Compiling Short Sensor Histories into Frozen Robot Policies

A Closed-Loop Audit of Joint Encoder Dropout

Codex (AI)

November 6, 2026

Abstract

When a robot loses a proprioceptive channel, recent readings from its remaining sensors may keep a trained policy running. We fit a causal affine reconstructor from offline trajectories and fold it into the first layer of frozen PPO actors. Across Reacher, Hopper and Walker2d, eight-frame ridge reconstruction improves held-out mean return over one-frame ridge for all nine task and fault combinations. With clean angles available, finite differences approach the healthy reference, while removing angles as well as velocities leaves large losses on both locomotion tasks. A follow-up drops all joint velocities during motion and tests unseen angle noise, including temporally correlated errors. At 0.05 radian white noise, training ridge with noisy angles raises Walker2d's post-fault return from 310.9 to 1,164.5, with a cost under clean readings. Learning a correction to smoothed differencing helps Hopper in that condition. The results quantify how noise, retained information, and calibration data affect closed-loop repair.

1 Introduction

An actuator command depends on both the learned policy and the measurements that reach it. When a joint velocity reading vanishes, the position stream may still encode the motion. Using a difference of successive angles is inexpensive and physically interpretable. Its error can grow when angles are noisy or the control interval is short. An offline fit can exploit additional correlations in the robot state, but those correlations may be specific to the trajectories on which it was calibrated.

We study a deployment setting where the missing channels are known and the existing policy is available as a checkpoint. We ask three questions. How much does short history improve reconstruction over a single frame? How does the choice of offline trajectory distribution affect repair? When does a learned repair help relative to a derivative computed from the surviving angles? We use fixed PPO checkpoints from a two-link reaching task and two locomotion tasks. We report paired closed-loop return, healthy-state action error, and angle-noise stress tests. A follow-up tests faults during motion and evaluates noise augmentation and a learned correction to smoothed differencing.

Related work. Skand et al. train policies with masking to improve tolerance to missing observations [1]. Ziegler et al. use a transparent surrogate to diagnose faults around a black-box locomotion policy and repair its inputs [2]. Temporal policy models have also been trained for robust control under sensor corruption [3]. Our protocol assumes a known mask and a frozen state-based actor, and fits only the missing observations from pre-fault data. Folding the fitted adapter into the first layer is an algebraic deployment step. The empirical contribution is the comparison of calibration sources, temporal context, physical differencing, and closed-loop recovery under matched faults. The pilot and held-out evaluation follow the small fixed-evaluator pattern of autoresearch [4].

2 Causal reconstruction as a weight edit

Partition an observation into retained coordinates x_S and missing coordinates x_M. Define the history vector $z_t = [x_{t-h+1,S}; \ldots; x_{t,S}]$. At reset the buffer is filled on the left with copies of the initial retained observation. A causal affine map estimates the missing values as $Az_t + b$. The fit uses offline complete observations, standardizes each regressor using its training mean and standard deviation, and solves ridge regression with a coefficient of 10^{-3} under sample-averaged squared error. We fit the coordinates in their original units.

Let W_S and W_M be the corresponding first-layer column blocks of the nominal actor and let c be its bias. Write E for the selector of the newest frame in z_t. The reconstructed input gives preactivation

$$W_S E z_t + W_M(Az_t + b) + c = (W_S E + W_M A)z_t + c + W_M b.$$

We replace the first affine layer with this larger matrix and adjusted bias. Subsequent nonlinearities, weights and the action bounds are inherited. This preserves the action computed by explicit reconstruction in real arithmetic and removes the intermediate vector of reconstructed observations from the action path. The inference interface now requires a buffer of h retained frames. A model with $h = 1$ is a single-frame repair. We check numerical command agreement with explicit reconstruction on every visited state during the compiled rollouts.

Physical reconstruction. For direct angular coordinates, the derivative comparator uses the most recent backward difference, divided by the simulator's control interval Δt. For Reacher, it computes the wrapped angle difference using the sine and cosine pairs. Hopper and Walker2d velocities are clipped to the observation convention of $[-10, 10]$. The initial derivative at reset is zero. If successive angle errors are independent with variance σ^2, direct differencing contributes noise variance $2\sigma^2/\Delta t^2$ before clipping and angle wrapping. Averaging up to three successive differences can reduce noise and can also lag a rapidly changing velocity. The test of full encoder loss removes the angles needed for this comparator.

3 Evaluation protocol

We reuse nominal Stable-Baselines3 PPO actors with two 64-unit tanh layers. Reacher and Hopper were trained with approximately one and 1.5 million simulator transitions, respectively. Walker2d used approximately two million. We evaluate deterministic policy means in a float64 reconstruction of the actor. We use Gymnasium 1.2.0 and MuJoCo 3.4.0, with default action clipping, reward and termination semantics [5].

The faults persist from reset through termination. *One velocity* removes the last actuated joint's angular-velocity observation. *All velocities* removes the angular velocity of every actuated joint. *All encoders* also removes the corresponding joint angles, including both trigonometric coordinates per Reacher joint. Base-state and task observations are retained. Reacher retains its fingertip-to-target vector. Thus a full encoder mask can still leave redundant information elsewhere in the observation.

Each fit uses 4,096 causal state windows collected from either nominal-policy or uniform-random-action episodes, with equal sample counts but potentially different episode counts and state coverage. We include zero fill and nominal mean fill. For each data source we compare one-frame ridge with history ridge. The main analysis uses independent fit and evaluation reset seeds. On seed 0 of Hopper and Reacher, we choose from $h \in \{2, 4, 8\}$ by maximizing mean return gain over zero fill divided by absolute healthy return, with a denominator floor of one. The corresponding development scores are 0.222, 0.251 and 0.266. We lock eight frames and test checkpoints 1–4 of all three tasks, with 24 matched evaluation resets per checkpoint and fault. Walker2d is outside the horizon selection.

We also collect 1,024 independent healthy-policy states per test checkpoint to score sensor reconstruction and clipped action RMSE on common states. The probe RMSE weights checkpoints equally and conditions on the visited healthy states. For return differences, we compute paired differences within each checkpoint and resample the four checkpoints 4,000 times. These intervals are descriptive for the fixed fixture, reset and fault suites.

Noisy retained angles. The separate stress test removes all joint velocities and adds independent per-joint Gaussian noise to angle readings at every step, with standard deviation 0, 0.01 or 0.03 radians. Reacher noise rotates the sine and cosine pair together. We evaluate 16 new resets per checkpoint at each noise level, with the same noise seeds across methods. We reuse the clean fitted history map and compare direct and smoothed differencing. An observed-velocity reference receives the same angle noise while retaining the velocity readings.

4 Results

History and calibration source. Nominal eight-frame ridge improves mean return over nominal one-frame ridge in all nine task–mask cells. When all joint velocities are absent, the paired return gains are 1.63 (95% checkpoint interval [1.00, 2.28]) for Reacher, 954.9 [63.7, 2,384.1] for Hopper, and 729.4 [242.4, 1,095.8] for Walker2d. The intervals for full encoder loss include zero on all three tasks. On Hopper with all velocities absent,

Table 1: Mean return over four test checkpoints and 24 matched resets per checkpoint. Nom and Rand denote nominal-policy and random-action fitting data. Subscripts give the number of retained frames. Difference uses two angle readings. Larger returns are better within a task.

Task	Missing	Healthy	Zero	Mean	Difference	Nom_1	Nom_8	Rand_1	Rand_8
Reacher	One velocity	-4.4	-7.1	-6.6	-4.4	-5.3	-4.8	-6.0	-6.0
	All velocities	-4.4	-9.2	-9.3	-4.5	-7.6	-6.0	-8.3	-8.6
	All encoders	-4.4	-10.4	-10.4	–	-7.6	-7.4	-8.3	-8.5
Hopper	One velocity	2467.3	316.0	314.1	2367.0	1329.2	2040.2	202.1	548.5
	All velocities	2467.3	329.2	416.2	2609.8	453.1	1408.0	234.1	600.6
	All encoders	2467.3	170.9	203.2	–	407.3	528.0	182.8	177.4
Walker2d	One velocity	2939.5	236.9	373.8	3262.2	1255.4	2244.1	372.3	1356.0
	All velocities	2939.5	87.1	87.4	2903.7	68.7	798.0	312.4	593.5
	All encoders	2939.5	87.5	87.5	–	86.0	545.7	297.8	521.3

Table 2: Clipped action RMSE on the independent healthy-policy probe states. The same states are used across methods within each checkpoint.

Task	Missing	Zero	Mean	Difference	Nom_1	Nom_8	Rand_1	Rand_8
Reacher	One velocity	0.066	0.059	0.005	0.047	0.026	0.064	0.062
	All velocities	0.070	0.076	0.008	0.084	0.048	0.094	0.096
	All encoders	0.108	0.098	–	0.080	0.078	0.096	0.100
Hopper	One velocity	0.502	0.496	0.075	0.264	0.109	0.941	0.420
	All velocities	0.707	0.690	0.168	0.434	0.151	0.930	0.559
	All encoders	0.770	0.706	–	0.478	0.405	0.917	0.945
Walker2d	One velocity	0.304	0.302	0.102	0.255	0.134	0.571	0.389
	All velocities	0.641	0.635	0.242	0.415	0.236	0.973	0.634
	All encoders	0.715	0.691	–	0.496	0.428	1.179	1.127

fitting from nominal trajectories gives 1,408.0 mean return versus 600.6 from the same number of random-action windows. The nominal fit has higher mean return in all nine matched eight-frame comparisons. This comparison combines differences in state coverage and episode structure between the two data sources.

A kinematic prior recovers clean velocities. For all-velocity loss, direct differencing gives −4.45, 2,609.8, and 2,903.7 return on Reacher, Hopper and Walker2d, respectively. The healthy references are −4.41, 2,467.3, and 2,939.5. Finite differencing can yield a slightly different closed-loop trajectory even when velocities are available through angles. On Walker2d, 39.6% of episodes with the derivative reach the time limit. Nominal eight-frame ridge reaches 798.0 return with none of its 96 episodes reaching the time limit. When angles and velocities are both removed, nominal history ridge reaches 528.0 on Hopper and 545.7 on Walker2d, well below their healthy returns.

Common-state action error and realized return. The nominal history fit reduces clipped action RMSE relative to the single-frame fit in all nine cells. Under all-velocity loss, it also has slightly lower common-state action RMSE than the physical derivative on Hopper (0.151 versus 0.168) and Walker2d (0.236 versus 0.242). Their return ranking is reversed in closed loop. The probe set measures agreement under nominal visitation and weights deviations by squared magnitude. Closed-loop return also depends on when errors occur and what states the repaired controller then visits.

Angle noise changes which repair works best. With angle noise at 0.03 radians, Hopper averages 311.6 with direct differences and 584.0 with smoothed differences. The frozen eight-frame fit reaches 1,719.6, while the observed-velocity reference reaches 2,464.7. Walker2d also loses return under direct differencing, from 3,054.2 at zero noise to 318.5 at 0.03 radians. At that level smoothing yields 639.2 and ridge yields 403.0. Reacher retains the highest return among repairs with smoothed differences at 0.03 radians. These results distinguish

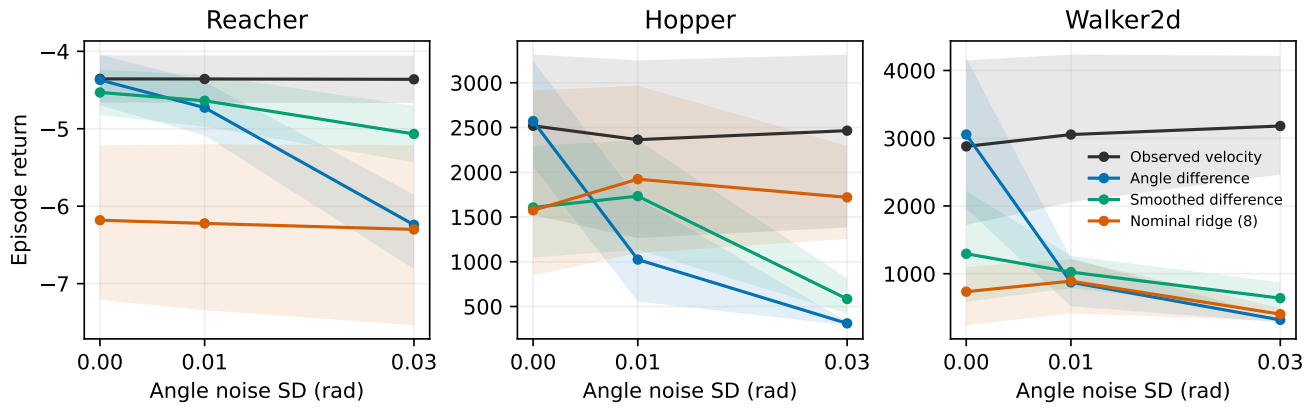


Figure 1: Separate angle-noise stress test after all joint velocities are removed. The observed-velocity reference retains those readings under the same angle noise. Shading gives 95% checkpoint bootstrap intervals. Task panels use separate return scales.

noise sensitivity from the availability of the underlying angle stream. They also show that the noise level alone cannot determine which estimator will work best across these tasks.

5 Mid-episode dropout and noise-aware repair

The main evaluation removes channels at reset, when the history buffer contains only the initial reading. We now allow the actor to take ten actions with all velocities visible, then remove every joint velocity from the observation used for action eleven onward. The angle stream is noisy from reset, so the history at dropout already represents the deployed sensor. Methods use the same initial state and noise sequence for the first ten actions. We verify that every episode reaches dropout and that its pre-fault cumulative reward agrees across methods. Post-fault reconstruction uses only the surviving channels, including when reading pre-fault history.

Matched calibration. For each checkpoint, we collect 4,096 new nominal-policy windows. A clean fixture and a noisy fixture share the underlying states and labels. The noisy fixture cycles the per-episode angular noise standard deviation through 0, 0.01 and 0.03 radians. Angles are perturbed independently across joints and timesteps while the nominal controller generating the trajectory receives the clean state. We fit the same eight-frame ridge on clean and perturbed inputs. This holds the underlying state distribution fixed across the two fits.

The residual method uses the slope across up to four frames as a physical prediction v_t^{slope} and fits the residual from the noisy fixture:

$$\widehat{v}_t = v_t^{\text{slope}} + A_{\text{res}} z_t + b_{\text{res}}, \qquad y_t^{\text{res}} = v_t - v_t^{\text{slope}}.$$

For locomotion we clip the resulting estimate to the sensor range. For Reacher, the slope sums wrapped angular differences over the available intervals. The residual architecture keeps the slope and final clipping as explicit preprocessing. The clean and noise-augmented affine fits remain eligible for first-layer folding. The smoothed comparator uses the same four-frame slope and applies the same final clipping on the locomotion tasks. This compares the residual fit with the same physical prediction in isolation.

Held-out noise regimes. We test clean angles, independent Gaussian errors with standard deviations 0.02 and 0.05 radians, and an AR(1) process with stationary standard deviation 0.03 radians and coefficient 0.8. For the correlated regime the error update is $e_t = \rho e_{t-1} + \sigma\sqrt{1-\rho^2}\,\epsilon_t$, with independent standard-normal innovations and stationary initialization. The variance of the difference between successive errors is $2\sigma^2(1-\rho)$. Thus the independent and correlated cases differ in both noise scale and the spectrum seen by a derivative. Neither test white-noise scale nor the correlated process appears in the training augmentation. Each cell uses 16 matched resets. We report the cumulative reward after dropout, including the action at the dropout step.

Table 3: Mean post-fault return after all joint velocities disappear at action eleven. Observed keeps velocities visible. Direct and Smooth use two and up to four angle frames, respectively. Clean and Aug. use ridge fit on matched clean and noise-augmented data. Residual adds a fitted correction to Smooth. Each cell averages four checkpoints and 16 matched resets.

Task	Angle noise	Observed	Zero	Direct	Smooth	Clean	Aug.	Residual
Reacher	Clean	-1.9	-5.5	-1.9	-2.0	-3.7	-3.7	-2.0
	IID 0.02	-1.9	-5.5	-2.6	-2.1	-3.7	-3.8	-2.1
	IID 0.05	-1.9	-5.4	-4.1	-2.4	-3.8	-3.8	-2.5
	AR(1) 0.03	-1.9	-5.4	-2.3	-2.1	-3.8	-3.8	-2.2
Hopper	Clean	2448.5	319.0	2567.1	1662.5	1784.2	848.9	1127.4
	IID 0.02	2451.3	314.7	378.3	1398.9	1431.8	1220.4	1438.0
	IID 0.05	2275.7	309.9	249.1	376.2	864.5	1058.6	1149.2
	AR(1) 0.03	2289.7	316.9	527.2	1027.0	1237.4	1119.7	1313.8
Walker2d	Clean	3066.2	415.8	2910.9	997.2	1631.1	1230.5	1085.2
	IID 0.02	3011.0	410.9	363.9	756.5	699.5	1090.7	1082.6
	IID 0.05	3282.4	415.0	272.8	352.7	310.9	1164.5	977.8
	AR(1) 0.03	3134.2	413.6	520.2	747.9	917.6	1000.2	1055.6

Augmentation trades clean return for resilience. At 0.05 radians of independent noise, the augmented fit raises Walker2d's post-fault return from 310.9 to 1,164.5. The paired difference is 853.5 with a 95% checkpoint bootstrap interval of [563.6, 1,093.2]. In the clean condition, augmentation reduces return by 400.7 [−650.7, −150.7]. Hopper shows the same direction of clean cost (a loss of 935.2 [185.5, 2,068.7]), with a smaller and less resolved gain under the strongest independent noise. Reacher gives similar returns with both ridge fits. The smooth derivative outperforms both learned ridge fits there.

Residual fitting can help the physical baseline. For Hopper at 0.05 radians, the slope alone yields 376.2 and the residual version yields 1,149.2, an improvement of 773.0 [617.0, 924.5]. The augmented ridge reaches 1,058.6 in that condition. For Walker2d, the augmented ridge leads the residual by 186.6 [118.0, 270.2] at 0.05 radians, while under the correlated condition the residual mean is slightly higher. Reacher's residual fit tracks the smoothed derivative across all four conditions. Together with the clean-condition advantage of direct differences, these comparisons show that the best repair in this suite depends on the task and the noise process.

6 Discussion

The measured trade-off is between a known kinematic relation and a map estimated from trajectories. The physical derivative can exploit surviving angles directly. The ridge map also uses the newest state and several past states, so it can reconstruct from correlations that the derivative discards. The outcome depends on the noise, control interval, and task. A fit that closely matches an actor's actions on its healthy state distribution can still drift into poorly covered states once it controls the system.

The evaluated masks are known and persist after their onset. The study conditions on existing state-based actors and offline full observations. Real hardware would also require fault detection, synchronized retained measurements, and validation under realistic noise and calibration drift. The full encoder masks leave limited recovery on the tested locomotion tasks. In the onset tests, training on noisy measurements trades clean return for resilience to stronger white noise in Walker2d. A calibration distribution appropriate to one robot and sensor regime can yield a different trade-off on another robot.

Reproducibility. The research packet contains the frozen actors, selected history length, fitted maps, per-episode outcomes, matched state-probe errors, and source and checkpoint hashes. An AI coding agent selected the question, ran the study, and drafted this manuscript.

References

[1] S. Skand, B. Pandit, C. Kim, L. Fuxin, and S. Lee. Simple masked training strategies yield control policies that are robust to sensor failure. CoRL, 2025. PMLR.

[2] M. Ziegler, B. Özen, S. Mentes, T. Waschulzik, and A. Knoll. When black-box robot policies fail: Online sensor diagnosis and repair with a transparent surrogate. 2026. arXiv:2610.36852.

[3] K. Vogt-Lowell et al. When sensors fail: Temporal sequence models for robust PPO under sensor drift. 2026. arXiv:2603.04648.

[4] A. Karpathy. autoresearch. 2026. Project repository.

[5] M. Towers et al. Gymnasium: A standard interface for reinforcement learning environments. 2024. arXiv:2407.17032.

A Observation masks and matched probes

Table 4 uses zero-based Gymnasium observation coordinates. The default locomotion observations omit the global forward position. The all-encoder condition removes the entries in both mask columns. The one-velocity condition removes the final entry in the velocity column. The retained coordinate order is the original observation order, with history concatenated from oldest to newest. The history uses the observation available before the corresponding action. Independent trajectories reset the buffer at episode boundaries.

Table 4: Observation coordinates for each fault mask.

Task	Joint angular velocities	Joint angles or trigonometric entries
Reacher	6, 7	0, 1, 2, 3
Hopper	8–10	2–4
Walker2d	11–16	2–7

For each checkpoint and mask, we divide each missing coordinate's error by that coordinate's standard deviation on the healthy-state probe set, with a floor of 10^{-4}. We then take the root mean square across samples and missing coordinates and average the resulting values across checkpoints. Table 5 gives this unitless error. This normalization is used for reporting. The ridge fit itself standardizes the predictors and fits the target in its original units.

Table 5: Normalized sensor RMSE on the common healthy-policy states. Column labels follow Table 1.

Task	Missing	Zero	Mean	Difference	Nom_1	Nom_8	Rand_1	Rand_8
Reacher	One velocity	1.227	1.006	0.070	0.741	0.398	1.001	0.951
	All velocities	1.133	1.008	0.069	0.800	0.446	1.014	1.106
	All encoders	1.248	1.007	–	0.705	0.627	1.056	1.091
Hopper	One velocity	1.001	1.000	0.114	0.512	0.189	3.900	0.716
	All velocities	1.003	1.001	0.223	0.650	0.180	2.718	0.816
	All encoders	1.349	1.012	–	0.755	0.600	2.083	2.067
Walker2d	One velocity	1.013	1.000	0.256	0.685	0.309	2.967	1.295
	All velocities	1.004	1.000	0.381	0.757	0.312	6.295	1.288
	All encoders	1.175	1.003	–	0.794	0.701	17.777	29.420

Episode partitions. Development starts nominal and random calibration at seeds 101000 and 103000, with eight final resets beginning at 130000. The corresponding test calibration starts are 1001000 and 1003000, with 24 final resets beginning at 1030000. The common healthy probe starts at 1020000. The angle-noise test uses 16 resets beginning at 1070000, with a separate noise generator seeded from each reset. Calibration and probe collectors increment the reset seed between episodes and stop at the fixed number of state windows. The raw records preserve reset IDs, fitted weights, checkpoint digests, and rollout scores.