

# Adaptive Spatial Calibration and Co-Adaptation Under Binary Feedback for Sensory Substitution

Ashwin Raj

The Hong Kong University of Science and Technology

18 October 2026

## Abstract

Sensory substitution aims to convey visual information through another sense. Proposed corneal interfaces, for example, would translate visual patterns into spatial tactile cues. Keeping these cues aligned with their intended locations may require recalibration as the device response and the user’s interpretation change. We investigate this problem in simulation using occasional left/right and above/below feedback. Six calibration methods are tuned on separate simulations and compared across 126 conditions, with 1,000 paired simulated users per condition. The main user-learning model receives the same noisy directional reports as the device. Across 20 core conditions, a dynamic logistic filter that tracks uncertainty reduced mean spatial error from 0.1131 to 0.1058 compared with fixed-size updates, an average reduction of 6.5%. It also increased the fraction of trials within an illustrative target radius of 0.10 from 50.6% to 54.7%. Its advantage persisted across the tested response-model mismatches, but reversed in four conditions with doubled drift. Allowing the simulated user and device to learn from the same noisy reports sometimes increased error. These findings show why calibration methods should be evaluated together with the feedback available to each learner and the speed of change. An expanded search over 111 candidates at 120 and 480 trials selected weak forgetting and slower decay for longer sessions, while faster drift still reversed some rankings. Together, the experiments provide a computational benchmark for selecting calibration policies in sensory-substitution research.

**Keywords:** sensory substitution; spatial calibration; binary feedback; dynamic logistic regression; co-adaptation; system identification; simulation.

## 1 Introduction

Sensory substitution seeks to convey visual information through a different sensory channel. Proposed corneal systems would encode camera images as patterns of tactile stimulation, using spatial cues to represent aspects of the visual scene [1]. Preliminary experiments have measured corneal tactile spatial discrimination and recognition of simple stimulated shapes [2]. These studies motivate an engineering question: how can a device keep a spatial cue aligned with its intended location when the response to stimulation and the user’s interpretation both change?

Consider a cue intended to represent a point to the left of a target, but interpreted as lying to its right. Initial calibration can reduce this mismatch, yet a changing spatial mapping may gradually invalidate it. A practical feedback task might ask only whether the cue lies left or right, and above or below, a reference. We study how calibration can use these occasional directional reports while the user also adapts. We represent this calibration task using a two-dimensional spatial map motivated by corneal sensory substitution.

This question combines two distinct sources of change: uncertainty remaining after a small calibration sample and evolution of the response channel. A large update can correct either, but can also amplify response noise. When user and device act on the same noisy report, their corrections can reinforce

each other. An evaluation that assumes an exactly known observation model or gives the user perfect continuous error information can therefore favour a policy for reasons that will not transfer to another feedback protocol.

We study a deliberately small spatial channel so that these mechanisms remain inspectable. The contribution is an experimentally controlled comparison: independent-seed hyperparameter selection, a full-covariance logistic estimation baseline, a user learner restricted to shared binary reports, and a complete grid of mismatch and sensitivity tests. Two elementary calculations clarify why shared report noise can amplify corrections and why unknown reporting bias cannot be separated from positional offset using this feedback alone. The resulting evidence identifies where a covariance-based estimator helps and where faster drift reverses that advantage. The expanded tuning study also tests how the preferred update schedule changes with session length.

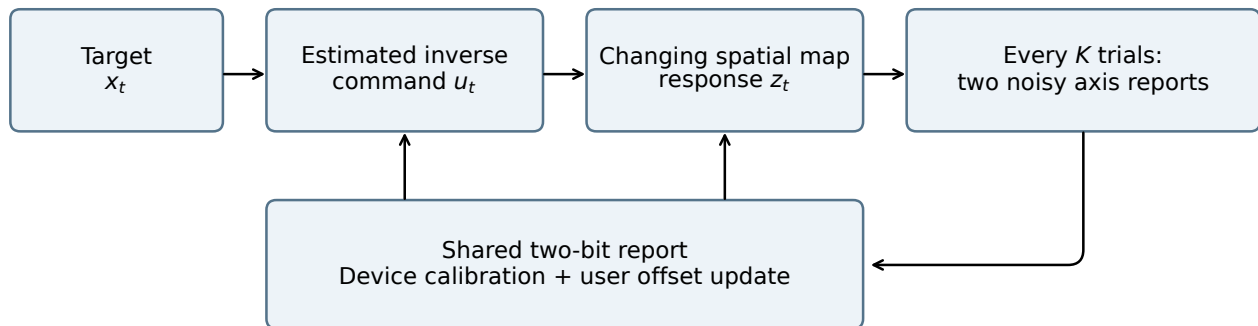


**Figure 1:** The primary closed loop. Every third or eighth trial supplies two binary axis comparisons against a jittered reference. Both device and binary user learner see the same report. Scoring uses the simulated latent response before either learner updates; that score is unavailable to both learners.

## 2 Related Work and Scope

Granley et al. [3] combine deep stimulus encoding with preferential Bayesian optimisation for visual prostheses. Their experiments already test response noise and forward-model misspecification. This establishes a foundation for limited-feedback personalisation and robustness testing. Their feedback consists of preferences between competing percepts; our protocol supplies two axis comparisons for one spatial command. The present benchmark examines directional spatial feedback within this broader personalisation problem.

Schoinas et al. [4] test that optimisation approach with 17 sighted participants viewing simulated prosthetic images. Their findings distinguish human preferences from the simulated decision rule. That distinction motivates an explicit mathematical sensitivity model for the information available to the user.

Co-adaptation has been modelled through joint encoder-decoder optimisation and multi-agent control [5, 6]. Those studies already show why device changes and user learning must be considered together. Our contribution is narrower: the consequences of giving both learners the same intermittent, noisy categorical observation, with a transparent comparison to a richer-information diagnostic.

Recursive estimation from quantised measurements is also established. Liu et al. [7] develop weighted least-squares identification under Gaussian inputs and fixed thresholds. Their assumptions differ from the bounded commands, changing references and policy-dependent trajectories here. McCormick et al. [8] describe dynamic logistic regression with covariance forgetting and model averaging. We use a single-model, fixed-forgetting, one-step Laplace update from that family. Work on online logistic regression

also connects second-order updates with extended Kalman filtering [9]. The present experiments evaluate this approach in a drifting, misspecified closed loop.

These distinctions define the research question: *after choosing settings on separate synthetic users, which empirical rankings transfer when user feedback, observation assumptions and drift rates change?* The study supplies a reproducible benchmark for this question.

### 3 Spatial Channel and Feedback

#### 3.1 Latent response and exogenous drift

Targets are  $x_t \in [-0.6, 0.6]^2$  and commands are  $u_{m,t} \in [-1, 1]^2$ . For policy  $m$ , the response before reporting or learning is

$$z_{m,t} = A_t u_{m,t} + b_{m,t} + \beta u_{m,t}^{\odot 3}, \quad A_t = R(\theta_t) \begin{bmatrix} \exp(\ell_{x,t}) & h_t \\ 0 & \exp(\ell_{y,t}) \end{bmatrix}. \quad (1)$$

The cubic term is zero in the core experiment. Initial angles are uniform on  $[-0.32, 0.32]$  radians, scales on  $[0.8, 1.2]$ , shear on  $[-0.1, 0.1]$ , and each offset on  $[-0.12, 0.12]$ . Independent Gaussian innovations produce the five scenarios below. Angles, log-scales and shear are clipped in the retained state to  $[-0.65, 0.65]$ ,  $[\log .7, \log 1.3]$  and  $[-0.15, 0.15]$ . Offsets have no exogenous clipping. Innovations occur before each scored test trial.

**Table 1:** Per-trial innovation standard deviations. Offset and log-scale values apply separately to each axis. Sensitivity runs multiply every nonzero value by 0.5 or 2.

| Scenario   | Offset | Angle | Log-scale | Shear |
|------------|--------|-------|-----------|-------|
| Stationary | 0      | 0     | 0         | 0     |
| Offset     | 0.008  | 0     | 0         | 0     |
| Rotation   | 0      | 0.008 | 0         | 0     |
| Shape      | 0      | 0     | 0.003     | 0.004 |
| Mixed      | 0.006  | 0.006 | 0.002     | 0.003 |

All coordinates and success tolerances are dimensionless. Table 1 defines the experimental drift magnitudes.

#### 3.2 Calibration and binary reports

Each synthetic user receives 25 calibration commands on a  $5 \times 5$  grid over  $[-0.65, 0.65]^2$ . Reference coordinates equal those commands plus independently uniform offsets in  $[-0.18, 0.18]$ . The grid is shuffled, although order does not affect the full-batch fit. At test, targets are uniform on  $[-.6, .6]^2$ , and references are  $q_t = x_t + \delta_t$ , with independent  $\delta_{t,a} \in \{-.16, +.16\}$ . Reports occur every  $K \in \{3, 8\}$  trials, producing 40 or 15 report occasions in 120 trials. There are 50 initial calibration bits and 80 or 30 test bits. The final test report does not influence a subsequent scored trial.

For axis  $a$ , a report is

$$r_{m,t,a} = \mathbb{I}[z_{m,t,a} - q_{t,a} + c_a + \epsilon_{t,a} > 0], \quad (2)$$

except that a lapse replaces the label by an independent fair binary draw. In core conditions  $c = 0$ , there are no lapses, and  $\epsilon$  is logistic with scale  $\tau = .12$ . The task supplies a spatial reference for each directional comparison.

All device models use an unbiased logistic observation with scale  $\tau_0 = 0.12$  and an affine response. The common initial estimate  $B = [\hat{A} \mid \hat{b}]$  starts at  $[I \mid 0]$  and takes 70 full-batch score-ascent steps

of size .025 on the calibration reports. A diagonal preconditioner  $D = \text{diag}(d_1, d_2, d_3)$ , with  $d_j = (25^{-1} \sum_k \phi_{k,j}^2 + 0.05)^{-1}$  and  $\phi_k = (u_k^T, 1)^T$ , is fixed for later gradient updates. Because all users receive the same grid, its feature moments are identical despite shuffling.

### 3.3 User learning with an explicit information budget

In the primary binary-user condition, the user offset is adjusted only after a report:

$$b_{m,t}^{\text{after}} = b_{m,t}^{\text{before}} - a(2r_{m,t} - 1), \quad a = .008. \quad (3)$$

The device and user use exactly the same noisy labels. The correction points toward the jittered reference; symmetry of the reference jitter centres its expected direction on the target in the unbiased model. The learner uses report signs and updates only on report trials.

We compare it to no user learning and to the rich-information diagnostic  $b^{\text{after}} = b^{\text{before}} - .02(z - x)$  after every trial. The latter receives exact continuous error. Its update frequency and gain are specified separately from Eq. 3. Binary-step sensitivity runs use  $a = .004$  and  $.016$ . All learner updates follow scoring and device feedback processing, and the next exogenous innovation precedes the next trial.

## 4 Policies and Independent Tuning

For  $\phi = (u^T, 1)^T$ , every learned model predicts

$$p_a = \sigma((B_a \phi - q_a)/\tau_0), \quad G_a = (r_a - p_a)\phi^T/\tau_0. \quad (4)$$

Frozen calibration leaves  $B$  fixed. Constant gain takes  $B^+ = B + \eta GD$ . Decaying gain replaces  $\eta$  by  $\eta/[1 + (n-1)/\kappa]$  at report  $n$ . Offset-only gain updates only the last column. The excess-loss gate uses the full step when the mean binary log loss minus predicted Bernoulli entropy exceeds  $H$ , and  $.175\eta$  otherwise. Subtracting entropy makes the excess exactly zero at  $p_a = 0.5$ .

### 4.1 Covariance-based logistic baseline

For each output axis, the filter maintains a  $3 \times 3$  covariance  $P_a$ . Its initial mean is the same calibration fit as the other policies, and

$$P_{a,0} = \left[ 4I + \sum_{k=1}^{25} p_{k,a}(1 - p_{k,a})\phi_k\phi_k^T/\tau_0^2 \right]^{-1}. \quad (5)$$

The ridge  $4I$  is a fixed covariance regulariser. At a report after  $K$  trials, set  $P_a^- = P_a/\lambda^K$ ,  $v_a = P_a^- \phi/\tau_0$ ,  $w_a = p_a(1 - p_a)$  and  $d_a = 1 + w_a \phi^T v_a/\tau_0$ . One Newton/Laplace step gives

$$B_a^{+T} = B_a^T + v_a(r_a - p_a)/d_a, \quad (6)$$

$$P_a^+ = P_a^- - w_a v_a v_a^T/d_a. \quad (7)$$

The filter can exploit cross-feature covariance and accumulated information. It therefore provides a substantive comparator to a fixed diagonal score update. The update uses a Gaussian approximation for the Bernoulli observation model.

All policies issue  $u = \text{clip}(\hat{A}^{-1}(x - \hat{b}), -1, 1)$ . A proposed estimate is rejected if its determinant magnitude is at most 0.18, condition number at least 6, or any parameter is nonfinite. A rejected filter update retains its prior covariance and previous mean. Numerical inversion failure returns a zero command. Clipping and rejection rates are recorded.

## 4.2 Selection and held-out design

A total of 44 candidate configurations are evaluated on 20 matched conditions: five scenarios, two reporting intervals and two primary learners. Each tuning condition uses three seed blocks of 128 users. We select one configuration per policy family by the equally weighted mean error over conditions. Selection has access to synthetic latent errors offline; during testing, learners receive only their defined reports.

The original grid used six gains,  $G = \{0.0005, 0.001, 0.002, 0.004, 0.008, 0.016\}$ . Decay counts were 6, 12 and 24 reports. Gate thresholds were 0.05, 0.15 and 0.30, paired with gains 0.002, 0.004 and 0.008. Filter forgetting factors were 0.98, 0.99, 0.995 and 1. These settings define the original 126-condition evaluation. Table 2 gives the original selections alongside the expanded follow-up grid and its horizon-specific selections.

**Table 2:** Original selections and expanded tuning at two horizons.  $G$  contains the six gains defined in the text. Decay entries list  $(\eta, \kappa)$  and gate entries  $(\eta, H)$ . A decay count of  $\infty$  is implemented exactly as constant gain;  $H = -0.7$  makes the gate always open.

| Policy   | Expanded grid                                                          | Original    | 120 trials    | 480 trials    |
|----------|------------------------------------------------------------------------|-------------|---------------|---------------|
| Constant | $\eta \in G$                                                           | 0.004       | 0.004         | 0.002         |
| Decay    | $\eta \in G; \kappa \in \{6, 12, 24, 48, 96, 192, \infty\}$            | 0.004, 24   | 0.004, 48     | 0.004, 96     |
| Offset   | $\eta \in G$                                                           | 0.002       | 0.002         | 0.002         |
| Gate     | $\eta \in G; H \in \{-0.7, -0.15, -0.05, 0, 0.025, 0.05, 0.15, 0.30\}$ | 0.004, 0.05 | 0.004, $-0.7$ | 0.002, $-0.7$ |
| Filter   | $\lambda \in \{0.90, 0.95, 0.98, 0.99, 0.995, 0.998, 0.999, 1\}$       | 1           | 0.999         | 0.998         |

The expanded study was specified after inspection of the original results. It evaluates 111 candidates separately at 120 and 480 trials, using the same 20 core conditions and three fresh blocks of 128 users per condition (seed split 2). Selection minimises equally weighted all-trial mean error. The decay grid includes its constant-gain limit, and the gate grid extends to its always-open endpoint because excess loss is bounded below by  $-\log 2$ . The filter grid tests both stronger forgetting and finer values near its natural endpoint  $\lambda = 1$ .

Both horizon-specific selections are frozen before testing on seed split 3. Each horizon uses five blocks of 200 users in 20 core and 16 doubled-drift conditions. The six new selections and the original decay, gate and filter are evaluated together, yielding 648,000 user-policy outcomes. Paired marginal bootstrap intervals use 2,000 resamples; these follow-up comparisons are exploratory. The two horizons are distinct simulation panels because array sizes change the random-stream allocation. The original results remain intact, allowing direct assessment of the wider search on new test users.

Test simulations use disjoint seed identifiers and five blocks of 200 users per condition. Independent random streams control initial maps, biases, response noise, lapses, targets, innovations and reference jitter. Within a scenario, condition variants replay the same underlying streams; across policies, commands and labels can differ even though random disturbances are paired. The test set includes 20 core conditions, 10 rich-user diagnostics, 56 observation/model mismatch conditions, 32 drift-amplitude conditions and eight binary-step conditions: 126 total, with 756,000 user-policy outcome rows. Condition variants reuse paired synthetic users.

Mismatch variants apply during both calibration and testing: true logistic scale .06 or .24; per-user, per-axis fixed bias uniform on  $[-0.08, 0.08]$ ; a 10% random-label lapse probability; cubic coefficient  $\beta = .2$ ; Gaussian noise with standard deviation  $.12\pi/\sqrt{3}$ ; and a combination of higher noise, bias, lapses and nonlinearity. Each is tested in stationary and mixed scenarios, both report intervals and both primary learners. Half/double drift uses the four nonstationary scenarios.

### 4.3 Outcomes and uncertainty

The primary outcome is each user’s mean pre-update latent error  $\bar{e} = T^{-1} \sum_{t=1}^T \|z_t - x_t\|_2$ , with  $T = 120$  in the original study and  $T \in \{120, 480\}$  in the tuning extension. Secondary outcomes are error over the last 30 trials, the fraction of trials within radii .05, .10 and .20, clipping/rejection rates and the fraction of users worse under a comparator. Tolerance success measures the fraction of trials inside each specified radius.

Paired differences resample users, retaining each complete trial trajectory. We report marginal percentile-bootstrap 95% intervals from 2,000 resamples. The primary comparison family consists of constant minus decay and filter minus constant in all 20 core conditions. Two-sided paired sign-flip tests use 9,999 randomisations and Holm adjustment across these 40 comparisons. Sign-flip validity assumes exchangeability of paired labels under the null (equivalently, symmetry for a difference-based randomisation interpretation). Marginal bootstrap intervals are not simultaneous intervals and need not agree with adjusted decisions. All other intervals and sensitivity comparisons are exploratory. Equal-condition averages and counts of winning conditions are descriptive, because condition variants are dependent.

The revised protocol was written before running these new experiments, after inspecting an earlier version of this simulation. This is exploratory research, not a preregistered confirmatory study. Seeds, selected settings, candidate surfaces, analysis code and protocol hashes are included for deterministic replay.

## 5 Results

### 5.1 Covariance estimation changes the primary comparison

Across the equally weighted core grid, mean errors were 0.1338, 0.1275, 0.1139, 0.1131, 0.1122 and 0.1058 for frozen, offset-only, gated, constant, decaying and filter policies, respectively. The filter–constant absolute reduction was 0.0073 (6.5% relative to constant gain). All 20 paired comparisons favoured the filter with Holm-adjusted  $p = .004$ ; this value reflects the finite resolution and shared lower bound of the randomisation estimates, not identical underlying effect sizes. The full core grid appears in Appendix A.

Constant gain had lower mean error than decay in eight of twenty conditions, but only three improvements survived multiplicity correction; eleven conditions favoured decay after correction. In mixed drift with  $K = 3$  and no user learning, constant gain improved by only 0.71% relative to the independently tuned decay schedule, with marginal interval  $[-0.0014, -0.0002]$  and adjusted  $p = 0.0708$ . The gate had no lower-mean core condition than decay; fifteen exploratory marginal intervals favoured decay.

Across core conditions, 14.2–42.4% of users had higher error under the filter than under constant gain. The fraction of trials within radius .10 averaged 54.7% for the filter versus 50.6% for constant gain, a gain of 4.1 percentage points. The numerical archive also reports results at radii 0.05 and 0.20.

### 5.2 The assumed user learner changes the interpretation

Figure 3 compares mixed drift at  $K = 3$ . With no user learning, constant and filter errors were 0.1080 and 0.1025. Shared-report binary learning increased these to 0.1116 and 0.1037, even though frozen calibration improved from 0.1446 to 0.1302. The rich exact-error learner instead gave constant and filter errors of 0.0934 and 0.0845. Thus a learner that helps an unchanging device can impair an already adapting device, and an exact-error model can obscure that interaction.

A local scalar calculation gives one plausible mechanism. Suppose the true and estimated slopes are both near one, hold the predicted probability fixed for a one-step approximation, and consider only offset

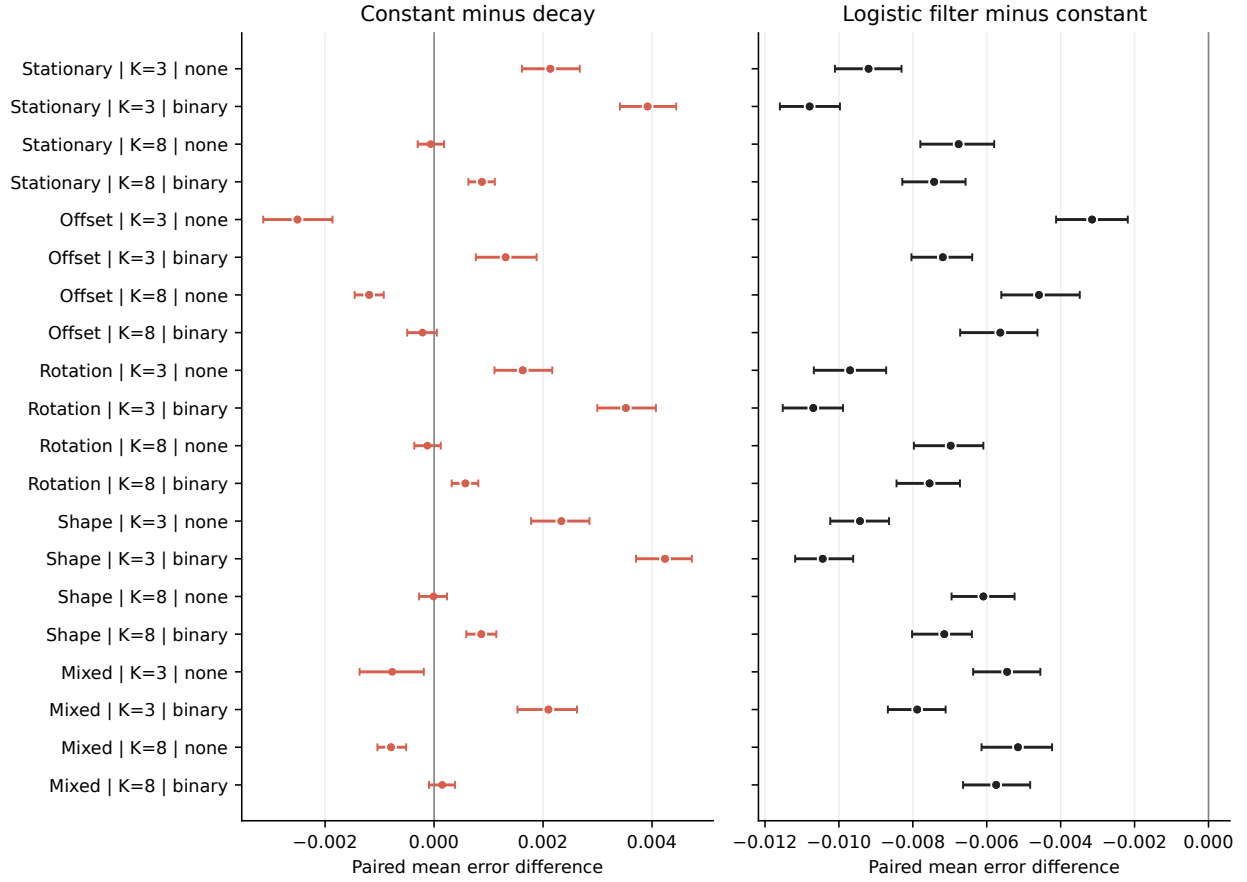


**Figure 2:** All primary paired differences with marginal bootstrap 95% intervals. Negative values favour the first policy. White-edged markers indicate Holm-adjusted  $p < .05$  in the 40-comparison primary family. Rows labelled “binary” use shared-report user learning.

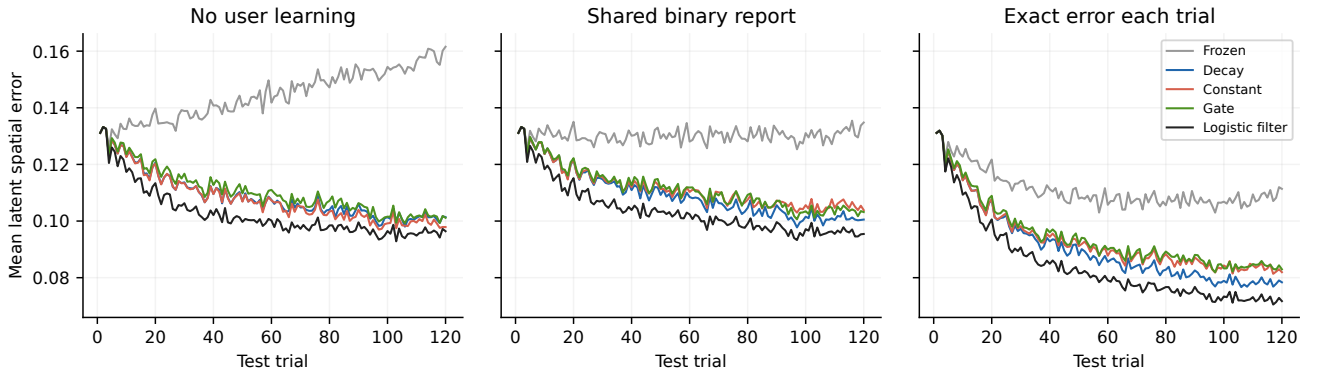


**Figure 3:** Mean trajectories for mixed drift and  $K = 3$ , using the same selected settings. The rich learner has more precise feedback, more frequent updates and a different gain; these panels compare complete learner assumptions rather than isolate a single causal factor.

corrections. A shared binary residual contributes device correction  $k(r - p)$ , with  $k = \eta D_{bb}/\tau_0$ , and user correction  $2a(r - p)$ . Their combined conditional noise variance is  $(k + 2a)^2 p(1 - p)$ . Separate independent reports with the same Bernoulli probability would give  $(k^2 + 4a^2)p(1 - p)$ ; sharing adds  $4akp(1 - p)$ . This one-step calculation identifies a source of reinforcing noise in simultaneous adaptation.

### 5.3 Mismatch and bias limit the meaning of success

The filter retained lower mean error than constant gain in every one of the 56 mismatch conditions, with exploratory marginal intervals below zero. Table 3 aggregates eight conditions per row for compact comparison; the corresponding individual results remain in the archive. Under the combined stress, mean errors increased to 0.1466 for constant and 0.1393 for the filter, and success within .10 was 33.9% and 36.6%. Absolute performance deteriorated despite the relative ranking.

**Table 3:** Mismatch summaries, averaged equally over stationary/mixed drift,  $K = 3/8$  and no/binary user learning. “Success” is the percentage of trials within radius .10. Rows are descriptive averages across the specified conditions.

| Variant         | Constant error | Filter error | Constant success | Filter success |
|-----------------|----------------|--------------|------------------|----------------|
| Low noise       | 0.1227         | 0.1069       | 48.3             | 55.6           |
| High noise      | 0.1335         | 0.1272       | 38.6             | 41.5           |
| Fixed bias      | 0.1310         | 0.1234       | 40.0             | 43.1           |
| Random lapses   | 0.1158         | 0.1094       | 48.5             | 52.0           |
| Cubic mapping   | 0.1127         | 0.1054       | 51.0             | 55.0           |
| Gaussian noise  | 0.1121         | 0.1051       | 50.8             | 54.8           |
| Combined stress | 0.1466         | 0.1393       | 33.9             | 36.6           |

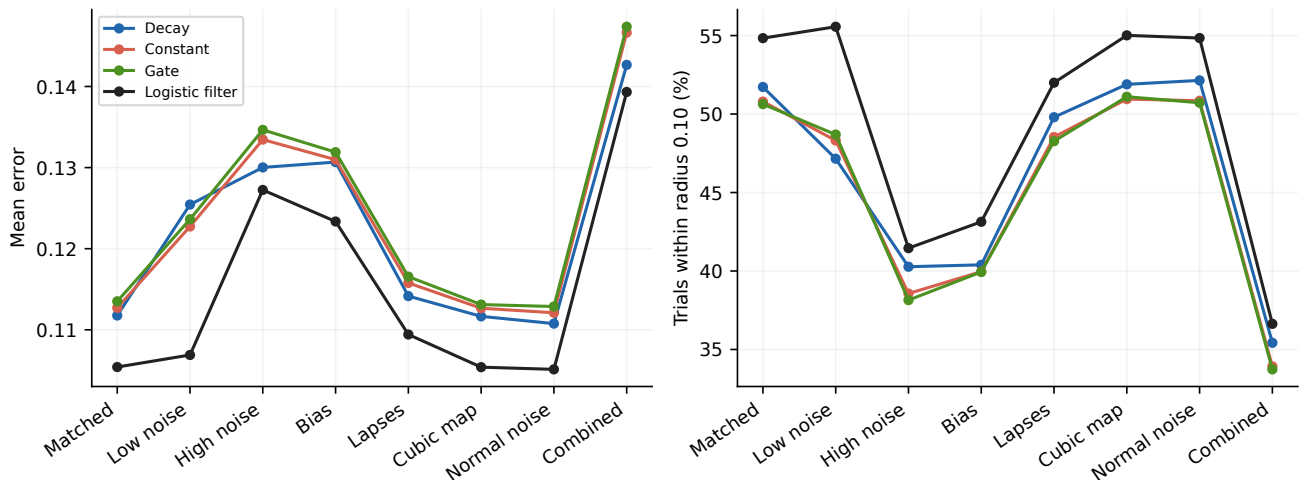


**Figure 4:** Observation and model mismatch, including the matched baseline. Methods retain settings selected on matched tuning data. Success uses the same dimensionless tolerance in every condition.

Lower true response noise did not uniformly reduce error. The estimator still assumed  $\tau_0 = .12$  when labels used .06, including during calibration; the slope of the fitted observation function and the finite-sample calibration fit were therefore misspecified. The low-noise panel demonstrates that less noisy feedback and a better matched estimator are different interventions.

Unknown reporting bias creates an additional identification limit. In the affine logistic model,

$$\Pr(r_a = 1 \mid u, q) = \sigma((A_a u + b_a + c_a - q_a)/\tau). \quad (8)$$

For any constant vector  $d$ , replacing  $b$  by  $b + d$  and  $c$  by  $c - d$  leaves every report probability unchanged while changing the latent response. Binary reports alone therefore identify  $b + c$ , not its two components. This is a direct invariance of the observation model. Better likelihood fitting cannot guarantee smaller latent error without an unbiased anchor or further assumptions. The bias and combined conditions expose a failure mode that cannot be solved simply by increasing the gain.

## 5.4 Sensitivity and counterexamples to dominance

The held-out gain sweep gives mean core errors of 0.1151, 0.1131 and 0.1255 for half, selected and double constant gains. Corresponding decay errors are 0.1173, 0.1122 and 0.1169. This supports the local usefulness of the selected grid point while showing substantial sensitivity to an overly large constant step.

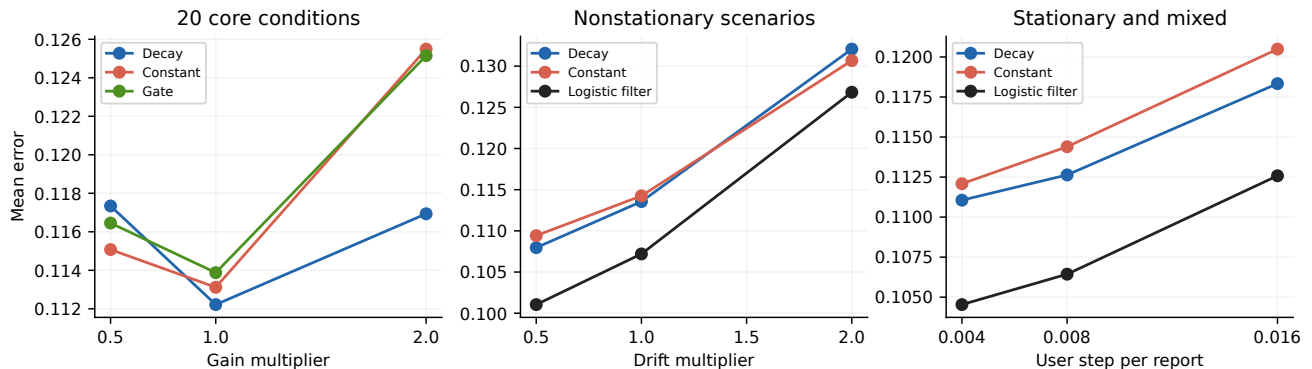


**Figure 5:** Descriptive held-out sensitivities. Left: half/base/double selected gains over 20 core conditions. Centre: drift amplitude over the four nonstationary scenarios. Right: binary-user step over stationary/mixed scenarios. The tuning choice is not revised using these panels.

Across all 126 conditions the filter had lower mean error than constant gain in 122. The four reversals all occurred under doubled drift. In doubled offset drift without user learning at  $K = 3$ , the filter’s error exceeded constant gain by 0.0076, with marginal interval  $[0.0064, 0.0088]$ . Doubled mixed drift at  $K = 3$  without learning also favoured constant gain; the other two reversals had intervals crossing zero. These exploratory counterexamples are consistent with the selected no-forgetting filter losing tracking ability when drift is stronger than in its tuning distribution.

## 5.5 Expanded tuning and longer sessions

The wider search moves the selected decay count from 24 reports to 48 at 120 trials and 96 at 480 trials, with initial gain  $\eta = 0.004$  in both cases. Constant gain remains 0.004 at 120 trials and changes to 0.002 at 480 trials. The filter selects  $\lambda = 0.999$  and 0.998, respectively. Both forgetting choices are interior points of the expanded grid. The gate selects  $H = -0.7$  at each horizon, with the same gain as constant updating. This is its exact always-open endpoint: the selected gate and constant policy have identical trajectories.

At 120 trials, the retuned filter reduces core mean error from 0.1134 to 0.1055 (7.0%) and has lower error in all 20 conditions, with all exploratory marginal intervals below zero. At this horizon the retuned filter’s average improvement over the original filter is less than 0.0001. Expanded decay has lower mean error than constant gain in 16 core conditions, with 13 marginal intervals below zero; its core average changes from 0.1126 for the original decay to 0.1125.

At 480 trials, the retuned filter reduces core average error from 0.1017 to 0.0945 (7.1%), improving 18

**Table 4:** Expanded selections on fresh held-out users, with equal weight per condition. Core panels contain 20 conditions and doubled-drift panels contain 16. Entries are all-trial mean spatial error. The original filter uses  $\lambda = 1$  on these same new users. The selected gate equals constant gain.

| Trials | Panel         | Constant | Decay  | Filter | Original filter |
|--------|---------------|----------|--------|--------|-----------------|
| 120    | Core          | 0.1134   | 0.1125 | 0.1055 | 0.1055          |
| 120    | Doubled drift | 0.1317   | 0.1323 | 0.1270 | 0.1273          |
| 480    | Core          | 0.1017   | 0.0994 | 0.0945 | 0.0953          |
| 480    | Doubled drift | 0.1415   | 0.1345 | 0.1364 | 0.1416          |

conditions and worsening two; all 20 marginal intervals exclude zero in their respective directions. Its core error is lower than the original filter’s 0.0953, with last-30-trial error changing from 0.0910 to 0.0887. Expanded decay also improves its core average from 0.1014 to 0.0994. Thus the wider grid changes the selected settings and strengthens the decay comparator while retaining an average filter advantage.

Transfer to faster drift remains condition dependent. At 120 trials, the retuned filter loses to constant gain in three of 16 doubled-drift conditions, with two marginal intervals above zero. At 480 trials it loses in four, with three intervals above zero, despite a lower panel average (0.1364 versus 0.1415). Expanded decay has the lowest average among the six newly selected policies in that panel (0.1345). The original gate also outperforms the newly selected always-open gate on average there (0.1346 versus 0.1415). Optimising the core distribution therefore leaves meaningful transfer failures even after the original boundary selections are resolved.

## 6 Discussion and Limitations

The experiments distinguish correction of calibration uncertainty from tracking of a changing spatial map. The covariance-based filter uses limited observations efficiently in the original short sessions. The expanded tuning results show how session length changes the preferred update schedule and quantify its transfer to faster drift. Selection on independent synthetic users makes these comparisons reproducible, while the original and follow-up evaluations expose the consequences of the tuning design.

The scope is a computational spatial channel motivated by sensory substitution. Its affine map, drift amplitudes and response noise are experimental settings rather than estimates from corneal physiology or a device population. Drift components are not matched in perceptual energy, so their rankings do not establish intrinsic geometric difficulty. The cubic distortion is one nonlinear family. Dimensionless error reductions and tolerance success require empirical calibration before they can support claims about perception, task performance or clinical benefit.

The user learner is an engineered sign-correction rule. A practical reference-comparison task would need to establish that users can interpret the reference and provide the assumed directional feedback. Actual learners may change strategy, fatigue or use other sensory cues. The rich-feedback diagnostic changes information, update frequency and gain together; it therefore measures the effect of that complete learner model. The one-step shared-noise calculation explains a local mechanism, rather than global stability of the affine closed loop.

The wider search includes the constant-gain and always-open endpoints for decay and gate, and tests forgetting over a fourfold longer horizon. Its equal weighting of scenarios and all-trial error objective remain design choices. Deployment frequencies, still longer sessions or a late-window objective could select different settings. The extension was prompted by the original results and is reported as an exploratory follow-up, with new tuning and test seeds. The fixed calibration optimiser, covariance regulariser and model noise scale are additional tuning dimensions. A diagonal-covariance ablation would help separate covariance structure from other differences between the filter and gradient updates.

Adaptive forgetting, particle methods, active reference selection and nonlinear estimators remain useful comparators.

The filter uses an approximate Gaussian update for binary observations. The gate’s entropy-adjusted loss is a heuristic signal rather than a calibrated drift detector. Likewise, paired bootstrap intervals and randomisation tests quantify variation within the simulator; condition variants share synthetic users and are not independent replications. These boundaries define the interpretation of the benchmark, while its archived condition-level outcomes allow other engineering objectives to be examined.

An empirical next step is a reference-comparison task that measures response noise, reporting bias and learning dynamics, followed by prospective evaluation of frozen calibration policies. That study could test shared-report adaptation, relate spatial tolerances to task outcomes, and determine which tuning horizon reflects actual use. The present benchmark supplies concrete policy comparisons and failure cases for that evaluation.

## 7 Conclusion

With independent tuning and a covariance-based comparator, sparse-feedback spatial calibration produced a different emphasis from a fixed-gain-only comparison. The logistic filter improved all 20 core contrasts against constant gain, but stronger drift provided counterexamples. Shared noisy user feedback could impair simultaneous adaptation, and unknown report bias remained inseparable from positional offset. These findings support evaluating calibration uncertainty, the information supplied to each learner, and mismatch transfer together. Expanded tuning selects weak forgetting, slower decay over longer sessions and an always-open gate; the filter retains lower average core error, while faster drift still changes the policy rankings. Together, these results provide a reproducible basis for choosing and testing spatial-calibration policies.

## Tool Use Statement

AI-assisted tools supported code analysis, copy-editing, and formatting. Responsibility for the scientific content and final submission rests with the author.

## A Complete Core Condition Grid

**Table 5:** Mean latent spatial error in all 20 core conditions, 1,000 paired users each. B denotes shared binary user learning; N denotes none. Settings were selected on disjoint tuning seeds.

| Drift      | $K$ | User | Frozen | Decay  | Constant | Offset | Gate   | Filter |
|------------|-----|------|--------|--------|----------|--------|--------|--------|
| Stationary | 3   | N    | 0.1318 | 0.0972 | 0.0993   | 0.1177 | 0.1012 | 0.0901 |
| Stationary | 3   | B    | 0.1196 | 0.1018 | 0.1057   | 0.1190 | 0.1047 | 0.0949 |
| Stationary | 8   | N    | 0.1318 | 0.1133 | 0.1132   | 0.1252 | 0.1150 | 0.1064 |
| Stationary | 8   | B    | 0.1262 | 0.1152 | 0.1161   | 0.1245 | 0.1159 | 0.1087 |
| Offset     | 3   | N    | 0.1556 | 0.1157 | 0.1132   | 0.1366 | 0.1159 | 0.1100 |
| Offset     | 3   | B    | 0.1378 | 0.1139 | 0.1152   | 0.1326 | 0.1150 | 0.1080 |
| Offset     | 8   | N    | 0.1556 | 0.1323 | 0.1311   | 0.1467 | 0.1335 | 0.1265 |
| Offset     | 8   | B    | 0.1474 | 0.1311 | 0.1308   | 0.1427 | 0.1314 | 0.1252 |
| Rotation   | 3   | N    | 0.1326 | 0.0991 | 0.1008   | 0.1211 | 0.1024 | 0.0911 |
| Rotation   | 3   | B    | 0.1229 | 0.1032 | 0.1068   | 0.1228 | 0.1061 | 0.0961 |
| Rotation   | 8   | N    | 0.1326 | 0.1149 | 0.1148   | 0.1271 | 0.1164 | 0.1078 |
| Rotation   | 8   | B    | 0.1282 | 0.1174 | 0.1180   | 0.1271 | 0.1178 | 0.1105 |
| Shape      | 3   | N    | 0.1279 | 0.0960 | 0.0984   | 0.1158 | 0.0999 | 0.0889 |
| Shape      | 3   | B    | 0.1177 | 0.1010 | 0.1052   | 0.1181 | 0.1038 | 0.0948 |
| Shape      | 8   | N    | 0.1279 | 0.1115 | 0.1114   | 0.1225 | 0.1129 | 0.1054 |
| Shape      | 8   | B    | 0.1235 | 0.1141 | 0.1149   | 0.1226 | 0.1143 | 0.1078 |
| Mixed      | 3   | N    | 0.1446 | 0.1088 | 0.1080   | 0.1286 | 0.1105 | 0.1025 |
| Mixed      | 3   | B    | 0.1302 | 0.1095 | 0.1116   | 0.1271 | 0.1116 | 0.1037 |
| Mixed      | 8   | N    | 0.1446 | 0.1243 | 0.1236   | 0.1369 | 0.1249 | 0.1184 |
| Mixed      | 8   | B    | 0.1379 | 0.1241 | 0.1242   | 0.1344 | 0.1244 | 0.1185 |

## B Reproduction and Verification Details

Run the archived simulation and analysis scripts, populate the manuscript with `build_manuscript.py`, and compile the resulting LaTeX source. The manuscript tables and numerical text are populated from the stored outputs. The original manifest identifies all 126 conditions, simulation constants, seeds and selected settings. The original per-user archive contains 756,000 rows, with 756 policy-condition means and 378 paired comparisons. The gain-sensitivity archive retains five seed-block means for each gain setting and core condition.

The separate expanded-tuning archive adds 13,320 candidate/condition/seed-block means, 648,000 user-policy outcomes, 648 summaries and 432 exploratory paired contrasts. It preserves both horizon-specific selections and the original-configuration comparators. Its manifest lists the two horizons, 36 evaluation conditions and distinct tuning/test seed splits. Independent validation recomputes family minima, summary and contrast means, checks exact gate equivalence and replays one simulation block at each horizon.

Tests check analytical score derivatives by finite differences, the covariance update against direct matrix inversion, covariance positivity, deterministic replay, policy-order invariance, disjoint tuning/test streams, identical trajectories for zero binary-user step and no learning, report counts and finite outputs in stress cases. The figures and manuscript are built from the same outputs. These checks establish computational consistency of the archived implementation.

## References

- [1] Z. Zalevsky, G. Elani, E. Azoulay, D. Ilani, Y. Beiderman and M. Belkin. Electromechanical tactile stimulation system for sensory vision substitution. *Optical Engineering* 52(2), 023202 (2013). [doi:10.1117/1.OE.52.2.023202](https://doi.org/10.1117/1.OE.52.2.023202).
- [2] Y. Beiderman, M. Belkin, Y. Rotenstreich and Z. Zalevsky. Experimental quantification of the tactile spatial responsivity of human cornea. *Journal of Medical Imaging* 2(1), 016002 (2015). [doi:10.1117/1.JMI.2.1.016002](https://doi.org/10.1117/1.JMI.2.1.016002).
- [3] J. Granley, T. Fauvel, M. Chalk and M. Beyeler. Human-in-the-Loop Optimization for Deep Stimulus Encoding in Visual Prostheses. *Advances in Neural Information Processing Systems* 36, 79376–79398 (2023). <https://arxiv.org/abs/2306.13104>.
- [4] E. Schoinas, A. Rastogi, A. Carter, J. Granley and M. Beyeler. Evaluating Deep Human-in-the-Loop Optimization for Retinal Implants Using Sighted Participants. *IEEE EMBC* (2025). [doi:10.1109/EMBC58623.2025.11253762](https://doi.org/10.1109/EMBC58623.2025.11253762). Open manuscript: <https://arxiv.org/abs/2502.00177>.
- [5] J. S. Merel, R. Fox, T. Jebara and L. Paninski. A multi-agent control framework for co-adaptation in brain-computer interfaces. *Advances in Neural Information Processing Systems* 26 (2013). [Proceedings paper](#).
- [6] J. Merel, D. M. Pianto, J. P. Cunningham and L. Paninski. Encoder-Decoder Optimization for Brain-Computer Interfaces. *PLOS Computational Biology* 11(6), e1004288 (2015). [doi:10.1371/journal.pcbi.1004288](https://doi.org/10.1371/journal.pcbi.1004288).
- [7] X. Liu, Y. Wang and Y. Zhao. A Unified Recursive Identification Algorithm with Quantized Observations Based on Weighted Least-Squares Type Criteria. *IEEE Transactions on Automatic Control*, accepted manuscript (2025). [doi:10.1109/TAC.2025.3620623](https://doi.org/10.1109/TAC.2025.3620623). <https://arxiv.org/abs/2502.20085>.
- [8] T. H. McCormick, A. E. Raftery, D. Madigan and R. S. Burd. Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. *Biometrics* 68, 23–30 (2012). [doi:10.1111/j.1541-0420.2011.01645.x](https://doi.org/10.1111/j.1541-0420.2011.01645.x).
- [9] J. de Villemarest and O. Wintenberger. Logarithmic Regret for parameter-free Online Logistic Regression. arXiv:1902.09803 (2019). <https://arxiv.org/abs/1902.09803>.