follows committed selection; reference guidance supplies coarse intervals, not true candidate returns. Base records original-candidate search, while Uniform and Local control for additional query budget and ordinary local refinement. DD2 and the mislocated arm examine allocation under an alternative readout and altered interval information, respectively. Table E1 lists the seven search rules.

Table E1: Seven fixed search rules. Uniform and Local are the primary equal-query controls. A fixed uniform-fill rule completes insufficient boundary proposals.

Rule	Candidate allocation	StickSlip / Puck Wall model queries
Base	Original candidates only	7 / 93
Uniform	Uniform additions over the admissible action region	15 / 237
Local	Refinement around the originally predicted best action	15 / 237
Reference boundary	Additions within/around privileged coarse intervals	15 / 237
FCRG guidance	Additions within/around estimated coarse intervals	15 / 237
DD2 guidance	Same allocation using the second-difference readout	15 / 237
Mislocated FCRG	Fixed two-bin rotation, keeping alarm decisions	15 / 237

StickSlipBlock targets $x_0 + 0.1$ with tolerance 0.05, zero initial velocity, and nonnegative actions. PuckWall2D targets $(x_{\text{wall}} - 0.1, y_0)$, with tolerance 0.1 in each coordinate. Terminal cost is the mean squared tolerance-normalized coordinate error. Success requires every coordinate to meet its tolerance, so two-dimensional success is not equivalent to a threshold of one on mean cost. Densified

arms use 7+8=15 and 93+144=237 complete eight-observation trajectory queries, respectively, in a single open-loop constant-action search.

Fresh-input evaluation uses eight original parameter environments per task, 32 inputs per environment, three training seeds for MLP and TrajWorld, and three existing readout seeds where applicable. StickSlip varies absolute position within translation-equivalent tasks; PuckWall varies wall distance and normal/tangential velocity. Methods share the inputs, with models, thresholds, candidate rules, and objectives fixed beforehand. The mislocated control rotates by two bins and can also alter width; it tests interval information rather than isolating position alone.

E.2 Complete Results and Candidate-Coverage Analysis

Table 4 reports the principal search rules; Table E2 gives complete candidate coverage, success, terminal cost, and auxiliary controls.

Table E2: Complete fresh-input action-search results.

Combination	Method	Coverage (%)	Success (%)	Terminal cost	Covered but failed (%)
StickSlip / MLP	Base	50.00 ± 53.45	46.48 ± 50.54	1.6127 ± 1.4787	3.52
StickSlip / MLP	Uniform	75.00 ± 46.29	67.06 ± 43.66	0.9809 ± 0.9006	7.94
StickSlip / MLP	Local	86.33 ± 19.27	82.81 ± 18.98	0.7585 ± 0.7391	3.52
StickSlip / MLP	Reference boundary	100.00 ± 0.00	95.44 ± 10.52	0.2658 ± 0.4187	4.56
StickSlip / MLP	FCRG guidance	100.00 ± 0.00	91.93 ± 11.93	0.3973 ± 0.4760	8.07
StickSlip / MLP	DD2 guidance	77.47 ± 42.04	70.27 ± 43.59	0.9429 ± 1.1809	7.20
StickSlip / MLP	Mislocated FCRG	52.91 ± 50.92	45.70 ± 49.47	1.6644 ± 1.4902	7.20
StickSlip / Traj World	Base	50.00 ± 53.45	30.08 ± 33.51	3.7960 ± 1.1476	19.92
StickSlip / Traj World	Uniform	75.00 ± 46.29	36.85 ± 28.09	4.6036 ± 1.3237	38.15
StickSlip / Traj World	Local	77.08 ± 24.52	40.76 ± 14.62	3.6423 ± 0.8262	36.33
StickSlip / Traj World	Reference boundary	100.00 ± 0.00	53.13 ± 18.93	3.3546 ± 1.6788	46.88
StickSlip / Traj World	FCRG guidance	97.74 ± 3.80	52.43 ± 19.89	3.3794 ± 1.7306	45.31
StickSlip / Traj World	DD2 guidance	96.35 ± 6.63	51.09 ± 19.63	3.3177 ± 1.4603	45.27
StickSlip / Traj World	Mislocated FCRG	52.34 ± 50.98	26.74 ± 28.48	4.7896 ± 1.6322	25.61
PuckWall / MLP	Base	33.20 ± 10.01	32.55 ± 9.89	7.0676 ± 2.2571	0.65
PuckWall / MLP	Uniform	37.50 ± 8.52	33.98 ± 8.94	6.3514 ± 2.1766	3.52
PuckWall / MLP	Local	35.16 ± 9.99	34.11 ± 9.91	6.8687 ± 2.2315	1.04
PuckWall / MLP	Reference boundary	37.50 ± 8.52	34.11 ± 8.85	6.3564 ± 2.1761	3.39

Continued on next page

Table E2 (continued)

Combination	Method	Coverage (%)	Success (%)	Terminal cost	Covered but failed (%)
PuckWall / MLP	FCRG guidance	37.50 ± 8.52	34.11 ± 8.85	6.3747 ± 2.1751	3.39
PuckWall / MLP	DD2 guidance	37.50 ± 8.52	34.38 ± 8.75	6.3722 ± 2.1825	3.13
PuckWall / MLP	Mislocated FCRG	37.50 ± 8.52	34.59 ± 8.66	6.3692 ± 2.1726	2.91
PuckWall / Traj World	Base	33.20 ± 10.01	28.52 ± 9.20	8.7767 ± 2.6836	4.69
PuckWall / Traj World	Uniform	37.50 ± 8.52	31.25 ± 6.77	NA	6.25
PuckWall / Traj World	Local	35.03 ± 10.34	29.69 ± 10.02	8.7766 ± 2.7736	5.34
PuckWall / Traj World	Reference boundary	37.50 ± 8.52	30.86 ± 7.06	8.2666 ± 2.7789	6.64
PuckWall / Traj World	FCRG guidance	36.46 ± 8.79	29.90 ± 6.23	8.3690 ± 2.8452	6.55
PuckWall / Traj World	DD2 guidance	33.98 ± 10.08	27.82 ± 8.72	9.6742 ± 3.0798	6.16
PuckWall / Traj World	Mislocated FCRG	35.63 ± 9.65	29.64 ± 8.49	8.2341 ± 2.7915	5.99

FCRG increases successful-candidate coverage on both StickSlip predictors, but the gap between coverage and realized success differs sharply. A successful candidate is present but a failing candidate is selected in 8.07% of MLP cases and 45.31% of TrajWorld cases under FCRG guidance. Candidate generation and consequence ranking are therefore distinct bottlenecks.

The complete paired environment and input confidence intervals remain in Table E4. The actual allocation between boundary-derived candidates and uniform fill remains in Table E5, including the observation that only 15.83 and 31.72 of the 144 added PuckWall candidates originate directly from FCRG boundary intervals for MLP and TrajWorld.

Table E3: StickSlip candidate quality and selection gap. True candidate costs are evaluated offline and never enter action selection.

Combination	Method	Best true candidate cost	Selection cost minus best cost
StickSlip / MLP	Base	1.0334	0.5793
StickSlip / MLP	Uniform	0.4664	0.5144
StickSlip / MLP	Local	0.3509	0.4076
StickSlip / MLP	Reference boundary	0.0532	0.2126
StickSlip / MLP	FCRG guidance	0.0705	0.3268
StickSlip / MLP	DD2 guidance	0.3546	0.5883
StickSlip / MLP	Mislocated FCRG	0.9975	0.6669
StickSlip / TrajWorld	Base	1.0334	2.7626
StickSlip / TrajWorld	Uniform	0.4664	4.1372
StickSlip / TrajWorld	Local	0.5043	3.1379
StickSlip / TrajWorld	Reference boundary	0.0532	3.3014
StickSlip / TrajWorld	FCRG guidance	0.0904	3.2890
StickSlip / TrajWorld	DD2 guidance	0.1317	3.1860
StickSlip / TrajWorld	Mislocated FCRG	0.9930	3.7966

The lower best-candidate costs under FCRG and reference guidance in Table E3 show that interval allocation finds not only more successful sets but also actions closer to the objective. The refer-

ence arm retains a large TrajWorld selection gap as well, reinforcing that candidate generation and consequence ranking are distinct stages.

E.3 All 24 Prespecified Paired Effects

Table E4: All fresh-input paired effects. Rate differences are in percentage points and cost differences retain cost units. Intervals are individual, not simultaneous. The machine-precision PuckWall/MLP success difference is displayed as zero.

Combination	Contrast	Metric	Effect	Environment 95% CI	Input 95% CI	Better/ tied/ worse env.
StickSlip / MLP	FCRG − Local	Coverage (pp)	13.67	[3.91, 26.56]	[12.89, 14.45]	3/5/0
StickSlip / MLP	FCRG − Local	Success (pp)	9.11	[1.04, 20.31]	[7.94, 10.33]	4/4/0
StickSlip / MLP	FCRG − Local	Terminal cost	-0.3612	[-0.8196, -0.0407]	[-0.4086, -0.3144]	4/4/0
StickSlip / MLP	FCRG − Uniform	Coverage (pp)	25.00	[0.00, 62.50]	[25.00, 25.00]	2/6/0
StickSlip / MLP	FCRG − Uniform	Success (pp)	24.87	[1.30, 54.56]	[23.70, 26.04]	4/4/0
StickSlip / MLP	FCRG − Uniform	Terminal cost	-0.5835	[-1.1051, -0.1414]	[-0.6316, -0.5354]	8/0/0
StickSlip / Traj World	FCRG − Local	Coverage (pp)	20.66	[5.43, 35.77]	[17.58, 23.74]	4/4/0
StickSlip / Traj World	FCRG − Local	Success (pp)	11.68	[4.12, 18.79]	[8.29, 15.06]	7/0/1
StickSlip / Traj World	FCRG − Local	Terminal cost	-0.2629	[-1.3759, 1.0270]	[-0.7759, 0.2324]	6/0/2
StickSlip / Traj World	FCRG − Uniform	Coverage (pp)	22.74	[-3.39, 58.69]	[21.61, 23.74]	2/4/2
StickSlip / Traj World	FCRG − Uniform	Success (pp)	15.58	[-0.95, 34.81]	[11.81, 19.27]	6/0/2
StickSlip / Traj World	FCRG − Uniform	Terminal cost	-1.2242	[-2.3010, -0.1465]	[-1.8385, -0.6181]	5/0/3
PuckWall / MLP	FCRG − Local	Coverage (pp)	2.34	[0.78, 4.30]	[0.39, 4.69]	4/4/0
PuckWall / MLP	FCRG − Local	Success (pp)	0.00	[-2.08, 2.08]	[-2.21, 2.34]	3/1/4
PuckWall / MLP	FCRG − Local	Terminal cost	-0.4939	[-0.6171, -0.3698]	[-0.6630, -0.3392]	8/0/0
PuckWall / MLP	FCRG − Uniform	Coverage (pp)	0.00	[0.00, 0.00]	[0.00, 0.00]	0/8/0
PuckWall / MLP	FCRG − Uniform	Success (pp)	0.13	[0.00, 0.39]	[0.00, 0.39]	1/7/0
PuckWall / MLP	FCRG − Uniform	Terminal cost	0.0234	[0.0062, 0.0440]	[0.0038, 0.0471]	1/0/7
PuckWall / Traj World	FCRG − Local	Coverage (pp)	1.43	[-0.43, 3.26]	[-0.52, 3.47]	5/1/2
PuckWall / Traj World	FCRG − Local	Success (pp)	0.22	[-2.73, 3.30]	[-1.74, 2.17]	4/0/4
PuckWall / Traj World	FCRG − Local	Terminal cost	-0.4076	[-0.8725, -0.0185]	[-0.8286, 0.0156]	6/0/2
PuckWall / Traj World	FCRG − Uniform	Coverage (pp)	-1.04	[-1.82, -0.30]	[-2.08, -0.26]	0/3/5

Continued on next page

Table E4 (continued)

Combination	Contrast	Metric	Effect	Environment 95% CI	Input 95% CI	Better/ tied/ worse env.
PuckWall / Traj World	FCRG − Uniform	Success (pp)	-1.35	[-2.95, 0.39]	[-2.95, 0.17]	2/1/5
PuckWall / Traj World	FCRG − Uniform	Terminal cost	NA	NA	NA	2/0/5

On StickSlip, FCRG–Local success effects remain positive for both predictors under both interval constructions, supporting the fixed rule on unseen positions. The full table also shows where effects are less clear: TrajWorld's environment-level success interval against Uniform and its cost intervals against Local cross zero. The primary positive finding is improved success and candidate coverage under equal-query search, not universal improvement of every outcome.

PuckWall/TrajWorld FCRG–Uniform terminal cost has an undefined original estimate. Although 3,455 environment and 3,607 input draws happen to omit the unknown entry, their finite-draw quantiles are not confidence intervals for the full contrast; Table E4 retains NA. The other 23 effects have 10,000 defined draws under both constructions.

E.4 How Interval Information Changes Candidate Allocation

Table E5: Boundary proposals and uniform-fill allocation.

Combination	Method	Boundary-added candidates	Uniform-fill candidates	Readout rays
StickSlip / MLP	Reference boundary	7.50	0.50	0
StickSlip / MLP	FCRG guidance	7.67	0.33	2
StickSlip / MLP	DD2 guidance	7.08	0.92	2
StickSlip / MLP	Mislocated FCRG	7.96	0.04	2
StickSlip / TrajWorld	Reference boundary	7.50	0.50	0
StickSlip / TrajWorld	FCRG guidance	7.38	0.62	2
StickSlip / TrajWorld	DD2 guidance	7.77	0.23	2
StickSlip / TrajWorld	Mislocated FCRG	7.94	0.06	2
PuckWall / MLP	Reference boundary	8.46	135.54	0
PuckWall / MLP	FCRG guidance	15.83	128.17	36
PuckWall / MLP	DD2 guidance	18.95	125.05	36
PuckWall / MLP	Mislocated FCRG	29.89	114.11	36
PuckWall / TrajWorld	Reference boundary	8.46	135.54	0
PuckWall / TrajWorld	FCRG guidance	31.72	112.28	36
PuckWall / TrajWorld	DD2 guidance	110.13	33.87	36
PuckWall / TrajWorld	Mislocated FCRG	50.85	93.15	36

Most StickSlip additions originate directly from boundary intervals, whereas FCRG relies more heavily on uniform fill in PuckWall. This implementation difference accompanies the difference in task benefit, motivating interpretation through the actual candidate set rather than a method label. DD2 allocates more boundary candidates in some PuckWall combinations without a corresponding success advantage, so candidate count alone is not an explanation of utility.

The practical conclusion is that first intervals can guide limited query allocation, but should be evaluated together with task objectives, the action space, and model-based ranking. Coverage–selection analysis identifies where this use of information already contributes and where subsequent stages limit the realized gain.

E.5 Development Results and Fresh-Input Evaluation

Table E6: Development and fresh-input results reported separately.

Combination	Method	Development success (%)	Fresh success (%)	Development cost	Fresh cost
StickSlip / MLP	FCRG guidance	89.93	91.93	0.4698	0.3973
StickSlip / MLP	DD2 guidance	69.10	70.27	0.9463	0.9429
StickSlip / MLP	Base	45.83	46.48	1.6316	1.6127
StickSlip / MLP	Local	82.29	82.81	0.7746	0.7585
StickSlip / MLP	Mislocated FCRG	46.88	45.70	1.6191	1.6644
StickSlip / MLP	Reference boundary	93.75	95.44	0.3289	0.2658
StickSlip / MLP	Uniform	67.71	67.06	0.9681	0.9809
PuckWall / MLP	FCRG guidance	46.88	34.11	2.7974	6.3747
PuckWall / MLP	DD2 guidance	46.88	34.38	2.8174	6.3722
PuckWall / MLP	Base	42.71	32.55	3.1961	7.0676
PuckWall / MLP	Local	45.83	34.11	3.0405	6.8687
PuckWall / MLP	Mislocated FCRG	48.61	34.59	2.7781	6.3692
PuckWall / MLP	Reference boundary	46.88	34.11	2.7955	6.3564
PuckWall / MLP	Uniform	47.92	33.98	2.8005	6.3514
PuckWall / Traj World	FCRG guidance	35.42	29.90	4.4010	8.3690
PuckWall / Traj World	DD2 guidance	34.38	27.82	4.4774	9.6742
PuckWall / Traj World	Base	32.29	28.52	4.3157	8.7767
PuckWall / Traj World	Local	37.50	29.69	4.1833	8.7766
PuckWall / Traj World	Mislocated FCRG	29.51	29.64	4.2991	8.2341
PuckWall / Traj World	Reference boundary	42.71	30.86	4.0492	8.2666
PuckWall / Traj World	Uniform	41.67	31.25	4.4471	NA
StickSlip / Traj World	FCRG guidance	50.00	52.43	2.8463	3.3794
StickSlip / Traj World	DD2 guidance	44.10	51.09	3.2695	3.3177
StickSlip / Traj World	Base	25.00	30.08	4.5956	3.7960
StickSlip / Traj World	Local	38.54	40.76	5.0273	3.6423

Continued on next page

Table E6 (continued)

Combination	Method	Development success (%)	Fresh success (%)	Development cost	Fresh cost
StickSlip / Traj World	Mislocated FCRG	20.49	26.74	5.7409	4.7896
StickSlip / Traj World	Reference boundary	51.04	53.13	2.7600	3.3546
StickSlip / Traj World	Uniform	33.33	36.85	5.7780	4.6036

Table E6 retains all seven rules and shows the persistence of the StickSlip mean benefit without pooling the two datasets for inference. StickSlip changes absolute position, whereas PuckWall changes wall-distance and velocity conditions; their new-input difficulty can therefore differ. Between-batch score changes are not interpreted as training improvement or deterioration.

E.6 Why Boundary Information Is Used for Search Rather Than Exclusion

Table E7: Development results for the earlier bin-exclusion rule. The earlier 64-input, five-rule study used a different candidate policy and remains a separate design comparison.

Combination	Base (%)	Reference (%)	FCRG (%)	Shuffled (%)	Fixed margin (%)
StickSlip / MLP	45.83	0.00	0.00	0.00	45.83
StickSlip / TrajWorld	25.00	0.00	0.69	1.39	25.00
PuckWall / MLP	42.71	42.71	42.01	41.32	21.88
PuckWall / TrajWorld	32.29	32.29	30.21	30.90	22.92

Table E7 evaluates an earlier rule that excluded the entire first-boundary bin and its neighborhood. On StickSlip, even reference boundaries removed actions that had succeeded under Base and reduced success to zero. The lesson concerns use, not measurability: physical mode changes are not task failures. Retaining candidates and using boundaries to decide where to refine responds to this mismatch, and the subsequent fresh-input experiment evaluates that fixed allocation rule.

E.7 Input Support and Unknown Outcomes

The 39,936 selection records are repeated evaluations of 256 inputs per task, not additional scenes created by seeds or methods. The 32 StickSlip positions per environment share a translation-equivalent physical task; after removing along-wall translation, PuckWall retains 32 distinct distance/velocity inputs per environment.

There are 107 invalid candidate-execution records associated with one PuckWall input across repeated blocks; one is selected by TrajWorld/Uniform. It counts as a success failure and has unknown terminal cost. Best-candidate cost is reported only when all candidate costs are known, and coverage counts verified successful actions. These rules preserve the shared input support rather than changing the evaluation target according to numerical availability.

F Robustness and Measurement Resolution

F.1 Readouts under Matched Noise Augmentation

Boundary measurement should respond to structural changes while remaining stable under irrelevant observation changes. Shared development responses are perturbed at σ=0.01 using three noise types and three realizations each. Table F1 compares augmentation-matched FCRG, Raw, MLP32,

and DD2-noise. Original system estimates are averaged equally over seven event systems; SmoothDuffing is separate.

Table F1: Main readout controls under matched noise augmentation.

Readout	**AP** ↑	**Joint NLL** ↓	**Exact-bin (%)** ↑	**Recall (%)** ↑	**FPR (%)** ↓	**nMAE** ↓
FCRG	0.9096	0.3386	70.76	90.09	9.56	0.1991
Raw	0.8767	0.4263	65.98	90.95	16.78	0.2040
MLP32	0.9078	0.3511	65.87	90.51	8.54	0.1991
DD2-noise	0.7204	0.7310	68.89	97.53	51.95	0.1910

FCRG improves interval accuracy and joint NLL over Raw under matched augmentation. DD2-noise retains more positive alarms but incurs substantially more false alarms, showing why robustness cannot be represented by recall alone. Complete system-level NLL, Brier, and interval-NLL results remain in Tables F2, F3, and F4.

Table F2: Per-system FCRG joint NLL and environment variation. Original mean ± environment SD, computed after within-environment averaging of noise realizations and seeds (ddof=1). Counts are metric-specific; small nonzero values retain scientific notation.

System	**Clean mean ± SD**	**Defined env.**	**Noisy mean ± SD**	**Defined env.**
PiecewiseAffine	0.3505 ± 0.1321	8	0.7232 ± 0.1693	8
HardStop1D	0.3601 ± 0.2464	8	0.4481 ± 0.2644	8
SaturationKink	0.0095 ± 0.0060	8	0.4899 ± 0.1046	8
BouncingBall	0.2830 ± 0.2474	8	0.2921 ± 0.2539	8
StickSlipBlock	0.0126 ± 0.0356	8	0.0730 ± 0.1293	8
PuckWall2D	0.0970 ± 0.0493	8	0.1251 ± 0.0570	8
TwoBody Collision	0.1283 ± 0.0760	8	0.2188 ± 0.0985	8
SmoothDuffing	$5.26 \times 10^{-10} \pm 8.50 \times 10^{-10}$	8	$1.74 \times 10^{-11} \pm 2.23 \times 10^{-11}$	8

Table F3: Per-system FCRG joint Brier and environment variation. Original mean ± environment SD, computed after within-environment averaging of noise realizations and seeds (ddof=1). Counts are metric-specific; small nonzero values retain scientific notation.

System	**Clean mean ± SD**	**Defined env.**	**Noisy mean ± SD**	**Defined env.**
PiecewiseAffine	0.1948 ± 0.0722	8	0.3637 ± 0.0872	8
HardStop1D	0.1786 ± 0.1316	8	0.2117 ± 0.1434	8
SaturationKink	0.0028 ± 0.0049	8	0.2664 ± 0.0648	8
BouncingBall	0.1110 ± 0.0926	8	0.1159 ± 0.0984	8
StickSlipBlock	0.0048 ± 0.0136	8	0.0241 ± 0.0365	8
PuckWall2D	0.0456 ± 0.0189	8	0.0540 ± 0.0222	8
TwoBody Collision	0.0554 ± 0.0366	8	0.0827 ± 0.0469	8
SmoothDuffing	$6.41 \times 10^{-17} \pm 1.29 \times 10^{-16}$	8	$7.56 \times 10^{-20} \pm 1.16 \times 10^{-19}$	8

Table F4: Per-system FCRG interval NLL and environment variation. Original mean ± environment SD, computed after within-environment averaging of noise realizations and seeds (ddof=1). Counts are metric-specific; small nonzero values retain scientific notation.

System	Clean mean ± SD	Defined env.	Noisy mean ± SD	Defined env.
PiecewiseAffine	0.5229 ± 0.1997	8	1.1671 ± 0.2445	8
HardStop1D	1.2023 ± 0.3906	8	1.4188 ± 0.3906	8
SaturationKink	0.0085 ± 0.0087	8	0.4684 ± 0.0963	8
BouncingBall	0.8890 ± 0.6829	8	0.9175 ± 0.7026	8
StickSlipBlock	0.0151 ± 0.0426	8	0.0866 ± 0.1303	8
PuckWall2D	1.0629 ± 0.4451	8	1.1858 ± 0.4352	8
TwoBodyCollision	1.5374 ± 1.3862	7	2.0602 ± 1.2570	7
SmoothDuffing	NA	0	NA	0

System-level statistics show comparatively little change in BouncingBall joint NLL at this noise level, whereas SaturationKink and PiecewiseAffine are more sensitive. These differences identify affected response structures instead of concealing environment heterogeneity in a macro average. Per-system environment SDs, seed SDs, and all 21 metrics for the 15 versions are retained in G.3–G.4 and the source table.

F.2 Boundary-Preserving Amplitude Transformations

Table F5: Native and half-amplitude readouts at all five distances. Distance 0.95 lies outside the 0.8 scan and is censored; event-conditional metrics are NA. Metric panel 1 of 2; identifying columns are repeated.

Readout	Distance	Native recall (%)	Half recall (%)	Native Exact (%)	Half Exact (%)
FCRG	0.05	93.75	100.00	18.75	6.25
FCRG	0.15	100.00	100.00	40.63	5.21
FCRG	0.3	100.00	100.00	100.00	100.00
FCRG	0.6	100.00	100.00	100.00	100.00
FCRG	0.95	NA	NA	NA	NA
FCRG-clean	0.05	72.92	98.96	23.96	9.38
FCRG-clean	0.15	94.79	95.83	88.54	81.25
FCRG-clean	0.3	100.00	100.00	100.00	100.00
FCRG-clean	0.6	100.00	95.83	100.00	95.83
FCRG-clean	0.95	NA	NA	NA	NA
DD2-noise	0.05	100.00	100.00	6.25	12.50
DD2-noise	0.15	100.00	100.00	3.13	3.13
DD2-noise	0.3	100.00	88.54	18.75	9.38
DD2-noise	0.6	37.50	25.00	37.50	25.00
DD2-noise	0.95	NA	NA	NA	NA

Table F5: Native and half-amplitude readouts at all five distances. Distance 0.95 lies outside the 0.8 scan and is censored; event-conditional metrics are NA. Metric panel 2 of 2; identifying columns are repeated.

Readout	Distance	Native FPR (%)	Half FPR (%)
FCRG	0.05	NA	NA
FCRG	0.15	NA	NA
FCRG	0.3	NA	NA

Continued on next page

Table F5 (continued)

Readout	Distance	Native FPR (%)	Half FPR (%)
FCRG	0.6	NA	NA
FCRG	0.95	0.00	0.00
FCRG-clean	0.05	NA	NA
FCRG-clean	0.15	NA	NA
FCRG-clean	0.3	NA	NA
FCRG-clean	0.6	NA	NA
FCRG-clean	0.95	0.00	0.00
DD2-noise	0.05	NA	NA
DD2-noise	0.15	NA	NA
DD2-noise	0.3	NA	NA
DD2-noise	0.6	NA	NA
DD2-noise	0.95	0.00	0.00

Table F5 compares native and half-amplitude responses; the transformation preserves the generator boundary. Performance at 0.3 and 0.6 is unchanged, whereas first-interval assignment deteriorates for the two nearest boundaries. This identifies a measurement sensitivity rather than a structural change in the target.

F.3 Smooth Negative Control

Across 512 fresh SmoothDuffing negative rays, FCRG produces no observed false alarms for any of the six world-model outputs. The full probability scores and alternative-readout results remain in Table F6. Together with deletion, displacement, and amplitude-preserving interventions, this control separates sensitivity to structural change from sensitivity to arbitrary trajectory error.

Table F6: All readouts on 512 fresh smooth negative rays.

Model	Readout	Scaled MSE	FPR (%)	Joint NLL	Binary Brier
MLP	FCRG	7.65×10^{-8}	0.00	1.75×10^{-8}	7.10×10^{-14}
MLP	FCRG-clean	7.65×10^{-8}	0.00	2.26×10^{-6}	2.18×10^{-9}
MLP	DD2-noise	7.65×10^{-8}	0.00	0.0128	0.0002
Ensemble	FCRG	5.26×10^{-8}	0.00	1.66×10^{-8}	6.57×10^{-14}
Ensemble	FCRG-clean	5.26×10^{-8}	0.00	5.97×10^{-7}	6.65×10^{-11}
Ensemble	DD2-noise	5.26×10^{-8}	0.00	0.0154	0.0003
DreamerV3	FCRG	0.1417	0.00	1.35×10^{-9}	1.16×10^{-15}
DreamerV3	FCRG-clean	0.1417	1.35	0.0380	0.0115
DreamerV3	DD2-noise	0.1417	0.07	0.0081	0.0007
TrajWorld	FCRG	0.0196	0.00	1.09×10^{-11}	4.70×10^{-20}
TrajWorld	FCRG-clean	0.0196	2.60	0.0533	0.0176
TrajWorld	DD2-noise	0.0196	0.30	0.0379	0.0029
WIMLE	FCRG	1.16×10^{-7}	0.00	1.71×10^{-8}	7.23×10^{-14}
WIMLE	FCRG-clean	1.16×10^{-7}	0.00	0.0001	1.17×10^{-6}
WIMLE	DD2-noise	1.16×10^{-7}	0.00	0.0047	6.82×10^{-5}
ADM-v2	FCRG	2.30×10^{-5}	0.00	3.03×10^{-9}	2.39×10^{-15}
ADM-v2	FCRG-clean	2.30×10^{-5}	0.00	0.0008	5.98×10^{-5}
ADM-v2	DD2-noise	2.30×10^{-5}	3.34	0.2208	0.0298

F.4 Observation Scope and Cause Support

Supported cause classes can explain interval outcomes, but coverage must be distinguished from accuracy. In the archived in-distribution StickSlip check, conditional start/stop cause accuracies

were approximately 100.00%/99.35%, with no in-distribution reverse support. A separate out-of-distribution physical stop↔reverse replacement was not detected under either anchor. These results distinguish supported cause readout from unsupported event changes; conditional accuracy is not general physical-type recognition.

The main fine-grained evidence therefore concerns ray-level presence, first intervals, and error decomposition, while joint cause modeling supplies probability information within supported labels. The witnesses in A.3, known interventions in D.1, and perturbation analyses here define a concrete interpretive scope for studying model and measurement differences separately.

G Module Contributions and Internal Representations

G.1 Module-Level Ablation Operations

Table G1: Main module comparisons and actual interventions.

Configuration	LCR-related information	SAE	REF	FBR
FCRG	Retained	Retained	Retained	Retained
w/o direct LCR vector	Direct I input removed; residual magnitudes retained for SAE	Retained	Retained	Retained
w/o SAE	Retained	Envelope and excess jointly removed	Retained	Retained

Module names must match the information actually removed. Without the direct I vector, SAE can still use LCR residual magnitudes; this tests the vector's incremental contribution beyond scale-aware evidence, not removal of all LCR computation. Removing SAE eliminates envelope and excess together. REF and FBR remain fixed in this comparison. Table G1 specifies the module interventions summarized in Table 5.

G.2 Complete Internal Feature Variants

Table G2: Internal feature comparisons under the final training recipe. Common development responses, mixed-noise training, and prescribed calibrations. Results are equal-system macro means; Exact-bin and FPR are percentages.

Configuration	Clean NLL	Clean Exact (%)	Clean FPR (%)	Noisy NLL	Noisy Exact (%)	Noisy FPR (%)
FCRG	0.1773	81.65	5.03	0.3386	70.76	9.56
w/o direct LCR-vector input	0.1799	81.01	4.93	0.3481	70.31	9.40
w/o SAE	0.3238	69.55	12.17	0.3308	68.90	12.43
w/o excess	0.3278	69.00	12.32	0.3344	68.43	12.48
w/o I vector and excess	0.4162	65.74	15.79	0.4164	65.85	15.80
Unit-reference log intensity	0.1748	81.58	5.53	0.3254	70.68	9.35
w/o direct envelope input	0.1744	82.38	5.65	0.3357	70.70	10.61

Table G2 shows that removing SAE or excess substantially lowers clean interval agreement, while direct-vector removal remains close to the full model. This identifies an increment from scale-organized change evidence rather than simply adding a vector input. Joint vector/excess removal still retains envelope and is not simultaneous removal of LCR and SAE.

Internal reference variants separate information content from implementation. Unit-reference log intensity gives clean Exact-bin 81.58% and noisy NLL 0.3254; because it retains logarithmic residual intensity, it is not SAE removal. Removing direct envelope input gives clean Exact-bin 82.38%, but clean/noisy FPRs rise to 5.65%/10.61%, compared with 5.03%/9.56% for FCRG. The results support structured scale information without identifying one reference form as uniformly optimal. Main-text comparisons remain at the module level; internal variants explain implementation trade-offs.

G.3 All 15 Versions, Aggregation, and Statistical Support

The per-system tables in Appendix G.4 summarize 5,040 metric records spanning 15 versions, eight systems, two observation conditions, and 21 metrics, with environment SDs and defined counts. These results use the historical development-validation set, separate from the fresh-input results in Appendix D. Table G3 distinguishes recipe and baseline controls from module ablations.

Table G3: Source versions and comparison targets. Version keys index the full matrices. A is not Raw, E is not a separate module, and E reuses the equivalent C record on SmoothDuffing.

Source key	Name	Comparison level	Operation
`C`	FCRG	Full method	Final mixed-noise-trained FCRG
`A`	FCRG-clean	Training-recipe control	Original full clean-trained FCRG, not Raw
`E`	Presence-weighted FCRG (E)	Training-objective control	Historical presence-weighted recipe, not a structural module
`raw_clean`	Raw-clean	Input/training control	Raw-response readout, clean training
`raw_noise`	Raw-noise	Input/training control	Raw-response readout, matched noise augmentation
`mlp32_clean`	MLP32-clean	Capacity/training control	32-hidden-unit raw-response readout, clean training
`mlp32_noise`	MLP32-noise	Capacity/training control	32-hidden-unit raw-response readout, matched noise augmentation
`dd2_clean`	DD2-clean	Difference-feature/ training control	Second-order differences with a linear readout, clean training
`dd2_noise`	DD2-noise	Difference-feature/ training control	Second-order differences with a linear readout, noise training
`no_i_vector`	w/o direct LCR-vector input	Main input ablation	Remove only direct I; SAE retains residual magnitudes
`innovation_only`	w/o SAE	Main module ablation	Remove joint envelope/ excess features; retain raw responses
`no_excess`	w/o excess	Internal SAE variant	Remove excess, retain envelope
`no_i_vector_no_excess`	w/o I vector and excess	Internal combined variant	Remove direct I and excess, retain envelope
`no_train_envelope`	Unit-reference log intensity	SAE reference variant	Unit-reference log intensity; not full SAE removal

Continued on next page

Table G3 (continued)

Source key	Name	Comparison level	Operation
no_envelope_input	w/o direct envelope input	SAE input variant	Remove direct envelope input; retain its role in the excess denominator

The macro estimator uses each system's historical_mean: for system estimate μ_s, $\bar{m} = \frac{1}{7}\sum_{s=1}^{7}\mu_s$, rather than pooling rays. Environment SD is computed within systems after averaging seeds and noise realizations in each environment (ddof=1). Event metrics such as Exact-bin usually have 55 defined environments, while joint NLL and FPR have 56; one TwoBodyCollision environment lacks event support. Equal-system means therefore differ from pooling all defined environments. Table G4 gives the clean macro means; its noisy-response panel reports the corresponding perturbed-response means.

Environment SD and seed SD describe heterogeneity and training-repeat variation, not macro-mean confidence intervals. The table does not contain paired environment-level scores across methods or cross-system seed covariances. Original SDs are retained without averaging them into a macro SD or creating new paired p-values or bootstrap intervals from marginal summaries. Per-system mean $\pm$ environment SD appears in G.4; seed SDs and counts remain in the source CSV.

Table G4: Clean responses. Seven-event-system macro means for all 15 versions. Clean and noisy conditions are separate. Recipe controls and feature ablations do not form a single architecture ranking. AP is on a 0–1 scale; detection and hit rates are percentages.

Source key	AP	Joint NLL	Exact (%)	Recall (%)	FPR (%)	nMAE
C	0.9350	0.1773	81.65	94.60	5.03	0.1512
A	0.9369	0.2462	78.42	89.09	4.32	0.1829
E	0.9329	0.1768	81.21	94.17	5.31	0.1551
raw_clean	0.8766	0.4260	66.07	90.91	16.79	0.2033
raw_noise	0.8764	0.4261	66.04	90.91	16.75	0.2034
mlp32_clean	0.9075	0.3500	65.95	90.45	8.36	0.1990
mlp32_noise	0.9074	0.3498	66.00	90.56	8.46	0.1988
dd2_clean	0.7571	0.4720	63.63	86.66	12.77	0.2282
dd2_noise	0.7602	0.4770	63.80	87.70	12.82	0.2336
no_i_vector	0.9429	0.1799	81.01	93.94	4.93	0.1528
innovation_only	0.9114	0.3238	69.55	90.56	12.17	0.2002
no_excess	0.9067	0.3278	69.00	90.39	12.32	0.2031
no_i_vector_no_excess	0.8724	0.4162	65.74	90.97	15.79	0.2017
no_train_envelope	0.9382	0.1748	81.58	94.98	5.53	0.1466
no_envelope_input	0.9372	0.1744	82.38	95.42	5.65	0.1450

Table G4: Noisy responses. Seven-event-system macro means for all 15 versions. Clean and noisy conditions are separate. Recipe controls and feature ablations do not form a single architecture ranking. AP is on a 0–1 scale; detection and hit rates are percentages.

Source key	AP	Joint NLL	Exact (%)	Recall (%)	FPR (%)	nMAE
C	0.9096	0.3386	70.76	90.09	9.56	0.1991
A	0.8967	0.8189	51.96	92.87	21.73	0.2261

Continued on next page

Table G4 (continued)

Source key	AP	Joint NLL	Exact (%)	Recall (%)	FPR (%)	nMAE
E	0.9073	0.3417	70.36	89.62	10.48	0.2022
raw_clean	0.8767	0.4262	66.03	90.96	16.80	0.2038
raw_noise	0.8767	0.4263	65.98	90.95	16.78	0.2040
mlp32_clean	0.9079	0.3516	65.90	90.47	8.48	0.1992
mlp32_noise	0.9078	0.3511	65.87	90.51	8.54	0.1991
dd2_clean	0.7163	0.8945	64.71	97.83	52.08	0.1975
dd2_noise	0.7204	0.7310	68.89	97.53	51.95	0.1910
no_i_vector	0.9072	0.3481	70.31	89.33	9.40	0.2024
innovation_only	0.9102	0.3308	68.90	90.31	12.43	0.2009
no_excess	0.9054	0.3344	68.43	90.15	12.48	0.2041
no_i_vector_no_excess	0.8729	0.4164	65.85	90.89	15.80	0.2021
no_train_envelope	0.9232	0.3254	70.68	91.25	9.35	0.1946
no_envelope_input	0.9219	0.3357	70.70	91.23	10.61	0.1950

Clean/noisy training controls separate adaptation from representation. FCRG-clean has higher noisy NLL and FPR than C, while the Raw and MLP32 training versions show smaller differences. This supports adaptation of the full recipe without assigning the A–C difference to one component. Complete E and other variant results are retained without replacing the main model based on selected favorable cells.

G.4 Complete Per-System Metrics and Environment Standard Deviations

All 5,040 rows are represented below as `historical_mean ± environment_sd`. Detection, hit, and directional error rates are percentages; their means and environment SDs are both multiplied by 100. AP, nAP, and calibration scores retain their defined scales. Eight defined environments are implicit; labels such as `[n=7]` denote other supports. `noise` is the prescribed three-type average at σ=0.01. Definitions follow B.2; Late / miss includes misses and is not accident risk, and False boundary maps to hallucination_rate. Macro means appear in Table G4 and its noisy-response panel.

NA follows the historical estimator rather than a finite cell-first subset. For example, noisy Smooth-Duffing/A precision remains NA even though a separate two-environment cell-first estimate is zero. Small probability losses use scientific notation rather than exact zeros.

G.4.1 PiecewiseAffine

Tables G5.1.1, G5.1.2, G5.1.3, and G5.1.4 report presence ranking and calibration, probability scores, fixed-threshold outcomes, and distance/direction metrics, respectively, for all versions and both observation conditions.

Table G5.1.1: Presence ranking and calibration.

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
C	clean	0.9806 ± 0.0155	0.9628 ± 0.0296	0.0507 ± 0.0318	0.0459 ± 0.0225	0.0038 ± 0.0336
C	noise	0.9439 ± 0.0372	0.8914 ± 0.0734	0.0983 ± 0.0351	0.0716 ± 0.0160	0.0233 ± 0.0352

Continued on next page

Table G5.1.1 (continued)

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
A	clean	0.9773 ± 0.0253	0.9566 ± 0.0488	0.0504 ± 0.0251	0.0798 ± 0.0127	-0.0207 ± 0.0333
A	noise	0.9070 ± 0.0541	0.8205 ± 0.1064	0.1876 ± 0.0254	0.2270 ± 0.0373	0.1792 ± 0.0466
E	clean	0.9798 ± 0.0159	0.9615 ± 0.0295	0.0470 ± 0.0292	0.0401 ± 0.0212	0.0088 ± 0.0346
E	noise	0.9461 ± 0.0384	0.8958 ± 0.0754	0.0991 ± 0.0355	0.0828 ± 0.0190	0.0441 ± 0.0341
raw_clean	clean	0.9267 ± 0.0516	0.8580 ± 0.1018	0.1121 ± 0.0419	0.0815 ± 0.0308	-0.0009 ± 0.0387
raw_clean	noise	0.9266 ± 0.0518	0.8579 ± 0.1023	0.1121 ± 0.0419	0.0804 ± 0.0305	-0.0011 ± 0.0388
raw_noise	clean	0.9267 ± 0.0516	0.8581 ± 0.1018	0.1121 ± 0.0419	0.0815 ± 0.0308	-0.0011 ± 0.0387
raw_noise	noise	0.9265 ± 0.0519	0.8578 ± 0.1023	0.1121 ± 0.0419	0.0799 ± 0.0301	-0.0013 ± 0.0388
mlp32_clean	clean	0.9484 ± 0.0345	0.8991 ± 0.0682	0.0929 ± 0.0358	0.0651 ± 0.0132	-0.0048 ± 0.0361
mlp32_clean	noise	0.9481 ± 0.0347	0.8987 ± 0.0687	0.0934 ± 0.0365	0.0654 ± 0.0122	-0.0052 ± 0.0361
mlp32_noise	clean	0.9477 ± 0.0356	0.8979 ± 0.0703	0.0939 ± 0.0362	0.0674 ± 0.0105	-0.0062 ± 0.0367
mlp32_noise	noise	0.9474 ± 0.0356	0.8975 ± 0.0704	0.0943 ± 0.0368	0.0648 ± 0.0110	-0.0066 ± 0.0367
dd2_clean	clean	0.9362 ± 0.0525	0.8793 ± 0.0935	0.0940 ± 0.0221	0.1660 ± 0.0167	0.0079 ± 0.0356
dd2_clean	noise	0.8217 ± 0.0763	0.6631 ± 0.1219	0.3312 ± 0.0432	0.3703 ± 0.0576	0.3668 ± 0.0588
dd2_noise	clean	0.9367 ± 0.0409	0.8789 ± 0.0731	0.1208 ± 0.0244	0.1868 ± 0.0260	-0.0747 ± 0.0395
dd2_noise	noise	0.8339 ± 0.0663	0.6859 ± 0.1051	0.2068 ± 0.0214	0.2212 ± 0.0421	0.1770 ± 0.0565
no_i_vector	clean	0.9780 ± 0.0185	0.9578 ± 0.0354	0.0532 ± 0.0331	0.0471 ± 0.0233	0.0032 ± 0.0338
no_i_vector	noise	0.9342 ± 0.0440	0.8729 ± 0.0868	0.1076 ± 0.0403	0.0724 ± 0.0172	0.0278 ± 0.0354

Continued on next page

Table G5.1.1 (continued)

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
innovation_only	clean	0.9706 ± 0.0219	0.9429 ± 0.0434	0.0698 ± 0.0263	0.0586 ± 0.0149	-0.0083 ± 0.0284
innovation_only	noise	0.9662 ± 0.0227	0.9343 ± 0.0450	0.0750 ± 0.0274	0.0567 ± 0.0100	-0.0092 ± 0.0294
no_excess	clean	0.9690 ± 0.0227	0.9397 ± 0.0450	0.0715 ± 0.0268	0.0582 ± 0.0135	-0.0077 ± 0.0289
no_excess	noise	0.9650 ± 0.0234	0.9320 ± 0.0463	0.0763 ± 0.0278	0.0562 ± 0.0105	-0.0086 ± 0.0298
no_i_vector_no_excess	clean	0.9267 ± 0.0520	0.8581 ± 0.1025	0.1119 ± 0.0419	0.0788 ± 0.0273	-0.0006 ± 0.0386
no_i_vector_no_excess	noise	0.9267 ± 0.0522	0.8580 ± 0.1029	0.1120 ± 0.0420	0.0796 ± 0.0297	-0.0009 ± 0.0387
no_train_envelope	clean	0.9809 ± 0.0154	0.9634 ± 0.0296	0.0511 ± 0.0317	0.0465 ± 0.0212	0.0029 ± 0.0341
no_train_envelope	noise	0.9437 ± 0.0364	0.8909 ± 0.0720	0.0983 ± 0.0349	0.0716 ± 0.0164	0.0214 ± 0.0350
no_envelope_input	clean	0.9807 ± 0.0156	0.9630 ± 0.0301	0.0514 ± 0.0319	0.0443 ± 0.0194	0.0031 ± 0.0343
no_envelope_input	noise	0.9427 ± 0.0367	0.8891 ± 0.0726	0.0996 ± 0.0352	0.0716 ± 0.0163	0.0239 ± 0.0352

Table G5.1.2: Joint and interval probabilities.

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
C	clean	0.3505 ± 0.1321	0.1948 ± 0.0722	0.5229 ± 0.1997	0.1832 ± 0.1350
C	noise	0.7232 ± 0.1693	0.3637 ± 0.0872	1.1671 ± 0.2445	0.3102 ± 0.1065
A	clean	0.3869 ± 0.0879	0.1985 ± 0.0540	0.6172 ± 0.1009	0.1727 ± 0.1081
A	noise	1.1111 ± 0.0912	0.5578 ± 0.0469	1.4329 ± 0.1302	0.8196 ± 0.0928
E	clean	0.3429 ± 0.1286	0.1874 ± 0.0677	0.4956 ± 0.1893	0.1937 ± 0.1453
E	noise	0.7265 ± 0.1636	0.3657 ± 0.0859	1.1314 ± 0.2286	0.3501 ± 0.1133
raw_clean	clean	0.8122 ± 0.1932	0.3946 ± 0.0904	1.3694 ± 0.2859	0.2949 ± 0.0967
raw_clean	noise	0.8128 ± 0.1928	0.3950 ± 0.0904	1.3711 ± 0.2855	0.2946 ± 0.0975

Continued on next page

Table G5.1.2 (continued)

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
raw_noise	clean	0.8125 ± 0.1932	0.3947 ± 0.0904	1.3703 ± 0.2857	0.2947 ± 0.0967
raw_noise	noise	0.8130 ± 0.1928	0.3952 ± 0.0904	1.3719 ± 0.2853	0.2944 ± 0.0975
mlp32_clean	clean	0.7486 ± 0.2233	0.3606 ± 0.0879	1.3006 ± 0.3658	0.2343 ± 0.0743
mlp32_clean	noise	0.7563 ± 0.2283	0.3644 ± 0.0906	1.3166 ± 0.3748	0.2351 ± 0.0776
mlp32_noise	clean	0.7474 ± 0.2173	0.3628 ± 0.0877	1.2971 ± 0.3531	0.2353 ± 0.0750
mlp32_noise	noise	0.7532 ± 0.2210	0.3657 ± 0.0897	1.3093 ± 0.3599	0.2357 ± 0.0779
dd2_clean	clean	0.7578 ± 0.1176	0.3907 ± 0.0641	1.2238 ± 0.1397	0.3276 ± 0.0829
dd2_clean	noise	1.5567 ± 0.0968	0.7418 ± 0.0312	1.2395 ± 0.0881	1.8356 ± 0.1355
dd2_noise	clean	0.8305 ± 0.1308	0.4327 ± 0.0714	1.4794 ± 0.1771	0.2334 ± 0.0425
dd2_noise	noise	1.1210 ± 0.0752	0.5616 ± 0.0405	1.3572 ± 0.1251	0.8980 ± 0.0805
no_i_vector	clean	0.3866 ± 0.1431	0.2113 ± 0.0776	0.5841 ± 0.2002	0.1960 ± 0.1460
no_i_vector	noise	0.7743 ± 0.1811	0.3846 ± 0.0930	1.2453 ± 0.2571	0.3363 ± 0.1141
innovation_only	clean	0.6137 ± 0.1557	0.3149 ± 0.0789	1.0752 ± 0.2481	0.1838 ± 0.0701
innovation_only	noise	0.6494 ± 0.1607	0.3350 ± 0.0813	1.1398 ± 0.2557	0.1936 ± 0.0749
no_excess	clean	0.6212 ± 0.1570	0.3179 ± 0.0792	1.0848 ± 0.2497	0.1893 ± 0.0721
no_excess	noise	0.6542 ± 0.1618	0.3367 ± 0.0817	1.1450 ± 0.2572	0.1982 ± 0.0767
no_i_vector_no_excess	clean	0.8117 ± 0.1935	0.3946 ± 0.0906	1.3679 ± 0.2863	0.2953 ± 0.0967
no_i_vector_no_excess	noise	0.8123 ± 0.1931	0.3950 ± 0.0906	1.3696 ± 0.2859	0.2950 ± 0.0975
no_train_envelope	clean	0.3539 ± 0.1344	0.1960 ± 0.0733	0.5337 ± 0.2067	0.1800 ± 0.1319
no_train_envelope	noise	0.7233 ± 0.1714	0.3635 ± 0.0872	1.1721 ± 0.2493	0.3058 ± 0.1054
no_envelope_input	clean	0.3532 ± 0.1346	0.1949 ± 0.0729	0.5307 ± 0.2068	0.1810 ± 0.1329
no_envelope_input	noise	0.7245 ± 0.1708	0.3643 ± 0.0871	1.1665 ± 0.2465	0.3132 ± 0.1076

Table G5.1.3: Fixed-threshold detection and interval hits.

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
C	clean	7.29 ± 6.49	91.72 ± 5.32	92.24 ± 6.84	77.04 ± 8.27	89.40 ± 5.73

Continued on next page

Table G5.1.3 (continued)

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
C	noise	5.81 ± 4.85	77.70 ± 6.44	92.43 ± 6.17	52.77 ± 9.68	72.52 ± 6.63
A	clean	4.56 ± 4.50	84.27 ± 7.44	94.42 ± 5.45	73.22 ± 8.36	83.16 ± 8.67
A	noise	15.72 ± 5.87	77.92 ± 8.13	82.16 ± 6.56	41.20 ± 7.27	68.00 ± 7.25
E	clean	6.91 ± 5.82	91.45 ± 4.08	92.57 ± 6.21	77.09 ± 8.32	89.07 ± 5.48
E	noise	4.95 ± 4.36	76.30 ± 6.79	93.30 ± 5.82	51.75 ± 9.78	71.38 ± 7.02
raw_clean	clean	16.86 ± 8.37	83.06 ± 8.46	81.80 ± 8.49	55.32 ± 9.38	77.11 ± 7.33
raw_clean	noise	16.84 ± 8.32	83.14 ± 8.48	81.85 ± 8.41	55.33 ± 9.38	77.01 ± 7.32
raw_noise	clean	16.80 ± 8.42	83.06 ± 8.36	81.88 ± 8.49	55.26 ± 9.10	77.11 ± 7.24
raw_noise	noise	16.80 ± 8.30	83.09 ± 8.49	81.88 ± 8.39	55.23 ± 9.33	76.97 ± 7.31
mlp32_clean	clean	7.49 ± 3.60	82.85 ± 8.04	90.70 ± 4.59	54.69 ± 9.50	77.39 ± 8.25
mlp32_clean	noise	7.98 ± 4.22	83.26 ± 7.89	90.29 ± 5.11	54.67 ± 9.71	77.52 ± 8.01
mlp32_noise	clean	7.92 ± 4.16	83.50 ± 8.18	90.32 ± 5.19	54.45 ± 9.71	77.54 ± 8.48
mlp32_noise	noise	8.32 ± 4.54	83.43 ± 8.13	89.92 ± 5.53	54.43 ± 9.18	77.45 ± 8.05
dd2_clean	clean	5.59 ± 3.35	80.71 ± 5.66	92.88 ± 4.55	46.43 ± 10.43	75.93 ± 7.62
dd2_clean	noise	89.92 ± 3.28	99.22 ± 0.82	50.19 ± 6.28	48.32 ± 6.22	83.74 ± 3.60
dd2_noise	clean	5.27 ± 3.37	82.55 ± 4.39	93.39 ± 4.11	49.74 ± 10.58	76.70 ± 6.73
dd2_noise	noise	79.30 ± 5.89	96.55 ± 2.26	52.62 ± 6.41	53.31 ± 7.68	85.44 ± 5.50
no_i_vector	clean	8.20 ± 6.75	91.99 ± 5.68	91.33 ± 6.76	76.01 ± 9.56	88.67 ± 6.68
no_i_vector	noise	7.87 ± 6.20	77.21 ± 7.53	89.92 ± 7.65	52.70 ± 9.79	72.17 ± 7.61
innovation_only	clean	3.17 ± 3.45	83.61 ± 4.02	96.04 ± 4.00	56.37 ± 8.95	76.60 ± 6.38
innovation_only	noise	4.45 ± 3.79	82.37 ± 4.89	94.49 ± 4.30	54.10 ± 9.01	75.26 ± 6.47
no_excess	clean	3.84 ± 3.44	83.54 ± 4.07	95.17 ± 4.01	56.16 ± 8.96	76.66 ± 6.64
no_excess	noise	4.68 ± 3.85	82.32 ± 4.75	94.21 ± 4.40	53.93 ± 9.16	75.20 ± 6.24
no_i_vector_no_excess	clean	16.79 ± 8.29	83.06 ± 8.46	81.86 ± 8.42	55.32 ± 9.53	77.11 ± 7.33
no_i_vector_no_excess	noise	16.84 ± 8.31	83.10 ± 8.57	81.84 ± 8.44	55.45 ± 9.65	76.93 ± 7.38
no_train_envelope	clean	7.22 ± 6.41	91.96 ± 5.37	92.35 ± 6.68	76.85 ± 8.53	89.43 ± 5.81

Continued on next page

Table G5.1.3 (continued)

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
no_train_envelope	noise	6.13 ± 4.93	78.45 ± 6.12	92.09 ± 6.19	53.15 ± 9.61	73.11 ± 6.40
no_envelope_input	clean	7.40 ± 6.63	91.80 ± 5.48	92.14 ± 6.98	76.97 ± 8.23	89.40 ± 5.82
no_envelope_input	noise	6.34 ± 5.14	78.56 ± 6.21	91.86 ± 6.43	53.17 ± 9.61	73.47 ± 6.54

Table G5.1.4: Distance and error direction. Metric panel 1 of 2; identifying columns are repeated.

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
C	clean	0.1112 ± 0.0219	0.1593 ± 0.0331	0.0468 ± 0.0352	55.74 ± 7.19
C	noise	0.2025 ± 0.0361	0.2882 ± 0.0464	0.1335 ± 0.0541	63.48 ± 9.79
A	clean	0.1406 ± 0.0322	0.2064 ± 0.0500	0.0829 ± 0.0444	59.75 ± 7.88
A	noise	0.2188 ± 0.0331	0.3076 ± 0.0436	0.1081 ± 0.0666	57.03 ± 11.72
E	clean	0.1179 ± 0.0244	0.1692 ± 0.0330	0.0536 ± 0.0382	56.00 ± 6.71
E	noise	0.2098 ± 0.0409	0.2973 ± 0.0520	0.1437 ± 0.0577	64.35 ± 10.07
raw_clean	clean	0.1753 ± 0.0399	0.2519 ± 0.0509	0.1091 ± 0.0570	64.20 ± 9.18
raw_clean	noise	0.1749 ± 0.0400	0.2515 ± 0.0514	0.1091 ± 0.0572	64.22 ± 9.50
raw_noise	clean	0.1754 ± 0.0392	0.2521 ± 0.0504	0.1088 ± 0.0566	64.19 ± 9.30
raw_noise	noise	0.1752 ± 0.0399	0.2518 ± 0.0514	0.1092 ± 0.0572	64.20 ± 9.53
mlp32_clean	clean	0.1703 ± 0.0395	0.2443 ± 0.0546	0.0954 ± 0.0572	60.72 ± 10.33
mlp32_clean	noise	0.1716 ± 0.0398	0.2453 ± 0.0538	0.0974 ± 0.0571	61.06 ± 9.60
mlp32_noise	clean	0.1698 ± 0.0414	0.2438 ± 0.0568	0.0951 ± 0.0581	60.81 ± 10.09
mlp32_noise	noise	0.1712 ± 0.0403	0.2454 ± 0.0545	0.0980 ± 0.0581	61.33 ± 9.88
dd2_clean	clean	0.1748 ± 0.0424	0.2446 ± 0.0539	0.1074 ± 0.0492	64.19 ± 6.83
dd2_clean	noise	0.1516 ± 0.0133	0.2049 ± 0.0223	-0.0564 ± 0.0193	41.66 ± 6.21
dd2_noise	clean	0.1782 ± 0.0462	0.2557 ± 0.0576	0.1149 ± 0.0520	65.49 ± 7.11
dd2_noise	noise	0.1409 ± 0.0177	0.1993 ± 0.0280	0.0256 ± 0.0302	54.05 ± 8.27
no_i_vector	clean	0.1111 ± 0.0265	0.1603 ± 0.0394	0.0434 ± 0.0359	55.70 ± 8.18
no_i_vector	noise	0.2086 ± 0.0405	0.2957 ± 0.0519	0.1498 ± 0.0545	66.16 ± 10.37

Continued on next page

Table G5.1.4 (continued)

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
innovation_only	clean	0.1776 ± 0.0338	0.2530 ± 0.0425	0.0970 ± 0.0459	59.21 ± 8.35
innovation_only	noise	0.1809 ± 0.0323	0.2571 ± 0.0431	0.0979 ± 0.0454	59.87 ± 8.06
no_excess	clean	0.1776 ± 0.0347	0.2530 ± 0.0435	0.0973 ± 0.0470	58.98 ± 8.38
no_excess	noise	0.1817 ± 0.0318	0.2585 ± 0.0421	0.0977 ± 0.0467	59.66 ± 8.06
no_i_vector_no_excess	clean	0.1753 ± 0.0404	0.2518 ± 0.0515	0.1091 ± 0.0576	64.20 ± 9.34
no_i_vector_no_excess	noise	0.1752 ± 0.0405	0.2517 ± 0.0521	0.1094 ± 0.0577	64.23 ± 9.48
no_train_envelope	clean	0.1097 ± 0.0237	0.1572 ± 0.0357	0.0447 ± 0.0354	55.85 ± 7.67
no_train_envelope	noise	0.1987 ± 0.0351	0.2835 ± 0.0448	0.1294 ± 0.0532	63.14 ± 9.75
no_envelope_input	clean	0.1101 ± 0.0221	0.1580 ± 0.0335	0.0453 ± 0.0349	55.45 ± 7.21
no_envelope_input	noise	0.1973 ± 0.0353	0.2818 ± 0.0453	0.1292 ± 0.0521	63.16 ± 9.82

Table G5.1.4: Distance and error direction. Metric panel 2 of 2; identifying columns are repeated.

Source key	Condition	Early (%)	Miss (%)	False boundary (%)
C	clean	44.26 ± 7.19	8.28 ± 5.32	7.29 ± 6.49
C	noise	36.52 ± 9.79	22.30 ± 6.44	5.81 ± 4.85
A	clean	40.25 ± 7.88	15.73 ± 7.44	4.56 ± 4.50
A	noise	42.97 ± 11.72	22.08 ± 8.13	15.72 ± 5.87
E	clean	44.00 ± 6.71	8.55 ± 4.08	6.91 ± 5.82
E	noise	35.65 ± 10.07	23.70 ± 6.79	4.95 ± 4.36
raw_clean	clean	35.80 ± 9.18	16.94 ± 8.46	16.86 ± 8.37
raw_clean	noise	35.78 ± 9.50	16.86 ± 8.48	16.84 ± 8.32
raw_noise	clean	35.81 ± 9.30	16.94 ± 8.36	16.80 ± 8.42
raw_noise	noise	35.80 ± 9.53	16.91 ± 8.49	16.80 ± 8.30
mlp32_clean	clean	39.28 ± 10.33	17.15 ± 8.04	7.49 ± 3.60
mlp32_clean	noise	38.94 ± 9.60	16.74 ± 7.89	7.98 ± 4.22
mlp32_noise	clean	39.19 ± 10.09	16.50 ± 8.18	7.92 ± 4.16
mlp32_noise	noise	38.67 ± 9.88	16.57 ± 8.13	8.32 ± 4.54
dd2_clean	clean	35.81 ± 6.83	19.29 ± 5.66	5.59 ± 3.35
dd2_clean	noise	58.34 ± 6.21	0.78 ± 0.82	89.92 ± 3.28
dd2_noise	clean	34.51 ± 7.11	17.45 ± 4.39	5.27 ± 3.37
dd2_noise	noise	45.95 ± 8.27	3.45 ± 2.26	79.30 ± 5.89
no_i_vector	clean	44.30 ± 8.18	8.01 ± 5.68	8.20 ± 6.75
no_i_vector	noise	33.84 ± 10.37	22.79 ± 7.53	7.87 ± 6.20
innovation_only	clean	40.79 ± 8.35	16.39 ± 4.02	3.17 ± 3.45
innovation_only	noise	40.13 ± 8.06	17.63 ± 4.89	4.45 ± 3.79
no_excess	clean	41.02 ± 8.38	16.46 ± 4.07	3.84 ± 3.44
no_excess	noise	40.34 ± 8.06	17.68 ± 4.75	4.68 ± 3.85
no_i_vector_no_excess	clean	35.80 ± 9.34	16.94 ± 8.46	16.79 ± 8.29

Continued on next page

Table G5.1.4 (continued)

Source key	Condition	Early (%)	Miss (%)	False boundary (%)
no_i_vector_no_excess	noise	35.77 ± 9.48	16.90 ± 8.57	16.84 ± 8.31
no_train_envelope	clean	44.15 ± 7.67	8.04 ± 5.37	7.22 ± 6.41
no_train_envelope	noise	36.86 ± 9.75	21.55 ± 6.12	6.13 ± 4.93
no_envelope_input	clean	44.55 ± 7.21	8.20 ± 5.48	7.40 ± 6.63
no_envelope_input	noise	36.84 ± 9.82	21.44 ± 6.21	6.34 ± 5.14

G.4.2 HARDSTOP1D

Tables G5.2.1, G5.2.2, G5.2.3, and G5.2.4 report presence ranking and calibration, probability scores, fixed-threshold outcomes, and distance/direction metrics, respectively, for all versions and both observation conditions.

Table G5.2.1: Presence ranking and calibration.

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
C	clean	0.9304 ± 0.0488	0.9020 ± 0.0769	0.0635 ± 0.0464	0.0817 ± 0.0542	-0.0321 ± 0.0675
C	noise	0.9226 ± 0.0606	0.8891 ± 0.0980	0.0686 ± 0.0459	0.1078 ± 0.0436	-0.0018 ± 0.0708
A	clean	0.9429 ± 0.0320	0.9193 ± 0.0579	0.0600 ± 0.0454	0.0810 ± 0.0485	-0.0260 ± 0.0645
A	noise	0.9170 ± 0.0447	0.8826 ± 0.0824	0.0730 ± 0.0457	0.1149 ± 0.0365	0.0059 ± 0.0721
E	clean	0.9318 ± 0.0469	0.9035 ± 0.0756	0.0614 ± 0.0438	0.0765 ± 0.0474	-0.0267 ± 0.0632
E	noise	0.9263 ± 0.0531	0.8942 ± 0.0893	0.0670 ± 0.0425	0.1038 ± 0.0363	0.0080 ± 0.0659
raw_clean	clean	0.8886 ± 0.0476	0.8507 ± 0.0677	0.0895 ± 0.0541	0.1175 ± 0.0489	-0.0365 ± 0.0805
raw_clean	noise	0.8875 ± 0.0493	0.8496 ± 0.0688	0.0895 ± 0.0541	0.1174 ± 0.0489	-0.0364 ± 0.0804
raw_noise	clean	0.8884 ± 0.0479	0.8506 ± 0.0679	0.0895 ± 0.0541	0.1175 ± 0.0489	-0.0365 ± 0.0805
raw_noise	noise	0.8875 ± 0.0493	0.8496 ± 0.0687	0.0896 ± 0.0541	0.1174 ± 0.0490	-0.0364 ± 0.0804

Continued on next page

Table G5.2.1 (continued)

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
mlp32_clean	clean	0.9206 ± 0.0353	0.8927 ± 0.0534	0.0760 ± 0.0459	0.1109 ± 0.0449	-0.0336 ± 0.0689
mlp32_clean	noise	0.9205 ± 0.0351	0.8925 ± 0.0533	0.0760 ± 0.0459	0.1109 ± 0.0449	-0.0336 ± 0.0689
mlp32_noise	clean	0.9206 ± 0.0353	0.8927 ± 0.0534	0.0760 ± 0.0459	0.1109 ± 0.0449	-0.0336 ± 0.0689
mlp32_noise	noise	0.9205 ± 0.0351	0.8925 ± 0.0533	0.0760 ± 0.0459	0.1110 ± 0.0449	-0.0336 ± 0.0689
dd2_clean	clean	0.6351 ± 0.1192	0.5227 ± 0.1464	0.1419 ± 0.0806	0.1437 ± 0.0886	-0.0816 ± 0.1114
dd2_clean	noise	0.6059 ± 0.1313	0.4872 ± 0.1483	0.1360 ± 0.0704	0.1538 ± 0.0708	-0.0324 ± 0.1118
dd2_noise	clean	0.6353 ± 0.1217	0.5240 ± 0.1484	0.1463 ± 0.0850	0.1485 ± 0.0948	-0.0974 ± 0.1126
dd2_noise	noise	0.6097 ± 0.1363	0.4930 ± 0.1555	0.1391 ± 0.0766	0.1563 ± 0.0814	-0.0560 ± 0.1132
no_i_vector	clean	0.9344 ± 0.0389	0.9097 ± 0.0587	0.0688 ± 0.0469	0.0889 ± 0.0562	-0.0428 ± 0.0711
no_i_vector	noise	0.9151 ± 0.0545	0.8812 ± 0.0860	0.0799 ± 0.0487	0.1230 ± 0.0448	-0.0221 ± 0.0769
innovation_only	clean	0.9178 ± 0.0610	0.8844 ± 0.0895	0.0758 ± 0.0505	0.1092 ± 0.0484	-0.0203 ± 0.0761
innovation_only	noise	0.9166 ± 0.0610	0.8832 ± 0.0884	0.0760 ± 0.0505	0.1097 ± 0.0485	-0.0202 ± 0.0762
no_excess	clean	0.9177 ± 0.0626	0.8842 ± 0.0919	0.0748 ± 0.0509	0.1096 ± 0.0481	-0.0196 ± 0.0758
no_excess	noise	0.9176 ± 0.0616	0.8841 ± 0.0904	0.0750 ± 0.0509	0.1096 ± 0.0482	-0.0195 ± 0.0759
no_i_vector_no_excess	clean	0.8942 ± 0.0487	0.8579 ± 0.0698	0.0917 ± 0.0526	0.1301 ± 0.0464	-0.0323 ± 0.0820
no_i_vector_no_excess	noise	0.8932 ± 0.0498	0.8569 ± 0.0704	0.0917 ± 0.0526	0.1300 ± 0.0464	-0.0322 ± 0.0820
no_train_envelope	clean	0.9289 ± 0.0495	0.9014 ± 0.0740	0.0649 ± 0.0448	0.0815 ± 0.0517	-0.0330 ± 0.0670
no_train_envelope	noise	0.9271 ± 0.0590	0.8963 ± 0.0922	0.0683 ± 0.0447	0.1083 ± 0.0428	-0.0011 ± 0.0713

Continued on next page

Table G5.2.1 (continued)

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
no_envelope_input	clean	0.9304 ± 0.0478	0.9027 ± 0.0730	0.0651 ± 0.0448	0.0842 ± 0.0506	-0.0332 ± 0.0672
no_envelope_input	noise	0.9283 ± 0.0597	0.8972 ± 0.0944	0.0687 ± 0.0449	0.1114 ± 0.0421	-0.0011 ± 0.0717

Table G5.2.2: Joint and interval probabilities.

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
C	clean	0.3601 ± 0.2464	0.1786 ± 0.1316	1.2023 ± 0.3906	0.0954 ± 0.0401
C	noise	0.4481 ± 0.2644	0.2117 ± 0.1434	1.4188 ± 0.3906	0.1420 ± 0.0448
A	clean	0.3517 ± 0.2446	0.1688 ± 0.1310	1.1387 ± 0.3834	0.0980 ± 0.0362
A	noise	0.5024 ± 0.2651	0.2404 ± 0.1444	1.5755 ± 0.3692	0.1585 ± 0.0407
E	clean	0.3541 ± 0.2398	0.1757 ± 0.1277	1.1681 ± 0.3730	0.0992 ± 0.0431
E	noise	0.4442 ± 0.2562	0.2087 ± 0.1383	1.3780 ± 0.3739	0.1512 ± 0.0491
raw_clean	clean	0.5682 ± 0.3228	0.2617 ± 0.1616	1.9805 ± 0.3735	0.1255 ± 0.0294
raw_clean	noise	0.5683 ± 0.3225	0.2617 ± 0.1616	1.9810 ± 0.3738	0.1256 ± 0.0294
raw_noise	clean	0.5683 ± 0.3228	0.2617 ± 0.1616	1.9806 ± 0.3734	0.1255 ± 0.0294
raw_noise	noise	0.5683 ± 0.3225	0.2618 ± 0.1616	1.9812 ± 0.3737	0.1256 ± 0.0294
mlp32_clean	clean	0.5236 ± 0.3159	0.2421 ± 0.1501	1.7971 ± 0.3221	0.1136 ± 0.0328
mlp32_clean	noise	0.5237 ± 0.3158	0.2422 ± 0.1501	1.7980 ± 0.3221	0.1137 ± 0.0328
mlp32_noise	clean	0.5236 ± 0.3159	0.2421 ± 0.1501	1.7973 ± 0.3221	0.1136 ± 0.0328
mlp32_noise	noise	0.5238 ± 0.3158	0.2422 ± 0.1501	1.7982 ± 0.3221	0.1137 ± 0.0328
dd2_clean	clean	0.6154 ± 0.3835	0.3344 ± 0.1896	2.1226 ± 0.5005	0.1309 ± 0.0495
dd2_clean	noise	0.6245 ± 0.3588	0.3256 ± 0.1771	1.9799 ± 0.5492	0.1935 ± 0.0509
dd2_noise	clean	0.6304 ± 0.4006	0.3405 ± 0.1951	2.2549 ± 0.5133	0.1115 ± 0.0430
dd2_noise	noise	0.6295 ± 0.3784	0.3298 ± 0.1850	2.1080 ± 0.5673	0.1613 ± 0.0442
no_i_vector	clean	0.3869 ± 0.2476	0.1917 ± 0.1320	1.3593 ± 0.3907	0.0886 ± 0.0276
no_i_vector	noise	0.4821 ± 0.2664	0.2324 ± 0.1459	1.6088 ± 0.3502	0.1347 ± 0.0322

Continued on next page

Table G5.2.2 (continued)

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
innovation_only	clean	0.4769 ± 0.2860	0.2276 ± 0.1515	1.5968 ± 0.4459	0.1241 ± 0.0380
innovation_only	noise	0.4777 ± 0.2869	0.2279 ± 0.1517	1.5986 ± 0.4468	0.1243 ± 0.0378
no_excess	clean	0.4721 ± 0.2842	0.2263 ± 0.1516	1.5924 ± 0.4580	0.1242 ± 0.0398
no_excess	noise	0.4729 ± 0.2851	0.2266 ± 0.1519	1.5941 ± 0.4591	0.1244 ± 0.0396
no_i_vector_no_excess	clean	0.5605 ± 0.2982	0.2654 ± 0.1582	1.9280 ± 0.3056	0.1372 ± 0.0303
no_i_vector_no_excess	noise	0.5606 ± 0.2980	0.2655 ± 0.1581	1.9283 ± 0.3058	0.1372 ± 0.0303
no_train_envelope	clean	0.3656 ± 0.2443	0.1820 ± 0.1300	1.2242 ± 0.3983	0.0955 ± 0.0385
no_train_envelope	noise	0.4524 ± 0.2640	0.2119 ± 0.1427	1.4315 ± 0.4013	0.1416 ± 0.0423
no_envelope_input	clean	0.3687 ± 0.2494	0.1823 ± 0.1310	1.2314 ± 0.4020	0.0958 ± 0.0379
no_envelope_input	noise	0.4583 ± 0.2702	0.2140 ± 0.1444	1.4492 ± 0.4129	0.1420 ± 0.0416

Table G5.2.3: Fixed-threshold detection and interval hits.

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
C	clean	6.33 ± 3.54	93.18 ± 12.63	81.49 ± 7.35	64.67 ± 16.38	88.63 ± 13.90
C	noise	9.28 ± 4.24	95.80 ± 7.84	74.96 ± 8.15	61.62 ± 16.18	89.33 ± 12.29
A	clean	7.57 ± 4.92	96.79 ± 6.62	79.63 ± 5.53	71.16 ± 12.05	92.65 ± 11.28
A	noise	13.63 ± 6.56	96.84 ± 6.35	67.02 ± 10.50	43.06 ± 9.51	91.39 ± 13.11
E	clean	6.24 ± 3.32	91.64 ± 11.69	81.31 ± 7.09	63.86 ± 14.85	87.13 ± 14.48
E	noise	9.40 ± 4.33	94.00 ± 7.92	74.45 ± 7.87	60.89 ± 16.27	88.55 ± 12.49
raw_clean	clean	17.42 ± 7.14	92.31 ± 7.81	60.31 ± 12.74	51.90 ± 17.45	87.86 ± 13.06
raw_clean	noise	17.44 ± 7.05	92.06 ± 7.82	60.15 ± 12.88	52.17 ± 17.19	87.34 ± 12.76
raw_noise	clean	17.42 ± 7.14	92.31 ± 7.81	60.31 ± 12.74	51.90 ± 17.45	87.86 ± 13.06
raw_noise	noise	17.43 ± 7.05	92.06 ± 7.82	60.18 ± 12.86	52.12 ± 17.23	87.34 ± 12.76
mlp32_clean	clean	13.95 ± 6.12	94.91 ± 5.21	66.10 ± 11.89	54.39 ± 17.10	90.00 ± 11.17
mlp32_clean	noise	14.03 ± 6.08	94.81 ± 5.35	65.95 ± 11.97	54.26 ± 17.15	89.91 ± 11.36
mlp32_noise	clean	13.95 ± 6.12	94.91 ± 5.21	66.10 ± 11.89	54.39 ± 17.10	90.00 ± 11.17

Continued on next page

Table G5.2.3 (continued)

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
mlp32_noise	noise	14.05 ± 6.11	94.86 ± 5.28	65.94 ± 11.96	54.31 ± 17.13	89.96 ± 11.29
dd2_clean	clean	21.79 ± 6.65	89.49 ± 14.48	52.93 ± 16.37	52.43 ± 22.81	77.73 ± 16.90
dd2_clean	noise	29.53 ± 6.67	94.08 ± 11.34	47.13 ± 15.65	54.32 ± 26.45	81.88 ± 14.44
dd2_noise	clean	22.22 ± 7.00	88.86 ± 14.15	52.47 ± 16.42	51.66 ± 22.26	77.33 ± 16.25
dd2_noise	noise	28.59 ± 7.00	93.98 ± 11.22	47.93 ± 15.99	54.41 ± 26.53	81.97 ± 14.56
no_i_vector	clean	9.86 ± 4.46	93.24 ± 9.88	72.56 ± 8.87	61.18 ± 14.65	89.59 ± 14.03
no_i_vector	noise	13.92 ± 4.88	95.45 ± 7.46	65.47 ± 11.34	59.10 ± 13.74	90.35 ± 13.17
innovation_only	clean	10.21 ± 4.19	90.84 ± 7.83	71.53 ± 7.97	55.37 ± 13.23	85.12 ± 12.21
innovation_only	noise	10.27 ± 4.16	91.10 ± 7.95	71.41 ± 8.09	55.43 ± 13.40	85.59 ± 12.16
no_excess	clean	10.46 ± 4.17	91.44 ± 8.52	71.03 ± 8.41	56.72 ± 13.55	85.80 ± 14.49
no_excess	noise	10.40 ± 4.00	91.77 ± 8.05	71.11 ± 8.69	56.80 ± 13.68	85.93 ± 13.23
no_i_vector_no_excess	clean	15.63 ± 6.70	92.31 ± 7.81	62.89 ± 11.81	53.10 ± 16.59	87.88 ± 12.41
no_i_vector_no_excess	noise	15.64 ± 6.60	92.06 ± 7.82	62.74 ± 11.97	52.89 ± 16.58	87.63 ± 12.39
no_train_envelope	clean	7.84 ± 3.76	93.18 ± 11.33	77.18 ± 8.38	63.14 ± 17.02	88.15 ± 12.42
no_train_envelope	noise	10.17 ± 3.68	95.84 ± 7.72	72.40 ± 9.52	59.24 ± 15.85	88.90 ± 11.37
no_envelope_input	clean	8.36 ± 4.03	93.41 ± 11.46	76.13 ± 10.27	63.74 ± 17.09	88.38 ± 12.45
no_envelope_input	noise	10.92 ± 3.45	96.23 ± 7.11	70.57 ± 9.92	58.50 ± 15.68	89.60 ± 10.67

Table G5.2.4: Distance and error direction. Metric panel 1 of 2; identifying columns are repeated.

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
C	clean	0.1633 ± 0.0740	0.2253 ± 0.1145	0.0375 ± 0.1228	51.15 ± 18.98
C	noise	0.1496 ± 0.0413	0.2081 ± 0.0662	0.0198 ± 0.0903	51.76 ± 19.41
A	clean	0.1314 ± 0.0403	0.1817 ± 0.0547	0.0114 ± 0.0666	50.54 ± 16.86
A	noise	0.1790 ± 0.0393	0.2428 ± 0.0485	-0.0152 ± 0.1004	47.79 ± 18.05
E	clean	0.1653 ± 0.0695	0.2260 ± 0.1062	0.0433 ± 0.1058	51.38 ± 15.46
E	noise	0.1518 ± 0.0444	0.2100 ± 0.0673	0.0261 ± 0.0752	51.73 ± 16.48

Continued on next page

Table G5.2.4 (continued)

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
raw_clean	clean	0.1805 ± 0.0667	0.2477 ± 0.0896	0.0711 ± 0.0841	61.19 ± 14.61
raw_clean	noise	0.1815 ± 0.0658	0.2491 ± 0.0880	0.0743 ± 0.0837	61.74 ± 14.81
raw_noise	clean	0.1805 ± 0.0667	0.2477 ± 0.0896	0.0711 ± 0.0841	61.19 ± 14.61
raw_noise	noise	0.1816 ± 0.0657	0.2492 ± 0.0880	0.0741 ± 0.0836	61.68 ± 14.76
mlp32_clean	clean	0.1662 ± 0.0538	0.2280 ± 0.0686	0.0558 ± 0.0670	59.27 ± 14.31
mlp32_clean	noise	0.1666 ± 0.0535	0.2288 ± 0.0688	0.0560 ± 0.0674	59.24 ± 14.32
mlp32_noise	clean	0.1662 ± 0.0538	0.2280 ± 0.0686	0.0558 ± 0.0670	59.27 ± 14.31
mlp32_noise	noise	0.1664 ± 0.0536	0.2284 ± 0.0687	0.0558 ± 0.0673	59.24 ± 14.32
dd2_clean	clean	0.2244 ± 0.0891	0.3046 ± 0.1359	0.0807 ± 0.1588	56.60 ± 20.87
dd2_clean	noise	0.2071 ± 0.0834	0.2811 ± 0.1284	0.0714 ± 0.1585	59.44 ± 25.85
dd2_noise	clean	0.2290 ± 0.0856	0.3138 ± 0.1297	0.0890 ± 0.1575	57.84 ± 21.39
dd2_noise	noise	0.2083 ± 0.0851	0.2831 ± 0.1319	0.0768 ± 0.1563	59.99 ± 24.25
no_i_vector	clean	0.1671 ± 0.0510	0.2288 ± 0.0779	0.0616 ± 0.0769	58.79 ± 15.82
no_i_vector	noise	0.1569 ± 0.0472	0.2172 ± 0.0637	0.0489 ± 0.0643	58.74 ± 16.86
innovation_only	clean	0.1641 ± 0.0447	0.2291 ± 0.0652	0.0726 ± 0.0591	60.32 ± 11.55
innovation_only	noise	0.1633 ± 0.0451	0.2278 ± 0.0663	0.0717 ± 0.0587	60.31 ± 11.35
no_excess	clean	0.1661 ± 0.0561	0.2294 ± 0.0858	0.0732 ± 0.0744	58.55 ± 12.20
no_excess	noise	0.1642 ± 0.0516	0.2276 ± 0.0778	0.0704 ± 0.0673	59.00 ± 11.50
no_i_vector_no_excess	clean	0.1808 ± 0.0674	0.2485 ± 0.0902	0.0793 ± 0.0854	62.52 ± 14.79
no_i_vector_no_excess	noise	0.1818 ± 0.0668	0.2498 ± 0.0893	0.0808 ± 0.0852	62.77 ± 14.89
no_train_envelope	clean	0.1644 ± 0.0596	0.2281 ± 0.0925	0.0364 ± 0.1110	51.98 ± 19.20
no_train_envelope	noise	0.1527 ± 0.0377	0.2126 ± 0.0580	0.0185 ± 0.0896	52.49 ± 19.47
no_envelope_input	clean	0.1615 ± 0.0608	0.2225 ± 0.0951	0.0328 ± 0.1146	51.64 ± 19.72
no_envelope_input	noise	0.1498 ± 0.0357	0.2086 ± 0.0508	0.0121 ± 0.0864	52.13 ± 19.57

Table G5.2.4: Distance and error direction. Metric panel 2 of 2; identifying columns are repeated.

Source key	Condition	Early (%)	Miss (%)	False boundary (%)
C	clean	48.85 ± 18.98	6.82 ± 12.63	6.33 ± 3.54
C	noise	48.24 ± 19.41	4.20 ± 7.84	9.28 ± 4.24
A	clean	49.46 ± 16.86	3.21 ± 6.62	7.57 ± 4.92
A	noise	52.21 ± 18.05	3.16 ± 6.35	13.63 ± 6.56
E	clean	48.62 ± 15.46	8.36 ± 11.69	6.24 ± 3.32
E	noise	48.27 ± 16.48	6.00 ± 7.92	9.40 ± 4.33
raw_clean	clean	38.81 ± 14.61	7.69 ± 7.81	17.42 ± 7.14
raw_clean	noise	38.26 ± 14.81	7.94 ± 7.82	17.44 ± 7.05
raw_noise	clean	38.81 ± 14.61	7.69 ± 7.81	17.42 ± 7.14
raw_noise	noise	38.32 ± 14.76	7.94 ± 7.82	17.43 ± 7.05
mlp32_clean	clean	40.73 ± 14.31	5.09 ± 5.21	13.95 ± 6.12
mlp32_clean	noise	40.76 ± 14.32	5.19 ± 5.35	14.03 ± 6.08
mlp32_noise	clean	40.73 ± 14.31	5.09 ± 5.21	13.95 ± 6.12
mlp32_noise	noise	40.76 ± 14.32	5.14 ± 5.28	14.05 ± 6.11
dd2_clean	clean	43.40 ± 20.87	10.51 ± 14.48	21.79 ± 6.65
dd2_clean	noise	40.56 ± 25.85	5.92 ± 11.34	29.53 ± 6.67
dd2_noise	clean	42.16 ± 21.39	11.14 ± 14.15	22.22 ± 7.00
dd2_noise	noise	40.01 ± 24.25	6.02 ± 11.22	28.59 ± 7.00
no_i_vector	clean	41.21 ± 15.82	6.76 ± 9.88	9.86 ± 4.46
no_i_vector	noise	41.26 ± 16.86	4.55 ± 7.46	13.92 ± 4.88
innovation_only	clean	39.68 ± 11.55	9.16 ± 7.83	10.21 ± 4.19
innovation_only	noise	39.69 ± 11.35	8.90 ± 7.95	10.27 ± 4.16
no_excess	clean	41.45 ± 12.20	8.56 ± 8.52	10.46 ± 4.17
no_excess	noise	41.00 ± 11.50	8.23 ± 8.05	10.40 ± 4.00
no_i_vector_no_excess	clean	37.48 ± 14.79	7.69 ± 7.81	15.63 ± 6.70
no_i_vector_no_excess	noise	37.23 ± 14.89	7.94 ± 7.82	15.64 ± 6.60
no_train_envelope	clean	48.02 ± 19.20	6.82 ± 11.33	7.84 ± 3.76
no_train_envelope	noise	47.51 ± 19.47	4.16 ± 7.72	10.17 ± 3.68
no_envelope_input	clean	48.36 ± 19.72	6.59 ± 11.46	8.36 ± 4.03
no_envelope_input	noise	47.87 ± 19.57	3.77 ± 7.11	10.92 ± 3.45

G.4.3 SaturationKink

Tables G5.3.1, G5.3.2, G5.3.3, and G5.3.4 report presence ranking and calibration, probability scores, fixed-threshold outcomes, and distance/direction metrics, respectively, for all versions and both observation conditions.

Table G5.3.1: Presence ranking and calibration.

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
C	clean	$1.0000 \pm 4.20 \times 10^{-17}$	$1.0000 \pm 1.68 \times 10^{-16}$	0.0011 ± 0.0025	0.0048 ± 0.0035	-0.0043 ± 0.0035
C	noise	0.9764 ± 0.0295	0.9264 ± 0.0841	0.0754 ± 0.0480	0.1012 ± 0.0203	0.0139 ± 0.0273
A	clean	$1.0000 \pm 4.20 \times 10^{-17}$	$1.0000 \pm 1.68 \times 10^{-16}$	0.0115 ± 0.0027	0.0823 ± 0.0086	-0.0143 ± 0.0120
A	noise	0.9853 ± 0.0176	0.9535 ± 0.0508	0.2196 ± 0.0581	0.2445 ± 0.0580	0.2397 ± 0.0589
E	clean	$1.0000 \pm 4.20 \times 10^{-17}$	$1.0000 \pm 1.68 \times 10^{-16}$	0.0011 ± 0.0027	0.0044 ± 0.0036	-0.0039 ± 0.0036
E	noise	0.9734 ± 0.0333	0.9173 ± 0.0944	0.0780 ± 0.0513	0.1047 ± 0.0247	0.0178 ± 0.0284
raw_clean	clean	0.9716 ± 0.0312	0.9098 ± 0.0861	0.0915 ± 0.0380	0.1083 ± 0.0173	-0.0182 ± 0.0370
raw_clean	noise	0.9714 ± 0.0313	0.9089 ± 0.0865	0.0916 ± 0.0380	0.1088 ± 0.0170	-0.0183 ± 0.0372
raw_noise	clean	0.9716 ± 0.0312	0.9098 ± 0.0862	0.0916 ± 0.0381	0.1093 ± 0.0180	-0.0181 ± 0.0370
raw_noise	noise	0.9713 ± 0.0314	0.9086 ± 0.0868	0.0917 ± 0.0380	0.1088 ± 0.0169	-0.0182 ± 0.0371
mlp32_clean	clean	0.9937 ± 0.0071	0.9798 ± 0.0208	0.0464 ± 0.0282	0.0477 ± 0.0204	-0.0061 ± 0.0322
mlp32_clean	noise	0.9936 ± 0.0075	0.9794 ± 0.0222	0.0468 ± 0.0284	0.0493 ± 0.0206	-0.0058 ± 0.0325
mlp32_noise	clean	0.9936 ± 0.0074	0.9794 ± 0.0215	0.0467 ± 0.0285	0.0505 ± 0.0189	-0.0061 ± 0.0321
mlp32_noise	noise	0.9935 ± 0.0075	0.9791 ± 0.0222	0.0472 ± 0.0287	0.0499 ± 0.0214	-0.0059 ± 0.0324
dd2_clean	clean	0.9996 ± 0.0011	0.9986 ± 0.0040	0.0394 ± 0.0192	0.1025 ± 0.0276	-0.0042 ± 0.0285
dd2_clean	noise	0.9246 ± 0.0719	0.7520 ± 0.2037	0.1476 ± 0.0606	0.1619 ± 0.0589	0.1545 ± 0.0608
dd2_noise	clean	0.9982 ± 0.0025	0.9939 ± 0.0089	0.0327 ± 0.0197	0.0662 ± 0.0187	-0.0135 ± 0.0232

Continued on next page

Table G5.3.1 (continued)

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
dd2_noise	noise	0.9321 ± 0.0722	0.7807 ± 0.2096	0.1468 ± 0.0645	0.1500 ± 0.0641	0.1475 ± 0.0647
no_i_vector	clean	1.0000 ± 4.20×10^{-17}	1.0000 ± 1.68×10^{-16}	0.0011 ± 0.0027	0.0045 ± 0.0036	-0.0040 ± 0.0036
no_i_vector	noise	0.9754 ± 0.0302	0.9229 ± 0.0868	0.0767 ± 0.0489	0.1043 ± 0.0214	0.0170 ± 0.0285
innovation_only	clean	0.9793 ± 0.0233	0.9333 ± 0.0711	0.0738 ± 0.0481	0.0955 ± 0.0283	-0.0235 ± 0.0256
innovation_only	noise	0.9791 ± 0.0238	0.9327 ± 0.0726	0.0743 ± 0.0485	0.0931 ± 0.0244	-0.0224 ± 0.0261
no_excess	clean	0.9793 ± 0.0233	0.9333 ± 0.0711	0.0738 ± 0.0481	0.0955 ± 0.0283	-0.0235 ± 0.0256
no_excess	noise	0.9791 ± 0.0238	0.9327 ± 0.0726	0.0743 ± 0.0485	0.0931 ± 0.0244	-0.0224 ± 0.0261
no_i_vector_no_excess	clean	0.9639 ± 0.0429	0.8863 ± 0.1227	0.0864 ± 0.0611	0.1148 ± 0.0341	-0.0181 ± 0.0335
no_i_vector_no_excess	noise	0.9637 ± 0.0430	0.8855 ± 0.1232	0.0864 ± 0.0611	0.1147 ± 0.0343	-0.0182 ± 0.0336
no_train_envelope	clean	0.9999 ± 0.0002	0.9998 ± 0.0006	0.0023 ± 0.0043	0.0066 ± 0.0034	0.0007 ± 0.0057
no_train_envelope	noise	0.9848 ± 0.0180	0.9519 ± 0.0537	0.0671 ± 0.0451	0.0847 ± 0.0255	0.0186 ± 0.0309
no_envelope_input	clean	1.0000 ± 4.20×10^{-17}	1.0000 ± 1.68×10^{-16}	0.0011 ± 0.0025	0.0048 ± 0.0035	-0.0043 ± 0.0035
no_envelope_input	noise	0.9764 ± 0.0295	0.9264 ± 0.0841	0.0754 ± 0.0480	0.1012 ± 0.0203	0.0139 ± 0.0273

Table G5.3.2: Joint and interval probabilities.

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
C	clean	0.0095 ± 0.0060	0.0028 ± 0.0049	0.0085 ± 0.0087	0.0013 ± 0.0005
C	noise	0.4899 ± 0.1046	0.2664 ± 0.0648	0.4684 ± 0.0963	0.5261 ± 0.2519
A	clean	0.3636 ± 0.0206	0.1225 ± 0.0102	0.2906 ± 0.0107	0.1542 ± 0.0169
A	noise	2.1141 ± 0.2147	0.8908 ± 0.0500	1.4347 ± 0.0548	3.2981 ± 0.6134

Continued on next page

Table G5.3.2 (continued)

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
E	clean	0.0092 ± 0.0063	0.0027 ± 0.0051	0.0080 ± 0.0092	0.0015 ± 0.0006
E	noise	0.4956 ± 0.1124	0.2698 ± 0.0690	0.4649 ± 0.0968	0.5536 ± 0.2781
raw_clean	clean	0.6813 ± 0.1090	0.3560 ± 0.0621	0.7220 ± 0.1482	0.5692 ± 0.1904
raw_clean	noise	0.6818 ± 0.1089	0.3564 ± 0.0619	0.7225 ± 0.1486	0.5695 ± 0.1905
raw_noise	clean	0.6817 ± 0.1090	0.3562 ± 0.0621	0.7222 ± 0.1482	0.5699 ± 0.1905
raw_noise	noise	0.6821 ± 0.1089	0.3566 ± 0.0619	0.7227 ± 0.1485	0.5702 ± 0.1906
mlp32_clean	clean	0.3949 ± 0.1118	0.2250 ± 0.0672	0.4551 ± 0.1549	0.2321 ± 0.1513
mlp32_clean	noise	0.3970 ± 0.1101	0.2272 ± 0.0660	0.4567 ± 0.1535	0.2356 ± 0.1540
mlp32_noise	clean	0.3948 ± 0.1112	0.2251 ± 0.0672	0.4541 ± 0.1535	0.2344 ± 0.1517
mlp32_noise	noise	0.3969 ± 0.1094	0.2273 ± 0.0661	0.4557 ± 0.1520	0.2378 ± 0.1543
dd2_clean	clean	0.6213 ± 0.0623	0.3185 ± 0.0364	0.3684 ± 0.0357	0.2748 ± 0.1225
dd2_clean	noise	1.6140 ± 0.2418	0.6767 ± 0.0496	0.9180 ± 0.0803	2.2540 ± 0.8445
dd2_noise	clean	0.5366 ± 0.1322	0.2981 ± 0.0652	0.3125 ± 0.0869	0.1773 ± 0.1167
dd2_noise	noise	1.3128 ± 0.2846	0.5316 ± 0.0452	0.4674 ± 0.0360	2.5435 ± 1.0642
no_i_vector	clean	0.0090 ± 0.0062	0.0026 ± 0.0052	0.0081 ± 0.0092	0.0014 ± 0.0005
no_i_vector	noise	0.4883 ± 0.1027	0.2671 ± 0.0643	0.4575 ± 0.0867	0.5467 ± 0.2627
innovation_only	clean	0.4527 ± 0.1066	0.2551 ± 0.0733	0.4839 ± 0.1062	0.3593 ± 0.2110
innovation_only	noise	0.4625 ± 0.1062	0.2623 ± 0.0725	0.4945 ± 0.1034	0.3653 ± 0.2150
no_excess	clean	0.4527 ± 0.1066	0.2551 ± 0.0733	0.4839 ± 0.1062	0.3593 ± 0.2110
no_excess	noise	0.4625 ± 0.1062	0.2623 ± 0.0725	0.4945 ± 0.1034	0.3653 ± 0.2150
no_i_vector_no_excess	clean	0.6417 ± 0.1517	0.3435 ± 0.0959	0.6814 ± 0.1618	0.5191 ± 0.3314
no_i_vector_no_excess	noise	0.6421 ± 0.1518	0.3440 ± 0.0957	0.6818 ± 0.1621	0.5196 ± 0.3317
no_train_envelope	clean	0.0156 ± 0.0119	0.0047 ± 0.0085	0.0110 ± 0.0162	0.0220 ± 0.0100
no_train_envelope	noise	0.4286 ± 0.1004	0.2404 ± 0.0638	0.4010 ± 0.0815	0.4773 ± 0.2681
no_envelope_input	clean	0.0095 ± 0.0060	0.0028 ± 0.0049	0.0085 ± 0.0087	0.0013 ± 0.0005
no_envelope_input	noise	0.4899 ± 0.1046	0.2664 ± 0.0648	0.4684 ± 0.0963	0.5261 ± 0.2519

Table G5.3.3: Fixed-threshold detection and interval hits.

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
C	clean	0.00 ± 0.00	99.73 ± 0.77	100.00 ± 0.00	99.73 ± 0.77	99.73 ± 0.77
C	noise	16.63 ± 15.73	90.64 ± 6.53	92.69 ± 7.72	79.46 ± 6.78	90.27 ± 6.16
A	clean	0.00 ± 0.00	79.98 ± 9.00	100.00 ± 0.00	79.98 ± 9.00	79.98 ± 9.00
A	noise	56.62 ± 17.66	100.00 ± 0.00	82.19 ± 6.99	22.48 ± 5.24	72.12 ± 3.45
E	clean	0.00 ± 0.00	99.73 ± 0.77	100.00 ± 0.00	99.73 ± 0.77	99.73 ± 0.77
E	noise	18.25 ± 17.29	90.50 ± 6.66	92.01 ± 8.41	79.45 ± 6.82	90.13 ± 6.29
raw_clean	clean	29.65 ± 11.79	93.45 ± 6.73	88.61 ± 6.95	77.78 ± 10.69	92.78 ± 6.14
raw_clean	noise	29.55 ± 11.79	93.43 ± 6.91	88.66 ± 6.93	77.43 ± 10.39	92.54 ± 6.10
raw_noise	clean	29.44 ± 11.92	93.45 ± 6.73	88.69 ± 6.99	77.60 ± 10.65	92.69 ± 6.03
raw_noise	noise	29.54 ± 11.82	93.42 ± 6.88	88.66 ± 6.93	77.39 ± 10.39	92.52 ± 6.06
mlp32_clean	clean	2.21 ± 4.45	88.94 ± 4.34	98.72 ± 2.53	77.16 ± 4.40	88.77 ± 4.30
mlp32_clean	noise	2.38 ± 4.71	88.87 ± 4.30	98.64 ± 2.65	77.08 ± 4.29	88.63 ± 4.14
mlp32_noise	clean	2.57 ± 5.09	89.12 ± 4.27	98.53 ± 2.86	77.70 ± 4.21	88.96 ± 4.22
mlp32_noise	noise	2.46 ± 4.76	88.93 ± 4.31	98.61 ± 2.69	77.07 ± 4.12	88.71 ± 4.16
dd2_clean	clean	0.00 ± 0.00	83.13 ± 5.59	100.00 ± 0.00	79.80 ± 5.01	83.13 ± 5.59
dd2_clean	noise	55.33 ± 18.21	99.87 ± 0.18	82.53 ± 7.06	54.96 ± 4.75	92.34 ± 3.45
dd2_noise	clean	0.00 ± 0.00	84.26 ± 6.22	100.00 ± 0.00	76.01 ± 4.45	83.74 ± 5.46
dd2_noise	noise	52.82 ± 19.73	99.71 ± 0.38	83.16 ± 7.64	81.05 ± 3.13	96.44 ± 1.64
no_i_vector	clean	0.00 ± 0.00	99.73 ± 0.77	100.00 ± 0.00	99.73 ± 0.77	99.73 ± 0.77
no_i_vector	noise	18.27 ± 17.21	91.25 ± 6.37	92.13 ± 8.17	79.93 ± 6.43	90.87 ± 6.00

Continued on next page

Table G5.3.3 (continued)

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
innovation_only	clean	21.44 ± 20.66	92.47 ± 6.21	90.98 ± 9.34	82.78 ± 6.84	91.71 ± 5.48
innovation_only	noise	21.74 ± 20.65	92.67 ± 6.14	90.91 ± 9.29	82.06 ± 6.38	91.68 ± 5.24
no_excess	clean	21.44 ± 20.66	92.47 ± 6.21	90.98 ± 9.34	82.78 ± 6.84	91.71 ± 5.48
no_excess	noise	21.74 ± 20.65	92.67 ± 6.14	90.91 ± 9.29	82.06 ± 6.38	91.68 ± 5.24
no_i_vector_no_excess	clean	27.66 ± 22.57	94.42 ± 5.53	89.31 ± 9.44	77.58 ± 9.84	93.66 ± 4.84
no_i_vector_no_excess	noise	27.65 ± 22.62	94.34 ± 5.35	89.32 ± 9.45	77.44 ± 9.52	93.37 ± 4.45
no_train_envelope	clean	0.00 ± 0.00	99.42 ± 1.09	100.00 ± 0.00	99.42 ± 1.09	99.42 ± 1.09
no_train_envelope	noise	7.78 ± 8.37	87.41 ± 6.29	96.37 ± 4.33	77.40 ± 6.65	87.14 ± 6.01
no_envelope_input	clean	0.00 ± 0.00	99.73 ± 0.77	100.00 ± 0.00	99.73 ± 0.77	99.73 ± 0.77
no_envelope_input	noise	16.63 ± 15.73	90.64 ± 6.53	92.69 ± 7.72	79.46 ± 6.78	90.27 ± 6.16

Table G5.3.4: Distance and error direction. Metric panel 1 of 2; identifying columns are repeated.

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
C	clean	0.0595 ± 0.0057	0.0771 ± 0.0088	0.0054 ± 0.0094	53.77 ± 5.89
C	noise	0.1285 ± 0.0347	0.1805 ± 0.0580	0.0698 ± 0.0375	56.66 ± 5.89
A	clean	0.1709 ± 0.0468	0.2340 ± 0.0667	0.1381 ± 0.0530	64.35 ± 6.95
A	noise	0.2045 ± 0.0189	0.2612 ± 0.0239	-0.1606 ± 0.0208	28.92 ± 5.29
E	clean	0.0595 ± 0.0057	0.0771 ± 0.0088	0.0054 ± 0.0094	53.77 ± 5.89
E	noise	0.1293 ± 0.0358	0.1822 ± 0.0601	0.0704 ± 0.0384	56.74 ± 6.04
raw_clean	clean	0.1189 ± 0.0421	0.1596 ± 0.0623	0.0534 ± 0.0450	56.28 ± 7.40
raw_clean	noise	0.1212 ± 0.0432	0.1623 ± 0.0639	0.0550 ± 0.0464	56.80 ± 7.24
raw_noise	clean	0.1195 ± 0.0415	0.1603 ± 0.0615	0.0540 ± 0.0444	56.37 ± 7.25

Continued on next page

Table G5.3.4 (continued)

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
raw_noise	noise	0.1213 ± 0.0429	0.1625 ± 0.0635	0.0552 ± 0.0459	56.75 ± 7.24
mlp32_clean	clean	0.1217 ± 0.0243	0.1652 ± 0.0300	0.0724 ± 0.0289	58.18 ± 8.62
mlp32_clean	noise	0.1213 ± 0.0227	0.1645 ± 0.0270	0.0731 ± 0.0279	58.95 ± 8.52
mlp32_noise	clean	0.1207 ± 0.0239	0.1639 ± 0.0292	0.0713 ± 0.0283	58.12 ± 8.32
mlp32_noise	noise	0.1208 ± 0.0226	0.1639 ± 0.0268	0.0729 ± 0.0277	58.97 ± 8.49
dd2_clean	clean	0.1142 ± 0.0234	0.1563 ± 0.0308	0.0930 ± 0.0227	66.43 ± 5.49
dd2_clean	noise	0.1313 ± 0.0187	0.1695 ± 0.0252	-0.0776 ± 0.0200	40.94 ± 4.79
dd2_noise	clean	0.1282 ± 0.0256	0.1794 ± 0.0343	0.1023 ± 0.0220	64.13 ± 5.49
dd2_noise	noise	0.0872 ± 0.0118	0.1145 ± 0.0133	0.0169 ± 0.0198	55.32 ± 6.03
no_i_vector	clean	0.0595 ± 0.0057	0.0771 ± 0.0088	0.0054 ± 0.0094	53.77 ± 5.89
no_i_vector	noise	0.1247 ± 0.0340	0.1750 ± 0.0573	0.0656 ± 0.0361	56.48 ± 5.43
innovation_only	clean	0.1182 ± 0.0326	0.1667 ± 0.0568	0.0555 ± 0.0392	56.74 ± 6.95
innovation_only	noise	0.1174 ± 0.0306	0.1659 ± 0.0543	0.0545 ± 0.0368	56.48 ± 6.27
no_excess	clean	0.1182 ± 0.0326	0.1667 ± 0.0568	0.0555 ± 0.0392	56.74 ± 6.95
no_excess	noise	0.1174 ± 0.0306	0.1659 ± 0.0543	0.0545 ± 0.0368	56.48 ± 6.27
no_i_vector_no_excess	clean	0.1046 ± 0.0242	0.1492 ± 0.0470	0.0362 ± 0.0269	54.61 ± 8.28
no_i_vector_no_excess	noise	0.1064 ± 0.0211	0.1517 ± 0.0430	0.0379 ± 0.0243	54.98 ± 8.23
no_train_envelope	clean	0.0595 ± 0.0057	0.0771 ± 0.0088	0.0067 ± 0.0104	54.08 ± 5.88
no_train_envelope	noise	0.1413 ± 0.0291	0.1981 ± 0.0491	0.0899 ± 0.0310	59.11 ± 5.58
no_envelope_input	clean	0.0595 ± 0.0057	0.0771 ± 0.0088	0.0054 ± 0.0094	53.77 ± 5.89
no_envelope_input	noise	0.1285 ± 0.0347	0.1805 ± 0.0580	0.0698 ± 0.0375	56.66 ± 5.89

Table G5.3.4: Distance and error direction. Metric panel 2 of 2; identifying columns are repeated.

Source key	Condition	Early (%)	Miss (%)	False boundary (%)
C	clean	46.23 ± 5.89	0.27 ± 0.77	0.00 ± 0.00
C	noise	43.34 ± 5.89	9.36 ± 6.53	16.63 ± 15.73
A	clean	35.65 ± 6.95	20.02 ± 9.00	0.00 ± 0.00
A	noise	71.08 ± 5.29	0.00 ± 0.00	56.62 ± 17.66
E	clean	46.23 ± 5.89	0.27 ± 0.77	0.00 ± 0.00

Continued on next page

Table G5.3.4 (continued)

Source key	Condition	Early (%)	Miss (%)	False boundary (%)
E	noise	43.26 ± 6.04	9.50 ± 6.66	18.25 ± 17.29
raw_clean	clean	43.72 ± 7.40	6.55 ± 6.73	29.65 ± 11.79
raw_clean	noise	43.20 ± 7.24	6.57 ± 6.91	29.55 ± 11.79
raw_noise	clean	43.63 ± 7.25	6.55 ± 6.73	29.44 ± 11.92
raw_noise	noise	43.25 ± 7.24	6.58 ± 6.88	29.54 ± 11.82
mlp32_clean	clean	41.82 ± 8.62	11.06 ± 4.34	2.21 ± 4.45
mlp32_clean	noise	41.05 ± 8.52	11.13 ± 4.30	2.38 ± 4.71
mlp32_noise	clean	41.88 ± 8.32	10.88 ± 4.27	2.57 ± 5.09
mlp32_noise	noise	41.03 ± 8.49	11.07 ± 4.31	2.46 ± 4.76
dd2_clean	clean	33.57 ± 5.49	16.87 ± 5.59	0.00 ± 0.00
dd2_clean	noise	59.06 ± 4.79	0.13 ± 0.18	55.33 ± 18.21
dd2_noise	clean	35.87 ± 5.49	15.74 ± 6.22	0.00 ± 0.00
dd2_noise	noise	44.68 ± 6.03	0.29 ± 0.38	52.82 ± 19.73
no_i_vector	clean	46.23 ± 5.89	0.27 ± 0.77	0.00 ± 0.00
no_i_vector	noise	43.52 ± 5.43	8.75 ± 6.37	18.27 ± 17.21
innovation_only	clean	43.26 ± 6.95	7.53 ± 6.21	21.44 ± 20.66
innovation_only	noise	43.52 ± 6.27	7.33 ± 6.14	21.74 ± 20.65
no_excess	clean	43.26 ± 6.95	7.53 ± 6.21	21.44 ± 20.66
no_excess	noise	43.52 ± 6.27	7.33 ± 6.14	21.74 ± 20.65
no_i_vector_no_excess	clean	45.39 ± 8.28	5.58 ± 5.53	27.66 ± 22.57
no_i_vector_no_excess	noise	45.02 ± 8.23	5.66 ± 5.35	27.65 ± 22.62
no_train_envelope	clean	45.92 ± 5.88	0.58 ± 1.09	0.00 ± 0.00
no_train_envelope	noise	40.89 ± 5.58	12.59 ± 6.29	7.78 ± 8.37
no_envelope_input	clean	46.23 ± 5.89	0.27 ± 0.77	0.00 ± 0.00
no_envelope_input	noise	43.34 ± 5.89	9.36 ± 6.53	16.63 ± 15.73

G.4.4 BOUNCINGBALL

Tables G5.4.1, G5.4.2, G5.4.3, and G5.4.4 report presence ranking and calibration, probability scores, fixed-threshold outcomes, and distance/direction metrics, respectively, for all versions and both observation conditions.

Table G5.4.1: Presence ranking and calibration.

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
C	clean	0.9498 ± 0.0663	0.9302 ± 0.0944	0.0424 ± 0.0462	0.0444 ± 0.0317	-0.0035 ± 0.0222
C	noise	0.9471 ± 0.0703	0.9264 ± 0.0997	0.0418 ± 0.0474	0.0498 ± 0.0379	0.0009 ± 0.0202
A	clean	0.9516 ± 0.0654	0.9326 ± 0.0933	0.0416 ± 0.0461	0.0463 ± 0.0325	-0.0020 ± 0.0221

Continued on next page

Table G5.4.1 (continued)

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
A	noise	0.9475 ± 0.0710	0.9269 ± 0.1008	0.0415 ± 0.0472	0.0497 ± 0.0383	-0.0002 ± 0.0201
E	clean	0.9453 ± 0.0663	0.9246 ± 0.0939	0.0444 ± 0.0445	0.0491 ± 0.0360	0.0029 ± 0.0187
E	noise	0.9453 ± 0.0706	0.9244 ± 0.0992	0.0439 ± 0.0458	0.0539 ± 0.0359	0.0112 ± 0.0164
raw_clean	clean	0.9440 ± 0.0735	0.9224 ± 0.1039	0.0408 ± 0.0471	0.0488 ± 0.0350	-0.0012 ± 0.0156
raw_clean	noise	0.9441 ± 0.0734	0.9225 ± 0.1038	0.0408 ± 0.0471	0.0491 ± 0.0352	-0.0012 ± 0.0155
raw_noise	clean	0.9440 ± 0.0735	0.9224 ± 0.1039	0.0408 ± 0.0471	0.0488 ± 0.0350	-0.0012 ± 0.0156
raw_noise	noise	0.9441 ± 0.0734	0.9225 ± 0.1038	0.0408 ± 0.0471	0.0491 ± 0.0353	-0.0012 ± 0.0155
mlp32_clean	clean	0.9482 ± 0.0632	0.9285 ± 0.0882	0.0447 ± 0.0382	0.0648 ± 0.0216	0.0278 ± 0.0190
mlp32_clean	noise	0.9479 ± 0.0635	0.9280 ± 0.0886	0.0446 ± 0.0382	0.0644 ± 0.0211	0.0278 ± 0.0190
mlp32_noise	clean	0.9482 ± 0.0632	0.9285 ± 0.0882	0.0446 ± 0.0382	0.0648 ± 0.0216	0.0278 ± 0.0190
mlp32_noise	noise	0.9479 ± 0.0635	0.9280 ± 0.0886	0.0446 ± 0.0382	0.0644 ± 0.0211	0.0278 ± 0.0190
dd2_clean	clean	0.5508 ± 0.0935	0.4243 ± 0.1038	0.1435 ± 0.0405	0.1220 ± 0.0252	-0.0408 ± 0.0731
dd2_clean	noise	0.5879 ± 0.0953	0.4701 ± 0.1201	0.1393 ± 0.0381	0.1403 ± 0.0182	-0.0055 ± 0.0736
dd2_noise	clean	0.5416 ± 0.1006	0.4133 ± 0.1090	0.1448 ± 0.0417	0.1245 ± 0.0277	-0.0516 ± 0.0728
dd2_noise	noise	0.5819 ± 0.1045	0.4637 ± 0.1248	0.1395 ± 0.0393	0.1386 ± 0.0214	-0.0175 ± 0.0733
no_i_vector	clean	0.9499 ± 0.0680	0.9304 ± 0.0968	0.0409 ± 0.0447	0.0495 ± 0.0300	-0.0031 ± 0.0185
no_i_vector	noise	0.9475 ± 0.0700	0.9271 ± 0.0993	0.0404 ± 0.0459	0.0507 ± 0.0329	0.0026 ± 0.0159
innovation_only	clean	0.9412 ± 0.0763	0.9184 ± 0.1075	0.0455 ± 0.0442	0.0660 ± 0.0287	0.0215 ± 0.0192

Continued on next page

Table G5.4.1 (continued)

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
innovation_only	noise	0.9414 ± 0.0762	0.9187 ± 0.1075	0.0456 ± 0.0442	0.0660 ± 0.0284	0.0215 ± 0.0193
no_excess	clean	0.9436 ± 0.0727	0.9217 ± 0.1029	0.0426 ± 0.0483	0.0508 ± 0.0431	-0.0014 ± 0.0204
no_excess	noise	0.9433 ± 0.0730	0.9214 ± 0.1033	0.0426 ± 0.0483	0.0504 ± 0.0428	-0.0014 ± 0.0204
no_i_vector_no_excess	clean	0.9470 ± 0.0689	0.9267 ± 0.0968	0.0389 ± 0.0428	0.0491 ± 0.0295	0.0010 ± 0.0161
no_i_vector_no_excess	noise	0.9468 ± 0.0690	0.9265 ± 0.0970	0.0389 ± 0.0428	0.0494 ± 0.0294	0.0010 ± 0.0160
no_train_envelope	clean	0.9492 ± 0.0678	0.9293 ± 0.0966	0.0421 ± 0.0470	0.0515 ± 0.0427	-0.0027 ± 0.0216
no_train_envelope	noise	0.9469 ± 0.0708	0.9261 ± 0.1005	0.0420 ± 0.0480	0.0514 ± 0.0400	0.0010 ± 0.0199
no_envelope_input	clean	0.9476 ± 0.0697	0.9270 ± 0.0993	0.0437 ± 0.0450	0.0563 ± 0.0366	0.0069 ± 0.0210
no_envelope_input	noise	0.9454 ± 0.0726	0.9241 ± 0.1031	0.0439 ± 0.0459	0.0584 ± 0.0330	0.0111 ± 0.0193

Table G5.4.2: Joint and interval probabilities.

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
C	clean	0.2830 ± 0.2474	0.1110 ± 0.0926	0.8890 ± 0.6829	0.0662 ± 0.0541
C	noise	0.2921 ± 0.2539	0.1159 ± 0.0984	0.9175 ± 0.7026	0.0693 ± 0.0583
A	clean	0.2758 ± 0.2504	0.1077 ± 0.0940	0.8540 ± 0.7020	0.0655 ± 0.0527
A	noise	0.3004 ± 0.2645	0.1218 ± 0.1061	0.9529 ± 0.7459	0.0676 ± 0.0566
E	clean	0.2896 ± 0.2406	0.1158 ± 0.0905	0.8795 ± 0.6454	0.0812 ± 0.0662
E	noise	0.2990 ± 0.2477	0.1209 ± 0.0966	0.8966 ± 0.6644	0.0878 ± 0.0706
raw_clean	clean	0.3171 ± 0.2632	0.1233 ± 0.1032	1.0051 ± 0.6613	0.0678 ± 0.0615
raw_clean	noise	0.3171 ± 0.2633	0.1233 ± 0.1032	1.0052 ± 0.6610	0.0679 ± 0.0617
raw_noise	clean	0.3170 ± 0.2632	0.1233 ± 0.1032	1.0049 ± 0.6611	0.0678 ± 0.0615
raw_noise	noise	0.3171 ± 0.2633	0.1233 ± 0.1032	1.0050 ± 0.6608	0.0679 ± 0.0617

Continued on next page

Table G5.4.2 (continued)

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
mlp32_clean	clean	0.3319 ± 0.2150	0.1385 ± 0.0927	0.9332 ± 0.4319	0.1185 ± 0.0750
mlp32_clean	noise	0.3320 ± 0.2150	0.1385 ± 0.0926	0.9335 ± 0.4315	0.1184 ± 0.0750
mlp32_noise	clean	0.3319 ± 0.2151	0.1385 ± 0.0927	0.9333 ± 0.4320	0.1184 ± 0.0750
mlp32_noise	noise	0.3320 ± 0.2150	0.1385 ± 0.0927	0.9336 ± 0.4315	0.1184 ± 0.0750
dd2_clean	clean	0.5932 ± 0.1847	0.3209 ± 0.0967	1.9960 ± 0.2411	0.1712 ± 0.0283
dd2_clean	noise	0.6020 ± 0.1757	0.3160 ± 0.0927	1.8896 ± 0.2380	0.2127 ± 0.0300
dd2_noise	clean	0.5946 ± 0.1878	0.3231 ± 0.0987	2.0485 ± 0.2401	0.1585 ± 0.0268
dd2_noise	noise	0.5977 ± 0.1780	0.3160 ± 0.0947	1.9244 ± 0.2366	0.1977 ± 0.0286
no_i_vector	clean	0.2842 ± 0.2437	0.1105 ± 0.0912	0.9032 ± 0.6523	0.0656 ± 0.0562
no_i_vector	noise	0.2929 ± 0.2487	0.1152 ± 0.0961	0.9267 ± 0.6633	0.0701 ± 0.0609
innovation_only	clean	0.2989 ± 0.2264	0.1235 ± 0.0983	0.8289 ± 0.5512	0.1076 ± 0.0669
innovation_only	noise	0.2992 ± 0.2265	0.1237 ± 0.0983	0.8295 ± 0.5513	0.1078 ± 0.0668
no_excess	clean	0.3070 ± 0.2595	0.1225 ± 0.1013	0.9838 ± 0.7014	0.0682 ± 0.0597
no_excess	noise	0.3073 ± 0.2598	0.1227 ± 0.1014	0.9844 ± 0.7018	0.0684 ± 0.0599
no_i_vector_no_excess	clean	0.3074 ± 0.2458	0.1209 ± 0.0957	1.0101 ± 0.6630	0.0659 ± 0.0538
no_i_vector_no_excess	noise	0.3075 ± 0.2458	0.1209 ± 0.0957	1.0101 ± 0.6626	0.0660 ± 0.0540
no_train_envelope	clean	0.2844 ± 0.2502	0.1106 ± 0.0936	0.8740 ± 0.6632	0.0665 ± 0.0578
no_train_envelope	noise	0.2951 ± 0.2560	0.1163 ± 0.0985	0.9077 ± 0.6769	0.0703 ± 0.0613
no_envelope_input	clean	0.2800 ± 0.2332	0.1116 ± 0.0926	0.8168 ± 0.6027	0.0833 ± 0.0587
no_envelope_input	noise	0.2956 ± 0.2408	0.1190 ± 0.0982	0.8667 ± 0.6194	0.0884 ± 0.0620

Table G5.4.3: Fixed-threshold detection and interval hits.

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
C	clean	11.55 ± 6.97	93.52 ± 7.41	71.65 ± 9.99	82.20 ± 10.57	89.97 ± 11.78
C	noise	12.33 ± 7.98	93.52 ± 7.35	70.89 ± 10.56	80.90 ± 10.22	89.35 ± 10.95
A	clean	10.82 ± 6.84	92.50 ± 8.59	72.91 ± 9.99	83.14 ± 11.47	90.04 ± 11.97

Continued on next page

Table G5.4.3 (continued)

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
A	noise	11.24 ± 7.42	92.82 ± 8.25	72.51 ± 10.83	78.07 ± 14.86	89.15 ± 11.01
E	clean	13.24 ± 7.79	93.41 ± 7.31	68.68 ± 11.13	81.55 ± 11.18	89.67 ± 12.62
E	noise	14.56 ± 9.50	93.85 ± 6.70	67.54 ± 11.99	80.49 ± 10.93	89.18 ± 11.03
raw_clean	clean	11.45 ± 8.45	91.75 ± 9.37	72.18 ± 13.02	77.74 ± 12.28	88.23 ± 12.26
raw_clean	noise	11.64 ± 8.69	91.82 ± 9.26	72.00 ± 13.32	77.90 ± 12.33	88.23 ± 12.26
raw_noise	clean	11.45 ± 8.45	91.75 ± 9.37	72.18 ± 13.02	77.74 ± 12.28	88.23 ± 12.26
raw_noise	noise	11.63 ± 8.68	91.85 ± 9.21	72.02 ± 13.30	77.90 ± 12.33	88.23 ± 12.26
mlp32_clean	clean	12.49 ± 10.40	92.49 ± 8.43	71.33 ± 12.62	77.38 ± 11.01	87.78 ± 12.88
mlp32_clean	noise	12.57 ± 10.41	92.49 ± 8.43	71.17 ± 12.42	77.25 ± 11.18	87.66 ± 13.10
mlp32_noise	clean	12.56 ± 10.32	92.49 ± 8.43	71.07 ± 12.27	77.38 ± 11.01	87.78 ± 12.88
mlp32_noise	noise	12.60 ± 10.37	92.49 ± 8.43	71.08 ± 12.29	77.25 ± 11.18	87.66 ± 13.10
dd2_clean	clean	47.92 ± 8.50	91.52 ± 7.03	35.00 ± 9.13	71.09 ± 7.90	79.49 ± 9.05
dd2_clean	noise	60.04 ± 9.50	99.57 ± 0.77	31.99 ± 8.81	70.80 ± 9.52	87.72 ± 6.70
dd2_noise	clean	47.81 ± 8.23	91.43 ± 7.73	35.07 ± 9.45	71.96 ± 7.65	79.72 ± 8.64
dd2_noise	noise	60.43 ± 9.42	99.69 ± 0.53	31.86 ± 8.76	73.16 ± 9.20	88.08 ± 6.52
no_i_vector	clean	9.02 ± 7.23	91.33 ± 9.94	77.29 ± 12.79	80.18 ± 10.97	89.03 ± 13.18
no_i_vector	noise	9.31 ± 7.64	91.30 ± 9.99	76.89 ± 13.39	78.97 ± 10.53	88.07 ± 12.43
innovation_only	clean	18.82 ± 9.89	94.85 ± 6.01	60.62 ± 10.30	80.19 ± 10.45	89.57 ± 10.63
innovation_only	noise	18.62 ± 9.79	95.18 ± 5.76	60.96 ± 10.29	80.07 ± 10.80	89.41 ± 10.92
no_excess	clean	16.05 ± 10.03	94.88 ± 5.72	65.20 ± 11.54	79.77 ± 11.26	89.15 ± 11.40
no_excess	noise	16.06 ± 9.89	94.88 ± 5.72	65.09 ± 11.47	79.90 ± 11.21	89.18 ± 11.35
no_i_vector_no_excess	clean	10.78 ± 8.78	91.43 ± 9.88	73.78 ± 14.14	77.64 ± 12.90	87.98 ± 12.50
no_i_vector_no_excess	noise	10.76 ± 8.78	91.43 ± 9.88	73.84 ± 14.15	77.79 ± 12.28	87.98 ± 12.50
no_train_envelope	clean	12.28 ± 8.12	93.41 ± 7.31	70.84 ± 11.65	81.79 ± 10.92	89.70 ± 12.77

Continued on next page

Table G5.4.3 (continued)

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
no_train_envelope	noise	13.12 ± 8.72	93.81 ± 6.83	69.76 ± 11.65	80.52 ± 10.70	88.85 ± 11.75
no_envelope_input	clean	12.81 ± 7.54	93.73 ± 7.04	69.13 ± 10.19	82.73 ± 11.02	90.12 ± 11.94
no_envelope_input	noise	13.49 ± 8.37	93.86 ± 6.94	68.45 ± 10.53	81.71 ± 11.11	89.60 ± 11.22

Table G5.4.4: Distance and error direction. Metric panel 1 of 2; identifying columns are repeated.

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
C	clean	0.1184 ± 0.0605	0.1592 ± 0.0843	0.0360 ± 0.0554	53.53 ± 9.74
C	noise	0.1199 ± 0.0607	0.1610 ± 0.0846	0.0360 ± 0.0580	53.29 ± 11.06
A	clean	0.1177 ± 0.0639	0.1572 ± 0.0882	0.0387 ± 0.0573	52.91 ± 8.78
A	noise	0.1229 ± 0.0683	0.1638 ± 0.0937	0.0411 ± 0.0647	53.31 ± 10.38
E	clean	0.1209 ± 0.0642	0.1622 ± 0.0886	0.0346 ± 0.0577	52.97 ± 9.97
E	noise	0.1213 ± 0.0621	0.1626 ± 0.0860	0.0306 ± 0.0589	51.96 ± 11.28
raw_clean	clean	0.1245 ± 0.0685	0.1661 ± 0.0929	0.0430 ± 0.0725	51.75 ± 11.17
raw_clean	noise	0.1244 ± 0.0682	0.1658 ± 0.0924	0.0431 ± 0.0718	51.91 ± 10.84
raw_noise	clean	0.1245 ± 0.0685	0.1661 ± 0.0929	0.0430 ± 0.0725	51.75 ± 11.17
raw_noise	noise	0.1243 ± 0.0680	0.1657 ± 0.0922	0.0430 ± 0.0716	51.91 ± 10.84
mlp32_clean	clean	0.1259 ± 0.0672	0.1690 ± 0.0915	0.0352 ± 0.0644	51.26 ± 9.72
mlp32_clean	noise	0.1263 ± 0.0677	0.1695 ± 0.0921	0.0347 ± 0.0645	51.17 ± 9.72
mlp32_noise	clean	0.1259 ± 0.0672	0.1690 ± 0.0915	0.0352 ± 0.0644	51.26 ± 9.72
mlp32_noise	noise	0.1263 ± 0.0677	0.1695 ± 0.0921	0.0347 ± 0.0645	51.17 ± 9.72
dd2_clean	clean	0.1820 ± 0.0477	0.2398 ± 0.0630	0.1186 ± 0.0529	67.83 ± 7.76
dd2_clean	noise	0.1614 ± 0.0333	0.2106 ± 0.0427	0.1120 ± 0.0551	69.70 ± 12.78
dd2_noise	clean	0.1735 ± 0.0423	0.2267 ± 0.0576	0.1088 ± 0.0515	66.46 ± 7.10
dd2_noise	noise	0.1547 ± 0.0306	0.2016 ± 0.0392	0.1027 ± 0.0507	68.34 ± 11.65
no_i_vector	clean	0.1245 ± 0.0683	0.1683 ± 0.0951	0.0444 ± 0.0653	53.78 ± 9.17
no_i_vector	noise	0.1261 ± 0.0682	0.1703 ± 0.0951	0.0428 ± 0.0688	53.03 ± 11.12

Continued on next page

Table G5.4.4 (continued)

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
innovation_only	clean	0.1167 ± 0.0555	0.1567 ± 0.0775	0.0253 ± 0.0606	51.57 ± 12.14
innovation_only	noise	0.1163 ± 0.0550	0.1562 ± 0.0770	0.0223 ± 0.0581	51.23 ± 12.12
no_excess	clean	0.1191 ± 0.0589	0.1597 ± 0.0818	0.0220 ± 0.0572	50.47 ± 11.18
no_excess	noise	0.1190 ± 0.0588	0.1595 ± 0.0817	0.0231 ± 0.0597	51.03 ± 12.25
no_i_vector_no_excess	clean	0.1270 ± 0.0718	0.1706 ± 0.0988	0.0418 ± 0.0730	51.89 ± 10.27
no_i_vector_no_excess	noise	0.1269 ± 0.0712	0.1705 ± 0.0980	0.0420 ± 0.0726	51.93 ± 10.22
no_train_envelope	clean	0.1202 ± 0.0634	0.1604 ± 0.0864	0.0334 ± 0.0569	52.56 ± 9.70
no_train_envelope	noise	0.1214 ± 0.0628	0.1618 ± 0.0856	0.0308 ± 0.0592	51.95 ± 11.41
no_envelope_input	clean	0.1177 ± 0.0615	0.1568 ± 0.0839	0.0331 ± 0.0559	52.45 ± 10.00
no_envelope_input	noise	0.1188 ± 0.0615	0.1582 ± 0.0839	0.0338 ± 0.0600	52.71 ± 11.71

Table G5.4.4: Distance and error direction. Metric panel 2 of 2; identifying columns are repeated.

Source key	Condition	Early (%)	Miss (%)	False boundary (%)
C	clean	46.47 ± 9.74	6.48 ± 7.41	11.55 ± 6.97
C	noise	46.71 ± 11.06	6.48 ± 7.35	12.33 ± 7.98
A	clean	47.09 ± 8.78	7.50 ± 8.59	10.82 ± 6.84
A	noise	46.69 ± 10.38	7.18 ± 8.25	11.24 ± 7.42
E	clean	47.03 ± 9.97	6.59 ± 7.31	13.24 ± 7.79
E	noise	48.04 ± 11.28	6.15 ± 6.70	14.56 ± 9.50
raw_clean	clean	48.25 ± 11.17	8.25 ± 9.37	11.45 ± 8.45
raw_clean	noise	48.09 ± 10.84	8.18 ± 9.26	11.64 ± 8.69
raw_noise	clean	48.25 ± 11.17	8.25 ± 9.37	11.45 ± 8.45
raw_noise	noise	48.09 ± 10.84	8.15 ± 9.21	11.63 ± 8.68
mlp32_clean	clean	48.74 ± 9.72	7.51 ± 8.43	12.49 ± 10.40
mlp32_clean	noise	48.83 ± 9.72	7.51 ± 8.43	12.57 ± 10.41
mlp32_noise	clean	48.74 ± 9.72	7.51 ± 8.43	12.56 ± 10.32
mlp32_noise	noise	48.83 ± 9.72	7.51 ± 8.43	12.60 ± 10.37
dd2_clean	clean	32.17 ± 7.76	8.48 ± 7.03	47.92 ± 8.50
dd2_clean	noise	30.30 ± 12.78	0.43 ± 0.77	60.04 ± 9.50
dd2_noise	clean	33.54 ± 7.10	8.57 ± 7.73	47.81 ± 8.23
dd2_noise	noise	31.66 ± 11.65	0.31 ± 0.53	60.43 ± 9.42
no_i_vector	clean	46.22 ± 9.17	8.67 ± 9.94	9.02 ± 7.23
no_i_vector	noise	46.97 ± 11.12	8.70 ± 9.99	9.31 ± 7.64
innovation_only	clean	48.43 ± 12.14	5.15 ± 6.01	18.82 ± 9.89
innovation_only	noise	48.77 ± 12.12	4.82 ± 5.76	18.62 ± 9.79
no_excess	clean	49.53 ± 11.18	5.12 ± 5.72	16.05 ± 10.03
no_excess	noise	48.97 ± 12.25	5.12 ± 5.72	16.06 ± 9.89
no_i_vector_no_excess	clean	48.11 ± 10.27	8.57 ± 9.88	10.78 ± 8.78

Continued on next page

Table G5.4.4 (continued)

Source key	Condition	Early (%)	Miss (%)	False boundary (%)
no_i_vector_no_excess	noise	48.07 ± 10.22	8.57 ± 9.88	10.76 ± 8.78
no_train_envelope	clean	47.44 ± 9.70	6.59 ± 7.31	12.28 ± 8.12
no_train_envelope	noise	48.05 ± 11.41	6.19 ± 6.83	13.12 ± 8.72
no_envelope_input	clean	47.55 ± 10.00	6.27 ± 7.04	12.81 ± 7.54
no_envelope_input	noise	47.29 ± 11.71	6.14 ± 6.94	13.49 ± 8.37

G.4.5 StickSlipBlock

Tables G5.5.1, G5.5.2, G5.5.3, and G5.5.4 report presence ranking and calibration, probability scores, fixed-threshold outcomes, and distance/direction metrics, respectively, for all versions and both observation conditions.

Table G5.5.1: Presence ranking and calibration. Metric panel 1 of 2; identifying columns are repeated.

Source key	Condition	AP	nAP	Binary Brier
C	clean	1.0000 ± 0.0000	$1.0000 \pm 4.20 \times 10^{-17}$	0.0009 ± 0.0027
C	noise	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0054 ± 0.0082
A	clean	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0041 ± 0.0020
A	noise	0.9929 ± 0.0130	0.9839 ± 0.0293	0.2308 ± 0.0896
E	clean	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0010 ± 0.0029
E	noise	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0054 ± 0.0084
raw_clean	clean	0.9972 ± 0.0047	0.9931 ± 0.0108	0.0169 ± 0.0153
raw_clean	noise	0.9972 ± 0.0047	0.9932 ± 0.0107	0.0169 ± 0.0154
raw_noise	clean	0.9972 ± 0.0048	0.9931 ± 0.0109	0.0169 ± 0.0153
raw_noise	noise	0.9972 ± 0.0047	0.9931 ± 0.0107	0.0169 ± 0.0154
mlp32_clean	clean	0.9996 ± 0.0008	0.9987 ± 0.0025	0.0066 ± 0.0109
mlp32_clean	noise	0.9996 ± 0.0008	0.9987 ± 0.0024	0.0066 ± 0.0110
mlp32_noise	clean	0.9996 ± 0.0008	0.9987 ± 0.0025	0.0066 ± 0.0109
mlp32_noise	noise	0.9996 ± 0.0008	0.9987 ± 0.0024	0.0066 ± 0.0110
dd2_clean	clean	0.9930 ± 0.0131	0.9843 ± 0.0296	0.0215 ± 0.0175
dd2_clean	noise	0.9729 ± 0.0300	0.9281 ± 0.0840	0.2385 ± 0.0622
dd2_noise	clean	0.9930 ± 0.0131	0.9843 ± 0.0296	0.0201 ± 0.0169
dd2_noise	noise	0.9726 ± 0.0308	0.9274 ± 0.0861	0.2017 ± 0.0574
no_i_vector	clean	1.0000 ± 0.0000	$1.0000 \pm 4.20 \times 10^{-17}$	0.0010 ± 0.0029
no_i_vector	noise	0.9999 ± 0.0003	0.9996 ± 0.0008	0.0083 ± 0.0106
innovation_only	clean	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0046 ± 0.0069
innovation_only	noise	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0048 ± 0.0071
no_excess	clean	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0046 ± 0.0069
no_excess	noise	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0048 ± 0.0071
no_i_vector_no_excess	clean	0.9974 ± 0.0038	0.9937 ± 0.0088	0.0173 ± 0.0153
no_i_vector_no_excess	noise	0.9975 ± 0.0038	0.9937 ± 0.0087	0.0172 ± 0.0154

Continued on next page

Table G5.5.1 (continued)

Source key	Condition	AP	nAP	Binary Brier
no_train_envelope	clean	1.0000 ± 0.0000	$1.0000 \pm 4.20 \times 10^{-17}$	$1.12 \times 10^{-7} \pm 1.92 \times 10^{-7}$
no_train_envelope	noise	$1.0000 \pm 7.25 \times 10^{-6}$	$1.0000 \pm 2.11 \times 10^{-5}$	0.0059 ± 0.0085
no_envelope_input	clean	1.0000 ± 0.0000	$1.0000 \pm 4.20 \times 10^{-17}$	0.0009 ± 0.0027
no_envelope_input	noise	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0054 ± 0.0082

Table G5.5.1: Presence ranking and calibration. Metric panel 2 of 2; identifying columns are repeated.

Source key	Condition	ECE	Calibration bias
C	clean	0.0013 ± 0.0036	0.0013 ± 0.0036
C	noise	0.0076 ± 0.0101	$3.61 \times 10^{-5} \pm 0.0129$
A	clean	0.0313 ± 0.0092	0.0123 ± 0.0078
A	noise	0.2631 ± 0.0814	0.2612 ± 0.0812
E	clean	0.0014 ± 0.0038	0.0014 ± 0.0038
E	noise	0.0081 ± 0.0105	0.0007 ± 0.0135
raw_clean	clean	0.0206 ± 0.0159	-0.0012 ± 0.0239
raw_clean	noise	0.0206 ± 0.0160	-0.0011 ± 0.0239
raw_noise	clean	0.0206 ± 0.0159	-0.0012 ± 0.0238
raw_noise	noise	0.0206 ± 0.0160	-0.0012 ± 0.0239
mlp32_clean	clean	0.0088 ± 0.0112	0.0054 ± 0.0111
mlp32_clean	noise	0.0088 ± 0.0113	0.0054 ± 0.0111
mlp32_noise	clean	0.0088 ± 0.0112	0.0054 ± 0.0110
mlp32_noise	noise	0.0088 ± 0.0113	0.0054 ± 0.0111
dd2_clean	clean	0.0669 ± 0.0289	0.0360 ± 0.0183
dd2_clean	noise	0.2520 ± 0.0637	0.2456 ± 0.0542
dd2_noise	clean	0.0571 ± 0.0244	-0.0091 ± 0.0147
dd2_noise	noise	0.2243 ± 0.0616	0.2116 ± 0.0447
no_i_vector	clean	0.0014 ± 0.0038	0.0014 ± 0.0039
no_i_vector	noise	0.0110 ± 0.0122	0.0032 ± 0.0159
innovation_only	clean	0.0063 ± 0.0083	0.0003 ± 0.0107
innovation_only	noise	0.0065 ± 0.0085	0.0005 ± 0.0109
no_excess	clean	0.0063 ± 0.0083	0.0003 ± 0.0107
no_excess	noise	0.0065 ± 0.0085	0.0005 ± 0.0109
no_i_vector_no_excess	clean	0.0208 ± 0.0154	-0.0018 ± 0.0244
no_i_vector_no_excess	noise	0.0209 ± 0.0155	-0.0017 ± 0.0244
no_train_envelope	clean	$1.95 \times 10^{-5} \pm 2.49 \times 10^{-5}$	$-1.33 \times 10^{-6} \pm 3.24 \times 10^{-5}$
no_train_envelope	noise	0.0078 ± 0.0099	-0.0001 ± 0.0128
no_envelope_input	clean	0.0013 ± 0.0036	0.0013 ± 0.0036
no_envelope_input	noise	0.0076 ± 0.0101	$3.61 \times 10^{-5} \pm 0.0129$

Table G5.5.2: Joint and interval probabilities.

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
C	clean	0.0126 ± 0.0356	0.0048 ± 0.0136	0.0151 ± 0.0426	0.0079 ± 0.0223
C	noise	0.0730 ± 0.1293	0.0241 ± 0.0365	0.0866 ± 0.1303	0.0573 ± 0.1541
A	clean	0.1160 ± 0.0123	0.0260 ± 0.0083	0.1074 ± 0.0092	0.0674 ± 0.0170
A	noise	1.2909 ± 0.3167	0.5529 ± 0.0740	0.6459 ± 0.0788	2.2103 ± 0.6776
E	clean	0.0081 ± 0.0197	0.0046 ± 0.0105	0.0080 ± 0.0172	0.0088 ± 0.0250
E	noise	0.0711 ± 0.1278	0.0243 ± 0.0379	0.0799 ± 0.1229	0.0647 ± 0.1709
raw_clean	clean	0.1778 ± 0.2320	0.0654 ± 0.0787	0.2043 ± 0.2169	0.1549 ± 0.3159
raw_clean	noise	0.1782 ± 0.2332	0.0655 ± 0.0790	0.2047 ± 0.2179	0.1555 ± 0.3171
raw_noise	clean	0.1779 ± 0.2320	0.0654 ± 0.0786	0.2045 ± 0.2172	0.1548 ± 0.3155
raw_noise	noise	0.1782 ± 0.2332	0.0655 ± 0.0790	0.2048 ± 0.2181	0.1553 ± 0.3166
mlp32_clean	clean	0.0586 ± 0.1074	0.0215 ± 0.0349	0.0359 ± 0.0517	0.1057 ± 0.2195
mlp32_clean	noise	0.0593 ± 0.1087	0.0218 ± 0.0356	0.0367 ± 0.0536	0.1062 ± 0.2198
mlp32_noise	clean	0.0586 ± 0.1070	0.0214 ± 0.0348	0.0360 ± 0.0515	0.1056 ± 0.2189
mlp32_noise	noise	0.0593 ± 0.1083	0.0217 ± 0.0355	0.0368 ± 0.0534	0.1061 ± 0.2192
dd2_clean	clean	0.3678 ± 0.2965	0.1658 ± 0.1357	0.2000 ± 0.1178	0.1606 ± 0.0566
dd2_clean	noise	1.2680 ± 0.3157	0.4554 ± 0.1084	0.2732 ± 0.0810	2.4174 ± 0.3855
dd2_noise	clean	0.4088 ± 0.3401	0.1852 ± 0.1600	0.3040 ± 0.1442	0.0768 ± 0.0301
dd2_noise	noise	0.9830 ± 0.3393	0.4218 ± 0.1454	0.2953 ± 0.1314	1.6369 ± 0.3012
no_i_vector	clean	0.0030 ± 0.0085	0.0021 ± 0.0060	0.0003 ± 0.0009	0.0082 ± 0.0231
no_i_vector	noise	0.0859 ± 0.1341	0.0336 ± 0.0485	0.0851 ± 0.1211	0.1011 ± 0.2395
innovation_only	clean	0.0589 ± 0.1022	0.0185 ± 0.0270	0.0683 ± 0.0999	0.0503 ± 0.1332
innovation_only	noise	0.0595 ± 0.1031	0.0189 ± 0.0277	0.0689 ± 0.1011	0.0511 ± 0.1340
no_excess	clean	0.0589 ± 0.1022	0.0185 ± 0.0270	0.0683 ± 0.0999	0.0503 ± 0.1332
no_excess	noise	0.0595 ± 0.1031	0.0189 ± 0.0277	0.0689 ± 0.1011	0.0511 ± 0.1340
no_i_vector_no_excess	clean	0.1796 ± 0.2385	0.0665 ± 0.0809	0.2041 ± 0.2223	0.1592 ± 0.3207
no_i_vector_no_excess	noise	0.1799 ± 0.2396	0.0666 ± 0.0812	0.2044 ± 0.2233	0.1597 ± 0.3218

Continued on next page

Table G5.5.2 (continued)

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
no_train_envelope	clean	0.0090 ± 0.0254	0.0033 ± 0.0093	0.0137 ± 0.0386	$2.85 \times 10^{-5} \pm 6.00 \times 10^{-5}$
no_train_envelope	noise	0.0725 ± 0.1121	0.0247 ± 0.0345	0.0876 ± 0.1183	0.0579 ± 0.1553
no_envelope_input	clean	0.0126 ± 0.0356	0.0048 ± 0.0136	0.0151 ± 0.0426	0.0079 ± 0.0223
no_envelope_input	noise	0.0730 ± 0.1293	0.0241 ± 0.0365	0.0866 ± 0.1303	0.0573 ± 0.1541

Table G5.5.3: Fixed-threshold detection and interval hits.

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
C	clean	0.00 ± 0.00	98.69 ± 1.90	100.00 ± 0.00	98.39 ± 2.59	98.39 ± 2.59
C	noise	0.00 ± 0.00	85.60 ± 7.86	100.00 ± 0.00	84.30 ± 5.90	85.01 ± 7.25
A	clean	0.00 ± 0.00	88.45 ± 7.60	100.00 ± 0.00	88.22 ± 7.25	88.45 ± 7.60
A	noise	35.88 ± 15.57	98.24 ± 3.33	83.13 ± 7.73	74.21 ± 7.50	92.01 ± 4.91
E	clean	0.00 ± 0.00	98.69 ± 1.90	100.00 ± 0.00	98.38 ± 2.43	98.49 ± 2.35
E	noise	0.00 ± 0.00	85.93 ± 8.11	100.00 ± 0.00	84.59 ± 6.10	85.34 ± 7.51
raw_clean	clean	1.14 ± 3.21	92.50 ± 6.52	99.43 ± 1.61	89.80 ± 5.19	91.31 ± 5.78
raw_clean	noise	1.05 ± 2.98	92.27 ± 6.62	99.47 ± 1.49	89.57 ± 5.17	91.08 ± 5.84
raw_noise	clean	1.14 ± 3.21	92.50 ± 6.52	99.43 ± 1.61	89.80 ± 5.19	91.31 ± 5.78
raw_noise	noise	1.05 ± 2.98	92.24 ± 6.62	99.47 ± 1.49	89.54 ± 5.15	91.05 ± 5.84
mlp32_clean	clean	0.57 ± 1.61	92.28 ± 4.91	99.70 ± 0.85	91.67 ± 4.40	91.78 ± 4.57
mlp32_clean	noise	0.44 ± 1.25	92.29 ± 5.00	99.77 ± 0.66	91.68 ± 4.50	91.79 ± 4.67
mlp32_noise	clean	0.38 ± 1.07	92.28 ± 4.91	99.80 ± 0.56	91.67 ± 4.40	91.78 ± 4.57
mlp32_noise	noise	0.44 ± 1.25	92.29 ± 5.01	99.77 ± 0.66	91.68 ± 4.51	91.79 ± 4.69
dd2_clean	clean	0.00 ± 0.00	82.71 ± 9.26	100.00 ± 0.00	76.43 ± 7.00	82.42 ± 9.07
dd2_clean	noise	30.14 ± 4.60	97.29 ± 2.95	85.47 ± 2.80	90.14 ± 4.70	96.97 ± 3.09
dd2_noise	clean	0.00 ± 0.00	84.17 ± 8.77	100.00 ± 0.00	76.60 ± 7.34	83.18 ± 8.72
dd2_noise	noise	28.15 ± 4.24	97.06 ± 2.92	86.36 ± 2.43	89.10 ± 7.87	96.04 ± 4.16

Continued on next page

Table G5.5.3 (continued)

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
no_i_vector	clean	0.00 ± 0.00	99.60 ± 0.82	100.00 ± 0.00	99.60 ± 0.82	99.60 ± 0.82
no_i_vector	noise	0.53 ± 1.49	85.79 ± 7.82	99.73 ± 0.77	84.19 ± 5.24	85.19 ± 6.75
innovation_only	clean	0.00 ± 0.00	89.47 ± 5.77	100.00 ± 0.00	88.63 ± 5.35	88.97 ± 5.84
innovation_only	noise	0.00 ± 0.00	89.43 ± 5.82	100.00 ± 0.00	88.49 ± 5.21	88.91 ± 5.89
no_excess	clean	0.00 ± 0.00	89.47 ± 5.77	100.00 ± 0.00	88.63 ± 5.35	88.97 ± 5.84
no_excess	noise	0.00 ± 0.00	89.43 ± 5.82	100.00 ± 0.00	88.49 ± 5.21	88.91 ± 5.89
no_i_vector_no_excess	clean	1.34 ± 3.18	92.85 ± 6.47	99.33 ± 1.59	90.15 ± 5.19	91.66 ± 5.81
no_i_vector_no_excess	noise	1.44 ± 3.21	92.78 ± 6.51	99.28 ± 1.60	90.08 ± 5.19	91.59 ± 5.83
no_train_envelope	clean	0.00 ± 0.00	96.34 ± 2.27	100.00 ± 0.00	96.04 ± 2.35	96.04 ± 2.35
no_train_envelope	noise	0.17 ± 0.48	89.32 ± 6.23	99.91 ± 0.24	88.08 ± 4.12	88.86 ± 5.63
no_envelope_input	clean	0.00 ± 0.00	98.69 ± 1.90	100.00 ± 0.00	98.39 ± 2.59	98.39 ± 2.59
no_envelope_input	noise	0.00 ± 0.00	85.60 ± 7.86	100.00 ± 0.00	84.30 ± 5.90	85.01 ± 7.25

Table G5.5.4: Distance and error direction. Metric panel 1 of 2; identifying columns are repeated.

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
C	clean	0.0693 ± 0.0193	0.0909 ± 0.0232	0.0228 ± 0.0349	56.76 ± 10.90
C	noise	0.1316 ± 0.0443	0.1968 ± 0.0657	0.0978 ± 0.0469	65.29 ± 10.39
A	clean	0.1127 ± 0.0408	0.1613 ± 0.0617	0.0803 ± 0.0367	63.42 ± 9.57
A	noise	0.1018 ± 0.0261	0.1414 ± 0.0522	-0.0210 ± 0.0256	53.75 ± 6.43
E	clean	0.0688 ± 0.0183	0.0902 ± 0.0221	0.0219 ± 0.0335	56.65 ± 10.90
E	noise	0.1323 ± 0.0487	0.1979 ± 0.0722	0.0969 ± 0.0507	64.74 ± 10.42
raw_clean	clean	0.0900 ± 0.0302	0.1273 ± 0.0519	0.0537 ± 0.0363	60.24 ± 10.17
raw_clean	noise	0.0914 ± 0.0310	0.1297 ± 0.0523	0.0551 ± 0.0368	60.25 ± 10.11
raw_noise	clean	0.0900 ± 0.0302	0.1273 ± 0.0519	0.0537 ± 0.0363	60.24 ± 10.17
raw_noise	noise	0.0915 ± 0.0310	0.1300 ± 0.0523	0.0552 ± 0.0368	60.26 ± 10.08
mlp32_clean	clean	0.0864 ± 0.0257	0.1211 ± 0.0405	0.0533 ± 0.0289	61.39 ± 10.09

Continued on next page

Table G5.5.4 (continued)

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
mlp32_clean	noise	0.0867 ± 0.0260	0.1213 ± 0.0408	0.0535 ± 0.0294	61.30 ± 10.18
mlp32_noise	clean	0.0864 ± 0.0257	0.1211 ± 0.0405	0.0533 ± 0.0289	61.39 ± 10.09
mlp32_noise	noise	0.0867 ± 0.0260	0.1213 ± 0.0407	0.0535 ± 0.0294	61.30 ± 10.18
dd2_clean	clean	0.1331 ± 0.0363	0.1987 ± 0.0520	0.1141 ± 0.0397	74.07 ± 8.02
dd2_clean	noise	0.0880 ± 0.0187	0.1263 ± 0.0343	0.0428 ± 0.0322	61.57 ± 12.71
dd2_noise	clean	0.1405 ± 0.0474	0.2118 ± 0.0643	0.1154 ± 0.0518	72.39 ± 9.39
dd2_noise	noise	0.0957 ± 0.0299	0.1357 ± 0.0423	0.0511 ± 0.0451	62.48 ± 14.38
no_i_vector	clean	0.0609 ± 0.0132	0.0804 ± 0.0180	0.0144 ± 0.0217	56.76 ± 10.90
no_i_vector	noise	0.1315 ± 0.0361	0.1859 ± 0.0540	0.0993 ± 0.0370	64.23 ± 9.12
innovation_only	clean	0.1158 ± 0.0343	0.1768 ± 0.0512	0.0745 ± 0.0456	63.36 ± 10.50
innovation_only	noise	0.1153 ± 0.0352	0.1763 ± 0.0525	0.0741 ± 0.0460	63.32 ± 10.35
no_excess	clean	0.1158 ± 0.0343	0.1768 ± 0.0512	0.0745 ± 0.0456	63.36 ± 10.50
no_excess	noise	0.1153 ± 0.0352	0.1763 ± 0.0525	0.0741 ± 0.0460	63.32 ± 10.35
no_i_vector_no_excess	clean	0.0887 ± 0.0298	0.1243 ± 0.0494	0.0522 ± 0.0357	59.89 ± 9.80
no_i_vector_no_excess	noise	0.0894 ± 0.0301	0.1252 ± 0.0496	0.0529 ± 0.0360	59.94 ± 9.80
no_train_envelope	clean	0.0683 ± 0.0177	0.0918 ± 0.0238	0.0305 ± 0.0265	59.41 ± 9.81
no_train_envelope	noise	0.1071 ± 0.0348	0.1631 ± 0.0546	0.0699 ± 0.0373	63.26 ± 10.19
no_envelope_input	clean	0.0693 ± 0.0193	0.0909 ± 0.0232	0.0228 ± 0.0349	56.76 ± 10.90
no_envelope_input	noise	0.1316 ± 0.0443	0.1968 ± 0.0657	0.0978 ± 0.0469	65.29 ± 10.39

Table G5.5.4: Distance and error direction. Metric panel 2 of 2; identifying columns are repeated.

Source key	Condition	Early (%)	Miss (%)	False boundary (%)
C	clean	43.24 ± 10.90	1.31 ± 1.90	0.00 ± 0.00
C	noise	34.71 ± 10.39	14.40 ± 7.86	0.00 ± 0.00
A	clean	36.58 ± 9.57	11.55 ± 7.60	0.00 ± 0.00
A	noise	46.25 ± 6.43	1.76 ± 3.33	35.88 ± 15.57
E	clean	43.35 ± 10.90	1.31 ± 1.90	0.00 ± 0.00
E	noise	35.26 ± 10.42	14.07 ± 8.11	0.00 ± 0.00
raw_clean	clean	39.76 ± 10.17	7.50 ± 6.52	1.14 ± 3.21
raw_clean	noise	39.75 ± 10.11	7.73 ± 6.62	1.05 ± 2.98
raw_noise	clean	39.76 ± 10.17	7.50 ± 6.52	1.14 ± 3.21

Continued on next page

Table G5.5.4 (continued)

Source key	Condition	Early (%)	Miss (%)	False boundary (%)
raw_noise	noise	39.74 ± 10.08	7.76 ± 6.62	1.05 ± 2.98
mlp32_clean	clean	38.61 ± 10.09	7.72 ± 4.91	0.57 ± 1.61
mlp32_clean	noise	38.70 ± 10.18	7.71 ± 5.00	0.44 ± 1.25
mlp32_noise	clean	38.61 ± 10.09	7.72 ± 4.91	0.38 ± 1.07
mlp32_noise	noise	38.70 ± 10.18	7.71 ± 5.01	0.44 ± 1.25
dd2_clean	clean	25.93 ± 8.02	17.29 ± 9.26	0.00 ± 0.00
dd2_clean	noise	38.43 ± 12.71	2.71 ± 2.95	30.14 ± 4.60
dd2_noise	clean	27.61 ± 9.39	15.83 ± 8.77	0.00 ± 0.00
dd2_noise	noise	37.52 ± 14.38	2.94 ± 2.92	28.15 ± 4.24
no_i_vector	clean	43.24 ± 10.90	0.40 ± 0.82	0.00 ± 0.00
no_i_vector	noise	35.77 ± 9.12	14.21 ± 7.82	0.53 ± 1.49
innovation_only	clean	36.64 ± 10.50	10.53 ± 5.77	0.00 ± 0.00
innovation_only	noise	36.68 ± 10.35	10.57 ± 5.82	0.00 ± 0.00
no_excess	clean	36.64 ± 10.50	10.53 ± 5.77	0.00 ± 0.00
no_excess	noise	36.68 ± 10.35	10.57 ± 5.82	0.00 ± 0.00
no_i_vector_no_excess	clean	40.11 ± 9.80	7.15 ± 6.47	1.34 ± 3.18
no_i_vector_no_excess	noise	40.06 ± 9.80	7.22 ± 6.51	1.44 ± 3.21
no_train_envelope	clean	40.59 ± 9.81	3.66 ± 2.27	0.00 ± 0.00
no_train_envelope	noise	36.74 ± 10.19	10.68 ± 6.23	0.17 ± 0.48
no_envelope_input	clean	43.24 ± 10.90	1.31 ± 1.90	0.00 ± 0.00
no_envelope_input	noise	34.71 ± 10.39	14.40 ± 7.86	0.00 ± 0.00

G.4.6 PUCKWALL2D

Tables G5.6.1, G5.6.2, G5.6.3, and G5.6.4 report presence ranking and calibration, probability scores, fixed-threshold outcomes, and distance/direction metrics, respectively, for all versions and both observation conditions.

Table G5.6.1: Presence ranking and calibration.

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
C	clean	0.9443 ± 0.0306	0.9399 ± 0.0337	0.0166 ± 0.0068	0.0147 ± 0.0057	-0.0038 ± 0.0107
C	noise	0.9179 ± 0.0339	0.9114 ± 0.0377	0.0191 ± 0.0071	0.0226 ± 0.0064	0.0135 ± 0.0101
A	clean	0.9345 ± 0.0286	0.9295 ± 0.0309	0.0189 ± 0.0056	0.0188 ± 0.0101	0.0114 ± 0.0149
A	noise	0.8835 ± 0.0468	0.8743 ± 0.0512	0.0306 ± 0.0092	0.0593 ± 0.0133	0.0586 ± 0.0139
E	clean	0.9393 ± 0.0384	0.9345 ± 0.0420	0.0161 ± 0.0077	0.0142 ± 0.0071	-0.0094 ± 0.0101

Continued on next page

Table G5.6.1 (continued)

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
E	noise	0.9066 ± 0.0400	0.8994 ± 0.0441	0.0180 ± 0.0080	0.0183 ± 0.0062	0.0056 ± 0.0096
raw_clean	clean	0.7831 ± 0.0930	0.7665 ± 0.1017	0.0303 ± 0.0137	0.0211 ± 0.0084	0.0004 ± 0.0073
raw_clean	noise	0.7833 ± 0.0927	0.7667 ± 0.1014	0.0304 ± 0.0137	0.0212 ± 0.0083	0.0004 ± 0.0072
raw_noise	clean	0.7821 ± 0.0921	0.7654 ± 0.1007	0.0304 ± 0.0137	0.0211 ± 0.0084	0.0004 ± 0.0073
raw_noise	noise	0.7833 ± 0.0927	0.7666 ± 0.1014	0.0304 ± 0.0137	0.0213 ± 0.0084	0.0004 ± 0.0072
mlp32_clean	clean	0.8938 ± 0.0480	0.8854 ± 0.0518	0.0237 ± 0.0074	0.0314 ± 0.0067	0.0123 ± 0.0082
mlp32_clean	noise	0.8940 ± 0.0466	0.8857 ± 0.0502	0.0237 ± 0.0074	0.0310 ± 0.0076	0.0122 ± 0.0081
mlp32_noise	clean	0.8941 ± 0.0480	0.8858 ± 0.0518	0.0237 ± 0.0074	0.0314 ± 0.0092	0.0123 ± 0.0082
mlp32_noise	noise	0.8940 ± 0.0467	0.8856 ± 0.0504	0.0237 ± 0.0074	0.0309 ± 0.0077	0.0122 ± 0.0081
dd2_clean	clean	0.7092 ± 0.1444	0.6862 ± 0.1581	0.0391 ± 0.0175	0.0356 ± 0.0153	0.0219 ± 0.0126
dd2_clean	noise	0.6877 ± 0.1492	0.6631 ± 0.1630	0.0526 ± 0.0173	0.1118 ± 0.0134	0.1096 ± 0.0140
dd2_noise	clean	0.7331 ± 0.1542	0.7116 ± 0.1683	0.0359 ± 0.0160	0.0339 ± 0.0097	0.0102 ± 0.0102
dd2_noise	noise	0.6945 ± 0.1541	0.6703 ± 0.1683	0.0424 ± 0.0159	0.0777 ± 0.0107	0.0705 ± 0.0112
no_i_vector	clean	0.9734 ± 0.0208	0.9713 ± 0.0224	0.0105 ± 0.0046	0.0113 ± 0.0061	-0.0038 ± 0.0078
no_i_vector	noise	0.9525 ± 0.0271	0.9488 ± 0.0294	0.0133 ± 0.0055	0.0137 ± 0.0069	-0.0048 ± 0.0056
innovation_only	clean	0.8926 ± 0.0563	0.8843 ± 0.0618	0.0185 ± 0.0093	0.0190 ± 0.0089	-0.0052 ± 0.0099
innovation_only	noise	0.8922 ± 0.0551	0.8839 ± 0.0605	0.0186 ± 0.0093	0.0189 ± 0.0090	-0.0053 ± 0.0099
no_excess	clean	0.8731 ± 0.0564	0.8635 ± 0.0615	0.0200 ± 0.0099	0.0191 ± 0.0092	-0.0065 ± 0.0105

Continued on next page

Table G5.6.1 (continued)

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
no_excess	noise	0.8731 ± 0.0553	0.8636 ± 0.0600	0.0201 ± 0.0099	0.0189 ± 0.0088	-0.0065 ± 0.0106
no_i_vector_no_excess	clean	0.7741 ± 0.1114	0.7568 ± 0.1196	0.0304 ± 0.0136	0.0208 ± 0.0083	-0.0035 ± 0.0081
no_i_vector_no_excess	noise	0.7756 ± 0.1099	0.7584 ± 0.1182	0.0304 ± 0.0136	0.0208 ± 0.0079	-0.0035 ± 0.0081
no_train_envelope	clean	0.9475 ± 0.0276	0.9433 ± 0.0304	0.0166 ± 0.0059	0.0172 ± 0.0100	0.0094 ± 0.0146
no_train_envelope	noise	0.9369 ± 0.0330	0.9319 ± 0.0361	0.0201 ± 0.0068	0.0369 ± 0.0108	0.0323 ± 0.0137
no_envelope_input	clean	0.9490 ± 0.0279	0.9450 ± 0.0307	0.0163 ± 0.0059	0.0167 ± 0.0098	0.0088 ± 0.0140
no_envelope_input	noise	0.9402 ± 0.0323	0.9354 ± 0.0356	0.0198 ± 0.0068	0.0379 ± 0.0106	0.0335 ± 0.0132

Table G5.6.2: Joint and interval probabilities.

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
C	clean	0.0970 ± 0.0493	0.0456 ± 0.0189	1.0629 ± 0.4451	0.0205 ± 0.0104
C	noise	0.1251 ± 0.0570	0.0540 ± 0.0222	1.1858 ± 0.4352	0.0407 ± 0.0130
A	clean	0.0875 ± 0.0327	0.0443 ± 0.0159	0.6655 ± 0.2806	0.0419 ± 0.0189
A	noise	0.1783 ± 0.0540	0.0801 ± 0.0272	1.1553 ± 0.2539	0.1024 ± 0.0250
E	clean	0.1022 ± 0.0578	0.0452 ± 0.0195	1.2334 ± 0.5889	0.0130 ± 0.0083
E	noise	0.1286 ± 0.0654	0.0535 ± 0.0233	1.3690 ± 0.5705	0.0300 ± 0.0118
raw_clean	clean	0.1881 ± 0.0964	0.0769 ± 0.0329	2.0212 ± 0.7902	0.0409 ± 0.0140
raw_clean	noise	0.1881 ± 0.0964	0.0769 ± 0.0329	2.0203 ± 0.7911	0.0409 ± 0.0139
raw_noise	clean	0.1881 ± 0.0964	0.0769 ± 0.0329	2.0212 ± 0.7900	0.0409 ± 0.0140
raw_noise	noise	0.1881 ± 0.0964	0.0769 ± 0.0329	2.0204 ± 0.7909	0.0409 ± 0.0139
mlp32_clean	clean	0.1630 ± 0.0641	0.0668 ± 0.0253	1.5806 ± 0.5184	0.0482 ± 0.0135
mlp32_clean	noise	0.1630 ± 0.0641	0.0668 ± 0.0253	1.5808 ± 0.5186	0.0481 ± 0.0134
mlp32_noise	clean	0.1630 ± 0.0641	0.0668 ± 0.0253	1.5800 ± 0.5183	0.0482 ± 0.0135

Continued on next page

Table G5.6.2 (continued)

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
mlp32_noise	noise	0.1630 ± 0.0641	0.0668 ± 0.0252	1.5803 ± 0.5184	0.0481 ± 0.0134
dd2_clean	clean	0.1867 ± 0.0663	0.0963 ± 0.0363	1.5223 ± 0.2169	0.0832 ± 0.0367
dd2_clean	noise	0.2731 ± 0.0660	0.1138 ± 0.0366	1.3161 ± 0.2001	0.1916 ± 0.0434
dd2_noise	clean	0.1795 ± 0.0612	0.0915 ± 0.0339	1.6836 ± 0.2716	0.0631 ± 0.0262
dd2_noise	noise	0.2328 ± 0.0595	0.0997 ± 0.0340	1.5043 ± 0.2343	0.1341 ± 0.0308
no_i_vector	clean	0.0765 ± 0.0519	0.0351 ± 0.0175	0.8358 ± 0.4424	0.0130 ± 0.0089
no_i_vector	noise	0.0976 ± 0.0608	0.0433 ± 0.0214	1.0908 ± 0.4941	0.0154 ± 0.0092
innovation_only	clean	0.1546 ± 0.0810	0.0561 ± 0.0226	1.8701 ± 0.6824	0.0193 ± 0.0075
innovation_only	noise	0.1548 ± 0.0809	0.0561 ± 0.0225	1.8733 ± 0.6834	0.0193 ± 0.0075
no_excess	clean	0.1659 ± 0.0832	0.0579 ± 0.0230	2.0357 ± 0.7558	0.0195 ± 0.0078
no_excess	noise	0.1662 ± 0.0832	0.0579 ± 0.0229	2.0393 ± 0.7571	0.0195 ± 0.0078
no_i_vector_no_excess	clean	0.1854 ± 0.0952	0.0758 ± 0.0317	2.0446 ± 0.8209	0.0365 ± 0.0129
no_i_vector_no_excess	noise	0.1854 ± 0.0952	0.0758 ± 0.0317	2.0435 ± 0.8217	0.0365 ± 0.0128
no_train_envelope	clean	0.0912 ± 0.0392	0.0448 ± 0.0190	0.7890 ± 0.3239	0.0365 ± 0.0183
no_train_envelope	noise	0.1247 ± 0.0456	0.0536 ± 0.0222	0.9013 ± 0.2904	0.0637 ± 0.0201
no_envelope_input	clean	0.0899 ± 0.0384	0.0442 ± 0.0185	0.7860 ± 0.3162	0.0354 ± 0.0169
no_envelope_input	noise	0.1245 ± 0.0448	0.0532 ± 0.0217	0.8889 ± 0.2810	0.0645 ± 0.0192

Table G5.6.3: Fixed-threshold detection and interval hits.

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
C	clean	4.29 ± 2.53	95.26 ± 5.36	64.98 ± 9.74	80.03 ± 9.97	90.18 ± 6.28
C	noise	8.64 ± 3.53	95.48 ± 5.16	49.02 ± 6.77	75.13 ± 11.73	90.15 ± 6.46
A	clean	3.10 ± 2.01	93.74 ± 7.82	71.67 ± 9.60	85.31 ± 9.51	90.96 ± 6.92
A	noise	6.70 ± 3.06	93.38 ± 8.43	53.95 ± 7.74	63.83 ± 12.52	90.41 ± 7.68
E	clean	5.99 ± 3.77	94.11 ± 6.65	57.13 ± 12.88	79.32 ± 12.73	89.04 ± 7.65
E	noise	13.21 ± 4.84	95.48 ± 6.16	36.63 ± 6.64	74.41 ± 12.57	90.39 ± 7.23

Continued on next page

Table G5.6.3 (continued)

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
raw_clean	clean	18.38 ± 4.25	94.74 ± 5.92	27.99 ± 3.81	66.76 ± 14.99	89.79 ± 8.67
raw_clean	noise	18.35 ± 4.24	94.74 ± 5.92	28.03 ± 3.89	66.76 ± 14.99	89.71 ± 8.74
raw_noise	clean	18.38 ± 4.25	94.74 ± 5.92	27.99 ± 3.81	66.76 ± 14.99	89.79 ± 8.67
raw_noise	noise	18.36 ± 4.22	94.74 ± 5.92	28.01 ± 3.87	66.76 ± 14.99	89.71 ± 8.74
mlp32_clean	clean	7.87 ± 3.27	93.95 ± 4.75	49.39 ± 11.79	65.40 ± 18.13	89.41 ± 7.45
mlp32_clean	noise	7.94 ± 3.24	93.95 ± 4.75	49.09 ± 11.82	65.40 ± 18.13	89.41 ± 7.45
mlp32_noise	clean	7.87 ± 3.27	93.95 ± 4.75	49.39 ± 11.79	65.40 ± 18.13	89.41 ± 7.45
mlp32_noise	noise	7.92 ± 3.22	93.95 ± 4.75	49.12 ± 11.77	65.40 ± 18.13	89.41 ± 7.45
dd2_clean	clean	6.26 ± 2.61	88.40 ± 8.43	52.82 ± 8.52	53.98 ± 17.11	75.32 ± 17.05
dd2_clean	noise	33.63 ± 3.03	94.76 ± 5.78	26.37 ± 4.06	61.19 ± 11.35	81.87 ± 12.21
dd2_noise	clean	6.61 ± 2.68	90.83 ± 7.42	52.09 ± 9.45	56.62 ± 14.74	79.87 ± 14.24
dd2_noise	noise	48.33 ± 2.83	95.70 ± 5.05	20.35 ± 3.18	63.37 ± 10.62	84.85 ± 11.73
no_i_vector	clean	1.30 ± 0.84	94.30 ± 5.59	85.39 ± 7.59	80.58 ± 11.64	91.43 ± 6.88
no_i_vector	noise	1.24 ± 0.89	92.89 ± 5.50	86.25 ± 8.25	75.77 ± 10.29	89.12 ± 5.77
innovation_only	clean	19.83 ± 4.78	94.63 ± 6.91	26.35 ± 3.23	69.38 ± 12.67	90.68 ± 7.85
innovation_only	noise	19.94 ± 4.55	94.35 ± 6.70	26.20 ± 3.78	68.30 ± 11.60	90.12 ± 7.73
no_excess	clean	19.34 ± 4.95	93.70 ± 7.14	26.74 ± 3.73	67.62 ± 11.80	89.29 ± 8.42
no_excess	noise	19.37 ± 4.75	93.42 ± 6.90	26.65 ± 4.12	67.15 ± 11.49	88.94 ± 8.11
no_i_vector_no_excess	clean	17.09 ± 4.69	93.22 ± 5.75	29.62 ± 4.64	68.24 ± 15.06	91.34 ± 6.96
no_i_vector_no_excess	noise	17.10 ± 4.65	93.22 ± 5.75	29.60 ± 4.59	68.37 ± 15.10	91.26 ± 6.95
no_train_envelope	clean	3.13 ± 1.84	95.34 ± 6.61	71.39 ± 8.73	77.71 ± 12.16	89.32 ± 8.39
no_train_envelope	noise	4.11 ± 2.51	96.37 ± 5.19	66.51 ± 9.44	75.11 ± 11.76	89.05 ± 8.53
no_envelope_input	clean	3.06 ± 1.80	95.34 ± 6.61	71.79 ± 8.76	78.55 ± 12.87	89.32 ± 8.39
no_envelope_input	noise	4.27 ± 2.39	96.93 ± 4.95	65.70 ± 8.82	75.79 ± 12.73	90.29 ± 8.54

Table G5.6.4: Distance and error direction. Metric panel 1 of 2; identifying columns are repeated.

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
C	clean	0.1000 ± 0.0190	0.1462 ± 0.0340	0.0443 ± 0.0497	51.27 ± 21.53
C	noise	0.1045 ± 0.0211	0.1542 ± 0.0383	0.0432 ± 0.0471	50.59 ± 20.21
A	clean	0.0993 ± 0.0244	0.1439 ± 0.0456	0.0491 ± 0.0525	52.46 ± 20.50
A	noise	0.1401 ± 0.0296	0.1985 ± 0.0425	0.0618 ± 0.0470	56.86 ± 11.34
E	clean	0.1095 ± 0.0412	0.1572 ± 0.0530	0.0608 ± 0.0614	53.43 ± 20.12
E	noise	0.1095 ± 0.0314	0.1595 ± 0.0434	0.0490 ± 0.0536	51.96 ± 18.02
raw_clean	clean	0.1266 ± 0.0619	0.1776 ± 0.0794	0.0555 ± 0.0841	48.57 ± 17.97
raw_clean	noise	0.1269 ± 0.0622	0.1781 ± 0.0799	0.0558 ± 0.0844	48.65 ± 18.06
raw_noise	clean	0.1266 ± 0.0619	0.1776 ± 0.0794	0.0555 ± 0.0841	48.57 ± 17.97
raw_noise	noise	0.1269 ± 0.0622	0.1781 ± 0.0799	0.0558 ± 0.0844	48.65 ± 18.06
mlp32_clean	clean	0.1071 ± 0.0529	0.1535 ± 0.0713	0.0337 ± 0.0647	45.47 ± 19.29
mlp32_clean	noise	0.1071 ± 0.0529	0.1535 ± 0.0713	0.0337 ± 0.0647	45.47 ± 19.29
mlp32_noise	clean	0.1071 ± 0.0529	0.1535 ± 0.0713	0.0337 ± 0.0647	45.47 ± 19.29
mlp32_noise	noise	0.1069 ± 0.0527	0.1534 ± 0.0711	0.0336 ± 0.0646	45.47 ± 19.29
dd2_clean	clean	0.1978 ± 0.0648	0.2781 ± 0.1003	0.0820 ± 0.0899	56.88 ± 17.96
dd2_clean	noise	0.1701 ± 0.0488	0.2324 ± 0.0677	0.0623 ± 0.0868	56.05 ± 16.42
dd2_noise	clean	0.1890 ± 0.0575	0.2638 ± 0.0868	0.0700 ± 0.0880	54.90 ± 16.57
dd2_noise	noise	0.1650 ± 0.0515	0.2233 ± 0.0696	0.0540 ± 0.0898	54.10 ± 17.01
no_i_vector	clean	0.0997 ± 0.0355	0.1431 ± 0.0521	0.0550 ± 0.0577	55.43 ± 19.62
no_i_vector	noise	0.1097 ± 0.0269	0.1600 ± 0.0413	0.0591 ± 0.0594	56.91 ± 19.08
innovation_only	clean	0.1165 ± 0.0500	0.1646 ± 0.0608	0.0511 ± 0.0808	48.24 ± 21.97
innovation_only	noise	0.1193 ± 0.0491	0.1685 ± 0.0592	0.0519 ± 0.0812	47.71 ± 22.34
no_excess	clean	0.1275 ± 0.0613	0.1784 ± 0.0743	0.0570 ± 0.0969	47.40 ± 23.28
no_excess	noise	0.1285 ± 0.0613	0.1798 ± 0.0737	0.0574 ± 0.0956	46.95 ± 22.93
no_i_vector_no_excess	clean	0.1232 ± 0.0574	0.1685 ± 0.0729	0.0587 ± 0.0770	51.82 ± 17.87
no_i_vector_no_excess	noise	0.1231 ± 0.0576	0.1682 ± 0.0735	0.0582 ± 0.0773	51.82 ± 17.87

Continued on next page

Table G5.6.4 (continued)

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
no_train_envelope	clean	0.1039 ± 0.0191	0.1530 ± 0.0438	0.0298 ± 0.0663	45.79 ± 25.30
no_train_envelope	noise	0.1040 ± 0.0238	0.1535 ± 0.0446	0.0274 ± 0.0713	46.12 ± 26.63
no_envelope_input	clean	0.1036 ± 0.0220	0.1527 ± 0.0477	0.0312 ± 0.0681	47.14 ± 25.26
no_envelope_input	noise	0.1017 ± 0.0259	0.1490 ± 0.0442	0.0256 ± 0.0726	46.68 ± 26.87

Table G5.6.4: Distance and error direction. Metric panel 2 of 2; identifying columns are repeated.

Source key	Condition	Early (%)	Miss (%)	False boundary (%)
C	clean	48.73 ± 21.53	4.74 ± 5.36	4.29 ± 2.53
C	noise	49.41 ± 20.21	4.52 ± 5.16	8.64 ± 3.53
A	clean	47.54 ± 20.50	6.26 ± 7.82	3.10 ± 2.01
A	noise	43.14 ± 11.34	6.62 ± 8.43	6.70 ± 3.06
E	clean	46.57 ± 20.12	5.89 ± 6.65	5.99 ± 3.77
E	noise	48.04 ± 18.02	4.52 ± 6.16	13.21 ± 4.84
raw_clean	clean	51.43 ± 17.97	5.26 ± 5.92	18.38 ± 4.25
raw_clean	noise	51.35 ± 18.06	5.26 ± 5.92	18.35 ± 4.24
raw_noise	clean	51.43 ± 17.97	5.26 ± 5.92	18.38 ± 4.25
raw_noise	noise	51.35 ± 18.06	5.26 ± 5.92	18.36 ± 4.22
mlp32_clean	clean	54.53 ± 19.29	6.05 ± 4.75	7.87 ± 3.27
mlp32_clean	noise	54.53 ± 19.29	6.05 ± 4.75	7.94 ± 3.24
mlp32_noise	clean	54.53 ± 19.29	6.05 ± 4.75	7.87 ± 3.27
mlp32_noise	noise	54.53 ± 19.29	6.05 ± 4.75	7.92 ± 3.22
dd2_clean	clean	43.12 ± 17.96	11.60 ± 8.43	6.26 ± 2.61
dd2_clean	noise	43.95 ± 16.42	5.24 ± 5.78	33.63 ± 3.03
dd2_noise	clean	45.10 ± 16.57	9.17 ± 7.42	6.61 ± 2.68
dd2_noise	noise	45.90 ± 17.01	4.30 ± 5.05	48.33 ± 2.83
no_i_vector	clean	44.57 ± 19.62	5.70 ± 5.59	1.30 ± 0.84
no_i_vector	noise	43.09 ± 19.08	7.11 ± 5.50	1.24 ± 0.89
innovation_only	clean	51.76 ± 21.97	5.37 ± 6.91	19.83 ± 4.78
innovation_only	noise	52.29 ± 22.34	5.65 ± 6.70	19.94 ± 4.55
no_excess	clean	52.60 ± 23.28	6.30 ± 7.14	19.34 ± 4.95
no_excess	noise	53.05 ± 22.93	6.58 ± 6.90	19.37 ± 4.75
no_i_vector_no_excess	clean	48.18 ± 17.87	6.78 ± 5.75	17.09 ± 4.69
no_i_vector_no_excess	noise	48.18 ± 17.87	6.78 ± 5.75	17.10 ± 4.65
no_train_envelope	clean	54.21 ± 25.30	4.66 ± 6.61	3.13 ± 1.84
no_train_envelope	noise	53.88 ± 26.63	3.63 ± 5.19	4.11 ± 2.51
no_envelope_input	clean	52.86 ± 25.26	4.66 ± 6.61	3.06 ± 1.80
no_envelope_input	noise	53.32 ± 26.87	3.07 ± 4.95	4.27 ± 2.39

G.4.7 TwoBodyCollision

Tables G5.7.1, G5.7.2, G5.7.3, and G5.7.4 report presence ranking and calibration, probability scores, fixed-threshold outcomes, and distance/direction metrics, respectively, for all versions and both observation conditions.

Table G5.7.1: Presence ranking and calibration.

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
C	clean	0.7399 ± 0.2535 [n=7]	0.7272 ± 0.2578 [n=7]	0.0252 ± 0.0153	0.0345 ± 0.0093 [n=7]	0.0264 ± 0.0137
C	noise	0.6593 ± 0.2636 [n=7]	0.6414 ± 0.2680 [n=7]	0.0382 ± 0.0168	0.0828 ± 0.0074 [n=7]	0.0786 ± 0.0123
A	clean	0.7522 ± 0.2600 [n=7]	0.7405 ± 0.2645 [n=7]	0.0255 ± 0.0170	0.0266 ± 0.0128 [n=7]	0.0152 ± 0.0129
A	noise	0.6436 ± 0.2658 [n=7]	0.6242 ± 0.2710 [n=7]	0.0369 ± 0.0181	0.0601 ± 0.0079 [n=7]	0.0536 ± 0.0093
E	clean	0.7341 ± 0.2546 [n=7]	0.7211 ± 0.2589 [n=7]	0.0260 ± 0.0156	0.0352 ± 0.0097 [n=7]	0.0268 ± 0.0144
E	noise	0.6536 ± 0.2675 [n=7]	0.6354 ± 0.2719 [n=7]	0.0404 ± 0.0169	0.0877 ± 0.0081 [n=7]	0.0837 ± 0.0127
raw_clean	clean	0.6248 ± 0.2644 [n=7]	0.6029 ± 0.2726 [n=7]	0.0362 ± 0.0191	0.0624 ± 0.0103 [n=7]	0.0544 ± 0.0074
raw_clean	noise	0.6270 ± 0.2650 [n=7]	0.6051 ± 0.2734 [n=7]	0.0362 ± 0.0192	0.0623 ± 0.0099 [n=7]	0.0544 ± 0.0074
raw_noise	clean	0.6249 ± 0.2644 [n=7]	0.6029 ± 0.2726 [n=7]	0.0362 ± 0.0191	0.0624 ± 0.0103 [n=7]	0.0544 ± 0.0074
raw_noise	noise	0.6270 ± 0.2649 [n=7]	0.6051 ± 0.2733 [n=7]	0.0362 ± 0.0192	0.0623 ± 0.0099 [n=7]	0.0543 ± 0.0074
mlp32_clean	clean	0.6485 ± 0.2479 [n=7]	0.6300 ± 0.2522 [n=7]	0.0347 ± 0.0163	0.0634 ± 0.0063 [n=7]	0.0596 ± 0.0083

Continued on next page

Table G5.7.1 (continued)

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
mlp32_clean	noise	0.6515 ± 0.2474 [n=7]	0.6330 ± 0.2519 [n=7]	0.0347 ± 0.0163	0.0635 ± 0.0064 [n=7]	0.0595 ± 0.0084
mlp32_noise	clean	0.6480 ± 0.2492 [n=7]	0.6294 ± 0.2536 [n=7]	0.0347 ± 0.0163	0.0634 ± 0.0063 [n=7]	0.0595 ± 0.0083
mlp32_noise	noise	0.6514 ± 0.2474 [n=7]	0.6330 ± 0.2519 [n=7]	0.0347 ± 0.0163	0.0634 ± 0.0064 [n=7]	0.0595 ± 0.0084
dd2_clean	clean	0.4761 ± 0.2646 [n=7]	0.4464 ± 0.2710 [n=7]	0.0414 ± 0.0237	0.0540 ± 0.0148 [n=7]	0.0172 ± 0.0319
dd2_clean	noise	0.4137 ± 0.1954 [n=7]	0.3803 ± 0.2003 [n=7]	0.0763 ± 0.0201	0.1676 ± 0.0316 [n=7]	0.1659 ± 0.0385
dd2_noise	clean	0.4837 ± 0.2740 [n=7]	0.4552 ± 0.2793 [n=7]	0.0391 ± 0.0257	0.0481 ± 0.0202 [n=7]	0.0036 ± 0.0325
dd2_noise	noise	0.4181 ± 0.2463 [n=7]	0.3858 ± 0.2504 [n=7]	0.0503 ± 0.0202	0.1152 ± 0.0112 [n=7]	0.0964 ± 0.0360
no_i_vector	clean	0.7649 ± 0.2572 [n=7]	0.7534 ± 0.2619 [n=7]	0.0225 ± 0.0143	0.0249 ± 0.0082 [n=7]	0.0202 ± 0.0140
no_i_vector	noise	0.6262 ± 0.2627 [n=7]	0.6057 ± 0.2683 [n=7]	0.0369 ± 0.0161	0.0790 ± 0.0064 [n=7]	0.0745 ± 0.0117
innovation_only	clean	0.6784 ± 0.2793 [n=7]	0.6608 ± 0.2854 [n=7]	0.0340 ± 0.0183	0.0662 ± 0.0121 [n=7]	0.0623 ± 0.0156
innovation_only	noise	0.6758 ± 0.2840 [n=7]	0.6580 ± 0.2902 [n=7]	0.0345 ± 0.0186	0.0670 ± 0.0114 [n=7]	0.0626 ± 0.0156
no_excess	clean	0.6642 ± 0.2831 [n=7]	0.6454 ± 0.2901 [n=7]	0.0329 ± 0.0188	0.0572 ± 0.0079 [n=7]	0.0487 ± 0.0135
no_excess	noise	0.6600 ± 0.2859 [n=7]	0.6409 ± 0.2930 [n=7]	0.0333 ± 0.0191	0.0563 ± 0.0090 [n=7]	0.0488 ± 0.0135

Continued on next page

Table G5.7.1 (continued)

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
no_i_vector_no_excess	clean	0.6034 ± 0.2436 [n=7]	0.5813 ± 0.2495 [n=7]	0.0332 ± 0.0189	0.0523 ± 0.0117 [n=7]	0.0428 ± 0.0090
no_i_vector_no_excess	noise	0.6071 ± 0.2450 [n=7]	0.5849 ± 0.2512 [n=7]	0.0332 ± 0.0189	0.0523 ± 0.0115 [n=7]	0.0427 ± 0.0091
no_train_envelope	clean	0.7610 ± 0.2670 [n=7]	0.7501 ± 0.2718 [n=7]	0.0211 ± 0.0140	0.0230 ± 0.0090 [n=7]	0.0157 ± 0.0138
no_train_envelope	noise	0.7234 ± 0.2661 [n=7]	0.7094 ± 0.2709 [n=7]	0.0292 ± 0.0150	0.0598 ± 0.0081 [n=7]	0.0567 ± 0.0119
no_envelope_input	clean	0.7527 ± 0.2613 [n=7]	0.7414 ± 0.2657 [n=7]	0.0216 ± 0.0140	0.0216 ± 0.0084 [n=7]	0.0159 ± 0.0138
no_envelope_input	noise	0.7204 ± 0.2649 [n=7]	0.7062 ± 0.2696 [n=7]	0.0300 ± 0.0151	0.0621 ± 0.0081 [n=7]	0.0590 ± 0.0118

Table G5.7.2: Joint and interval probabilities.

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
C	clean	0.1283 ± 0.0760	0.0554 ± 0.0366	1.5374 ± 1.3862 [n=7]	0.0641 ± 0.0243
C	noise	0.2188 ± 0.0985	0.0827 ± 0.0469	2.0602 ± 1.2570 [n=7]	0.1300 ± 0.0273
A	clean	0.1416 ± 0.1014	0.0573 ± 0.0402	1.9288 ± 1.8209 [n=7]	0.0586 ± 0.0395
A	noise	0.2351 ± 0.1317	0.0881 ± 0.0525	2.7780 ± 1.5905 [n=7]	0.1106 ± 0.0397
E	clean	0.1314 ± 0.0774	0.0565 ± 0.0367	1.5433 ± 1.3623 [n=7]	0.0671 ± 0.0281
E	noise	0.2268 ± 0.0984	0.0856 ± 0.0465	2.0452 ± 1.2341 [n=7]	0.1391 ± 0.0308
raw_clean	clean	0.2374 ± 0.1356	0.0847 ± 0.0567	3.0000 ± 1.5343 [n=7]	0.1019 ± 0.0287
raw_clean	noise	0.2374 ± 0.1358	0.0847 ± 0.0568	2.9977 ± 1.5296 [n=7]	0.1018 ± 0.0287
raw_noise	clean	0.2374 ± 0.1356	0.0847 ± 0.0567	3.0001 ± 1.5347 [n=7]	0.1019 ± 0.0287
raw_noise	noise	0.2374 ± 0.1358	0.0846 ± 0.0568	2.9978 ± 1.5299 [n=7]	0.1018 ± 0.0287
mlp32_clean	clean	0.2295 ± 0.1224	0.0818 ± 0.0522	2.6728 ± 1.2299 [n=7]	0.1050 ± 0.0236

Continued on next page

Table G5.7.2 (continued)

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
mlp32_clean	noise	0.2295 ± 0.1226	0.0818 ± 0.0523	2.6706 ± 1.2245 [n=7]	0.1050 ± 0.0236
mlp32_noise	clean	0.2295 ± 0.1224	0.0818 ± 0.0522	2.6728 ± 1.2294 [n=7]	0.1050 ± 0.0236
mlp32_noise	noise	0.2295 ± 0.1226	0.0818 ± 0.0523	2.6706 ± 1.2240 [n=7]	0.1049 ± 0.0236
dd2_clean	clean	0.1616 ± 0.1003	0.0846 ± 0.0532	1.9065 ± 0.5816 [n=7]	0.0700 ± 0.0311
dd2_clean	noise	0.3232 ± 0.0830	0.1324 ± 0.0446	1.4609 ± 0.5999 [n=7]	0.2597 ± 0.0476
dd2_noise	clean	0.1584 ± 0.1041	0.0814 ± 0.0569	2.3036 ± 0.5729 [n=7]	0.0495 ± 0.0177
dd2_noise	noise	0.2404 ± 0.0875	0.0963 ± 0.0487	1.8475 ± 0.6356 [n=7]	0.1557 ± 0.0250
no_i_vector	clean	0.1127 ± 0.0748	0.0495 ± 0.0358	1.4439 ± 1.4822 [n=7]	0.0529 ± 0.0217
no_i_vector	noise	0.2155 ± 0.0979	0.0808 ± 0.0472	2.1402 ± 1.1491 [n=7]	0.1236 ± 0.0257
innovation_only	clean	0.2108 ± 0.1089	0.0792 ± 0.0495	2.2423 ± 1.2977 [n=7]	0.1097 ± 0.0309
innovation_only	noise	0.2124 ± 0.1103	0.0799 ± 0.0499	2.2651 ± 1.3323 [n=7]	0.1106 ± 0.0316
no_excess	clean	0.2166 ± 0.1219	0.0790 ± 0.0525	2.6829 ± 1.5237 [n=7]	0.0955 ± 0.0296
no_excess	noise	0.2177 ± 0.1228	0.0795 ± 0.0528	2.6974 ± 1.5443 [n=7]	0.0961 ± 0.0301
no_i_vector_no_excess	clean	0.2273 ± 0.1341	0.0806 ± 0.0557	3.1333 ± 1.5455 [n=7]	0.0867 ± 0.0239
no_i_vector_no_excess	noise	0.2273 ± 0.1342	0.0806 ± 0.0558	3.1308 ± 1.5409 [n=7]	0.0866 ± 0.0239
no_train_envelope	clean	0.1039 ± 0.0762	0.0488 ± 0.0352	1.3356 ± 1.5091 [n=7]	0.0474 ± 0.0216
no_train_envelope	noise	0.1809 ± 0.0988	0.0699 ± 0.0445	1.8620 ± 1.4211 [n=7]	0.0991 ± 0.0247
no_envelope_input	clean	0.1068 ± 0.0765	0.0499 ± 0.0351	1.3748 ± 1.4959 [n=7]	0.0485 ± 0.0216
no_envelope_input	noise	0.1842 ± 0.0981	0.0711 ± 0.0444	1.8725 ± 1.4109 [n=7]	0.1021 ± 0.0241

Table G5.7.3: Fixed-threshold detection and interval hits.

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
C	clean	5.75 ± 2.06	90.14 ± 16.77 [n=7]	38.55 ± 21.04	69.50 ± 24.27 [n=7]	77.81 ± 17.65 [n=7]
C	noise	14.26 ± 4.95	91.88 ± 16.29 [n=7]	22.19 ± 12.68	61.13 ± 26.98 [n=7]	80.44 ± 15.98 [n=7]

Continued on next page

Table G5.7.3 (continued)

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
A	clean	4.19 ± 1.85	87.87 ± 21.03 [n=7]	48.67 ± 30.65	67.91 ± 24.89 [n=7]	76.85 ± 22.11 [n=7]
A	noise	12.32 ± 4.43	90.92 ± 20.61 [n=7]	25.70 ± 14.25	40.88 ± 6.37 [n=7]	76.37 ± 20.36 [n=7]
E	clean	4.78 ± 1.40	90.14 ± 16.77 [n=7]	42.22 ± 22.53	68.55 ± 24.53 [n=7]	77.81 ± 17.65 [n=7]
E	noise	12.99 ± 4.28	91.26 ± 17.19 [n=7]	23.70 ± 13.63	60.94 ± 27.21 [n=7]	79.73 ± 16.81 [n=7]
raw_clean	clean	22.62 ± 5.28	88.57 ± 20.30 [n=7]	16.16 ± 10.22	43.19 ± 17.66 [n=7]	70.24 ± 18.31 [n=7]
raw_clean	noise	22.69 ± 5.32	89.28 ± 18.52 [n=7]	16.22 ± 10.11	43.04 ± 17.96 [n=7]	70.72 ± 17.53 [n=7]
raw_noise	clean	22.62 ± 5.28	88.57 ± 20.30 [n=7]	16.16 ± 10.22	43.19 ± 17.66 [n=7]	70.24 ± 18.31 [n=7]
raw_noise	noise	22.69 ± 5.31	89.28 ± 18.52 [n=7]	16.22 ± 10.12	42.91 ± 18.18 [n=7]	70.59 ± 17.73 [n=7]
mlp32_clean	clean	13.97 ± 4.91	87.70 ± 24.50 [n=7]	22.12 ± 13.35	40.98 ± 19.32 [n=7]	63.50 ± 17.62 [n=7]
mlp32_clean	noise	13.98 ± 4.91	87.62 ± 24.51 [n=7]	22.04 ± 13.32	40.98 ± 19.32 [n=7]	63.65 ± 17.41 [n=7]
mlp32_noise	clean	13.97 ± 4.91	87.70 ± 24.50 [n=7]	22.12 ± 13.35	40.98 ± 19.32 [n=7]	63.50 ± 17.62 [n=7]
mlp32_noise	noise	14.00 ± 4.92	87.62 ± 24.51 [n=7]	22.02 ± 13.30	40.98 ± 19.32 [n=7]	63.65 ± 17.41 [n=7]
dd2_clean	clean	7.81 ± 3.65	90.65 ± 15.80 [n=7]	38.58 ± 29.91	65.23 ± 38.05 [n=7]	84.52 ± 20.21 [n=7]
dd2_clean	noise	65.99 ± 4.47	100.00 ± 0.00 [n=7]	10.86 ± 7.77	73.22 ± 26.58 [n=7]	91.47 ± 10.65 [n=7]
dd2_noise	clean	7.81 ± 3.65	91.84 ± 16.20 [n=7]	38.85 ± 30.04	64.00 ± 36.56 [n=7]	83.02 ± 19.36 [n=7]
dd2_noise	noise	66.00 ± 3.92	100.00 ± 0.00 [n=7]	11.10 ± 7.80	67.84 ± 30.42 [n=7]	88.53 ± 10.33 [n=7]
no_i_vector	clean	6.09 ± 2.32	87.41 ± 25.15 [n=7]	39.50 ± 23.99	69.76 ± 31.43 [n=7]	76.05 ± 29.43 [n=7]

Continued on next page

Table G5.7.3 (continued)

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
no_i_vector	noise	14.66 ± 4.89	91.42 ± 20.78 [n=7]	22.88 ± 13.19	61.50 ± 27.05 [n=7]	78.87 ± 19.18 [n=7]
innovation_only	clean	11.73 ± 4.33	88.05 ± 24.30 [n=7]	25.20 ± 15.10	54.13 ± 35.84 [n=7]	69.09 ± 24.27 [n=7]
innovation_only	noise	12.00 ± 4.14	87.05 ± 23.53 [n=7]	24.66 ± 14.99	53.83 ± 35.43 [n=7]	68.66 ± 24.06 [n=7]
no_excess	clean	15.12 ± 4.59	87.26 ± 24.15 [n=7]	22.13 ± 13.34	51.30 ± 32.21 [n=7]	70.00 ± 24.57 [n=7]
no_excess	noise	15.13 ± 4.72	86.58 ± 23.50 [n=7]	22.05 ± 13.38	50.67 ± 31.98 [n=7]	69.77 ± 24.12 [n=7]
no_i_vector_no_excess	clean	21.19 ± 5.59	89.52 ± 20.36 [n=7]	17.04 ± 10.48	38.14 ± 19.29 [n=7]	69.78 ± 18.76 [n=7]
no_i_vector_no_excess	noise	21.20 ± 5.69	89.29 ± 20.34 [n=7]	16.99 ± 10.41	38.90 ± 18.96 [n=7]	70.05 ± 18.80 [n=7]
no_train_envelope	clean	8.24 ± 2.71	95.24 ± 12.60 [n=7]	33.04 ± 19.18	76.12 ± 22.64 [n=7]	84.44 ± 13.68 [n=7]
no_train_envelope	noise	23.96 ± 5.63	97.53 ± 6.53 [n=7]	17.72 ± 10.45	61.23 ± 30.14 [n=7]	85.14 ± 11.70 [n=7]
no_envelope_input	clean	7.94 ± 2.77	95.24 ± 12.60 [n=7]	33.41 ± 19.11	76.52 ± 22.81 [n=7]	85.12 ± 13.46 [n=7]
no_envelope_input	noise	22.64 ± 5.55	96.83 ± 8.40 [n=7]	18.02 ± 10.65	61.93 ± 28.92 [n=7]	83.69 ± 13.44 [n=7]

Table G5.7.4: Distance and error direction. Metric panel 1 of 2; identifying columns are repeated.

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
C	clean	0.1496 ± 0.0835 [n=7]	0.2005 ± 0.1114 [n=7]	0.1140 ± 0.0718 [n=7]	76.03 ± 20.84 [n=7]
C	noise	0.1532 ± 0.0723 [n=7]	0.2051 ± 0.0964 [n=7]	0.1055 ± 0.0963 [n=7]	75.44 ± 20.74 [n=7]
A	clean	0.1451 ± 0.0733 [n=7]	0.1962 ± 0.1034 [n=7]	0.1150 ± 0.0881 [n=7]	75.63 ± 20.58 [n=7]
A	noise	0.1897 ± 0.0540 [n=7]	0.2672 ± 0.0585 [n=7]	0.0902 ± 0.1190 [n=7]	68.28 ± 20.85 [n=7]

Continued on next page

Table G5.7.4 (continued)

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
E	clean	0.1524 ± 0.0845 [n=7]	0.2041 ± 0.1121 [n=7]	0.1168 ± 0.0744 [n=7]	76.03 ± 20.84 [n=7]
E	noise	0.1538 ± 0.0737 [n=7]	0.2059 ± 0.0986 [n=7]	0.1077 ± 0.0981 [n=7]	75.99 ± 20.97 [n=7]
raw_clean	clean	0.2142 ± 0.0792 [n=7]	0.2926 ± 0.0981 [n=7]	0.0986 ± 0.1168 [n=7]	64.35 ± 23.55 [n=7]
raw_clean	noise	0.2119 ± 0.0774 [n=7]	0.2901 ± 0.0959 [n=7]	0.0953 ± 0.1113 [n=7]	64.21 ± 23.51 [n=7]
raw_noise	clean	0.2142 ± 0.0792 [n=7]	0.2926 ± 0.0981 [n=7]	0.0986 ± 0.1168 [n=7]	64.35 ± 23.55 [n=7]
raw_noise	noise	0.2122 ± 0.0779 [n=7]	0.2908 ± 0.0970 [n=7]	0.0949 ± 0.1111 [n=7]	64.08 ± 23.54 [n=7]
mlp32_clean	clean	0.2263 ± 0.0742 [n=7]	0.3121 ± 0.0974 [n=7]	0.0856 ± 0.1276 [n=7]	63.79 ± 25.03 [n=7]
mlp32_clean	noise	0.2260 ± 0.0747 [n=7]	0.3118 ± 0.0977 [n=7]	0.0854 ± 0.1276 [n=7]	63.82 ± 25.04 [n=7]
mlp32_noise	clean	0.2263 ± 0.0742 [n=7]	0.3121 ± 0.0974 [n=7]	0.0856 ± 0.1276 [n=7]	63.79 ± 25.03 [n=7]
mlp32_noise	noise	0.2260 ± 0.0747 [n=7]	0.3118 ± 0.0977 [n=7]	0.0854 ± 0.1276 [n=7]	63.82 ± 25.04 [n=7]
dd2_clean	clean	0.1318 ± 0.0966 [n=7]	0.1753 ± 0.1315 [n=7]	0.0567 ± 0.1061 [n=7]	72.29 ± 21.66 [n=7]
dd2_clean	noise	0.1185 ± 0.0690 [n=7]	0.1579 ± 0.0918 [n=7]	0.0522 ± 0.0562 [n=7]	66.62 ± 19.32 [n=7]
dd2_noise	clean	0.1385 ± 0.0912 [n=7]	0.1837 ± 0.1245 [n=7]	0.0612 ± 0.1064 [n=7]	70.76 ± 20.65 [n=7]
dd2_noise	noise	0.1351 ± 0.0730 [n=7]	0.1796 ± 0.0996 [n=7]	0.0619 ± 0.0687 [n=7]	67.41 ± 19.19 [n=7]
no_i_vector	clean	0.1560 ± 0.1093 [n=7]	0.2114 ± 0.1534 [n=7]	0.1087 ± 0.1066 [n=7]	73.57 ± 17.24 [n=7]
no_i_vector	noise	0.1568 ± 0.0750 [n=7]	0.2123 ± 0.1035 [n=7]	0.1137 ± 0.0924 [n=7]	73.18 ± 21.04 [n=7]
innovation_only	clean	0.1867 ± 0.0945 [n=7]	0.2547 ± 0.1300 [n=7]	0.1201 ± 0.0861 [n=7]	71.86 ± 22.17 [n=7]
innovation_only	noise	0.1865 ± 0.0942 [n=7]	0.2547 ± 0.1293 [n=7]	0.1208 ± 0.0880 [n=7]	72.06 ± 21.68 [n=7]

Continued on next page

Table G5.7.4 (continued)

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
no_excess	clean	0.1874 ± 0.0923 [n=7]	0.2580 ± 0.1274 [n=7]	0.1027 ± 0.0968 [n=7]	66.98 ± 20.38 [n=7]
no_excess	noise	0.1898 ± 0.0933 [n=7]	0.2611 ± 0.1274 [n=7]	0.1037 ± 0.0990 [n=7]	67.07 ± 20.36 [n=7]
no_i_vector_no_excess	clean	0.2144 ± 0.0687 [n=7]	0.2993 ± 0.0899 [n=7]	0.0763 ± 0.1314 [n=7]	61.35 ± 25.88 [n=7]
no_i_vector_no_excess	noise	0.2131 ± 0.0689 [n=7]	0.2974 ± 0.0912 [n=7]	0.0748 ± 0.1329 [n=7]	60.74 ± 26.30 [n=7]
no_train_envelope	clean	0.1150 ± 0.0646 [n=7]	0.1585 ± 0.0873 [n=7]	0.0887 ± 0.0534 [n=7]	74.16 ± 20.24 [n=7]
no_train_envelope	noise	0.1401 ± 0.0677 [n=7]	0.1896 ± 0.0885 [n=7]	0.0725 ± 0.0671 [n=7]	70.86 ± 17.60 [n=7]
no_envelope_input	clean	0.1140 ± 0.0650 [n=7]	0.1568 ± 0.0879 [n=7]	0.0883 ± 0.0529 [n=7]	74.56 ± 20.41 [n=7]
no_envelope_input	noise	0.1401 ± 0.0682 [n=7]	0.1899 ± 0.0905 [n=7]	0.0718 ± 0.0684 [n=7]	71.38 ± 18.31 [n=7]

Table G5.7.4: Distance and error direction. Metric panel 2 of 2; identifying columns are repeated.

Source key	Condition	Early (%)	Miss (%)	False boundary (%)
C	clean	23.97 ± 20.84 [n=7]	9.86 ± 16.77 [n=7]	5.75 ± 2.06
C	noise	24.56 ± 20.74 [n=7]	8.12 ± 16.29 [n=7]	14.26 ± 4.95
A	clean	24.37 ± 20.58 [n=7]	12.13 ± 21.03 [n=7]	4.19 ± 1.85
A	noise	31.72 ± 20.85 [n=7]	9.08 ± 20.61 [n=7]	12.32 ± 4.43
E	clean	23.97 ± 20.84 [n=7]	9.86 ± 16.77 [n=7]	4.78 ± 1.40
E	noise	24.01 ± 20.97 [n=7]	8.74 ± 17.19 [n=7]	12.99 ± 4.28
raw_clean	clean	35.65 ± 23.55 [n=7]	11.43 ± 20.30 [n=7]	22.62 ± 5.28
raw_clean	noise	35.79 ± 23.51 [n=7]	10.72 ± 18.52 [n=7]	22.69 ± 5.32
raw_noise	clean	35.65 ± 23.55 [n=7]	11.43 ± 20.30 [n=7]	22.62 ± 5.28
raw_noise	noise	35.92 ± 23.54 [n=7]	10.72 ± 18.52 [n=7]	22.69 ± 5.31
mlp32_clean	clean	36.21 ± 25.03 [n=7]	12.30 ± 24.50 [n=7]	13.97 ± 4.91

Continued on next page

Table G5.7.4 (continued)

Source key	Condition	Early (%)	Miss (%)	False boundary (%)
mlp32_clean	noise	36.18 ± 25.04 [n=7]	12.38 ± 24.51 [n=7]	13.98 ± 4.91
mlp32_noise	clean	36.21 ± 25.03 [n=7]	12.30 ± 24.50 [n=7]	13.97 ± 4.91
mlp32_noise	noise	36.18 ± 25.04 [n=7]	12.38 ± 24.51 [n=7]	14.00 ± 4.92
dd2_clean	clean	27.71 ± 21.66 [n=7]	9.35 ± 15.80 [n=7]	7.81 ± 3.65
dd2_clean	noise	33.38 ± 19.32 [n=7]	0.00 ± 0.00 [n=7]	65.99 ± 4.47
dd2_noise	clean	29.24 ± 20.65 [n=7]	8.16 ± 16.20 [n=7]	7.81 ± 3.65
dd2_noise	noise	32.59 ± 19.19 [n=7]	0.00 ± 0.00 [n=7]	66.00 ± 3.92
no_i_vector	clean	26.43 ± 17.24 [n=7]	12.59 ± 25.15 [n=7]	6.09 ± 2.32
no_i_vector	noise	26.82 ± 21.04 [n=7]	8.58 ± 20.78 [n=7]	14.66 ± 4.89
innovation_only	clean	28.14 ± 22.17 [n=7]	11.95 ± 24.30 [n=7]	11.73 ± 4.33
innovation_only	noise	27.94 ± 21.68 [n=7]	12.95 ± 23.53 [n=7]	12.00 ± 4.14
no_excess	clean	33.02 ± 20.38 [n=7]	12.74 ± 24.15 [n=7]	15.12 ± 4.59
no_excess	noise	32.93 ± 20.36 [n=7]	13.42 ± 23.50 [n=7]	15.13 ± 4.72
no_i_vector_no_excess	clean	38.65 ± 25.88 [n=7]	10.48 ± 20.36 [n=7]	21.19 ± 5.59
no_i_vector_no_excess	noise	39.26 ± 26.30 [n=7]	10.71 ± 20.34 [n=7]	21.20 ± 5.69
no_train_envelope	clean	25.84 ± 20.24 [n=7]	4.76 ± 12.60 [n=7]	8.24 ± 2.71
no_train_envelope	noise	29.14 ± 17.60 [n=7]	2.47 ± 6.53 [n=7]	23.96 ± 5.63
no_envelope_input	clean	25.44 ± 20.41 [n=7]	4.76 ± 12.60 [n=7]	7.94 ± 2.77
no_envelope_input	noise	28.62 ± 18.31 [n=7]	3.17 ± 8.40 [n=7]	22.64 ± 5.55

G.4.8 SMOOTHDUFFING

Tables G5.8.1, G5.8.2, G5.8.3, and G5.8.4 report presence ranking and calibration, probability scores, fixed-threshold outcomes, and distance/direction metrics, respectively, for all versions and both observation conditions.

Table G5.8.1: Presence ranking and calibration.

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
C	clean	NA	NA	4.40×10^{-17} ± 8.86×10^{-17}	NA	5.26×10^{-10} ± 8.50×10^{-10}

Continued on next page

Table G5.8.1 (continued)

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
C	noise	NA	NA	$5.17 \times 10^{-20} \pm 7.95 \times 10^{-20}$	NA	$1.74 \times 10^{-11} \pm 2.23 \times 10^{-11}$
A	clean	NA	NA	$5.19 \times 10^{-19} \pm 7.58 \times 10^{-19}$	NA	$8.39 \times 10^{-11} \pm 9.29 \times 10^{-11}$
A	noise	NA	NA	0.0010 ± 0.0012	NA	0.0049 ± 0.0043
E	clean	NA	NA	$4.40 \times 10^{-17} \pm 8.86 \times 10^{-17}$	NA	$5.26 \times 10^{-10} \pm 8.50 \times 10^{-10}$
E	noise	NA	NA	$5.17 \times 10^{-20} \pm 7.95 \times 10^{-20}$	NA	$1.74 \times 10^{-11} \pm 2.23 \times 10^{-11}$
raw_clean	clean	NA	NA	$1.20 \times 10^{-8} \pm 1.45 \times 10^{-8}$	NA	$1.56 \times 10^{-5} \pm 1.44 \times 10^{-5}$
raw_clean	noise	NA	NA	$1.19 \times 10^{-8} \pm 1.44 \times 10^{-8}$	NA	$1.56 \times 10^{-5} \pm 1.44 \times 10^{-5}$
raw_noise	clean	NA	NA	$1.20 \times 10^{-8} \pm 1.45 \times 10^{-8}$	NA	$1.56 \times 10^{-5} \pm 1.44 \times 10^{-5}$
raw_noise	noise	NA	NA	$1.19 \times 10^{-8} \pm 1.44 \times 10^{-8}$	NA	$1.56 \times 10^{-5} \pm 1.44 \times 10^{-5}$
mlp32_clean	clean	NA	NA	$1.21 \times 10^{-17} \pm 3.18 \times 10^{-17}$	NA	$2.18 \times 10^{-10} \pm 3.90 \times 10^{-10}$
mlp32_clean	noise	NA	NA	$1.26 \times 10^{-17} \pm 3.31 \times 10^{-17}$	NA	$2.22 \times 10^{-10} \pm 3.99 \times 10^{-10}$
mlp32_noise	clean	NA	NA	$1.22 \times 10^{-17} \pm 3.19 \times 10^{-17}$	NA	$2.18 \times 10^{-10} \pm 3.90 \times 10^{-10}$
mlp32_noise	noise	NA	NA	$1.26 \times 10^{-17} \pm 3.31 \times 10^{-17}$	NA	$2.22 \times 10^{-10} \pm 4.00 \times 10^{-10}$
dd2_clean	clean	NA	NA	$0.0181 \pm 4.11 \times 10^{-6}$	NA	$0.1345 \pm 1.55 \times 10^{-5}$
dd2_clean	noise	NA	NA	0.1516 ± 0.0210	NA	0.1595 ± 0.0209
dd2_noise	clean	NA	NA	$0.0009 \pm 8.20 \times 10^{-5}$	NA	0.0300 ± 0.0015
dd2_noise	noise	NA	NA	0.0000 ± 0.0000	NA	0.0000 ± 0.0000
no_i_vector	clean	NA	NA	$4.40 \times 10^{-17} \pm 8.87 \times 10^{-17}$	NA	$5.26 \times 10^{-10} \pm 8.50 \times 10^{-10}$
no_i_vector	noise	NA	NA	$5.14 \times 10^{-20} \pm 7.91 \times 10^{-20}$	NA	$1.73 \times 10^{-11} \pm 2.22 \times 10^{-11}$
innovation_only	clean	NA	NA	$2.77 \times 10^{-17} \pm 4.29 \times 10^{-17}$	NA	$4.44 \times 10^{-10} \pm 5.70 \times 10^{-10}$
innovation_only	noise	NA	NA	$2.76 \times 10^{-17} \pm 4.27 \times 10^{-17}$	NA	$4.42 \times 10^{-10} \pm 5.69 \times 10^{-10}$

Continued on next page

Table G5.8.1 (continued)

Source key	Condition	AP	nAP	Binary Brier	ECE	Calibration bias
no_excess	clean	NA	NA	$2.77 \times 10^{-17} \pm 4.29 \times 10^{-17}$	NA	$4.44 \times 10^{-10} \pm 5.70 \times 10^{-10}$
no_excess	noise	NA	NA	$2.76 \times 10^{-17} \pm 4.27 \times 10^{-17}$	NA	$4.42 \times 10^{-10} \pm 5.69 \times 10^{-10}$
no_i_vector_no_excess	clean	NA	NA	$2.76 \times 10^{-17} \pm 4.26 \times 10^{-17}$	NA	$4.42 \times 10^{-10} \pm 5.69 \times 10^{-10}$
no_i_vector_no_excess	noise	NA	NA	$2.74 \times 10^{-17} \pm 4.23 \times 10^{-17}$	NA	$4.41 \times 10^{-10} \pm 5.67 \times 10^{-10}$
no_train_envelope	clean	NA	NA	$5.84 \times 10^{-30} \pm 1.04 \times 10^{-29}$	NA	$1.51 \times 10^{-16} \pm 2.30 \times 10^{-16}$
no_train_envelope	noise	NA	NA	$3.36 \times 10^{-16} \pm 5.58 \times 10^{-16}$	NA	$1.34 \times 10^{-9} \pm 1.77 \times 10^{-9}$
no_envelope_input	clean	NA	NA	$4.40 \times 10^{-17} \pm 8.86 \times 10^{-17}$	NA	$5.26 \times 10^{-10} \pm 8.50 \times 10^{-10}$
no_envelope_input	noise	NA	NA	$5.17 \times 10^{-20} \pm 7.95 \times 10^{-20}$	NA	$1.74 \times 10^{-11} \pm 2.23 \times 10^{-11}$

Table G5.8.2: Joint and interval probabilities.

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
C	clean	$5.26 \times 10^{-10} \pm 8.50 \times 10^{-10}$	$6.41 \times 10^{-17} \pm 1.29 \times 10^{-16}$	NA	$5.26 \times 10^{-10} \pm 8.50 \times 10^{-10}$
C	noise	$1.74 \times 10^{-11} \pm 2.23 \times 10^{-11}$	$7.56 \times 10^{-20} \pm 1.16 \times 10^{-19}$	NA	$1.74 \times 10^{-11} \pm 2.23 \times 10^{-11}$
A	clean	$8.39 \times 10^{-11} \pm 9.29 \times 10^{-11}$	$7.76 \times 10^{-19} \pm 1.12 \times 10^{-18}$	NA	$8.39 \times 10^{-11} \pm 9.29 \times 10^{-11}$
A	noise	0.0056 ± 0.0051	0.0014 ± 0.0016	NA	0.0056 ± 0.0051
E	clean	$5.26 \times 10^{-10} \pm 8.50 \times 10^{-10}$	$6.41 \times 10^{-17} \pm 1.29 \times 10^{-16}$	NA	$5.26 \times 10^{-10} \pm 8.50 \times 10^{-10}$
E	noise	$1.74 \times 10^{-11} \pm 2.23 \times 10^{-11}$	$7.56 \times 10^{-20} \pm 1.16 \times 10^{-19}$	NA	$1.74 \times 10^{-11} \pm 2.23 \times 10^{-11}$
raw_clean	clean	$1.56 \times 10^{-5} \pm 1.44 \times 10^{-5}$	$1.66 \times 10^{-8} \pm 2.01 \times 10^{-8}$	NA	$1.56 \times 10^{-5} \pm 1.44 \times 10^{-5}$

Continued on next page

Table G5.8.2 (continued)

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
raw_clean	noise	$1.56 \times 10^{-5} \pm 1.44 \times 10^{-5}$	$1.66 \times 10^{-8} \pm 2.00 \times 10^{-8}$	NA	$1.56 \times 10^{-5} \pm 1.44 \times 10^{-5}$
raw_noise	clean	$1.56 \times 10^{-5} \pm 1.44 \times 10^{-5}$	$1.66 \times 10^{-8} \pm 2.01 \times 10^{-8}$	NA	$1.56 \times 10^{-5} \pm 1.44 \times 10^{-5}$
raw_noise	noise	$1.56 \times 10^{-5} \pm 1.44 \times 10^{-5}$	$1.66 \times 10^{-8} \pm 2.00 \times 10^{-8}$	NA	$1.56 \times 10^{-5} \pm 1.44 \times 10^{-5}$
mlp32_clean	clean	$2.18 \times 10^{-10} \pm 3.90 \times 10^{-10}$	$1.57 \times 10^{-17} \pm 4.10 \times 10^{-17}$	NA	$2.18 \times 10^{-10} \pm 3.90 \times 10^{-10}$
mlp32_clean	noise	$2.22 \times 10^{-10} \pm 3.99 \times 10^{-10}$	$1.63 \times 10^{-17} \pm 4.25 \times 10^{-17}$	NA	$2.22 \times 10^{-10} \pm 3.99 \times 10^{-10}$
mlp32_noise	clean	$2.18 \times 10^{-10} \pm 3.90 \times 10^{-10}$	$1.57 \times 10^{-17} \pm 4.10 \times 10^{-17}$	NA	$2.18 \times 10^{-10} \pm 3.90 \times 10^{-10}$
mlp32_noise	noise	$2.22 \times 10^{-10} \pm 4.00 \times 10^{-10}$	$1.63 \times 10^{-17} \pm 4.25 \times 10^{-17}$	NA	$2.22 \times 10^{-10} \pm 4.00 \times 10^{-10}$
dd2_clean	clean	$0.1445 \pm 1.78 \times 10^{-5}$	$0.0228 \pm 6.27 \times 10^{-5}$	NA	$0.1445 \pm 1.78 \times 10^{-5}$
dd2_clean	noise	1.6029 ± 0.2856	0.3031 ± 0.0419	NA	1.6029 ± 0.2856
dd2_noise	clean	0.0305 ± 0.0015	0.0012 ± 0.0001	NA	0.0305 ± 0.0015
dd2_noise	noise	0.0000 ± 0.0000	$4.33 \times 10^{-188} \pm 0.0000$	NA	0.0000 ± 0.0000
no_i_vector	clean	$5.26 \times 10^{-10} \pm 8.50 \times 10^{-10}$	$6.41 \times 10^{-17} \pm 1.29 \times 10^{-16}$	NA	$5.26 \times 10^{-10} \pm 8.50 \times 10^{-10}$
no_i_vector	noise	$1.73 \times 10^{-11} \pm 2.22 \times 10^{-11}$	$7.52 \times 10^{-20} \pm 1.15 \times 10^{-19}$	NA	$1.73 \times 10^{-11} \pm 2.22 \times 10^{-11}$
innovation_only	clean	$4.44 \times 10^{-10} \pm 5.70 \times 10^{-10}$	$4.08 \times 10^{-17} \pm 6.26 \times 10^{-17}$	NA	$4.44 \times 10^{-10} \pm 5.70 \times 10^{-10}$
innovation_only	noise	$4.42 \times 10^{-10} \pm 5.69 \times 10^{-10}$	$4.06 \times 10^{-17} \pm 6.24 \times 10^{-17}$	NA	$4.42 \times 10^{-10} \pm 5.69 \times 10^{-10}$
no_excess	clean	$4.44 \times 10^{-10} \pm 5.70 \times 10^{-10}$	$4.08 \times 10^{-17} \pm 6.26 \times 10^{-17}$	NA	$4.44 \times 10^{-10} \pm 5.70 \times 10^{-10}$
no_excess	noise	$4.42 \times 10^{-10} \pm 5.69 \times 10^{-10}$	$4.06 \times 10^{-17} \pm 6.24 \times 10^{-17}$	NA	$4.42 \times 10^{-10} \pm 5.69 \times 10^{-10}$
no_i_vector_no_excess	clean	$4.42 \times 10^{-10} \pm 5.69 \times 10^{-10}$	$4.06 \times 10^{-17} \pm 6.23 \times 10^{-17}$	NA	$4.42 \times 10^{-10} \pm 5.69 \times 10^{-10}$
no_i_vector_no_excess	noise	$4.41 \times 10^{-10} \pm 5.67 \times 10^{-10}$	$4.03 \times 10^{-17} \pm 6.18 \times 10^{-17}$	NA	$4.41 \times 10^{-10} \pm 5.67 \times 10^{-10}$

Continued on next page

Table G5.8.2 (continued)

Source key	Condition	Joint NLL	Joint Brier	Interval NLL	Survival NLL
no_train_envelope	clean	$1.51 \times 10^{-16} \pm 2.30 \times 10^{-16}$	$8.26 \times 10^{-30} \pm 1.47 \times 10^{-29}$	NA	$1.51 \times 10^{-16} \pm 2.30 \times 10^{-16}$
no_train_envelope	noise	$1.34 \times 10^{-9} \pm 1.77 \times 10^{-9}$	$4.91 \times 10^{-16} \pm 8.12 \times 10^{-16}$	NA	$1.34 \times 10^{-9} \pm 1.77 \times 10^{-9}$
no_envelope_input	clean	$5.26 \times 10^{-10} \pm 8.50 \times 10^{-10}$	$6.41 \times 10^{-17} \pm 1.29 \times 10^{-16}$	NA	$5.26 \times 10^{-10} \pm 8.50 \times 10^{-10}$
no_envelope_input	noise	$1.74 \times 10^{-11} \pm 2.23 \times 10^{-11}$	$7.56 \times 10^{-20} \pm 1.16 \times 10^{-19}$	NA	$1.74 \times 10^{-11} \pm 2.23 \times 10^{-11}$

Table G5.8.3: Fixed-threshold detection and interval hits.

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
C	clean	0.00 ± 0.00	NA	NA	NA	NA
C	noise	0.00 ± 0.00	NA	NA	NA	NA
A	clean	0.00 ± 0.00	NA	NA	NA	NA
A	noise	0.07 ± 0.18	NA	NA	NA	NA
E	clean	0.00 ± 0.00	NA	NA	NA	NA
E	noise	0.00 ± 0.00	NA	NA	NA	NA
raw_clean	clean	0.00 ± 0.00	NA	NA	NA	NA
raw_clean	noise	0.00 ± 0.00	NA	NA	NA	NA
raw_noise	clean	0.00 ± 0.00	NA	NA	NA	NA
raw_noise	noise	0.00 ± 0.00	NA	NA	NA	NA
mlp32_clean	clean	0.00 ± 0.00	NA	NA	NA	NA
mlp32_clean	noise	0.00 ± 0.00	NA	NA	NA	NA
mlp32_noise	clean	0.00 ± 0.00	NA	NA	NA	NA
mlp32_noise	noise	0.00 ± 0.00	NA	NA	NA	NA
dd2_clean	clean	0.00 ± 0.00	NA	NA	NA	NA
dd2_clean	noise	15.81 ± 2.05	NA	0.00 ± 0.00	NA	NA
dd2_noise	clean	0.00 ± 0.00	NA	NA	NA	NA

Continued on next page

Table G5.8.3 (continued)

Source key	Condition	FPR (%)	Recall (%)	Precision (%)	Exact-bin (%)	Within-one-bin (%)
dd2_noise	noise	0.00 ± 0.00	NA	NA	NA	NA
no_i_vector	clean	0.00 ± 0.00	NA	NA	NA	NA
no_i_vector	noise	0.00 ± 0.00	NA	NA	NA	NA
innovation_only	clean	0.00 ± 0.00	NA	NA	NA	NA
innovation_only	noise	0.00 ± 0.00	NA	NA	NA	NA
no_excess	clean	0.00 ± 0.00	NA	NA	NA	NA
no_excess	noise	0.00 ± 0.00	NA	NA	NA	NA
no_i_vector_no_excess	clean	0.00 ± 0.00	NA	NA	NA	NA
no_i_vector_no_excess	noise	0.00 ± 0.00	NA	NA	NA	NA
no_train_envelope	clean	0.00 ± 0.00	NA	NA	NA	NA
no_train_envelope	noise	0.00 ± 0.00	NA	NA	NA	NA
no_envelope_input	clean	0.00 ± 0.00	NA	NA	NA	NA
no_envelope_input	noise	0.00 ± 0.00	NA	NA	NA	NA

Table G5.8.4: Distance and error direction. Metric panel 1 of 2; identifying columns are repeated.

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
C	clean	NA	NA	NA	NA
C	noise	NA	NA	NA	NA
A	clean	NA	NA	NA	NA
A	noise	NA	NA	NA	NA
E	clean	NA	NA	NA	NA
E	noise	NA	NA	NA	NA
raw_clean	clean	NA	NA	NA	NA
raw_clean	noise	NA	NA	NA	NA
raw_noise	clean	NA	NA	NA	NA
raw_noise	noise	NA	NA	NA	NA
mlp32_clean	clean	NA	NA	NA	NA
mlp32_clean	noise	NA	NA	NA	NA
mlp32_noise	clean	NA	NA	NA	NA
mlp32_noise	noise	NA	NA	NA	NA
dd2_clean	clean	NA	NA	NA	NA
dd2_clean	noise	NA	NA	NA	NA
dd2_noise	clean	NA	NA	NA	NA
dd2_noise	noise	NA	NA	NA	NA
no_i_vector	clean	NA	NA	NA	NA

Continued on next page

Table G5.8.4 (continued)

Source key	Condition	Distance MAE	nMAE	Signed bias	Late / miss (%)
no_i_vector	noise	NA	NA	NA	NA
innovation_only	clean	NA	NA	NA	NA
innovation_only	noise	NA	NA	NA	NA
no_excess	clean	NA	NA	NA	NA
no_excess	noise	NA	NA	NA	NA
no_i_vector_no_excess	clean	NA	NA	NA	NA
no_i_vector_no_excess	noise	NA	NA	NA	NA
no_train_envelope	clean	NA	NA	NA	NA
no_train_envelope	noise	NA	NA	NA	NA
no_envelope_input	clean	NA	NA	NA	NA
no_envelope_input	noise	NA	NA	NA	NA

Table G5.8.4: Distance and error direction. Metric panel 2 of 2; identifying columns are repeated.

Source key	Condition	Early (%)	Miss (%)	False boundary (%)
C	clean	NA	NA	0.00 ± 0.00
C	noise	NA	NA	0.00 ± 0.00
A	clean	NA	NA	0.00 ± 0.00
A	noise	NA	NA	0.07 ± 0.18
E	clean	NA	NA	0.00 ± 0.00
E	noise	NA	NA	0.00 ± 0.00
raw_clean	clean	NA	NA	0.00 ± 0.00
raw_clean	noise	NA	NA	0.00 ± 0.00
raw_noise	clean	NA	NA	0.00 ± 0.00
raw_noise	noise	NA	NA	0.00 ± 0.00
mlp32_clean	clean	NA	NA	0.00 ± 0.00
mlp32_clean	noise	NA	NA	0.00 ± 0.00
mlp32_noise	clean	NA	NA	0.00 ± 0.00
mlp32_noise	noise	NA	NA	0.00 ± 0.00
dd2_clean	clean	NA	NA	0.00 ± 0.00
dd2_clean	noise	NA	NA	15.81 ± 2.05
dd2_noise	clean	NA	NA	0.00 ± 0.00
dd2_noise	noise	NA	NA	0.00 ± 0.00
no_i_vector	clean	NA	NA	0.00 ± 0.00
no_i_vector	noise	NA	NA	0.00 ± 0.00
innovation_only	clean	NA	NA	0.00 ± 0.00
innovation_only	noise	NA	NA	0.00 ± 0.00
no_excess	clean	NA	NA	0.00 ± 0.00
no_excess	noise	NA	NA	0.00 ± 0.00
no_i_vector_no_excess	clean	NA	NA	0.00 ± 0.00
no_i_vector_no_excess	noise	NA	NA	0.00 ± 0.00
no_train_envelope	clean	NA	NA	0.00 ± 0.00

Continued on next page

Table G5.8.4 (continued)

Source key	Condition	Early (%)	Miss (%)	False boundary (%)
`no_train_envelope`	noise	NA	NA	0.00 ± 0.00
`no_envelope_input`	clean	NA	NA	0.00 ± 0.00
`no_envelope_input`	noise	NA	NA	0.00 ± 0.00

H Computation and Numerical Materials

H.1 Query Budgets and Readout Overhead

Table H1: Readout and model-query costs in the measured batches.

Combination	Method	Model trajectories	Readout rays	Amortized readout (μs)	Reference-label queries
StickSlip / MLP	FCRG guidance	15	2	50.3	0
StickSlip / MLP	DD2 guidance	15	2	398.7	0
StickSlip / Traj World	FCRG guidance	15	2	50.9	0
StickSlip / Traj World	DD2 guidance	15	2	403.5	0
PuckWall / MLP	FCRG guidance	237	36	397.5	0
PuckWall / MLP	DD2 guidance	237	36	6908.9	0
PuckWall / Traj World	FCRG guidance	237	36	399.3	0
PuckWall / Traj World	DD2 guidance	237	36	6930.9	0

These measurements are batch-amortized readout costs and exclude world-model forward passes and reference-label generation. Probe responses are already included among the original candidate predictions, so FCRG does not require additional probe trajectories. Table H1 reports the query budgets and readout overhead.

H.2 Reference Information and Full Execution Cost

Reference-boundary guidance requires privileged simulator information rather than deployable free input. It uses an average of 888.00 reference-label trajectory queries for StickSlip and 14,109.50 for PuckWall. These costs are kept separate from world-model queries and readout computation. Complete cached-union execution costs and post-selection simulator evaluations remain in Table H2.

Table H2: Actual model execution over the cross-method candidate union. Cached, deduplicated batch execution, not the online cost of an individual method.

Stage	Task	Predictor	Union trajectories	Time (s)	Jobs
base	PuckWall2D	MLP	71424	1.55	24
base	PuckWall2D	TrajWorld	71424	2561.00	24

Continued on next page

Table H2 (continued)

Stage	Task	Predictor	Union trajectories	Time (s)	Jobs
base	StickSlip Block	MLP	5376	1.94	24
base	StickSlip Block	TrajWorld	5376	152.56	24
expanded	PuckWall2D	MLP	335726	8.84	24
expanded	PuckWall2D	TrajWorld	426917	14942.50	24
expanded	StickSlip Block	MLP	27543	2.31	24
expanded	StickSlip Block	TrajWorld	27128	542.37	24

All 817,314 true candidate executions follow committed selections and evaluate coverage, terminal outcomes, and best-candidate cost only. Separating these expenditures from model queries and readout clarifies which information deployment requires, without treating evaluation expense as a method advantage or disadvantage.

H.3 Numerical Data and Main-Text Reference Index

Percentages and percentage-point differences are displayed to two decimal places; AP, NLL, nMAE, Brier scores, and other non-percentage scores are displayed to four decimal places. Means, standard deviations, and confidence-interval endpoints follow the same metric-specific units and precision. Nonzero magnitudes below one unit in the last displayed decimal place use scientific notation with three significant digits. Counts and source identifiers remain integers where applicable; exact experimental settings use their necessary precision. Readout latency is reported in microseconds to one decimal place, total execution time in seconds to two decimal places, and averaged query counts to two decimal places. Displayed values use the reported estimates and metric-specific precision. Tables and figures organize existing results without new training, simulation, calibration, or significance tests. Table H3 links the main claims to their supporting appendix tables.

Table H3: Main-text and appendix evidence cross-reference.

Main text	Appendix destination	Supporting tables
5.1	A.1–A.3; B.1–B.3	Tables A1, A2, and B2
5.2 error matching	C.1; D.1	Tables C1, D1, and D2
5.2 task comparison	C.2	Tables C2, C3a, and C3b
5.2 geometry; Fig. C1	C.3	Tables C4–C8
5.2 high-accuracy example	C.4; D.3	Tables D3 and D4
5.3 measurement and diagnosis	D.1–D.6	Tables D3–D7
5.4 decision utility	E.1–E.7	Tables E1, E2, and E4
5.5 robustness	F.1–F.4	Tables F1–F6
5.6 module ablations	G.1–G.4	Tables G1–G4; Appendix G.4
5.4 computation	H.1–H.2	Tables H1 and H2