FROM NOISE BELIEFS TO LOGICAL DECISIONS: SURFACE-CODE DECODING WITH TEMPORAL SIDE INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Noise in quantum processors has memory: defects drift, error rates switch between regimes, and bursts persist over many syndrome rounds. A decoder that tracks this hidden process should make better corrections, but in continuous operation it must commit each correction before the full record is available. We ask when temporal memory actually changes a logical correction under this constraint. Our decoder couples a causal two-state hidden Markov filter over auxiliary observations to rolling minimum-weight matching: the filter's belief sets the edge weights, and a carried residual-syndrome frontier keeps every committed correction consistent with the detector record. Because the filter, the physical hypotheses, and the commit rule are separate modules, each can be compared against matched controls. Across 45 noise conditions and 92,160 trajectories, filtering reduces state-estimation error by up to 93% and reduces exactly to the current-observation rule under IID noise. In a held-out confirmation with 4,096 rolling streams per code distance, the filter lowers logical failures from 507 to 433 at $d = 3$ and from 158 to 131 at $d = 5$ relative to current-only calibration of the same decoder (Holm-adjusted $p = .0002$ and $.0049$). Three independently trained GRUs match these error counts but spend 13–15$\times$ longer per belief update. On shared terminal records the belief reduces failures by 6–9% for each of four solver families, and fixed-budget candidate interventions show that better physical hypotheses can lower the error rate even while support coverage shrinks. A replay under frozen calibrations from two cloud quantum processors (6,144 returned shots) preserves the same ordering.

1 INTRODUCTION

Fault-tolerant quantum memories run continuously: stabilizers are measured round after round, and a decoder turns each syndrome record into a correction (Gottesman, 1997; Calderbank & Shor, 1996; Steane, 1996; Dennis et al., 2002; Fowler et al., 2012). How well the decoder performs depends on its noise model, which is usually held fixed. Hardware noise, however, has memory. Two-level-system defects drift in and out of resonance and make qubit relaxation rates fluctuate (Klimov et al., 2018), cosmic-ray impacts cause correlated error bursts (McEwen et al., 2022), and calibrations drift between recalibration cycles. Now that surface-code memories operate below threshold (Google Quantum AI and Collaborators, 2025), a decoder faces a slowly changing hidden process rather than a fixed channel, and the simplest faithful model of such a process is a hidden Markov chain over noise regimes.

Tracking the environment is a natural remedy. Error rates can be estimated from syndrome statistics (Huo & Li, 2017; Spitz et al., 2018; Sivak et al., 2024), decoding graphs can be reweighted under drift (Wang et al., 2023), and hidden Markov models can flag persistent conditions such as leakage (Varbanov et al., 2020). What remains unclear is *when a better estimate of the hidden noise state actually changes a logical correction*. A matching decoder sees the estimate only through its edge weights, and only records near a decision boundary are affected, so a large gain in state estimation can leave nearly every decision unchanged. Continuous operation adds a second constraint: a streaming decoder must commit corrections for old rounds while new rounds keep arriving (Skoric et al., 2023; Tan et al., 2023), so any benefit of memory has to survive an irreversible commit.

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.

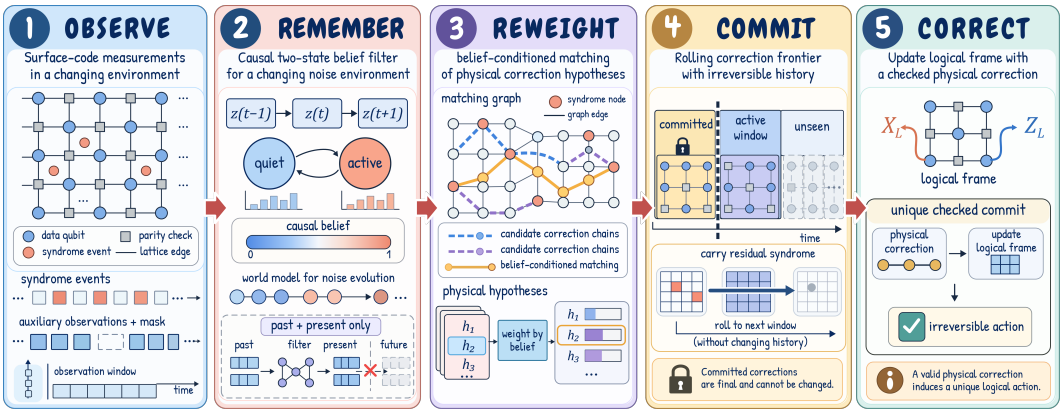

Figure 1: **Overview of belief-guided rolling decoding.** (1) Syndrome events arrive together with auxiliary observations and validity masks. (2) A causal two-state filter turns past and present observations into a belief about the hidden noise regime. (3) The belief sets the weights of the matching graph and hence the physical correction hypotheses. (4) A rolling window commits the correction edges of its oldest rounds and carries the residual syndrome forward; committed history is never rewritten. (5) Each checked commit updates the logical frame. Bottom: the three matched controls used throughout the paper, namely identical observations for every estimator, an identical frontier budget for every decoder, and a replay under frozen device calibrations.

We study this question with *belief-guided rolling decoding* (Figure 1). A causal two-state filter turns auxiliary observations into a belief about the hidden noise regime, the belief sets posterior-mean fault rates on the matching graph, and a rolling decoder commits the correction edges of its oldest rounds while carrying their boundary effect forward as a residual-syndrome frontier. Each stage has a fixed interface, so we can replace the belief estimator (static, current-only, EMA, GRU, Markov), the solver (MWPM, BP+OSD, BP+LSD, BP+UF), or the candidate set while everything else stays the same. The current-only estimator is the key control: it uses the same observations as the filter but no temporal law, so the gap between the two measures the value of memory itself.

This design gives a consistent picture. Filtering improves state estimates across a broad map of noise conditions, and the gain reaches the logical layer when the environment is persistent and the observations are noisy or missing. There, on held-out streams, current-only calibration of the same MWPM decoder makes 74 more errors at $d = 3$ and 27 more at $d = 5$ than the Markov filter, and three GRU filters match Markov only at more than ten times the update cost. Under matched IID noise the two decoders make identical decisions, so their shared gain over a static decoder comes from the observations, not from memory. The belief also lowers failures for four solver families, and the physical hypotheses matter more than the selector that chooses among them. Our contributions are summarized as follows:

- **A rolling decoder with a causal belief and a checked frontier.** An exact two-state filter drives rolling matching through a carried residual syndrome. We prove that every commit preserves syndrome consistency and never repeats an edge, and all 2,107,136 executed commits pass these checks.
- **A decision-level account of temporal information** that links memory decay, the decision boundary, correlations lost under posterior-mean weights, candidate support, and disagreement-only gradients to measurable quantities, each tested by a matched intervention.
- **Controlled evidence at several levels:** a 45-condition state map, a held-out $d = 3/5$ confirmation with recurrent controls, four solver families, fixed-budget candidate interventions, a retrained circuit-level reference, and a frozen-calibration device replay.

2 Related Work

Time-varying noise and adaptive decoding. Practical decoders estimate their noise model from data. Syndrome statistics can track time-dependent error rates (Huo & Li, 2017; Spitz et al., 2018),

decoder priors can be optimized directly on experimental records (Sivak et al., 2024), and decoding graphs can be reweighted under drift and correlated noise (Wang et al., 2023). Temporally correlated and heavy-tailed noise can change logical error rates substantially (Clader et al., 2021), and hidden Markov models can detect persistent hidden conditions such as leakage (Varbanov et al., 2020). We use a standard Bayesian filter (Rabiner, 1989); our focus is not a new estimator but the path from its belief to an irreversible correction, with controls that separate access to observations from use of their temporal law.

Learned and ensemble decoders. Neural decoders learn directly from syndrome data (Torlai & Melko, 2017; Baireuther et al., 2018), and AlphaQubit and its successor learn from rich hardware records at scale (Bausch et al., 2024; Senior et al., 2025). Ensembles select among complementary decoders (Sheth et al., 2020), harmonization and matching synthesis combine several matching solutions (Shutty et al., 2026; Jones, 2024), and neural MWPM learns weights inside a matching pipeline (Peled et al., 2026). Our fixed-K experiments ask whether a gain comes from the physical hypotheses or from the selector that chooses among them, and our learned temporal baseline is a GRU (Cho et al., 2014) with the same inputs as the filter.

Windowed and streaming decoding. Sliding and parallel window decoders bound latency by decoding overlapping blocks and committing only their oldest part (Dennis et al., 2002; Skoric et al., 2023; Tan et al., 2023), and real-time operation adds further hardware constraints (Battistel et al., 2023). Our sequential adapter follows the same commit principle but carries the full boundary effect of each committed chain as an explicit frontier, which lets us state and test a consistency invariant. Stim, PyMatching, and Sparse Blossom provide the circuit and matching tools (Gidney, 2021; Higgott, 2022; Higgott & Gidney, 2025), and the `ldpc` package provides BP+OSD, BP+LSD, and BP-guided union-find (Roffe et al., 2020; Hillmann et al., 2025; Delfosse et al., 2022). We use surface-code graphs so that changes in temporal modeling are not confounded with changes of code family.

3 Belief-guided rolling decoding

3.1 A hidden Markov model of noise with memory

A record contains detector events $D_{0:T}$, auxiliary observations $Y_{0:T}$ with validity masks $M_{0:T}$, and a terminal logical label L. A hidden environment $Z_t \in \{0, 1\}$ (quiet or active) sets the fault probabilities in round t and follows a symmetric Markov chain with switch probability

$$a = \tfrac{1}{2}(1 - e^{-2\lambda\Delta}), \qquad Y_t \mid Z_t \sim \mathcal{N}(2Z_t - 1, \sigma^2). \tag{1}$$

This is the discrete-time view of a random telegraph process with rate λ sampled once per syndrome round. By default $\lambda = .12$, $\Delta = 1$ round, and $\sigma = .9$, so $a \simeq .1067$. Let $\mathcal{F}_t^Y = \sigma(Y_{0:t}, M_{0:t})$ and $b_t = \mathbb{P}(Z_t = 1 \mid \mathcal{F}_t^Y)$. Hidden states are used only for evaluation. Starting from $\bar{b}_0 = 1/2$, the filter predicts and updates as

$$\bar{b}_t = a + (1 - 2a)b_{t-1}, \qquad b_t = \begin{cases} \mathrm{sigmoid}(\mathrm{logit}\,\bar{b}_t + 2Y_t/\sigma^2), & M_t = 1, \\ \bar{b}_t, & M_t = 0, \end{cases} \tag{2}$$

so a missing observation simply skips the likelihood update (Figure 2a). We compare the filter with three estimators that receive exactly the same observations and calibration: a *current-only* rule that resets the prior to $\bar{b}_t = 1/2$ in every round, an EMA that mixes the previous estimate with the current-observation posterior, and a *static* rule with $b_t = 1/2$. A causal GRU trained on the same inputs serves as the learned alternative.

3.2 Noise models

The *continuous study*, which supplies all rolling results, uses a phenomenological rotated surface code whose parity-check and logical operators are extracted by propagating single-qubit faults through ideal Stim circuits. Given Z_t, data and measurement faults are independent with

$$p_{\mathrm{data}}(z) = 2(.002 + .010z), \qquad p_{\mathrm{meas}}(z) = 2(.020 - .017z). \tag{3}$$

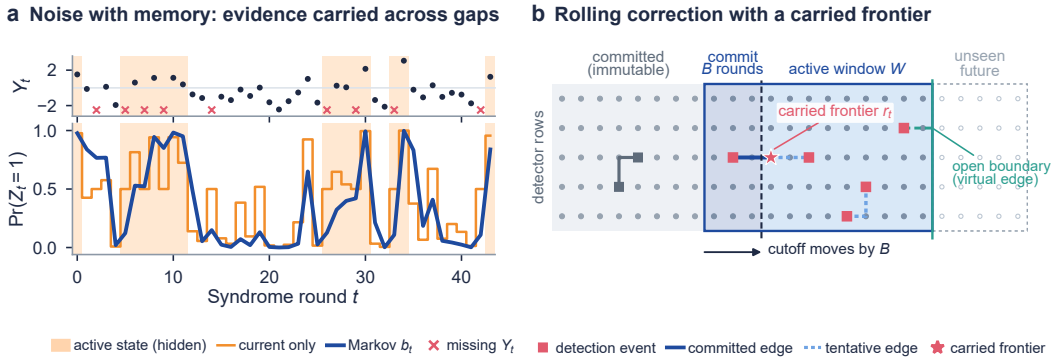

Figure 2: **Belief and frontier.** (a) A retained hidden-process trajectory with 30% missing observations. The Markov belief b_t carries evidence through gaps, whereas the current-only posterior returns to $1/2$ whenever Y_t is missing. (b) The implemented commit rule: real edges in the oldest B rounds are committed, an edge that crosses the cutoff leaves its endpoint in the carried frontier, and virtual edges to the open future boundary are never committed (schematic).

The factor two is a fixed stress setting. The active state raises data errors but lowers measurement errors, so no single rescaling of the weights is right in both regimes. Each stream ends with one perfect data readout. The *circuit-level reference* is a rotated surface-code memory circuit with gate/reset rate $.002 + .010z$ and measurement-flip rate $.020 - .017z$; its corrected version (V2) fixes a terminal-transition convention of the original study and retrains every affected model (Appendix I).

3.3 Rolling corrections with a carried frontier

Let H and O be the binary detector-incidence and logical-action matrices of the fault graph. After committing corrections c^{com}, the decoder holds the visible residual syndrome

$$r_t = s_{\le t} \oplus (Hc^{\mathrm{com}})_{\le t}. \tag{4}$$

The online state stores at most W detector rounds, their beliefs, the residual frontier, and the committed logical frame. At each step, MWPM explains the visible residual with posterior-mean fault rates $p_e(b_t) = (1 - b_t)\, p_e(0) + b_t\, p_e(1)$, and an open temporal boundary lets chains leave through the unobserved future. When the window is full, the decoder commits the real edges incident on its oldest B rounds, XORs their full boundary into the residual, and releases those rounds once their residual is zero (Figure 2b). An edge that crosses the cutoff moves its unmatched endpoint into the retained frontier instead of being clipped, virtual boundary edges are never committed, and global edge identifiers prevent duplicate commits. The frame is updated by $O\Delta c$ rather than by combining predictions from overlapping windows, and the terminal readout closes the boundary and flushes the remaining constraints. With the default $W = 4d$ and $B = d$, a round waits $W - B$ to $W - 1$ rounds before it is committed.

Invariant. Each commit preserves Equation (4): released rows are satisfied, retained rows contain the full boundary of every committed edge, and later updates cannot change committed actions. By induction, the final flush satisfies $Hc^{\mathrm{com}} = D$. This is a correctness guarantee for the short-range graphs used here, not a claim of optimal decoding; Appendix B gives the proof and the tests.

3.4 Terminal candidates and learned selection

To study physical hypotheses without window effects, terminal experiments decode complete records with a committee of $K = 7$ matching corrections. The *fixed* committee uses seven environmental templates. The *joint-posterior* committee uses a mean profile, filtered soft and hard profiles, and four trajectories drawn by backward sampling from $P(Z_{0:T} \mid Y_{0:T}, M_{0:T})$ (Carter & Kohn, 1994); a *matched-marginal* control replaces these four trajectories by independent draws from their exact smoothed marginals. Each candidate is a valid matching correction c_k with $Hc_k = D$

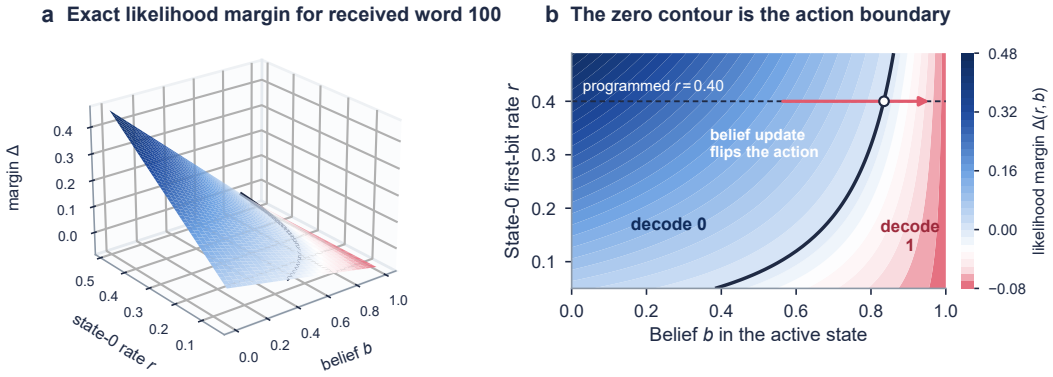

Figure 3: **A belief matters only when it crosses the action boundary.** Exact likelihood margin $\Delta(r, b)$ between logical 0 and 1 for received word 100 of the three-qubit repetition probe, as a function of the belief b and the state-0 first-bit flip rate r (Appendix H.1). (a) Surface view. (b) The zero contour separates the two decisions; at the programmed rate $r = .40$, a belief update that crosses $b \approx .84$ flips the correction. This is an analytic function, not an additional measurement.

and logical class $\ell_k = Oc_k$, and a selector outputs

$$\pi_\theta(l \mid x) = \sum_{k=1}^{K} \mathrm{softmax}(g_\theta(x))_k \mathbf{1}[\ell_k = l]. \tag{5}$$

A 2×2 design crosses fixed versus posterior candidates with and without explicit belief features, using the same two-hidden-layer MLP and the same raw observations in every cell.

4 When temporal information changes a decision

Four elementary facts explain where the gains in Section 5 can come from (proofs in Appendix N).

Memory decays at a known rate. During g consecutive missing observations with ordinary transitions,

$$b_t - \tfrac{1}{2} = (1-2a)^g (b_{t-g} - \tfrac{1}{2}), \qquad \tau_{\mathrm{env}} = -1/\log(1-2a). \tag{6}$$

Old evidence fades geometrically. At $a = .01$ the memory time is $\tau_{\mathrm{env}} \simeq 49.5$ rounds, at the default a it is 4.17 rounds, and in the IID case $a = .5$ one prediction step resets the prior, so the filter coincides exactly with the current-observation estimator. Gains are therefore expected under slow switching and long gaps, and none under IID noise.

Better state estimates are guaranteed; better decisions are not. For $X_t = \mathbf{1}[Z_t = 1]$ and correctly specified conditional means $b_t = \mathbb{E}[X_t \mid \mathcal{F}_t^Y]$ and $m_t = \mathbb{E}[X_t \mid Y_t, M_t]$,

$$\mathbb{E}[(X_t - m_t)^2] - \mathbb{E}[(X_t - b_t)^2] = \mathbb{E}[(b_t - m_t)^2] \geq 0. \tag{7}$$

This projection identity orders the state Brier score, not the logical error rate. A decision changes only when the belief crosses an action boundary. Figure 3 makes this concrete for the three-qubit repetition probe used in our device replay: the likelihood margin between the two logical classes is linear in b, and its zero contour separates the two decisions. An improvement in b that stays on one side of the contour lowers the Brier score but leaves the correction unchanged. Part of the remaining gap is physical. For distinct faults that are conditionally independent given Z_t, with $\delta_i = p_i(1) - p_i(0)$,

$$\mathrm{Cov}(E_i, E_j \mid \mathcal{F}_t^Y) = \delta_i \delta_j b_t (1 - b_t), \tag{8}$$

and posterior-mean edge weights discard this shared-environment covariance. Joint environmental trajectories target this approximation directly, which motivates the matched-marginal candidate control.

Proposition 1 (Support and selection). *For $S(x) = \{\ell_k : 1 \le k \le K\}$ and $\widehat{L}(x) \in S(x)$,*

$$\mathbf{1}[\widehat{L} \neq L] = \mathbf{1}[L \notin S(x)] + \mathbf{1}[L \in S(x), \widehat{L} \neq L]. \tag{9}$$

Consequently $R = r_{\text{shared}} + p_{\text{dis}} r_{\text{sel}}$ in the binary task, with $r_{\text{shared}} = \mathbb{P}(L \notin S)$, $p_{\text{dis}} = \mathbb{P}(|S| > 1)$, and $r_{\text{sel}} = \mathbb{P}(\widehat{L} \neq L \,|\, |S| > 1)$. Enlarging the support can only reduce the first term; it need not reduce the total risk.

Proposition 2 (Disagreement-dependent supervision). *For $q = \sum_k w_k \ell_k$ with softmax weights and logits g_j,*

$$\frac{\partial q}{\partial g_j} = w_j(\ell_j - q). \tag{10}$$

If all candidates agree, every differentiable loss of a smoothed q has zero direct gradient with respect to the candidate weights.

For two selectors restricted to the same candidates,

$$|\mathcal{R}(A) - \mathcal{R}(B)| \le \mathbb{P}(A(x) \neq B(x)) \le \mathbb{P}(|S(x)| > 1). \tag{11}$$

Disagreement records are thus the only place where such selectors can differ and the only place where a selector receives a direct training signal. A less obvious consequence, tested below, is that better physical hypotheses can reduce observed errors even while they reduce diversity and enlarge the shared-error term.

5 Experiments

5.1 Design and statistical units

We evaluate state inference, terminal correction, and rolling correction separately, each with its own unit of analysis. The *state map* covers 45 conditions (switch probability $a \in \{.01, .05, .107, .25, .5\}$, observation noise $\sigma \in \{.4, .9, 1.5\}$, independent missingness in $\{0, .3, .6\}$), each with 2,048 independent length-256 trajectories. The *focused confirmation* tests rolling decisions at $d = 3$ and $d = 5$ with $H = 65$, $W = 12$, $B = 3$, slow switching ($a = .01$), noisy observations ($\sigma = 1.5$), and 30% missingness; we chose this setting on 1,536 separate development streams before generating any confirmation data (Appendix D), and each distance has 4,096 held-out streams in 2,048 independent environment profiles. The *long-horizon confirmation* fixes $d = 5$, $H = 257$, $W = 20$, $B = 5$ and four conditions, each with 1,024 streams in 256 profiles. A *terminal study* at $d = 5$, $H = 65$ uses 512 test records in 128 profiles, shared by four official solver families and the candidate interventions, with hyperparameters chosen on separate development records (Appendix A).

All logical comparisons are paired: methods decode the same records, and inference resamples environment profiles, the independent units, rather than streams or commits. The focused study uses 10,000 profile-bootstrap and sign-randomization draws with Holm correction (Holm, 1979) over its two primary contrasts; the other studies use 20,000 stratified draws with Holm families of 12 (long-horizon), four (solvers), five (candidates), and eight (circuit) comparisons. Table 1 collects the main logical results.

5.2 Where memory improves state estimates

Figure 4 maps where history helps estimate the environment. The gain grows with persistence and with observational ambiguity: at $a = .01$ the Markov filter removes 66–93% of the current-only Brier error, and at the default $a = .107$ it removes 15–51%. Missing observations hurt both estimators, but the filter degrades much less because it carries evidence across gaps (Figure 4b). At $a = .5$ the gain is exactly zero, as Equation (6) predicts. A fixed-coefficient EMA lies between the two in 8 of the 9 cells at $a = .01$ but is worse than the current-only rule in all 18 cells with $a \ge .25$, because its fixed smoothing carries stale evidence across frequent switches.

5.3 Memory changes logical decisions in held-out rolling streams

At the focused setting, the Markov filter removes 74 of the 507 current-only failures at $d = 3$ (1.81 percentage points, 95% CI $[1.07, 2.54]$, Holm $p = .0002$) and 27 of 158 at $d = 5$ (0.66 points, CI

Table 1: **Logical failures across studies.** Within each column, all rows decode identical records with the same solver and differ only in the belief estimator. Long-horizon conditions: default persistence, 30% missingness, slow switching with 16-round gaps, and matched IID noise ($H = 257$); terminal solvers: MWPM, BP+OSD, BP+LSD, BP+UF ($H = 65$). Red marks the lowest count per column; Holm p compares Markov with current-only beliefs within each study (bold: $p < .05$). EMA is validation-tuned in the focused study.

	Focused rolling		Long-horizon rolling, $d = 5$				Terminal solvers, $d = 5$			
Estimator	$d = 3$	$d = 5$	Persist.	Miss.	Gap	IID	MWPM	OSD	LSD	UF
	of 4,096			of 1,024				of 512		
Static	543	181	475	**448**	407	495	119	124	126	126
Current only	507	158	443	450	393	**435**	104	107	108	107
EMA	488	145	465	450	389	464	113	–	–	–
GRU ×3	434–435	**129**–130	–	–	–	–	–	–	–	–
Markov (ours)	**433**	131	**441**	460	**382**	**435**	**97**	**101**	**98**	**99**
vs. current	−14.6%	−17.1%	−0.5%	+2.2%	−2.8%	0.0%	−6.7%	−5.6%	−9.3%	−7.5%
Holm p	**.0002**	**.0049**	1	1	1	1	.51	.51	.30	.51
True state	–	–	–	–	–	–	96	–	–	–

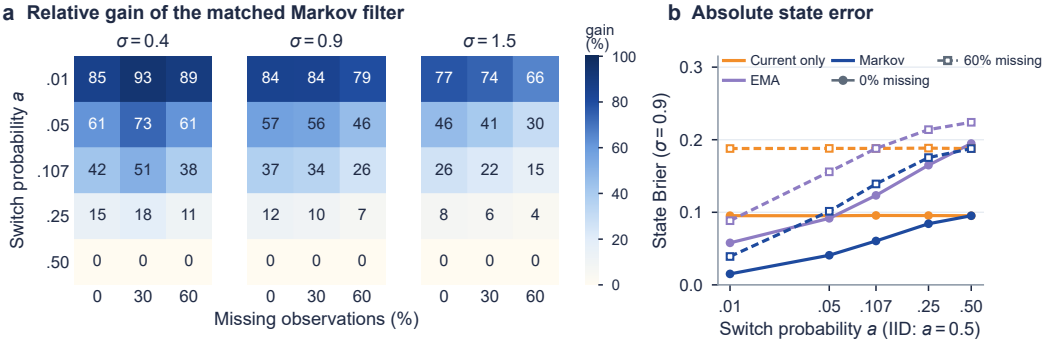

Figure 4: **Where memory improves state estimates.** (a) Relative reduction in state Brier score of the matched Markov filter over the current-only rule for all 45 conditions (2,048 independent 256-round trajectories each); the zero row at $a = .5$ is an exact identity. (b) Absolute Brier scores at $\sigma = 0.9$ with complete (solid) or 60%-missing (dashed) observations; the static rule scores 0.25 throughout.

$[0.22, 1.12]$, Holm $p = .0049$; Table 1). Logical failures follow the same order as state Brier scores, from static through current-only and EMA to Markov and the GRUs (Figure 5a), so here better beliefs translate directly into better corrections. The gain is the net of many changed decisions: at $d = 3$, switching to Markov repairs 163 current-only errors and introduces 89 (Figure 5c). The three GRUs make 434/435/434 errors at $d = 3$ and 130/129/129 at $d = 5$, all within two errors of Markov, but each belief update costs 145–172 μs against 11.2 μs for the two-state recursion (Table 3). Without a prespecified equivalence margin, we read this as parity in observed errors, not formal equivalence.

Where the memory gain vanishes. The long-horizon confirmation marks the limits of this effect (Table 1). At $H = 257$, terminal failure rates of 37–48% compress local improvements, and none of the four Markov-versus-current-only differences is significant after Holm correction despite large gains in state estimation. Matched IID noise gives the clearest contrast. Markov and current-only decoding make identical decisions (435 errors), and both beat the static rule (495 errors; 5.86 points, Holm $p = .0228$), so this gain comes from the observations, not from memory.

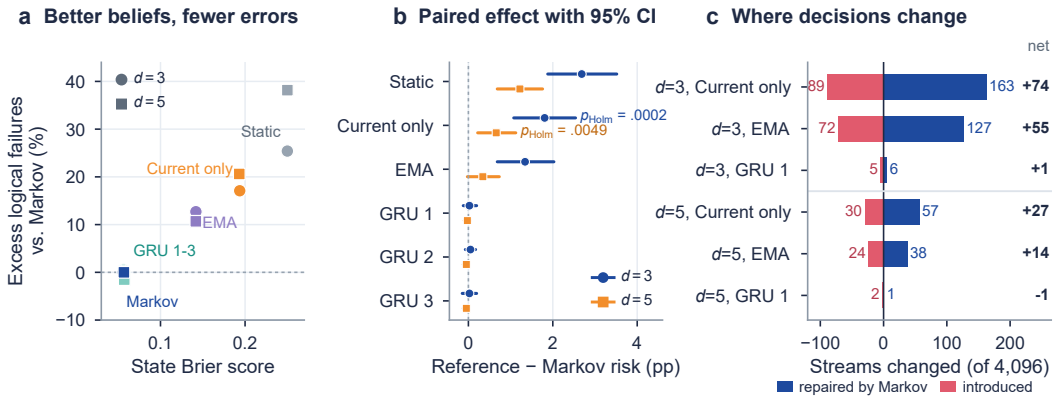

Figure 5: **Held-out $d = 3/5$ confirmation.** All seven estimators feed the same MWPM decoder on 4,096 streams per distance. (a) Excess logical failures relative to Markov follow the state Brier score. (b) Paired risk differences with 95% profile-bootstrap intervals; the two primary current-only contrasts are Holm-adjusted. (c) Streams on which switching to Markov repairs (blue) or introduces (red) an error, with the net change on the right.

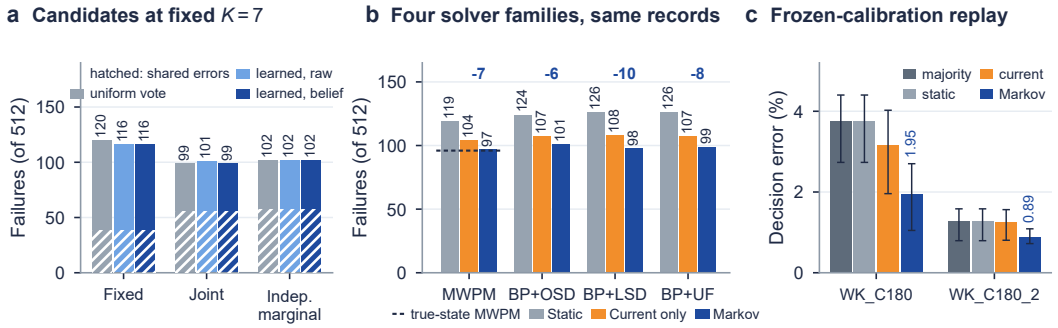

Figure 6: **Grouped comparisons on shared records.** (a) Failures of all nine candidate–selector pairs on 512 terminal records; hatching marks shared errors, which no selector can fix. (b) Static, current-only, and Markov-conditioned versions of four solver families on the same records; numbers above give Markov minus current-only failures, and the dashed line marks MWPM given the true hidden states. (c) Expected decision error on two Origin Quantum backends under one frozen pilot calibration; bars are means over three fresh acquisition repeats and whiskers their range.

5.4 Physical hypotheses matter more than the selector

The fixed-K experiment changes the physical hypotheses while keeping seven candidates and the same selector architecture (Figure 6a). With explicit beliefs, joint-posterior candidates give 99 failures versus 116 for fixed templates, a 14.7% relative reduction (3.32 points; unadjusted 95% CI $[-0.20, 6.84]$, Holm $p = .457$). The same drop appears with uniform voting (120 to 99) and with the raw-observation selector (116 to 101), whereas adding explicit beliefs to the fixed committee changes nothing (116 in both cases): the candidates, not the aggregator, drive the change. The mechanism follows Proposition 1: joint hypotheses have *more* shared errors, where every candidate is wrong (55 versus 38), but far fewer selection errors (44 versus 78), and their mean number of distinct physical chains falls from 4.73 to 2.50. A committee can thus make better decisions while its support-oracle bound gets worse. The matched-marginal control makes 102 errors, so this sample shows no separate benefit of joint path correlations (Appendix G).

Table 2: **Circuit-level reference (V2).** Failures on 32,768 reused test records; Holm p versus the belief Transformer.

Aggregator (V2)	Failures	Holm p
Belief Transformer	**1,891**	–
Local MLP, belief	**1,891**	1
GRU, belief	1,898	1
No-belief Transformer	1,902	1
Logistic stacking	1,907	1
Direct Transformer	1,930	.115
Best fixed template	1,932	.210
Uniform vote	1,945	**.033**
Markov fusion	1,947	**.030**

Table 3: **Cost at $d = 5$.** Belief update and complete per-round service in sequential CPU replay.

	Update (μs)		Service (ms)	
Estimator	mean	p99	p50	p99
Static	1.04	1.9	.67	4.59
Current only	12.89	38.4	.77	4.66
EMA	3.84	7.8	.54	3.20
GRU 1	152.95	358.6	.74	4.48
GRU 2	171.53	405.3	.99	4.92
GRU 3	144.94	356.4	.71	4.27
Markov (ours)	11.22	42.5	.59	4.33

5.5 Solver families, the circuit-level reference, cost, and a device replay

The same beliefs transfer across physical solvers (Figure 6b). Markov-conditioned MWPM, BP+OSD, BP+LSD, and BP+UF make 97, 101, 98, and 99 errors, compared with 104–108 for their current-only versions and 119–126 with static weights. The improvement of 6–10 errors appears in all four families, although with 512 records no single comparison is significant after Holm correction ($p = .30$–$.51$). For MWPM, the filter recovers 7 of the 8 errors that separate current-only weights from the true-state diagnostic.

In the circuit-level reference, retraining all 15 affected models under the corrected boundary convention preserves the original conclusion on the same 32,768 test records (Table 2): the belief Transformer makes 1,891 failures, 2.88% fewer than Markov fusion and 2.78% fewer than uniform voting (Holm $p = .030$ and .033), and the belief-fed local MLP matches it; because these test records were used before, this is a robustness check rather than a new confirmation.

Cost separates the estimators more than accuracy does (Table 3). The two-state recursion is 13–15$\times$ cheaper per update than the GRUs, while complete service time, dominated by graph construction and matching, differs much less. At $d = 5$, $W = 20$, $B = 5$, the rolling decoder needs 0.479 ms per arriving round on average (p99 2.69 ms), and its serialized frontier stays below 1.3 kB with at most 660 active edges for streams of up to 1,025 rounds (Appendix L).

Device-calibrated replay. To check that device-specific calibrations fit the same pipeline, we ran three-qubit repetition-code circuits with programmed flip rates on two Origin Quantum backends (6,144 returned shots, no failed jobs) and froze a likelihood fitted on pilot repeats before acquiring fresh repeats. With synthetic auxiliary histories replayed through these frozen likelihoods, the Markov, current-only, and static rules give mean decision errors of 1.95/3.16/3.75% on `WK_C180` and 0.89/1.26/1.27% on `WK_C180_2`, and Markov is best on all six repeats (Figure 6c). This is a classical replay, not error correction or live feedback on hardware (Appendix H).

6 Conclusion

Noise with memory can be turned into better corrections, but only through a chain of steps: an observation updates a belief, the belief changes a physical hypothesis, and a checked commit turns that hypothesis into an irreversible action. Belief-guided rolling decoding makes each step a testable module. In persistent, partially observed noise, an exact two-state Markov filter lowers logical failures at two code distances on held-out streams, matches three recurrent networks at a fraction of their update cost, and improves four solver families on shared records. Under IID noise it reduces exactly to the current-observation rule, which separates the value of the observations from the value of memory. These results come from simulations with a known two-state environment: the memory gain is confirmed at a development-selected operating point, is small under long-horizon stress, and is consistent but not individually significant in the 512-record comparisons, and the device study is a replay of frozen calibrations. Learning the noise model online and using real device telemetry are the natural next steps.

AI USE STATEMENT

In this work, generative AI tools were used only to aid and polish the writing, that is, to improve the grammar, wording, and readability of the manuscript text and the visual layout of figures produced from our own records. All scientific content, including the methods, experiments, analyses, results, and conclusions, was produced by the authors. All numerical values are computed from retained experimental records, and every quantitative figure is generated programmatically from those records. The authors reviewed all AI-assisted edits and take full responsibility for the content of this paper.

ETHICS STATEMENT

This work uses simulated quantum-error-correction data, synthetic auxiliary observations, an attributed public simulated-data subset released under CC-BY-4.0, and aggregate outputs returned by public cloud quantum services. It involves no human subjects or personal data. Credentials, task identifiers, and raw service responses are excluded from the submission. We report environmental state estimates and logical-error rates separately and use cluster-aware inference for records that share a noise trajectory.

REPRODUCIBILITY STATEMENT

Section 5 defines every study, its independent units, and its statistical procedure. Appendix A lists seeds, counts, and metrics; Appendix B gives the commit rule, its proof, and the executed correctness tests; Appendix D documents the development selection, estimator training, frozen confirmation, and timing protocol; Appendix N contains all proofs; and Appendices H and H.1 describe the device runs and the frozen-calibration replay. The supplementary material contains the complete TeX source, all figures and tables, and an anonymous implementation of the filter, the rolling decoder, and the simulator with smoke tests. Training seeds, overlapping windows, and synthetic replay histories are never counted as independent samples.

REFERENCES

Paul Baireuther, Thomas E. O'Brien, Brian Tarasinski, and Carlo W. J. Beenakker. Machine-learning-assisted correction of correlated qubit errors in a topological code. *Quantum*, 2:48, 2018. doi: 10.22331/q-2018-01-29-48.

F. Battistel, C. Chamberland, K. Johar, R. W. J. Overwater, F. Sebastiano, L. Skoric, Y. Ueno, and M. Usman. Real-time decoding for fault-tolerant quantum computing: progress, challenges and outlook. *Nano Futures*, 7(3):032003, 2023. doi: 10.1088/2399-1984/aceba6.

Johannes Bausch, Andrew W. Senior, Francisco J. H. Heras, Thomas Edlich, Alex Davies, Michael Newman, Cody Jones, Kevin Satzinger, et al. Learning high-accuracy error decoding for quantum processors. *Nature*, 635(8040):834–840, 2024. doi: 10.1038/s41586-024-08148-8.

A. Robert Calderbank and Peter W. Shor. Good Quantum Error-Correcting Codes Exist. *Physical Review A*, 54(2):1098–1105, 1996. doi: 10.1103/PhysRevA.54.1098.

C. K. Carter and R. Kohn. On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553, 1994. doi: 10.1093/biomet/81.3.541.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179.

B. D. Clader, Colin J. Trout, Jeff P. Barnes, Kevin Schultz, Gregory Quiroz, and Paraj Titum. Impact of correlations and heavy tails on quantum error correction. *Physical Review A*, 103(5):052428, 2021. doi: 10.1103/PhysRevA.103.052428.

Nicolas Delfosse, Vivien Londe, and Michael E. Beverland. Toward a Union-Find decoder for quantum LDPC codes. *IEEE Transactions on Information Theory*, 68(5):3187–3199, 2022. doi: 10.1109/TIT.2022.3143452. URL https://doi.org/10.1109/TIT.2022.3143452.

Eric Dennis, Alexei Kitaev, Andrew Landahl, and John Preskill. Topological Quantum Memory. *Journal of Mathematical Physics*, 43(9):4452–4505, 2002. doi: 10.1063/1.1499754.

Austin G. Fowler, Matteo Mariantoni, John M. Martinis, and Andrew N. Cleland. Surface codes: Towards practical large-scale quantum computation. *Physical Review A*, 86(3):032324, 2012. doi: 10.1103/PhysRevA.86.032324.

Craig Gidney. Stim: a fast stabilizer circuit simulator. *Quantum*, 5:497, 2021. doi: 10.22331/q-2021-07-06-497.

Google Quantum AI and Collaborators. Quantum error correction below the surface code threshold. *Nature*, 638(8052):920–926, 2025. doi: 10.1038/s41586-024-08449-y.

Daniel Gottesman. *Stabilizer Codes and Quantum Error Correction*. PhD thesis, California Institute of Technology, 1997. URL https://arxiv.org/abs/quant-ph/9705052. arXiv:quant-ph/9705052.

Oscar Higgott. PyMatching: A Python package for decoding quantum codes with minimum-weight perfect matching. *ACM Transactions on Quantum Computing*, 3(3):1–16, 2022. doi: 10.1145/3505637. URL https://doi.org/10.1145/3505637.

Oscar Higgott and Craig Gidney. Sparse Blossom: correcting a million errors per core second with minimum-weight matching. *Quantum*, 9:1600, 2025. doi: 10.22331/q-2025-01-20-1600.

Timo Hillmann, Lucas Berent, Armanda O. Quintavalle, Jens Eisert, Robert Wille, and Joschka Roffe. Localized statistics decoding for quantum low-density parity-check codes. *Nature Communications*, 16(1):8214, 2025. doi: 10.1038/s41467-025-63214-7.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Ming-Xia Huo and Ying Li. Learning time-dependent noise to reduce logical errors: Real time error rate estimation in quantum error correction. *New Journal of Physics*, 19(12):123032, 2017. doi: 10.1088/1367-2630/aa916e.

Cody Jones. Improved accuracy for decoding surface codes with matching synthesis. arXiv preprint arXiv:2408.12135, 2024. URL https://arxiv.org/abs/2408.12135.

P. V. Klimov, J. Kelly, Z. Chen, M. Neeley, A. Megrant, B. Burkett, R. Barends, K. Arya, B. Chiaro, Yu Chen, A. Dunsworth, A. Fowler, et al. Fluctuations of Energy-Relaxation Times in Superconducting Qubits. *Physical Review Letters*, 121(9):090502, 2018. doi: 10.1103/PhysRevLett.121.090502.

Matt McEwen, Lara Faoro, Kunal Arya, Andrew Dunsworth, Trent Huang, Seon Kim, Brian Burkett, Austin Fowler, et al. Resolving catastrophic error bursts from cosmic rays in large arrays of superconducting qubits. *Nature Physics*, 18(1):107–111, 2022. doi: 10.1038/s41567-021-01432-8.

Yotam Peled, David Zenati, and Eliya Nachmani. Neural Minimum Weight Perfect Matching for Quantum Error Codes. arXiv preprint arXiv:2601.00242, 2026. URL https://arxiv.org/abs/2601.00242v2.

Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. doi: 10.1109/5.18626.

Joschka Roffe, David R. White, Simon Burton, and Earl T. Campbell. Decoding across the quantum low-density parity-check code landscape. *Physical Review Research*, 2(4):043423, 2020. doi: 10.1103/PhysRevResearch.2.043423.

Andrew W. Senior, Thomas Edlich, Francisco J. H. Heras, Lei M. Zhang, Oscar Higgott, James S. Spencer, Taylor Applebaum, Sam Blackwell, et al. A scalable and real-time neural decoder for topological quantum codes. arXiv preprint arXiv:2512.07737, 2025. URL https://arxiv.org/abs/2512.07737v2. arXiv:2512.07737v2, revised March 11, 2026.

Milap Sheth, Sara Zafar Jafarzadeh, and Vlad Gheorghiu. Neural ensemble decoding for topological quantum error-correcting codes. *Physical Review A*, 101(3):032338, 2020. doi: 10.1103/physreva.101.032338.

Noah Shutty, Michael Newman, and Benjamin Villalonga. Efficient near-optimal decoding of the surface code through ensembling. *Physical Review Letters*, 136(7):070603, 2026. doi: 10.1103/77j6-32xx. URL https://doi.org/10.1103/77j6-32xx.

Volodymyr Sivak, Michael Newman, and Paul Klimov. Optimization of Decoder Priors for Accurate Quantum Error Correction. *Physical Review Letters*, 133(15):150603, 2024. doi: 10.1103/PhysRevLett.133.150603.

Luka Skoric, Dan E. Browne, Kenton M. Barnes, Neil I. Gillespie, and Earl T. Campbell. Parallel window decoding enables scalable fault tolerant quantum computation. *Nature Communications*, 14(1):7040, 2023. doi: 10.1038/s41467-023-42482-1. URL https://arxiv.org/abs/2209.08552.

Stephen T. Spitz, Brian Tarasinski, Carlo W. J. Beenakker, and Thomas E. O'Brien. Adaptive Weight Estimator for Quantum Error Correction in a Time-Dependent Environment. *Advanced Quantum Technologies*, 1(1):1800012, 2018. doi: 10.1002/qute.201800012.

Andrew M. Steane. Error Correcting Codes in Quantum Theory. *Physical Review Letters*, 77(5):793–797, 1996. doi: 10.1103/PhysRevLett.77.793.

Xinyu Tan, Fang Zhang, Rui Chao, Yaoyun Shi, and Jianxin Chen. Scalable Surface-Code Decoders with Parallelization in Time. *PRX Quantum*, 4(4):040344, 2023. doi: 10.1103/PRXQuantum.4.040344.

Giacomo Torlai and Roger G. Melko. Neural Decoder for Topological Codes. *Physical Review Letters*, 119(3):030501, 2017. doi: 10.1103/PhysRevLett.119.030501.

B. M. Varbanov, F. Battistel, B. M. Tarasinski, V. P. Ostroukh, T. E. O'Brien, L. DiCarlo, and B. M. Terhal. Leakage detection for a transmon-based surface code. *npj Quantum Information*, 6(1):102, 2020. doi: 10.1038/s41534-020-00330-w.

Hanrui Wang, Pengyu Liu, Yilian Liu, Jiaqi Gu, Jonathan Baker, Frederic T. Chong, and Song Han. DGR: Tackling Drifted and Correlated Noise in Quantum Error Correction via Decoding Graph Re-weighting. arXiv preprint arXiv:2311.16214, 2023. URL https://arxiv.org/abs/2311.16214.

A PROTOCOL FOR THE EXPANDED DYNAMIC-MEMORY STUDY

The expanded study has three empirical units. A hidden-process trajectory evaluates environmental inference, a terminal QEC record evaluates a correction computed from a complete detector record, and a rolling stream evaluates irreversible commits followed by one real terminal readout. These units are never added together into one accuracy sample count. The 512-record terminal test is shared by the external baselines and the candidate intervention and is counted once.

A.1 FROZEN MATRIX AND INDEPENDENCE

Table 4: **Expanded evaluation matrix.** Counts refer to distinct records, not method evaluations; rows that reuse test records say so explicitly.

Study	Physical / temporal scope	Independent sampling structure
State map	45 conditions, 256 rounds	2,048 trajectories each; 92,160 total
Long-horizon confirmation	$d = 5, H = 257$, four conditions	256 profiles $\times$ 4 shots each; 4,096 streams
Scale diagnostics	$d = 3, 7$; $H = 65, 257, 1025$; two observation conditions	32 profiles $\times$ 2 shots in each of 12 cells; 768 streams
Dynamic diagnostics	Four gap lengths, drift, semi-Markov; $d = 5, H = 65$	64 profiles $\times$ 2 shots per condition; 768 streams
Window sweep	$W = 10, 20, 40$; $B = 1, 5$	The same 64 streams for all six budgets
Terminal physical test	$d = 5, H = 65$; four decoder families	128 profiles $\times$ 4 shots; 512 records
Candidate intervention	Same terminal test; fixed $K = 7$	1,024 training, 256 validation; 512 shared test records
Boundary consistency	Original circuit records, corrected filtering	15 new training runs; original 49,152 test records reused

The four long-horizon conditions are default persistence with no missingness, default persistence with 30% independent missingness, $a = .01$ with deterministic 16-round missing blocks, and fully regenerated matched IID noise with $a = .5$. Each has equal X and Z strata. Hidden state trajectories, physical faults, auxiliary noise, and validity masks use separate child random-number streams. Shots within a profile share the environment but draw their own faults and observations, so the profile is the unit resampled for logical-risk inference.

The confirmation fixes $W = 4d$ and $B = d$ without selecting a window on the test set; the scale and window sweeps are additional diagnostics, not searches for a favorable confirmation cell. The noise scale in Equation (3) is fixed at two for all new QEC conditions. At $H = 1025$ this stress setting can approach terminal parity saturation, and we report it rather than discarding it after seeing the result.

A.2 SEEDS AND ARTIFACT IDENTITY

The study root seed is 2026092200. State-map cells consume a fixed, ordered random-number stream. Rolling tasks use 2026092300 plus their position in the stored 46-task manifest; the task order is saved, so a later experiment cannot change the existing seed mapping. Correctness tests use the separate 912000 series. Terminal development, test, training, and validation use 2026093200, 2026093300, 2026093400, and 2026093500 plus the basis index, and candidate sampling adds an offset of 10000. Selector training seeds are 2026094101–2026094103. Circuit V2 training keeps the original seeds 940301–940303 to isolate the boundary convention.

The manifest records configuration, code, and result hashes. A test-driven fix aligned the initial EMA batch update with its online update before the state atlas and terminal estimates were finalized; online confirmation filtering was unaffected, and the state atlas was regenerated with the aligned rule. No model, physical rate, confirmation condition, or statistical endpoint was chosen from confirmation outcomes.

A.3 METRICS AND FAILURE ACCOUNTING

A rolling failure is a mismatch between the final decoded logical frame and the true logical parity at the real terminal readout. We report $P_{\text{fail}}(H)$ rather than $P_{\text{fail}}(H)/H$, because parity errors can cancel and a per-round logical hazard would require a different measurement protocol. State Brier and log loss average state-estimation errors, and their observations are not added to the logical sample count. Log loss clips probabilities to $[10^{-7}, 1 - 10^{-7}]$.

Invalid physical corrections would count as failures and be reported separately; none occurred in the external baseline study. The rolling implementation raises an error immediately if candidate syndrome validity, residual deletion, or the unique-commit rule fails. All planned tasks completed, so no failed request was removed from any denominator. Trial execution times in the parallel experiment logs are not used as service benchmarks.

B PHYSICAL FRONTIER, COMMIT RULE, AND CORRECTNESS

B.1 FROM A SURFACE CODE TO A FAULT GRAPH

For each basis and distance, an ideal one-round Stim memory circuit defines the data qubits, the stabilizer detector rows, and the terminal logical observable. We inject one X fault for a Z-memory task, or one Z fault for an X-memory task, at a known pre-round location, and the detector error model gives that fault's detector incidence and logical action. Repeating this over data qubits constructs the spatial parity matrix H_s and the row vector O_s. The code geometry is thus verified by propagation rather than by a hand-coded coordinate convention.

The phenomenological simulator samples independent single-basis data faults before each syndrome round and independent measurement faults in that round, conditional on the hidden state. With data-fault vector e_t and measurement-fault vector v_t,

$$D_t = H_s e_t \oplus v_t \oplus v_{t-1}, \qquad v_{-1} = 0.$$

The perfect terminal readout adds no new data or measurement fault, so $D_H = v_{H-1}$, and the true logical label is $L = \bigoplus_{t=0}^{H-1} O_s e_t$. The terminal measurement happens once per experiment, not once per decoding window.

Two data qubits with identical detector incidence and logical action are equivalent for this binary decoding problem. Their graph edges are merged using the exact odd-parity probability

$$p_{\text{odd}} = \tfrac{1}{2}\left[1 - \prod_{j=1}^{m}(1 - 2p_j)\right],$$

so a grouped edge represents an equivalence class of physical corrections with the same (H, O) action. We check explicitly that no grouping joins equal detector incidence with different logical actions. Grouping does not assert independence after marginalizing the shared environment; posterior-rate matching remains the approximation discussed around Equation (8).

B.2 OPEN TEMPORAL BOUNDARY AND SAFE SPLITTING

The active graph contains spatial edges within each retained round and measurement edges between adjacent rounds, so its maximum temporal span is one round. At the newest non-terminal layer, an explicit boundary node with zero-cost open edges represents unconstrained future matching. These edges are boundary conditions, not real faults, and there are no such edges at the true terminal layer.

For a cutoff c, let J_c contain the selected real edges whose earliest detector endpoint lies before c. The decoder commits exactly these edges, including their full incidence on later retained endpoints, and excludes all virtual open-boundary edges. Because the committed block ends strictly before the open temporal boundary, every selected edge incident on a row to be deleted is a real edge in J_c, and after the residual update each deleted row is zero. A temporal edge across c contributes its second endpoint to the retained frontier; it is not clipped.

Proposition 3 (Residual preservation for the implemented adapter)**.** *Assume the active graph has temporal span one, the candidate satisfies every visible detector row, $1 \le B < W$, and only real edges with earliest endpoint before the cutoff are committed. Then (i) all released detector rows are satisfied, (ii) the retained residual equals the original syndrome XOR the full boundary of all committed corrections, and (iii) no physical edge is committed twice.*

Proof. For (i), every selected edge incident on an old row has an endpoint before the cutoff and is committed, while no virtual edge reaches those rows; candidate validity then makes their updated residual zero. For (ii), XOR the full incidence of each committed edge into every endpoint, including retained endpoints; this is linearity of H over GF(2). For (iii), after the left boundary advances, no newly constructed real edge has its earliest time in a previously committed block, and the implementation checks this monotonicity with stable global identifiers. Induction gives all three properties for every commit. At a closed terminal flush all rows are released, so $Hc^{\mathrm{com}} = D$, and the associated frame is Oc^{com} by linearity. □

B.3 Algorithm and retained state

Table 5: **Sequential rolling correction.** Every observation is incorporated exactly once, and candidate reconstruction does not run the filter a second time.

Step	Operation
1	Receive one detector layer and one auxiliary value with its mask; increment the round counter.
2	Predict and update the environmental belief, using the missing-value branch when necessary.
3	Append the new detector residual and belief to the bounded ring state.
4	If fewer than W layers are retained and no terminal readout has arrived, return.
5	Build posterior-rate spatial and temporal edges; add an open future boundary only for a non-terminal window.
6	Solve the visible residual with MWPM and verify its full detector incidence.
7	Select the real correction edges incident on the oldest B layers, or all real edges at the terminal flush.
8	XOR every selected edge's full boundary into the residual and its logical action into the committed frame.
9	Verify that released rows are zero and identifiers are new; write the commit audit and advance the cutoff.
10	Keep the updated frontier; after a terminal flush, close the stream and return its logical frame.

The online state contains the scalar filter, round and cutoff counters, at most W residual layers and beliefs, the frame, and fixed-size performance counters. It does not keep the complete detector history or the list of all committed edges. Audit examples stream to disk, and verification counters do not become an in-memory history. Transient matching graphs have $O(Wd^2)$ vertices and edges for this short-range geometry. Serialized counter lengths grow logarithmically with total time, so "bounded state" means a fixed number of physical layers and scalar counters rather than a constant byte string for unbounded time.

B.4 Executable correctness coverage

The circuit-level repository contributes another 36 tests. These checks cover a specified implementation and environment; they do not certify arbitrary future graph models. Long-range faults, noisy final readout, or circuit-level hyperedge decomposition would require extending the adapter and its invariants before the same streaming guarantee applies.

Table 6: **Correctness contract.** Fifteen parameterized tests cover the ten required mechanisms plus exact filtering and path-sampling checks. These tests establish correctness, not performance.

Mechanism	Executed assertion
No and single faults	Both logical bases; all nine $d = 3$ data qubits and measurement locations near seven commit/terminal positions.
Cross-window chain	A measurement chain spans several windows; committed global identifiers are unique and the final residual is empty.
Future invariance	Identical visible prefixes give identical states before divergent continuations.
Chunk invariance	Single-round and chunked execution agree under the same commit schedule.
State continuation	Restoring the JSON state at an interior round gives the same final state as uninterrupted execution.
Full-window limit	A window covering the full record matches the corresponding terminal solver, including the logical frame.
Filter correctness	Missing-value propagation, terminal identity transition, matched IID, and online/batch agreement.
Micrograph enumeration	All 2^{11} fault configurations of a two-layer $d = 3$ graph satisfy the decoded detector incidence.
State bound	Fixed d, W, B at 65, 257, and 1,025 rounds retain the same graph and layer bound.
Posterior sampling	A four-step exhaustive HMM posterior agrees with smoothing; 30,000 sampled paths match its marginals within tolerance.

C Temporal inference, gap law, and the 45-condition map

C.1 Exact gap and covariance calculations

Writing $u_t = b_t - 1/2$, the no-observation recursion is $u_t = (1 - 2a)u_{t-1}$, and iterating proves Equation (6). If $a = .5$, one prediction gives zero centered belief; if $a = 0$, the previous belief is preserved. The logarithmic time constant is used only for $0 < a < .5$.

With information held fixed at time s, the symmetric chain satisfies

$$\mathbb{E}[Z_{s+k} - \tfrac{1}{2} \mid Z_s] = (1 - 2a)^k (Z_s - \tfrac{1}{2}), \qquad \mathrm{Cov}(Z_s, Z_{s+k} \mid \mathcal{F}_s^Y) = (1 - 2a)^k b_s (1 - b_s).$$

This conditional covariance excludes later observations. Substituting a smoothed posterior that has already absorbed future data would change the conditioning problem, and this formula is not the one used to create the terminal posterior-path samples.

C.2 Complete state scores

Tables 7–9 list every frozen cell. "Difference" is current-only minus Markov Brier, with a paired standard error across independent trajectories; these are descriptive summaries, not 45 separately selected significance claims. The static Brier score is .25 throughout, and the EMA baseline uses coefficient .25 without label-dependent test adaptation. Under IID noise, the Markov/current-only identity holds numerically even when observations are missing.

Table 7: **State map,** $\sigma = .4$**.** Each row has 2,048 independent trajectories.

a	Missing (%)	Current	EMA	Markov	Difference	Paired SE
0.0100	0	0.00475	0.01419	0.00071	0.00405	0.00007
0.0100	30	0.07822	0.02297	0.00541	0.07282	0.00019
0.0100	60	0.15209	0.04292	0.01673	0.13536	0.00035
0.0500	0	0.00481	0.05066	0.00186	0.00296	0.00006
0.0500	30	0.07851	0.07456	0.02130	0.05721	0.00024
0.0500	60	0.15216	0.11771	0.05858	0.09357	0.00039
0.1067	0	0.00471	0.08492	0.00272	0.00199	0.00005
0.1067	30	0.07838	0.11361	0.03822	0.04016	0.00024
0.1067	60	0.15190	0.15883	0.09378	0.05811	0.00035
0.2500	0	0.00474	0.12985	0.00404	0.00070	0.00003
0.2500	30	0.07823	0.15638	0.06390	0.01434	0.00016
0.2500	60	0.15209	0.19401	0.13477	0.01733	0.00019
0.5000	0	0.00481	0.16230	0.00481	0.00000	0.00000
0.5000	30	0.07843	0.18188	0.07843	0.00000	0.00000
0.5000	60	0.15177	0.20877	0.15177	0.00000	0.00000

Table 8: **State map,** $\sigma = .9$**.** Each row has 2,048 independent trajectories.

a	Missing (%)	Current	EMA	Markov	Difference	Paired SE
0.0100	0	0.09532	0.05791	0.01511	0.08021	0.00030
0.0100	30	0.14199	0.06815	0.02295	0.11904	0.00035
0.0100	60	0.18786	0.08843	0.03911	0.14875	0.00053
0.0500	0	0.09524	0.09168	0.04080	0.05444	0.00027
0.0500	30	0.14217	0.11780	0.06313	0.07903	0.00036
0.0500	60	0.18803	0.15578	0.10143	0.08661	0.00049
0.1067	0	0.09564	0.12330	0.06049	0.03515	0.00023
0.1067	30	0.14180	0.15230	0.09290	0.04890	0.00030
0.1067	60	0.18800	0.18792	0.13908	0.04892	0.00036
0.2500	0	0.09550	0.16494	0.08421	0.01130	0.00013
0.2500	30	0.14176	0.18747	0.12702	0.01475	0.00017
0.2500	60	0.18840	0.21386	0.17538	0.01302	0.00017
0.5000	0	0.09526	0.19476	0.09526	0.00000	0.00000
0.5000	30	0.14164	0.20700	0.14164	0.00000	0.00000
0.5000	60	0.18778	0.22402	0.18778	0.00000	0.00000

Table 9: **State map,** $\sigma = 1.5$**.** Each row has 2,048 independent trajectories.

a	Missing (%)	Current	EMA	Markov	Difference	Paired SE
0.0100	0	0.16972	0.13181	0.03865	0.13107	0.00049
0.0100	30	0.19404	0.13890	0.05128	0.14276	0.00060
0.0100	60	0.21789	0.15318	0.07504	0.14285	0.00084
0.0500	0	0.16968	0.15358	0.09136	0.07832	0.00041
0.0500	30	0.19377	0.17166	0.11512	0.07865	0.00048
0.0500	60	0.21829	0.19734	0.15281	0.06548	0.00058
0.1067	0	0.16985	0.17457	0.12491	0.04494	0.00031
0.1067	30	0.19384	0.19408	0.15090	0.04294	0.00034
0.1067	60	0.21769	0.21604	0.18412	0.03357	0.00036
0.2500	0	0.16997	0.20162	0.15711	0.01287	0.00015
0.2500	30	0.19387	0.21593	0.18223	0.01165	0.00016
0.2500	60	0.21818	0.23072	0.20980	0.00838	0.00014
0.5000	0	0.16952	0.22124	0.16952	0.00000	0.00000
0.5000	30	0.19380	0.22764	0.19380	0.00000	0.00000
0.5000	60	0.21783	0.23652	0.21783	0.00000	0.00000

D INDEPENDENT DISTANCE AND RECURRENT-CONTROL CONFIRMATION

This study tests whether temporal information changes rolling logical decisions in a persistent, partially observed regime. The primary endpoint is terminal logical risk at each of two distances, and the primary contrast is current-only minus Markov risk. Development, estimator training, estimator validation, and confirmation use disjoint random seeds. The protocol, the selected operating point, and the estimator calibration files were hashed before any confirmation data were generated, and the retained freeze and shard checks make this separation verifiable.

D.1 OPERATING-POINT SELECTION AND INDEPENDENT UNITS

Development crosses channel scales $s \in \{.25, .5, 1\}$ with horizons $H \in \{33, 65\}$. Every point uses $d = 3$, $W = 12$, $B = 3$, switch probability .01, observation noise 1.5, and independent missingness .30. Each basis has 64 independent profiles with two fault/observation realizations per profile, giving 256 streams per point. Selection minimizes the pooled Markov-minus-current-only failure count, with ties resolved by grid order; the selected point is $s = 1$, $H = 65$. Table 10 lists all six points, so the favorable regime is fixed by development data rather than by confirmation outcomes.

Table 10: **Complete development selection.** Failures per 256 streams at $d = 3$. The final column is the selection criterion, not a held-out treatment effect.

Scale	Horizon	Static	Current only	Markov	Markov − current
.25	33	0	1	1	0
.25	65	5	5	3	−2
.50	33	5	3	1	−2
.50	65	11	9	8	−1
1.00	33	26	21	20	−1
1.00	65	24	25	20	−5

Confirmation keeps $H = 65$, $W = 12$, $B = 3$ at both distances. Each distance has 1,024 independent profiles per basis and two realizations per profile, giving 2,048 clusters and 4,096 streams. All seven estimators receive identical detector records, observation realizations, masks, channel calibration, and MWPM implementation. Each basis is generated in 16 shards of 64 profiles, with confirmation shard seeds $933100 + 1000d + 100b + j$ for basis index $b \in \{0, 1\}$ and shard index $j \in \{0, \ldots, 15\}$. Development seeds are $931100 + 10i + b$ for grid index i; training and validation use 932100 and 932200. Neither distance was used to retune the selected point or any estimator.

D.2 RECURRENT AND SMOOTHING CONTROLS

The GRU is a causal belief estimator that feeds the same MWPM decoder. It takes the observation and its mask as inputs, has a 32-dimensional recurrent state, and ends in a linear scalar output with a sigmoid, for 3,489 trainable parameters. Training uses 2,048 labeled calibration profiles of 65 noisy rounds plus the terminal observation, and validation uses 512 separate profiles. Adam minimizes the state binary cross-entropy for 30 epochs with batch size 128 and learning rate .003. Validation Brier selects the checkpoint independently for seeds 301, 302, and 303 (epochs 30, 13, and 20). Test hidden states are used only to score state estimates, never as decoder inputs, and prefix-invariance tests confirm causality for all three checkpoints.

The EMA control selects its coefficient from $\{.05, .15, .25, .5, .8\}$ on the same validation profiles. The validation Brier scores are .18692, .15246, .14265, .14065, and .15316, so the retained coefficient is .5. The GRU validation Brier scores are .05910, .05892, and .05896. All three GRUs are reported individually; we neither pick a test seed nor average their predictions. They are same-observation recurrent controls, not parameter-matched models or reproductions of published large neural decoders.

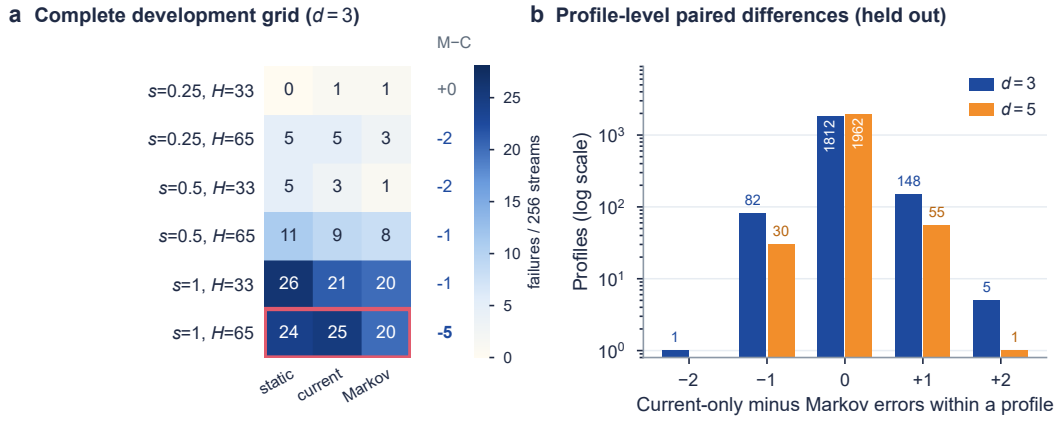

Figure 7: **Development selection and held-out paired structure.** (a) Every development cell for the three selection-stage estimators; the outlined row was selected before confirmation, and the right column gives the Markov-minus-current selection score. (b) Distribution of current-only minus Markov errors within each held-out profile (two streams per profile), on a log scale. Most profiles agree; among those that differ, profiles favoring Markov outnumber the reverse by 153 to 83 at $d = 3$ and by 56 to 30 at $d = 5$.

D.3 Paired inference and interpretation

For each reference estimator, we average its paired error difference against Markov over the two realizations within a profile. We then draw 10,000 pooled-profile bootstrap samples for percentile 95% intervals and 10,000 independent profile-level sign randomizations for two-sided tests, with Monte Carlo correction $(1 + N_{\text{extreme}})/(1 + N_{\text{draws}})$ and generator seed 934100. Holm correction is applied only to the two primary current-only comparisons, one per distance. These resampling settings differ from the 20,000-draw stratified procedure of the long-horizon study.

At $d = 3$, Markov repairs 163 current-only errors and introduces 89, a net reduction of 74 of 4,096 streams. At $d = 5$, it repairs 57 and introduces 30, a net reduction of 27. The absolute reductions are 1.8066 and .6592 percentage points, with intervals $[1.0742, 2.5391]$ and $[.2197, 1.1230]$ and Holm $p = .000200$ and $.004900$. The raw failure rates are 10.5713% versus 12.3779% at $d = 3$ and 3.1982% versus 3.8574% at $d = 5$. These results establish a memory-specific effect at the selected persistent-noise setting; the long-horizon conditions keep their own endpoints and results.

All GRU counts are within two errors of Markov on each 4,096-stream test set, and their paired intervals are shown in Figure 5b. No equivalence margin was specified, so near-equal counts are not an equivalence-test conclusion. Across the seven estimators, each distance executes 544,768 checked commits, for 1,089,536 additional commit checks; commits measure implementation coverage rather than independent observations.

D.4 Estimator cost versus complete service cost

Sequential CPU replay times four streams per basis after one warm-up stream per basis, giving 528 arriving rounds for each estimator and distance. The replay includes input conversion, belief updates, graph construction, matching, and commit checks, and excludes model loading. At $d = 5$, the mean Markov belief update takes 11.22 μs and the three GRU means are 152.95, 171.53, and 144.94 μs. This 12.9–15.3-fold ratio applies to the estimator, not to the whole decoder: Figure 8 shows that complete per-round service is dominated by graph construction and matching. The measurements describe the retained CPU implementation and do not specify a quantum-hardware feedback deadline.

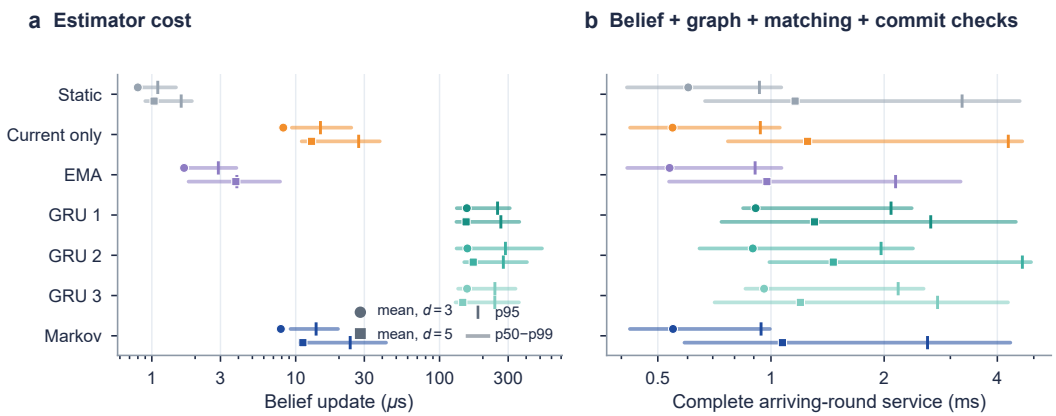

Figure 8: **Cost of the estimator and of the complete service.** (a) Belief-update time and (b) complete arriving-round service for all seven estimators on the same sequential replay. Markers show means (circles $d = 3$, squares $d = 5$), bars span p50–p99, and ticks mark p95. Both axes are logarithmic; timings add no QEC accuracy samples.

E COMPLETE ROLLING CONFIRMATION AND STATISTICAL INTERPRETATION

Table 11: **Long-horizon rolling confirmation.** Failures on the same records at $d = 5$, $H = 257$, $W = 20$, $B = 5$; each row has 1,024 streams in 256 independent profiles. Final risk uses the real terminal readout and the complete committed frame.

Condition	n	Static	Current only	EMA	Markov
Persistent	1024	475	443	465	441
Missing 30%	1024	448	450	450	460
Slow + gap 16	1024	407	393	389	382
Matched IID	1024	495	435	464	435

Table 12: **All 12 long-horizon rolling comparisons.** Positive values favor Markov. Each condition has 1,024 streams in 256 profiles. Intervals are unadjusted profile-bootstrap intervals; Holm correction spans all 12 tests.

Condition	Reference	Reduction (pp)	95% CI	Holm p
Persistent	Current only	+0.195	[-2.344, +2.734]	1.0000
Persistent	Static	+3.320	[-0.586, +7.227]	1.0000
Persistent	EMA	+2.344	[-0.684, +5.371]	1.0000
Missing 30%	Current only	-0.977	[-3.906, +1.953]	1.0000
Missing 30%	Static	-1.172	[-5.078, +2.734]	1.0000
Missing 30%	EMA	-0.977	[-4.590, +2.539]	1.0000
Slow + gap 16	Current only	+1.074	[-2.051, +4.199]	1.0000
Slow + gap 16	Static	+2.441	[-1.074, +5.859]	1.0000
Slow + gap 16	EMA	+0.684	[-2.051, +3.418]	1.0000
Matched IID	Current only	+0.000	[+0.000, +0.000]	1.0000
Matched IID	Static	+5.859	[+2.344, +9.277]	0.0228
Matched IID	EMA	+2.832	[-0.488, +6.152]	1.0000

A profile difference is the sum of reference-minus-candidate error indicators over the profile's four records. The test statistic sums these differences and divides by the total number of records. Sign randomization flips complete profile differences rather than individual correlated records, and the bootstrap resamples profiles within the X/Z strata while keeping all of their records. The sign test assumes that paired profile differences are symmetric under the null; it is not an exact distribution-free test for every heterogeneous environment model. We use 20,000 draws, a plus-one Monte Carlo correction, and the same statistic in every cell.

The memory-specific contrasts are the Markov versus current-only comparisons. Their absolute reductions are $+.195$, $-.977$, $+1.074$, and 0 percentage points in the persistent, missing, slow-gap, and IID conditions, and all Holm values equal one. The static comparison under IID has an adjusted $p = .0228$, but it is not evidence for memory, because the Markov and current-only decoders are identical there. The IID control is designed to separate the value of an observable environmental signal from the added value of its temporal law.

At these horizons, terminal failure probabilities of 37–48% compress relative effects and make long-horizon parity an insensitive measure of local improvements. We report the stress conditions in full. The focused experiment in Appendix D instead uses a development-selected finite horizon and new confirmation seeds; its terminal-risk endpoint is distinct from a first-passage or threshold analysis.

F TERMINAL EXTERNAL BASELINES AND MATCHED INPUT ACCESS

The four executed families use PyMatching 2.4.0 or `ldpc` 2.4.1 with identical detector constraints and logical operator. Their static, current-only, and Markov variants differ only in the channel probabilities passed to the solver. BP-based methods receive the graph parity matrix, independent-column probabilities, and the syndrome, and the posterior-rate approximation is common to every family.

For BP+OSD, BP+LSD, and BP+UF, development selection compares a 30-iteration minimum-sum configuration with a 60-iteration product-sum configuration. Both use a parallel schedule and one OpenMP thread. The OSD/LSD order is zero in the first configuration and four with combination-sweep search in the second, and BP+UF uses inversion-based union-find with one bit per growth step. Development uses 16 records per basis and no confirmation labels, and ties favor the lower-index setting. The selected setting is shared by the static, current-only, and Markov versions and is not retuned to favor the filter.

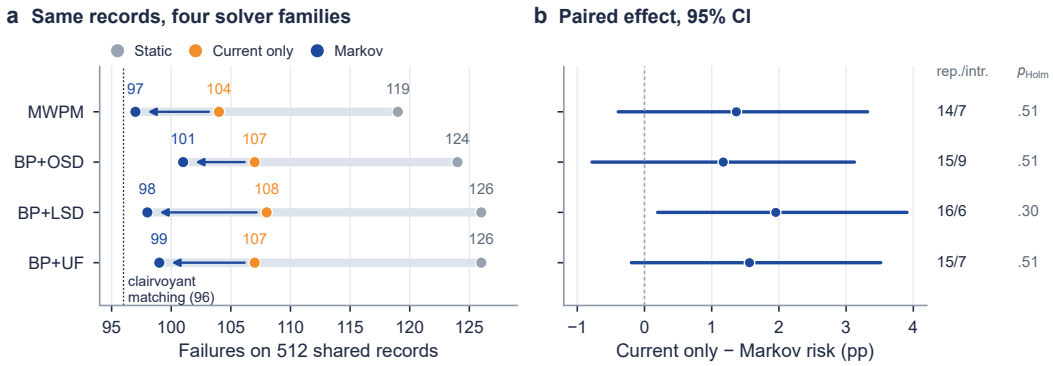

Figure 9: **External solver families on the same records.** (a) Failures of the static, current-only, and Markov-conditioned versions of each family on 512 terminal records; arrows show the change from current-only to Markov weights, and the dotted line marks MWPM given the true hidden states. (b) Paired current-only minus Markov risk with unadjusted profile-bootstrap intervals, the paired repaired/introduced counts, and Holm-adjusted p-values over the four families.

For the four memory-specific family comparisons, the paired repairs/introductions are $14/7$, $15/9$, $16/6$, and $15/7$. The confidence intervals reflect profile variability rather than repeated runs of the same decoder. These terminal solvers do not face the rolling decoder's 15–19-round steady-state waiting budget, so their accuracy should not be combined with the rolling service measurements as a single operating point.

Figure 10 makes the paired structure explicit. In the candidate-support flow, the fixed committee has $(263, 211, 38)$ all-correct, mixed, and all-wrong records, whereas the joint committee has $(336, 121, 55)$. The largest transition moves 79 records from mixed to all-correct; the same intervention also moves 20 mixed records to all-wrong. For the external families, the two-by-two panels give the complete current-only-to-Markov record flow: Markov yields net reductions of 7, 6, 10, and 8 errors for MWPM, BP+OSD, BP+LSD, and BP+UF. These flows are descriptive paired evidence

Table 13: **All external variants and diagnostic matching references.** Every method uses 512 records and returns zero invalid corrections. Clairvoyant matching receives the hidden state; it is a non-deployable diagnostic, not an exact logical decoder.

Method	Static	Current only	Markov
MWPM	119	104	97
BP+OSD	124	107	101
BP+LSD	126	108	98
BP+UF	126	107	99
MWPM, EMA observations		113	
MWPM, clairvoyant state		96	

of where decisions change, not a multiplicity-adjusted claim that the filter improves every physical decoder.

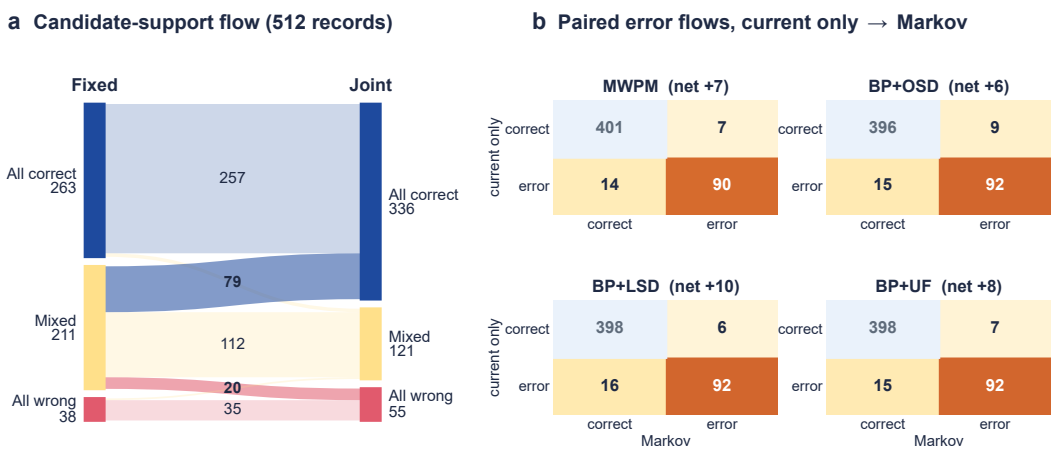

Figure 10: **Where decisions change on the same records.** (a) Exact fixed-to-joint candidate-support transitions over 512 terminal records; the two highlighted flows start from mixed support. The categories describe committee support, not logical correctness. (b) Paired current-only-to-Markov error flows for four external solver families; rows give the current-only outcome and columns the Markov outcome.

DGR, harmonization, Libra, AlphaQubit, AlphaQubit 2, and neural MWPM are discussed as related work rather than executed as baselines, because this package does not reproduce their original training and data interfaces. The comparison supported here is among the stated physical solver families and information pathways, not a state-of-the-art ranking on an official public benchmark.

G Fixed-budget posterior hypotheses and selector interventions

G.1 Joint and matched-marginal sampling

Let b_t be the forward filtering probability and $A_{ij} = P(Z_{t+1} = j \mid Z_t = i)$. We sample Z_T from b_T and then move backward using

$$P(Z_t = 1 \mid Z_{t+1} = j, Y_{0:T}, M_{0:T}) = \frac{b_t A_{1j}}{b_t A_{1j} + (1 - b_t) A_{0j}}.$$

Future observations cancel from this conditional once the next state is fixed, so the recursion samples the exact hidden-state path posterior of the specified HMM (Carter & Kohn, 1994). It does not additionally condition on the detector syndrome. A backward smoothing recursion computes $\gamma_t =$

$P(Z_t = 1 \mid Y_{0:T}, M_{0:T})$, and the independent control draws each state from $\mathrm{Bernoulli}(\gamma_t)$, which preserves the marginals but not the joint dependence.

Both dynamic committees contain the same mean, filtered-soft, and filtered-hard candidates plus four sampled candidates. The soft and hard candidates use forward filtering rather than smoothed means. Seven physical graph constructions are attempted in every condition, and duplicates are measured rather than replenished. Every candidate correction is checked for detector validity. Chain diversity counts unique full fault-edge vectors, not just binary logical labels.

G.2 Selection features and optimization

The common MLP has hidden widths 64 and 32, GELU activations, and seven output logits. Its raw input concatenates normalized observations, masks, per-round detector density, candidate logical bits, and normalized matching costs. The explicit-belief variant also receives per-round filtering beliefs and candidate compatibility features, while the raw variant holds these channels constant, so the architecture and input dimension are the same across the 2×2 intervention. Standardization is fitted on training data.

AdamW uses learning rate .001 and weight decay .0001 for 40 epochs, with eight shuffled batches per epoch. The loss is binary cross-entropy on $.001 + .998q$. Each of three seeds selects its checkpoint by validation cross-entropy, and the deployed probability is the mean over the three members rather than the best test seed. This is terminal supervision; we do not assume a correspondence to a window-by-window optimal action.

Table 14: **All candidate/selector combinations.** Shared and selection errors sum to total errors. Disagreement counts and mean numbers of unique chains refer to the same 512 records.

Candidates	Selector	Errors	Shared	Selected	Disagree	Chains
Fixed	Vote	120	38	82	211	4.73
Fixed	Raw	116	38	78	211	4.73
Fixed	Belief	116	38	78	211	4.73
Joint	Vote	99	55	44	121	2.50
Joint	Raw	101	55	46	121	2.50
Joint	Belief	99	55	44	121	2.50
Independent	Vote	102	57	45	114	2.50
Independent	Raw	102	57	45	114	2.50
Independent	Belief	102	57	45	114	2.50

Table 15: **Predeclared candidate contrasts.** Positive values favor the named candidate. Intervals are unadjusted; the five p-values are Holm-adjusted.

Reference	Candidate	Reduction (pp)	95% CI	Holm p
fixed raw	fixed belief	+0.000	[-1.172, +1.172]	1.000
joint raw	joint belief	+0.391	[+0.000, +0.977]	1.000
fixed raw	joint raw	+2.930	[-0.391, +6.445]	0.509
fixed belief	joint belief	+3.320	[-0.195, +6.836]	0.457
independent belief	joint belief	+0.586	[-0.586, +1.758]	1.000

The paired interaction asks whether adding an explicit belief has a different effect under joint and fixed candidates. Its absolute difference-in-differences is .391 percentage points, with unadjusted 95% interval $[-.781, 1.563]$ and raw $p = .755$; it is an exploratory interaction, not an additional significant result. Joint and independent-marginal learned candidates differ by three errors with adjusted $p = 1$, so we do not claim an independent advantage from trajectory correlations.

The fixed-to-joint explicit-belief comparison repairs 49 former errors and introduces 32. Expanded support does not explain this net change, because shared errors rise by 17; instead, conditional selection errors fall by 34. In the earlier circuit committee, adding two candidates changed support on only one record. Together, these interventions show why candidate count, coverage, physical diversity, and decision quality should be reported as distinct quantities.

The exact support-state transitions show where physical hypotheses change the available decisions (Figure 11a). With rows for fixed-committee states and columns for joint-posterior states, the counts are $(257, 6, 0)$, $(79, 112, 20)$, and $(0, 3, 35)$. The joint committee keeps $257/263$ all-correct records and converts 79 mixed records to all-correct, but it also turns 20 mixed records into all-wrong records while rescuing only three previously all-wrong records. Its shared-error count therefore rises by 17, from 38 to 55. An exploratory representation projection had a local support-state purity of .507 against a chance level of .498; because it showed no separation, we report the operational counts instead.

Changing the physical hypotheses changes both the quality of each candidate and the support available to a selector, whereas explicit beliefs change only how the selector chooses within that support. In this sample, the reduction in selection errors more than offsets the increase in shared errors.

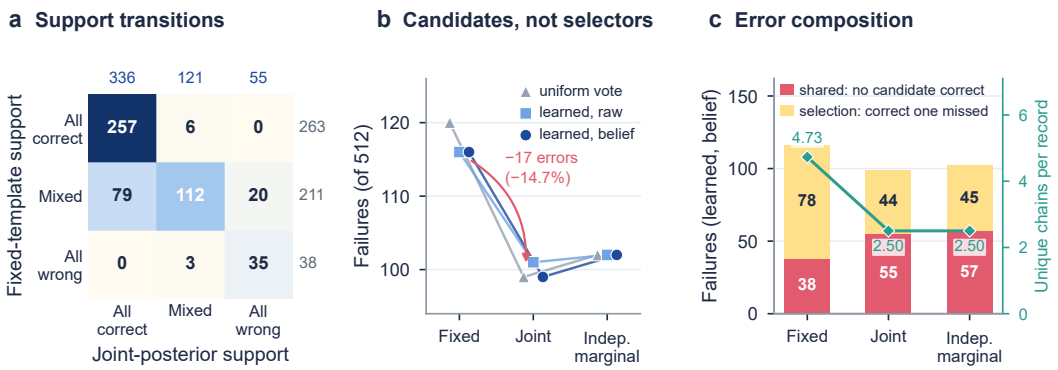

Figure 11: **Better hypotheses, less coverage.** All panels use the same 512 terminal records and $K = 7$ candidates. (a) Transitions of each record's support state from fixed to joint-posterior candidates, with marginal totals. (b) Failures for all nine candidate–selector pairs. (c) Failures of the belief selector split into shared errors (no candidate correct) and selection errors, with the mean number of distinct physical chains per record (right axis).

H Device execution checks and frozen-calibration replay

We first ran a redacted Tianyan calibration preflight for device `tianyan176` on September 24, 2026. The snapshot exposes 66 qubits and 110 couplers, of which two qubits and fourteen couplers are disabled; Figure 12 shows this calibration topology only. We then executed two independent batches of a deliberately small smoke test on four enabled couplers, using Bell-style and all-zero baseline circuits with 128 shots per circuit per batch. All eight QCIS checks passed in each batch and all sixteen result batches returned, for 2,048 shots in total. This validates the path from local API to device to returned results; it is not a decoder, logical-error, gate-fidelity, or threshold benchmark.

Aggregated over the two batches, the Bell-style circuits give correlated outcomes (00/11) with fractions 0.965, 0.930, 0.879, and 0.898 on G0–G3, with odd-parity fractions (01/10) of 0.035, 0.070, 0.121, and 0.102. The all-zero baselines return 00 with fractions 0.992, 0.957, 0.938, and 0.980. A matched Origin validation then ran three two-qubit circuits on two independently exposed backends: `WK_C180` (180 qubits, 168 topology edges) and `WK_C180_2` (180 qubits, 20 topology edges). The primary characterization used eight repeats of 512 shots for all-zero, Bell, and product-plus circuits on each backend, 24,576 returned shots in total. An independent shot-allocation replication used four repeats of 1,024 shots per circuit, again 24,576 shots, and reproduced the same qualitative output patterns. The returned probabilities are normalized to within 3×10^{-7}, and no credential, task identifier, or raw service response is packaged. These device outputs are never pooled with synthetic QEC records or used as decoder-performance evidence.

H.1 Frozen calibration and independent Origin confirmation

The hardware confirmation tests one narrow question: whether a device-specific calibration can enter the same temporal inference and correction pipeline used in simulation. We submitted 12

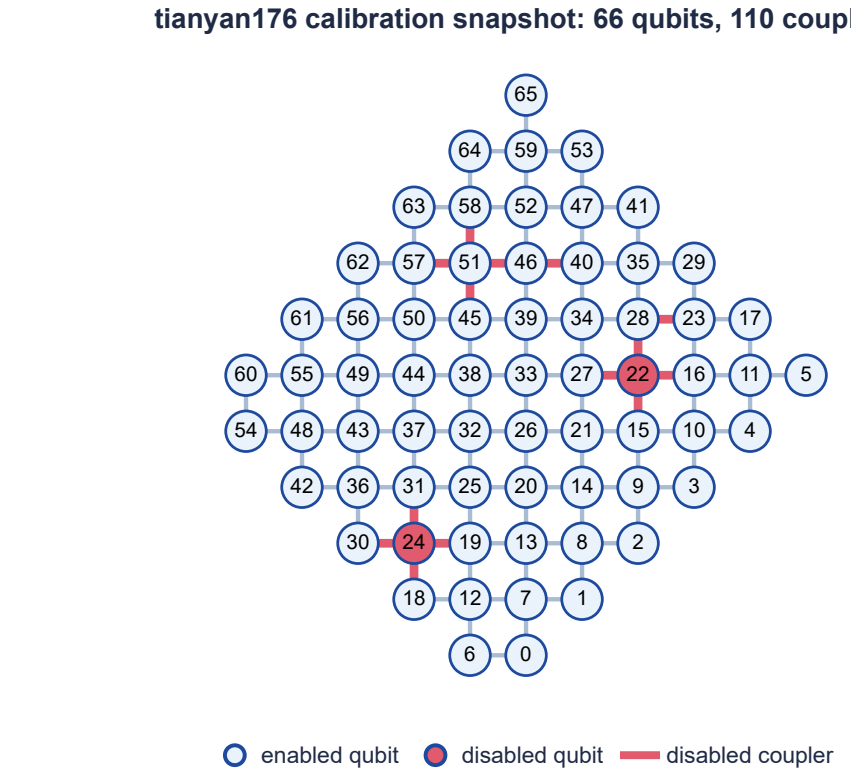

Figure 12: **Tianyan calibration topology (not live-shot data).** Connectivity reconstructed from the redacted calibration snapshot of `tianyan176`; rose marks disabled qubits and couplers. Coordinates are a schematic embedding of the returned topology, not die geometry or a statement about current availability.

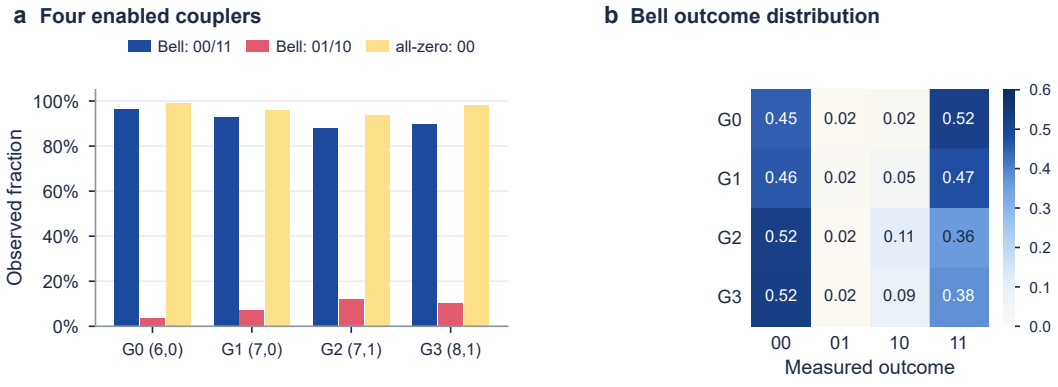

Figure 13: **Tianyan smoke test.** Four enabled couplers, each with one Bell-style circuit and one all-zero baseline, in two independent batches of 128 shots per circuit (256 aggregate shots per circuit). (a) Correlated Bell outcomes, odd-parity outcomes, and baseline 00 fractions. (b) Outcome distributions of the Bell-style circuits. The figure documents live execution and returned counts only.

Table 16: **De-identified Tianyan validation counts.** Circuit labels, selected couplers, shot counts, and returned computational-basis counts only.

Coupler	Qubits	Circuit	Shots	00	01	10	11
G0	6,0	Bell	256	115	4	5	132
G0	6,0	Baseline	256	254	2	0	0
G1	7,0	Bell	256	117	5	13	121
G1	7,0	Baseline	256	245	1	10	0
G2	7,1	Bell	256	133	4	27	92
G2	7,1	Baseline	256	240	11	5	0
G3	8,1	Bell	256	134	4	22	96
G3	8,1	Baseline	256	251	4	1	0

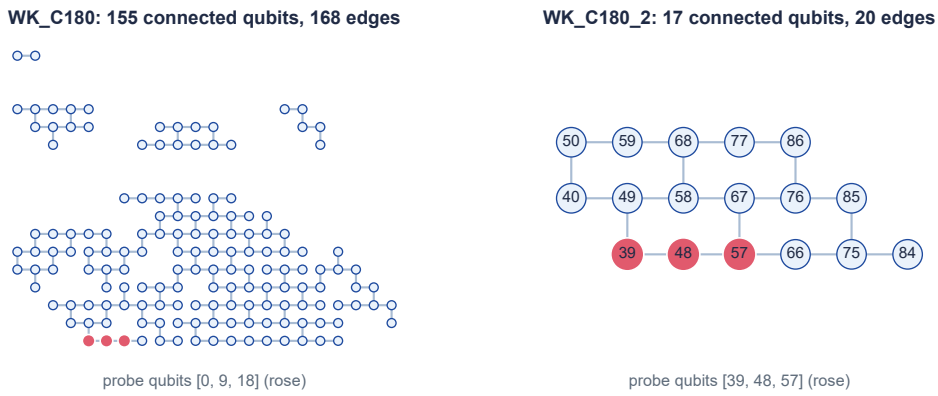

Figure 14: **Origin Quantum topology snapshots for the two validation backends.** Returned connectivity of `WK_C180` and `WK_C180_2`; rose nodes mark the qubits used by the validation circuits. Coordinates are a schematic lattice embedding, not die geometry or a statement about current availability.

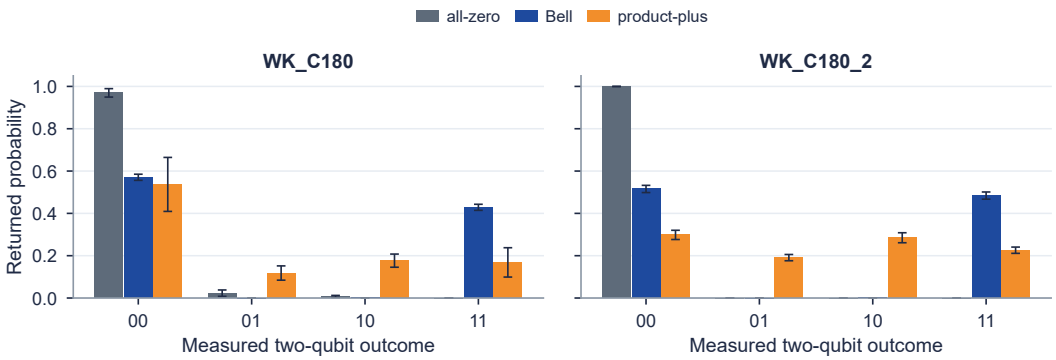

Figure 15: **Origin Quantum returned outcome probabilities.** Bars show the mean of eight 512-shot repeats for the all-zero, Bell, and product-plus circuits on each backend; whiskers show the across-repeat standard error. The panel documents live execution and output distributions, not logical error rates.

Table 17: **De-identified Origin Quantum validation probabilities.** Returned computational-basis probabilities for six backend–circuit pairs, each averaged over eight 512-shot repeats; "same" is $p_{00} + p_{11}$ and "opposite" is $p_{01} + p_{10}$.

Backend	Circuit	p_{00}	p_{01}	p_{10}	p_{11}	same	opposite
WK_C180	all_zero	0.970	0.023	0.007	0.000	0.970	0.030
WK_C180	bell	0.571	0.000	0.001	0.429	0.999	0.001
WK_C180	product_plus	0.537	0.118	0.177	0.168	0.705	0.295
WK_C180_2	all_zero	1.000	0.000	0.000	0.000	1.000	0.000
WK_C180_2	bell	0.515	0.000	0.000	0.484	1.000	0.000
WK_C180_2	product_plus	0.298	0.191	0.285	0.226	0.524	0.476

circuit records per backend, with three repeats and 256 requested shots per circuit, for 6,144 returned shots across `WK_C180` and `WK_C180_2`. All records completed, and the retained evidence lists no failures. A pilot calibration was frozen before the replay; its SHA-256 digest is `df2cabdf532b843fc6e49904a948998769797043ae511500f07b5f1be1df524a`, and the preflight script checks that it matches the calibration-freeze artifact.

The replay uses codewords 000 and 111, 1,024 synthetic causal profiles, and 65 rounds. Each state/codeword circuit prepares a computational-basis product state and applies independent R_Y rotations with programmed flip marginals $(.40, .025, .025)$ or $(.025, .30, .30)$. For each backend, the pilot repeats are averaged to obtain $P(y \mid z, \ell)$, where z is the engineered state, ℓ the logical bit, and y a three-bit outcome. Adding 10^{-4} to each outcome probability and renormalizing fixes the decoder likelihood before the independent confirmation records are acquired. All three estimators use this same likelihood and a uniform logical prior. The synthetic auxiliary histories have switch probability .01, observation noise 1.5, and 30% missingness, with replay seed 935300.

For each fresh repeat, the returned probabilities are used to evaluate the already-frozen decisions by summing over all eight outcomes and both logical bits and then averaging over the synthetic state histories. The confirmation histograms are therefore evaluation distributions, not refitted likelihoods, and all compared decisions share the same frozen calibration and auxiliary observations. Synthetic histories are not additional QPU shots, and returned probabilities are not converted into integer counts. Repeat-to-repeat spread describes acquisition variability; it is not a hardware confidence interval or a measurement of native temporal noise.

The calibrated replay gives Markov/current-only/static mean risks of 1.95/3.16/3.75% on `WK_C180` and 0.89/1.26/1.27% on `WK_C180_2`. The current-only minus Markov difference is positive on every repeat, at 0.91, 1.32, and 1.40 percentage points on `WK_C180` and 0.55, 0.08, and 0.47 points on `WK_C180_2`. The size of the gain differs between backends, as expected for device-specific calibrations, and we report the two backends separately rather than pooling them.

Table 18: **Independent Origin confirmation under one frozen calibration.** Descriptive replay risks in percent for three repeats on each backend; the final column is the current-only minus Markov difference. No synthetic history is counted as a QPU shot.

Backend	Repeat	Markov	Current-only	Static	Current–Markov
`WK_C180`	1	1.05	1.96	2.73	0.91
	2	2.70	4.03	4.40	1.32
	3	2.09	3.49	4.10	1.40
	mean	1.95	3.16	3.75	1.21
`WK_C180_2`	1	0.85	1.41	1.43	0.55
	2	0.72	0.80	0.79	0.08
	3	1.09	1.56	1.58	0.47
	mean	0.89	1.26	1.27	0.37

The acquisition supplies device-specific output distributions rather than a continuously streamed syndrome process. Its contribution is a frozen, auditable calibration-to-decision path; the temporal histories and correction decisions remain classical replay.

Why history can change the correction. A short calculation shows the decision boundary behind this probe. For the received word 100 and a uniform codeword prior, let r be the first-bit flip

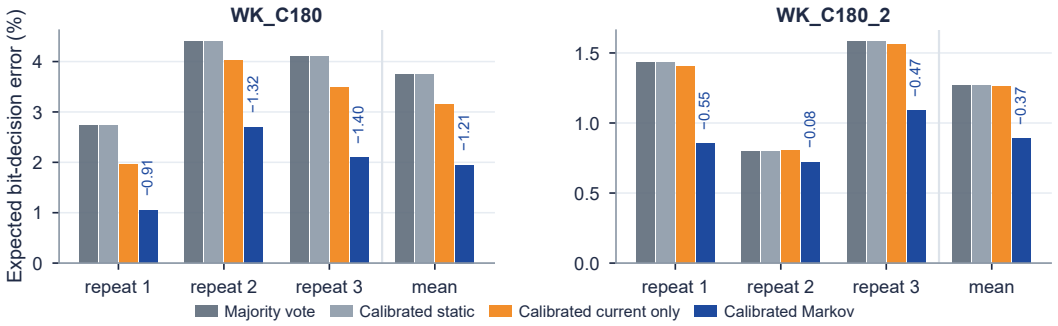

Figure 16: **Frozen-calibration replay on two Origin backends.** All three acquisition repeats and their mean for four decision rules, including uncalibrated majority vote; rotated labels give the reduction of Markov relative to current-only decoding, in percentage points. The plotted quantity is the expected classical bit-decision error under fresh returned probabilities and synthetic auxiliary histories, not live quantum feedback.

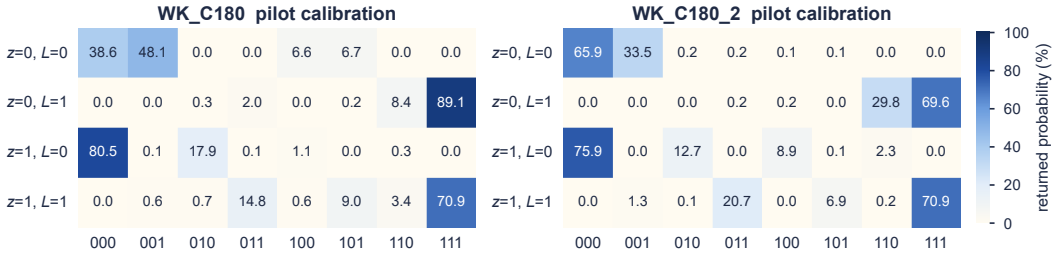

Figure 17: **Complete device-specific pilot calibration.** Each backend contributes a 4×8 probability map: four engineered-state/codeword combinations by all eight three-bit outcomes. Cells show mean returned probabilities in percent over three pilot repeats, on a common color scale.

probability in state zero while all other programmed marginals stay fixed. The logical likelihood difference is

$$\Delta(r,b) = (1-b)\big[r(.975)^2 - (1-r)(.025)^2\big] + b\big[.025(.7)^2 - .975(.3)^2\big]. \tag{12}$$

The decoder chooses logical zero when $\Delta > 0$ and logical one when $\Delta < 0$. Figure 3 shows this function as a surface and as a contour map: a change in belief matters when it crosses the zero contour, not simply when its state-estimation loss decreases. This exact independent-flip calculation explains the programmed mechanism; the empirical replay uses the frozen device likelihoods instead.

A separate cross-edge characterization sampled the first four usable topology edges returned by each backend, with two repeats of 256 shots for each of the three circuits, adding 12,288 returned shots across eight edges. The spread is descriptive: device outputs vary with both backend and edge, which is why we do not pool these records into a hardware-wide decoder claim. A fresh low-load rerun on the same two backends then submitted two 256-shot repeats per circuit, adding 3,072 returned shots with no failed backend record. The shot-allocation replication, the cross-edge sweep, and the low-load rerun are retained as separately hashed evidence objects.

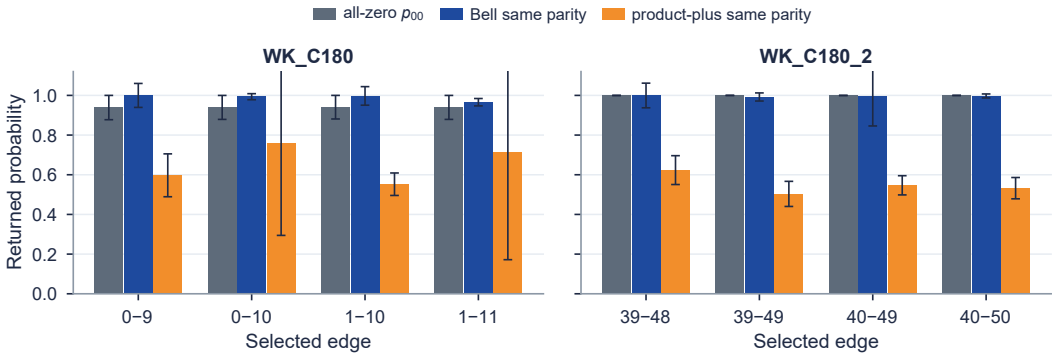

Figure 18: **Cross-edge Origin Quantum characterization.** For the first four usable edges of each backend, bars show the all-zero p_{00}, the Bell same-parity probability, and the product-plus same-parity probability; whiskers are across-repeat standard errors from two 256-shot repeats. This is a device-output variability check, not a decoder benchmark.

Table 19: **De-identified cross-edge Origin Quantum characterization.** Each row is one selected edge averaged over two repeats per circuit; "same" is $p_{00} + p_{11}$ and "opposite" is $p_{01} + p_{10}$.

Backend	Edge	All-zero p_{00}	Bell same	Bell opposite	Product same	Product opposite
WK_C180	0–9	0.939	1.000	0.000	0.597	0.403
WK_C180	0–10	0.939	0.993	0.007	0.757	0.243
WK_C180	1–10	0.941	0.997	0.003	0.552	0.448
WK_C180	1–11	0.939	0.966	0.034	0.714	0.286
WK_C180_2	39–48	1.000	1.000	0.000	0.623	0.377
WK_C180_2	39–49	1.000	0.992	0.008	0.503	0.497
WK_C180_2	40–49	1.000	0.994	0.006	0.547	0.453
WK_C180_2	40–50	1.000	0.997	0.003	0.532	0.468

I BOUNDARY-CONSISTENT CIRCUIT RETRAINING

The historical circuit generator copies its final hidden state. V0 applies a regular transition at that final filtering step. V1 uses an identity transition in the belief and compatibility features but keeps the V0 weights and feature scaler. V2 uses the same corrected inputs, refits the training feature standardization, and retrains five neural backbones at the three original seeds. The seven physical templates are unchanged, input-dependent logistic stacking and Markov-fusion temperature selection are repeated on the corrected training and validation inputs, and no new circuit test examples are generated.

The larger nine-candidate historical extension is not relabeled as V2; its records and original results remain a separate experiment. The V2 comparison covers the original seven-candidate interface, and posterior-candidate reconstruction under the revised circuit model would be another versioned experiment.

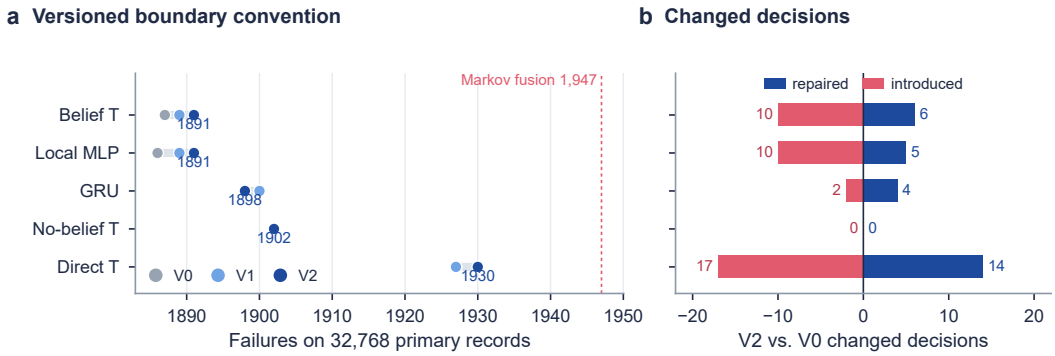

Figure 19: **A consistent terminal convention across estimation and learning.** (a) Failures of each backbone under V0 (historical model), V1 (corrected filtering input only), and V2 (retrained with corrected inputs) on the same 32,768 records; the dashed line marks Markov fusion under V2. (b) Decisions repaired and introduced by V2 relative to V0; their difference gives the net change.

Table 20: **Every boundary-version condition and backbone.** Repairs and introduced errors compare V2 with V0. The original test data are reused for this sensitivity analysis and are not counted as a new confirmation.

Condition	Method	V0	V1	V2	Repaired	Introduced
primary	Direct T	1927	1927	1930	14	17
primary	GRU	1900	1900	1898	4	2
primary	Local MLP	1886	1889	1891	5	10
primary	Belief T	1887	1889	1891	6	10
primary	No-belief T	1902	1902	1902	0	0
distance7	Direct T	147	147	146	3	2
distance7	GRU	149	151	150	1	2
distance7	Local MLP	154	154	154	1	1
distance7	Belief T	149	149	149	0	0
distance7	No-belief T	150	150	150	0	0
long25	Direct T	506	505	505	9	8
long25	GRU	513	511	511	2	0
long25	Local MLP	512	511	512	2	2
long25	Belief T	515	514	516	2	3
long25	No-belief T	518	518	518	0	0
missing30	Direct T	179	179	177	3	1
missing30	GRU	190	189	190	1	1
missing30	Local MLP	188	188	188	0	0
missing30	Belief T	190	191	193	0	3
missing30	No-belief T	192	192	192	0	0
drift	Direct T	168	168	167	4	3
drift	GRU	169	169	169	0	0
drift	Local MLP	168	168	168	0	0
drift	Belief T	172	172	172	0	0
drift	No-belief T	169	169	169	0	0

On the primary records, the V2 Transformer and the local MLP each make 1,891 errors. V2 GRU makes 1,898, the no-belief Transformer 1,902, the direct Transformer 1,930, corrected logistic stacking 1,907, the best fixed template 1,932, uniform voting 1,945, and Markov fusion 1,947. The eight-comparison Holm family gives $p = .0300$ against fusion and .0325 against voting, and the other comparisons are not significant. The corrected models thus keep the advantage of learned aggrega-

tion over fixed fusion, and the boundary correction does not create an architectural advantage of its own.

J TRANSITIONS, OBSERVATION GAPS, AND MODEL MISMATCH

Transition-aligned state curves use every hidden switch whose full $[-8, +16]$ window lies inside a confirmation record. Nearby switches may share rounds and shots can share a hidden profile, so we show descriptive means rather than confidence bands that would treat overlapping events as independent. The slow-switching filter is very accurate between switches but has a transient error peak when the environment actually changes. The same persistence prior that helps across gaps therefore delays adaptation to a new state.

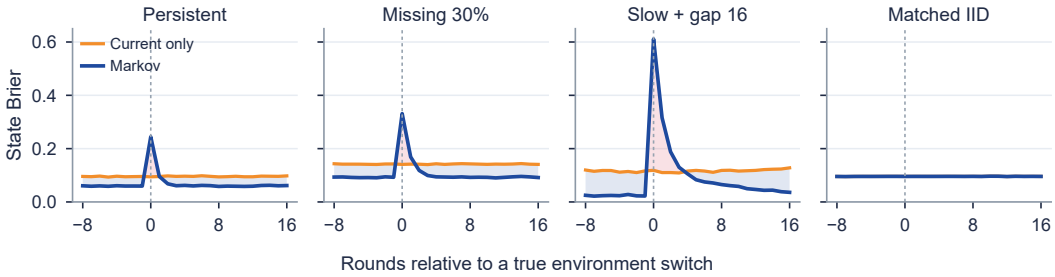

Figure 20: **Response to a state switch.** Mean state Brier score aligned to true environmental transitions for the four long-horizon conditions; blue shading marks rounds where Markov is better and red shading where it is worse. Both filters see the same observations, and the matched-IID curves coincide. Overlapping event windows are descriptive.

Missing blocks start at a deterministic quarter-horizon location and recur at a fixed spacing from the stored configuration, with gap lengths of 0, 4, 16, and 32 rounds under slow switching. The drift diagnostic adds a linear observation offset that reaches .75 at the final round while keeping the original filter calibration. The semi-Markov diagnostic replaces geometric switching by alternating state durations drawn uniformly from a fixed bounded range. Neither altered generator is used to tune the confirmation model.

The event plots measure state response, not the time of a logical failure or recovery, which would require additional logical measurements or a clearly specified posterior target. The current study also assumes known emission and transition parameters in its matched cells. Calibration from an unknown device, observation-dependent missingness, and online parameter learning remain separate questions.

Table 21: **Complete dynamic diagnostic counts.** Each condition uses 128 records in 64 profile clusters. Brier refers to states; errors refer to terminal logical parity.

Condition / method	n	Errors	Brier	Max edges
block0, Static	128	31	0.25000	660
block0, Current only	128	22	0.09849	660
block0, EMA	128	25	0.06471	660
block0, Markov	128	21	0.01989	660
block16, Static	128	22	0.25000	660
block16, Current only	128	15	0.16822	660
block16, EMA	128	20	0.08386	660
block16, Markov	128	21	0.04308	660
block32, Static	128	27	0.25000	660
block32, Current only	128	23	0.17144	660
block32, EMA	128	25	0.10191	660
block32, Markov	128	29	0.05990	660
block4, Static	128	24	0.25000	660
block4, Current only	128	17	0.11135	660
block4, EMA	128	18	0.06275	660
block4, Markov	128	16	0.01872	660
drift075, Static	128	31	0.25000	660
drift075, Current only	128	22	0.11526	660
drift075, EMA	128	24	0.13994	660
drift075, Markov	128	21	0.09026	660
semi markov, Static	128	39	0.25000	660
semi markov, Current only	128	25	0.09312	660
semi markov, EMA	128	28	0.13772	660
semi markov, Markov	128	28	0.06121	660

K LENGTH, CODE DISTANCE, AND COMMIT-BUDGET DIAGNOSTICS

The scale sweep crosses $d = 3, 7$, $H = 65, 257, 1025$, and complete versus 30%-missing observations. Every cell contains 64 records in 32 independent profiles, evenly split by logical basis. These cells have fewer clusters than the confirmation and are not used to claim asymptotic suppression. Every horizon has its own terminal readout and independently generated environments; we never treat prefixes of one record as independent experiments.

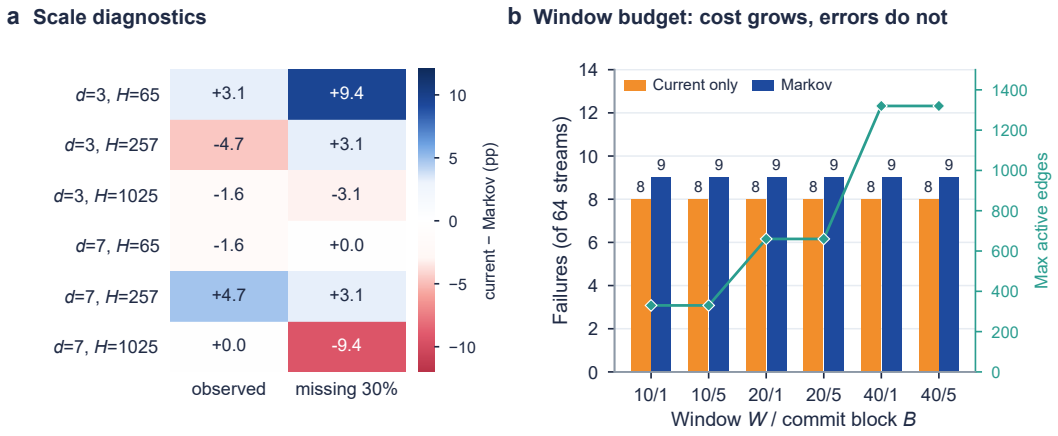

Figure 21: **Scale and waiting budgets.** (a) Descriptive current-only minus Markov terminal risk in percentage points on 64 records per cell, on a color scale symmetric around zero. (b) The same 64 records decoded with six (W, B) budgets: failure counts do not change, while the largest active matching graph grows with W. Equal error counts do not imply identical correction chains or service costs.

Table 22: **Every scale diagnostic.** Max edges is the largest active matching graph, which depends on distance and window rather than on the full horizon.

Condition / method	n	Errors	Brier	Max edges
missing30 d3 h1025, Current only	64	31	0.14188	132
missing30 d3 h1025, Markov	64	33	0.09239	132
missing30 d3 h257, Current only	64	30	0.14327	132
missing30 d3 h257, Markov	64	28	0.09757	132
missing30 d3 h65, Current only	64	23	0.13966	132
missing30 d3 h65, Markov	64	17	0.08978	132
missing30 d7 h1025, Current only	64	25	0.14238	1876
missing30 d7 h1025, Markov	64	31	0.09271	1876
missing30 d7 h257, Current only	64	24	0.14208	1876
missing30 d7 h257, Markov	64	22	0.09496	1876
missing30 d7 h65, Current only	64	6	0.14217	1876
missing30 d7 h65, Markov	64	6	0.09945	1876
persistent d3 h1025, Current only	64	31	0.09591	132
persistent d3 h1025, Markov	64	32	0.06021	132
persistent d3 h257, Current only	64	28	0.09283	132
persistent d3 h257, Markov	64	31	0.05906	132
persistent d3 h65, Current only	64	14	0.09205	132
persistent d3 h65, Markov	64	12	0.06039	132
persistent d7 h1025, Current only	64	35	0.09576	1876
persistent d7 h1025, Markov	64	35	0.06066	1876
persistent d7 h257, Current only	64	18	0.09576	1876
persistent d7 h257, Markov	64	15	0.06045	1876
persistent d7 h65, Current only	64	6	0.09279	1876
persistent d7 h65, Markov	64	7	0.05932	1876

Table 23: **Every window/commit setting.** All settings share the same 64-record test and the same fixed scheduling rule; no window is selected on these records.

Condition / method	n	Errors	Brier	Max edges
memoryless w10 b1	64	8	0.09138	330
markov w10 b1	64	9	0.05574	330
memoryless w10 b5	64	8	0.09138	330
markov w10 b5	64	9	0.05574	330
memoryless w20 b1	64	8	0.09138	660
markov w20 b1	64	9	0.05574	660
memoryless w20 b5	64	8	0.09138	660
markov w20 b5	64	9	0.05574	660
memoryless w40 b1	64	8	0.09138	1320
markov w40 b1	64	9	0.05574	1320
memoryless w40 b5	64	8	0.09138	1320
markov w40 b5	64	9	0.05574	1320

L ONLINE RESOURCE ACCOUNTING AND QUEUE REPLAY

After the parallel scientific runs finished, the resource experiment replays the first 32 streams from one confirmation basis sequentially on CPU, with one warm-up stream before each method. For every arriving round, timing starts before the belief update and ends after any matching, commit, and invariant checks, so it includes graph construction and native matching but excludes acquisition, audit-disk writes, and quantum feedback. The measured frame agrees with the saved confirmation frame on all 64 method-record replays.

Most rounds perform no matching, so their small median service time should not be mistaken for the cost of a commit. Each stream has 258 incoming detector layers, giving 8,256 timed requests per method. Markov/current-only means are $.47906/.47919$ ms, p50 values $.0082/.0082$ ms, p95 values $2.4593/2.4607$ ms, and p99 values $2.6941/2.6654$ ms. The decoder also waits 15–19 physical rounds before a steady-state block commit; algorithmic and computational latency are different quantities.

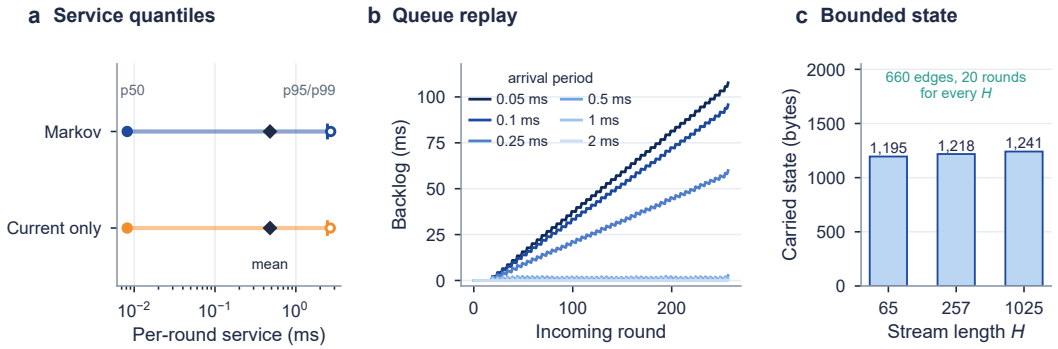

Figure 22: **Service cost and carried state of the rolling decoder.** The same $d = 5$, $W = 20$, $B = 5$ implementation measured sequentially on CPU. (a) p50, p95, p99, and mean per-round service, including graph construction, on a logarithmic axis. (b) Queue backlog under six synthetic arrival periods. (c) Largest serialized carried state for three stream lengths; transient graphs and process memory are excluded.

Queue replay combines the measured service sequence with synthetic arrival intervals of .05, .1, .25, .5, 1, and 2 ms. With arrivals a_i and service costs s_i, departures follow $f_i = \max(a_i, f_{i-1}) + s_i$, and the plotted backlog is $\max(0, f_i - a_i - \Delta)$. This isolates the queueing consequence of the measured software workload; it is not a device benchmark or a proof that real arrival and service processes match this trace.

The retained state is measured after every round by its serialized size. On the timed records, the maxima are 1,256 and 1,267 bytes for current-only and Markov decoding. The zero-fault length test keeps $W = 20$ and at most 660 graph edges while the total horizon grows to 1,025. Python objects, library allocations, and native graph storage are not included in the serialized byte count; the bounded-state claim is the structural $O(Wd^2)$ bound given by the ring buffer and the local graph, with explicit counter semantics.

M EXPANDED ARTIFACT AND CLAIM AUDIT

The aggregate evidence records all conditions, paired intervals, corrected p-values, physical candidate errors, training runs, source hashes, and service scopes. Figure and table scripts read this evidence, and every quantitative plot is generated programmatically from it. The overview in Figure 1 and panel (b) of Figure 2 are schematics and encode no unmeasured comparison.

The new implementation adds 15 parameterized correctness tests to 36 existing tests. It executes 46 rolling tasks over 5,696 distinct streams; 1,017,600 is the number of verified commits across all methods and settings, not the number of independent streams. The 18 terminal selector training runs and 15 corrected circuit runs give 33 new neural runs. The external and candidate terminal

comparisons share their 512 test examples. The historical 53,248 circuit QEC records, the new phenomenological QEC records, and the 92,160 state trajectories have separate provenance.

The finite confirmation does not establish a general temporal-Markov logical gain, a new optimal decoder, a threshold, or a real-time quantum service. Its positive technical result is an executable, physically checked connection from causal auxiliary inference to continuous correction. Its empirical result is a set of controlled distinctions: observation access versus temporal history, physical hypotheses versus selector features, support oracle versus actual selection, and carried state versus full service cost.

HISTORICAL CIRCUIT-LEVEL REFERENCE AND RETAINED DIAGNOSTICS

The following appendices document the original V0 circuit-level experiments, the public simulated-data prefix, the nine-candidate extension, and the implementation measurements. In these sections, "primary" and "base model" refer to the frozen V0 study unless V2 is stated explicitly. They are not results of the rolling decoder, which are reported above together with the corrected seven-candidate V2 results. Keeping the versions separate makes the boundary audit and the added evidence traceable.

CLAIM-TO-EVIDENCE MATRIX

Table 24 links each headline claim to its evaluation unit, the evidence behind it, and its scope.

N FILTERING AND DECISION IDENTITIES

N.1 TWO-STATE TRANSITION AND MISSING OBSERVATIONS

The continuous-time generator is

$$Q = \begin{pmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{pmatrix}.$$

Its eigenvalues are 0 and -2λ, giving $P(\Delta) = \exp(Q\Delta)$ with diagonal entries $1-a$ and off-diagonal entries $a = (1 - e^{-2\lambda\Delta})/2$. The stationary distribution is $(1/2, 1/2)$. Given a prior probability r and observation y, the Gaussian log-likelihood ratio is

$$\log \frac{p(y \mid Z = 1)}{p(y \mid Z = 0)} = \frac{(y+1)^2 - (y-1)^2}{2\sigma^2} = \frac{2y}{\sigma^2}.$$

Bayes' rule yields Equation (2). The mask is independent of the state in the missing-observation generator; a missing observation supplies no likelihood ratio. State-dependent missingness would require modeling its likelihood. At $a = 1/2$, each predictive prior is $1/2$, so the filter equals the memoryless estimator. This is an exact IID degeneration rather than a fitted approximation.

N.2 BRIER PROJECTION IDENTITY

Let $\mathcal{G}_t = \sigma(Y_t, M_t) \subseteq \mathcal{F}_t$, $X = \mathbf{1}[Z_t = 1]$, $m = \mathbb{E}[X \mid \mathcal{G}_t]$, and $b = \mathbb{E}[X \mid \mathcal{F}_t]$. Expanding $X - m = (X - b) + (b - m)$ gives

$$\mathbb{E}[(X-m)^2] = \mathbb{E}[(X-b)^2] + \mathbb{E}[(b-m)^2] + 2\mathbb{E}[(X-b)(b-m)].$$

The last term vanishes because $b - m$ is $\mathcal{F}_t$-measurable and $\mathbb{E}[X - b \mid \mathcal{F}_t] = 0$. This proves Equation (7); the gain is strict exactly when $b \neq m$ on a set of positive probability. The identity assumes actual conditional expectations. It does not compare misspecified fixed filters, nor does it transform a state Brier improvement into a logical-error improvement.

N.3 FILTER STABILITY IS NOT A DECODER-RISK GUARANTEE

For $0 < a < 1/2$ and log odds $u = \operatorname{logit}(b)$, define the predictive map

$$f(u) = \log \frac{(1-a)e^u + a}{ae^u + 1 - a}.$$

Claim	Evaluation unit	Evidence	Scope
Rolling correction preserves the carried frontier.	stream / commit	5,696 long-horizon and diagnostic streams and 8,192 focused confirmation streams; 2,107,136 checked commits.	Implementation coverage; commits and methods are not independent samples.
Temporal observations improve environmental inference.	hidden-state trajectory	45 conditions; 92,160 length-256 trajectories; Markov, current-only, EMA, and static estimators.	State-estimation benefit, reported separately from logical risk.
Observation access and temporal law are separable.	profile cluster	Four held-out conditions; 4,096 long-horizon streams; 12 cluster-aware comparisons.	The matched-IID control identifies the value of observations without attributing it to memory.
Memory changes logical decisions in a persistent regime.	profile cluster	4,096 streams at each of $d = 3, 5$; paired reductions of 1.81/.66 percentage points; Holm $p = .00020/.00490$.	Held-out confirmation of a development-selected setting; all three GRUs reported.
The belief interface feeds several physical solvers.	terminal record	512 shared records; MWPM, BP+OSD, BP+LSD, and BP+UF.	Consistent direction across solver families; not a universal decoder ranking.
Candidate support and decision quality differ.	fixed-budget record	512 test records; 18 selectors; fixed, joint, and matched-marginal candidates.	Mechanism analysis rather than a standalone superiority claim.
Boundary-consistent aggregation survives retraining.	circuit record	32,768 reused records; 15 corrected circuit models; three seeds each.	Versioned sensitivity study on a fixed test archive.
The pipeline executes on two device providers.	device execution	Tianyan: 2,048 shots. Origin: 24,576-shot primary run, 24,576-shot allocation replication, 12,288-shot cross-edge characterization, and a 3,072-shot low-load rerun.	Execution and returned-output validation, kept separate from logical-error claims.
A frozen calibration transfers to fresh Origin acquisitions.	acquisition repeat	6,144 returned shots on two backends; three repeats each; fixed pilot likelihood.	Classical replay with synthetic histories, not hardware QEC.

Table 24: **Claim-to-evidence matrix.** Evaluation units and evidence separate state inference, logical decisions, solver transfer, and device-calibrated replay.

Its derivative is

$$f'(u) = \frac{(1-2a)e^u}{[(1-a)e^u + a][ae^u + 1 - a]}.$$

The denominator minus e^u equals $a(1-a)(e^u-1)^2 \geq 0$, hence $0 \leq f'(u) \leq 1-2a$. Two filters exposed to the same observations add identical log-likelihood increments, so their log-odds difference contracts by at most $1-2a$ each ordinary step. The sigmoid is $1/4$-Lipschitz, yielding

$$|b_t - b'_t| \leq \tfrac{1}{4}(1-2a)^W |u_{t-W} - u'_{t-W}|.$$

This is stability of a specified HMM with respect to initialization. It neither selects the attention window nor bounds logical risk. A copied terminal state corresponds to an identity transition and does not contract at that step. Moreover, the attention tokens contain recurrent beliefs and cumulative detector features, so an eight-token neural window is not an eight-round total information horizon.

N.4 Proof of the support decomposition

For any $\widehat{L}(x) \in S(x)$, the event $\{L \notin S(x)\}$ implies $\widehat{L} \neq L$. Partition the error event by whether $L \in S(x)$ to obtain Equation (9). The events are disjoint and exhaustive. If $S_1(x) \subseteq S_2(x)$, then $\mathbf{1}[L \notin S_2(x)] \leq \mathbf{1}[L \notin S_1(x)]$ pointwise. No ordering follows for the second term because the decision rule may change. In this binary task, disagreement means both labels are available, and support absence occurs precisely when all candidates agree on the wrong label. The label-informed oracle is a diagnostic, not an implementable decoder.

N.5 Proof of disagreement-dependent gradients

With $w_k = e^{g_k} / \sum_i e^{g_i}$, the softmax derivative is $\partial w_k / \partial g_j = w_k(\mathbf{1}[k=j] - w_j)$. Therefore

$$\frac{\partial q}{\partial g_j} = \sum_k \ell_k w_k(\mathbf{1}[k=j] - w_j) = w_j \ell_j - w_j \sum_k w_k \ell_k = w_j(\ell_j - q).$$

If all $\ell_k = c$, then $q = c$ for every finite logit vector and every derivative is zero. For the smoothed probability $q_\epsilon = \epsilon + (1-2\epsilon)q$, $0 < \epsilon < 1/2$, the chain rule preserves the zero derivative for finite differentiable losses. Weight decay and finite-precision residuals are not supervised candidate-discrimination signals. Conversely, when labels differ and all logits are finite, $0 < q < 1$ and the probability has nonzero derivatives for both label groups. This characterizes where the logical loss can directly train relative candidate weights.

N.6 Environmental covariance and action sensitivity

Conditional on Z_t, the distinct fault indicators have zero covariance under the stated independence assumption. The law of total covariance therefore gives

$$\mathrm{Cov}(E_i, E_j \mid \mathcal{F}_t) = \mathrm{Cov}(p_i(0) + \delta_i Z_t, p_j(0) + \delta_j Z_t \mid \mathcal{F}_t) = \delta_i \delta_j \,\mathrm{Var}(Z_t \mid \mathcal{F}_t).$$

The variance of the Bernoulli state is $b_t(1-b_t)$, proving Equation (8). For two gate channels, $\delta_i \delta_j = 0.010^2 > 0$; for a gate and measurement channel it is $0.010(-0.017) < 0$. At $b_t = 1/2$ these environmental covariance terms are 2.5×10^{-5} and -4.25×10^{-5}, respectively. They are exact illustrative consequences of the specified affine model, not newly measured experimental effects. Additional intrinsic correlations within a physical fault require their own conditional-covariance term.

For Equation (11), define $\mathcal{R}(A) = \mathbb{P}(A(x) \neq L)$. Pointwise, $|\mathbf{1}[A(x) \neq L] - \mathbf{1}[B(x) \neq L]| \leq \mathbf{1}[A(x) \neq B(x)]$. If both actions belong to a singleton candidate support, they agree. Taking expectations and applying the triangle inequality yields both inequalities. The bound compares selectors over the same input-dependent candidate set; a method that changes the physical support is not confined to the same disagreement event.

N.7 A standard interface to logical decision risk

For a true logical distribution p, approximate distribution q, and their maximizing actions a_p, a_q, the conditional excess risk obeys

$$p(a_p) - p(a_q) \leq |p(a_p) - q(a_p)| + |p(a_q) - q(a_q)| \leq 2\,\mathrm{TV}(p,q) \leq \sqrt{2\,\mathrm{KL}(p\|q)}.$$

The first inequality uses $q(a_q) \geq q(a_p)$; the last uses Pinsker's inequality. This standard inequality makes the required connection explicit: controlling environmental estimation is insufficient unless it controls the logical distribution. The present experiments measure logical outcomes directly; they do not establish a general KL bound for candidate compression.

O REPRODUCIBLE EXPERIMENTAL PROTOCOL

O.1 DATA GENERATION AND SPLITS

Stim's rotated-memory circuits supply the gate schedule and detector coordinates. Each sampled environmental profile instantiates round-dependent gate/reset and measurement probabilities. For each configuration, 32 profiles are sampled and each generates multiple independent fault realizations conditional on that profile. Auxiliary Gaussian observations are drawn per record. Thus records sharing a profile are conditionally independent draws given an environment, but are not interchangeable unconditional independent samples for cross-environment inference.

Training configurations $(X, 3), (X, 5), (Z, 3), (Z, 5)$ use seeds 940100–940103; validation uses 940200–940203. The eight primary strata use two seeds per configuration, 941000/941001, 941010/941011, 941020/941021, and 941030/941031. Secondary-condition seeds are 942000/942001 ($d = 7$), 942010/942011 ($T = 25$), 942020/942021 (missing), and 942030/942031 (drift). Each secondary basis has 2,048 records and 32 profiles. The candidate-extension test uses 950100–950103, each with 1,024 records and 32 profiles; its training and validation records reuse the original development splits.

The drift condition adds an independent Ornstein–Uhlenbeck offset with stationary standard deviation 1.2 and time constant 30 to the auxiliary observations. The model retains its original emission calibration. Physical circuits remain classical Pauli-noise simulations; they do not simulate general coherent errors or continuous weak-measurement backaction.

The hidden-process study uses seeds 960100–960104 for test and 960200–960204 for development, in the order Markov, missing, IID, semi-Markov, shuffled. Every test condition has 2,048 trajectories of length 128; development has 512. Semi-Markov states switch after durations drawn uniformly from integers 4 through 12. Shuffling jointly permutes each trajectory's states, observations, and masks, preserving the single-time observation relationship while disrupting order. The fixed HMM rate remains unchanged in these diagnostic controls. These data are not added to the QEC sample count.

O.2 CANDIDATE FEATURES AND MODEL TRAINING

For template values $s_{k,t} \in [0, 1]$, set $\widetilde{s}_{k,t} = 0.1 + 0.8 s_{k,t}$. The compatibility feature is

$$v_k = \frac{1}{\sqrt{T+1}} \sum_{t=0}^{T} \left[b_t \log \widetilde{s}_{k,t} + (1 - b_t) \log(1 - \widetilde{s}_{k,t}) \right].$$

The candidate head receives K labels, K normalized matching costs, and K compatibility features. Standardization is fitted on the training split. Side features are telemetry divided by 3, observation mask, belief, and time index divided by 16. Matching candidates are terminal features and are not presented as information available to earlier neural time steps.

All base neural models use training seeds 940301–940303, batch size 512, 30 epochs, AdamW learning rate 0.001, weight decay 0.0001, gradient clipping at norm 1, and float32 without mixed precision. The selected checkpoint minimizes validation BCE. The three candidate-extension backbones use seeds 950301–950303 with the same schedule. Across base and extension studies there are 24 formal training runs. A separately retained, same-seed full retraining reproduces one base checkpoint and is not a new test replicate.

Validation selects symmetric smoothing $\epsilon \in \{0, 0.001, 0.003, 0.01, 0.03, 0.1\}$ by BCE. All values are below $1/2$ and preserve the threshold action. The best fixed template minimizes validation error. Markov fusion chooses temperature from $\{1, 2, 4, 8\}$ using validation BCE. Logistic regression uses $C = 1$, maximum 1,000 iterations, and training-standardized candidate features plus temporal/spatial summary statistics. Model selection does not consult test labels.

Table 25: **Base model training footprint.** Times are measured per training run in the retained environment, averaged over three seeds; they exclude circuit generation and candidate construction. Selected epochs correspond to seeds 940301–940303.

Backbone	Parameters	Mean seconds	Selected epochs
BELIEF-T	13306	9.84	10, 13, 9
Belief local MLP	5818	7.96	10, 19, 21
Belief GRU	11098	8.55	10, 13, 13
No-belief Transformer	13306	10.07	2, 1, 9
Direct Transformer	12982	10.11	26, 23, 24

The retained compute environment uses Python 3.11.9, PyTorch 2.11.0+cu128, and an NVIDIA GeForce RTX 5080 Laptop GPU. Deterministic algorithms are enabled; TF32 is disabled. Training and replay use the recorded package versions, while plotting and PDF generation use a separate document environment. Local replay is not a substitute for an independent clean-environment reproduction.

O.3 INFORMATION ACCESS AND PHYSICAL BASELINES

Table 26: **Information pathways.** A dash indicates the pathway is not used. All methods receive the appropriate terminal detector record. "Raw" means observations and masks; "belief" means the recursive HMM feature, not access to hidden labels.

Method family	Auxiliary input	Belief use	Physical candidates
Fixed matching / BP	—	—	Fixed calibrated model
Memoryless BP	Raw	Per-round prior reset	Per-record rates
Markov BP	Raw	Rate construction	Per-record rates
BELIEF-T / GRU / MLP	Raw + belief	Aggregation	Seven fixed
No-belief Transformer	Raw	—	Seven fixed
Posterior matching	Raw + belief	Rate construction	One posterior profile
Nine-candidate models	Raw + belief	Construction + aggregation	Seven fixed + two new

BP+OSD and BP+LSD run through the native `ldpc` package. Static baselines use detector records with a fixed mean-channel DEM. The adaptive variants rebuild the DEM from either the Markov or the memoryless soft profile. Decoder hyperparameters are selected on validation through a retained 12-setting search; the selected settings are shared between the two adaptive variants. Syndrome-validity checks accompany decoded corrections. Comparisons between a neural method with auxiliary observations and a static physical decoder are therefore not information-matched.

O.4 PAIRED INFERENCE AND MULTIPLICITY

For a reference and candidate, let $e^{\mathrm{ref}}_{g,i}, e^{\mathrm{cand}}_{g,i}$ be paired error indicators in profile g. Define the profile total difference $\delta_g = \sum_i (e^{\mathrm{ref}}_{g,i} - e^{\mathrm{cand}}_{g,i})$ and estimate the record-weighted risk difference by $\sum_g \delta_g / \sum_g n_g$. A two-sided sign-randomization test independently flips profile totals 20,000 times, with the plus-one correction for a finite Monte Carlo p-value. Its null calibration assumes sign symmetry of independent profile differences within the fixed configuration–seed mixture. It is not an unrestricted exact test for every heterogeneous random-effects distribution.

Confidence intervals resample profiles within configuration–seed strata, retaining each selected profile's records, for 20,000 bootstrap draws. Main-family significance uses Holm adjustment over eight comparisons; posterior expansion uses seven; the Markov-versus-memoryless external pair uses two. Intervals themselves are unadjusted and need not agree with family-adjusted binary decisions. Hidden-process intervals resample complete trajectories for 2,000 draws. No training-seed repetitions or repeated time points are added to independent sample sizes. Public-prefix records lack a verified cluster structure and receive descriptive, not confirmatory, comparisons.

P COMPLETE RESULTS AND CALIBRATION

Table 27: **All primary methods and probability metrics.** Each row has 32,768 records. BCE uses probability clipping at 10^{-7} for evaluation; ECE uses 15 equal-width bins. Calibration scores do not alter the interpretation of action error.

Method	Failures	BCE	Brier	ECE_{15}
BELIEF-T	1887	0.19964	0.04959	0.02045
Belief GRU	1900	0.19954	0.04956	0.02050
Belief local MLP	1886	0.19967	0.04958	0.02050
No-belief Transformer	1902	0.19962	0.04968	0.02090
Direct Transformer	1927	0.17787	0.04835	0.00852
Linear stacking	1912	0.19068	0.04959	0.00693
Best fixed template	1932	0.23541	0.05632	0.02896
Uniform vote	1945	0.20078	0.05022	0.02250
Markov fusion	1947	0.20105	0.05031	0.02296
Ordinary MWPM	2163	0.25040	0.06281	0.04958
Correlated mean matching	1962	0.23692	0.05790	0.05386
correlated_cold	1932	0.23541	0.05632	0.02896
correlated_hot	2044	0.24242	0.05990	0.05586
correlated_cold_hot	1997	0.23927	0.05875	0.05295
correlated_hot_cold	2022	0.24094	0.05937	0.05269
correlated_pulse	2017	0.24061	0.05924	0.05305
correlated_valley	1998	0.23933	0.05878	0.05405

The direct Transformer has a lower primary BCE than the constrained candidate model while making more logical mistakes. These measure different properties: a distribution can improve average probability assignment without changing enough threshold decisions in the favorable direction. Similarly, validation smoothing changes confidence but not logical class. This is why the primary endpoint is a paired action risk, with probability metrics serving as diagnostics.

Table 28: **All primary training seeds.** Counts use the same 32,768 test records, not three independent test sets. Seeds 1–3 are 940301–940303. The final column is probability averaging, not the best member.

Backbone	Seed 1	Seed 2	Seed 3	Ensemble
BELIEF-T	1886	1889	1894	1887
Belief local MLP	1883	1884	1898	1886
Belief GRU	1902	1898	1896	1900
No-belief Transformer	1917	1927	1889	1902
Direct Transformer	1935	1929	1946	1927

Table 29: **Full condition-level failure counts.** T denotes BELIEF-T, MLP and GRU share its input interface, Best is the validation-selected fixed template, and Oracle is the label-informed candidate-support lower bound. Secondary conditions are not pooled into the primary significance family.

Condition	n	T	MLP	GRU	Best	Vote	Oracle
Primary	32768	1887	1886	1900	1932	1945	1305
$d = 7$	4096	149	154	149	164	149	85
$T = 25$	4096	515	512	513	529	505	308
Missing 30%	4096	190	188	190	183	187	117
Drift	4096	172	168	169	180	171	103

Table 30: **Hidden-process state-estimation results.** Each condition has 2,048 independent length-128 trajectories. Relative reduction compares the fixed Markov filter to the current-observation estimator. Intervals are unadjusted absolute Brier reductions, not relative-percent intervals.

Process	Current only	Markov	Reduction (%)	Absolute 95% CI
Markov	0.09538	0.06028	+36.80	[0.0344, 0.0358]
Missing 30%	0.14162	0.09307	+34.28	[0.0477, 0.0493]
IID	0.09600	0.13456	-40.16	[-0.0392, -0.0379]
Semi-Markov	0.09595	0.06994	+27.11	[0.0254, 0.0266]
Shuffled	0.09494	0.12677	-33.53	[-0.0326, -0.0311]

Q POSTERIOR CANDIDATES AND SAME-INFORMATION CONTROLS

Table 31: **All extension and external methods on the same 4,096 records.** Auxiliary access is shown explicitly. Three neural training seeds are probability-averaged. Static and adaptive BP variants have different information and construction cost.

Method	Failures / 4,096	Input
7-candidate Transformer	248	Detector + auxiliary
9-candidate Transformer	242	Detector + auxiliary
9-candidate GRU	242	Detector + auxiliary
9-candidate local MLP	244	Detector + auxiliary
Posterior matching	233	Detector + auxiliary
Hard-state matching	234	Detector + auxiliary
9-candidate vote	236	Detector + auxiliary
Best fixed template	251	Detector
BP+OSD	230	Detector
BP+LSD	297	Detector
Markov BP+OSD	207	Detector + auxiliary
Markov BP+LSD	298	Detector + auxiliary
Memoryless BP+OSD	211	Detector + auxiliary
Memoryless BP+LSD	296	Detector + auxiliary

Table 32: **All seven extension comparisons.** Positive reductions favor the nine-candidate Transformer. Intervals are profile-bootstrap intervals over 128 profiles, unadjusted for multiplicity; p-values are Holm-adjusted.

Reference	Reduction (pp)	95% CI	Holm p
9-candidate GRU	+0.000	[-0.146, +0.122]	1.000
9-candidate local MLP	+0.049	[-0.073, +0.195]	1.000
Posterior matching	-0.220	[-0.513, +0.049]	1.000
Hard-state matching	-0.195	[-0.488, +0.122]	1.000
9-candidate vote	-0.146	[-0.415, +0.122]	1.000
Best fixed template	+0.220	[-0.122, +0.586]	1.000
7-candidate Transformer	+0.146	[-0.073, +0.366]	1.000

The same-information BP+OSD comparison gives a Markov risk reduction of 0.0977 percentage points, unadjusted interval $[-0.1465, 0.3418]$, and raw sign-randomization $p = 0.5601$. BP+LSD gives -0.0488 percentage points, interval $[-0.4883, 0.3906]$, and raw $p = 0.9147$. Both Holm values are 1. These are prespecified subsequent mechanism comparisons on an already available test, not an independently untouched replication of the primary result. Their role is attribution, not inflating the total evaluation sample.

Table 33: **All candidate-extension training seeds.** Counts use the same 4,096 test records. Seeds 1–3 are 950301–950303. Ensemble values average probabilities before thresholding; the best individual seed is not selected on the test set.

Backbone	Seed 1	Seed 2	Seed 3	Ensemble
9-candidate Transformer	237	246	243	242
9-candidate GRU	238	241	239	242
9-candidate local MLP	237	242	247	244

R LEARNING SIGNAL AND PROFILE HETEROGENEITY

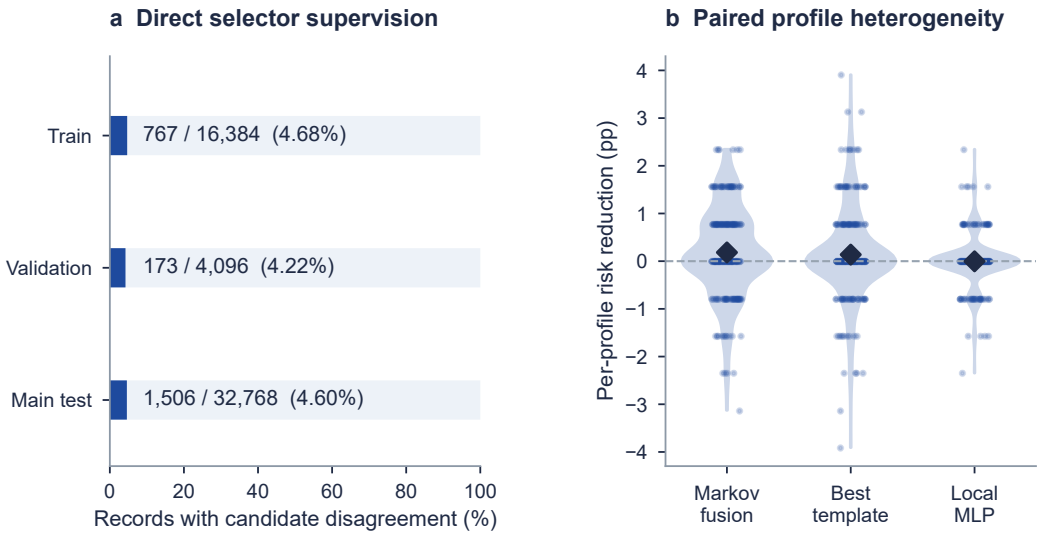

Figure 23: **Support determines where reweighting can learn.** (a) Candidate-disagreement records provide the direct relative-weight supervision characterized by Proposition 2. The remainder agree, whether correctly or incorrectly. (b) Descriptive distributions of paired per-profile risk reductions in the primary test, relative to BELIEF-T; each dot is one of 256 profiles, and diamonds are profile means. Violin densities are visual summaries, not additional inferential tests.

The support counts are 649 errors in training, 170 in validation, and 1,305 in the primary test. These values differ from disagreement counts: unanimous wrong predictions have no candidate-selection gradient but a nonzero irreducible error for the fixed interface. In the primary test, 29,957 records are unanimous and correct; 1,305 are unanimous and incorrect; 924 disagreement records are selected correctly; and 582 are selected incorrectly. The four counts sum to 32,768. These disjoint cells make the risk and supervision analysis reproducible without interpreting attention weights as causal explanations.

S PUBLIC SIMULATED-DATA TRANSFER

The public Pauli+ object is a released simulated-data source associated with AlphaQubit, used under CC-BY-4.0. The bounded download contains a fixed 32 MiB prefix, from which the first 4,096 complete CRC-validated records are parsed. Official coordinates and deterministic logical-line relations define the detector and logical-boundary conversion. Re-parsing verifies 8,192 boundary parity identities. These checks validate the implemented mapping, not identity of the training and public noise models.

The model receives hard-thresholded records and no telemetry; the belief is fixed at $1/2$. Available soft/leakage fields are retained in the archive but are not modeled in this transfer. The assumed

circuit noise is an approximate SI1000-like model rather than the released generator. No fine-tuning or test-based choice of coordinate orientation, template, seed, or threshold is performed.

Table 34: **Public-prefix transfer.** All methods use the same 4,096 prefix records. These are descriptive results, not a full official AlphaQubit benchmark or a hardware experiment.

Method	Failures / 4,096	Risk (%)
BELIEF-T	705	17.212
Belief GRU	723	17.651
Belief local MLP	714	17.432
No-belief Transformer	690	16.846
Direct Transformer	799	19.507
Linear stacking	733	17.896
Best fixed template	694	16.943
Uniform vote	738	18.018
Markov fusion	738	18.018
Ordinary MWPM	822	20.068
Correlated mean matching	733	17.896
correlated_cold	694	16.943
correlated_hot	834	20.361
correlated_cold_hot	756	18.457
correlated_hot_cold	782	19.092
correlated_pulse	772	18.848
correlated_valley	784	19.141

Static BP+OSD and BP+LSD each make 811 errors on the same public prefix. Official AlphaQubit/AlphaQubit 2, harmonization, Libra, and Neural MWPM are not executed in this artifact; their literature results are not inserted into these tables. The transfer establishes a functioning input/output mapping and a measured boundary of the calibrated auxiliary-information setting.

T TERMINAL BOUNDARY AND MODEL MISMATCH

The last environmental state in the QEC generator is copied, while the frozen HMM predicts a transition. Replacing only the final predictive step by identity yields maximum per-record posterior differences of 0.687, 0.786, 0.858, and 0.709 across the four extension configurations. Mean absolute differences are approximately 0.041–0.049. Corresponding final-state Brier scores change from 0.03365 to 0.02344, 0.03783 to 0.03270, 0.04022 to 0.02681, and 0.04184 to 0.03179. These are posterior diagnostics on retained records. They are not new logical-error results because the neural models and physical candidates have not been retrained or rescored as a new corrected system.

This convention matters when interpreting the filter as a conditional expectation. Ordinary-step matched-model statements remain well-defined, and the separate length-128 process study has no terminal copying. The paper therefore distinguishes exact filtering identities, the frozen approximate circuit-level feature, and the empirically measured decoder. The same distinction applies to observation drift and unavailable public telemetry.

U RESOURCE MEASUREMENTS AND NUMERICAL CHECKS

Cache tests cover $d \in \{3, 5, 7\}$ and $T \in \{9, 25, 65, 129\}$. Neural key/value state remains 2,048 bytes for one model; full-system memory also includes parameters, spatial features, detector records, and matching graphs. The batch-local and step-cache outputs differ by at most 1.79×10^{-7} in the retained comparison. The Python step-cache implementation does not outperform batch-network processing of complete records; cache size and complete-record throughput are different measurements.

The output-preserving feature optimization compares identical single-model batch services. Median times change from 1.975 to 1.081 ms ($d = 5, T = 9$), 3.289 to 1.455 ms ($d = 7, T = 9$), 19.666 to 6.344 ms ($d = 5, T = 129$), and 39.396 to 12.702 ms ($d = 7, T = 129$). Each condition has 100 measured requests and 20 warm-up requests; all 480 reference/optimized probabilities, including warm-ups, agree bitwise.

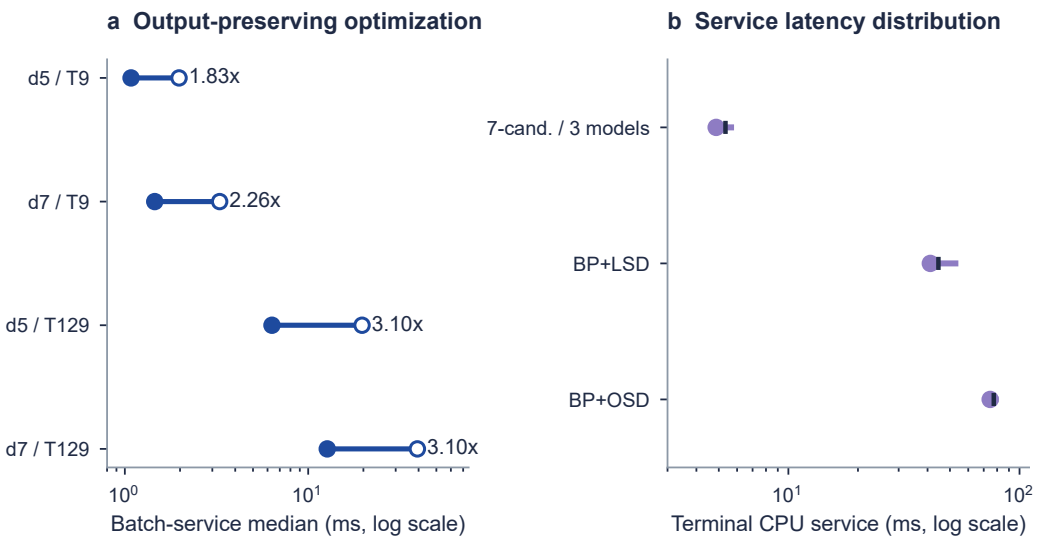

Figure 24: **Implementation and service measurements have different scopes.** (a) Paired feature implementations use the same seven matchers, one model, and batch network; open markers are the reference implementation and filled markers the optimized one. (b) A separate same-process interleaved experiment compares terminal services at $d = 5, T = 9$; markers are medians, ticks p95, and interval ends p99. The three-model neural service has auxiliary observations while static BP services do not. Both panels exclude acquisition, queueing, and preloading.

The separate three-model terminal service has p50/p95/p99 of 4.870/5.344/5.649 ms. Static BP+OSD gives 74.638/77.380/78.779 ms and BP+LSD 41.121/44.576/52.776 ms under the same-process interleaved timing protocol. The neural service includes seven matches, features, three forward passes, and averaging. These latency results should not be combined with the nine-candidate accuracy into a single operating point.

CPU single-record replay differs from the original GPU batch probabilities by at most 9.48×10^{-6} on the 120 checked requests. All logical actions agree, but the original 2×10^{-6} probability tolerance is not met. Timing quantiles and action agreement are therefore reported separately from probability equivalence. One hundred measured requests provide little information about extreme tail behavior; p99 is a sample quantile, not a sustained-service guarantee.

Posterior-driven candidates reconstruct the circuit, DEM, and matcher per record. On the $d = 5, T = 9$ extension test, each additional soft or hard candidate has roughly 15 ms median construction-and-decoding time in its recorded run. This cost is not the precompiled seven-candidate service cost. Full streaming matching, window-boundary accuracy, backlogs, and closed-loop QPU feedback are outside these measurements.

V Artifact contents and verification scope

The accompanying source package contains the unmodified official ICLR 2027 style files, complete manuscript sources, references, vector/raster figure exports, aggregate evidence, and scripts that regenerate all numerical plots and tables. Conceptual diagrams contain no empirical measurements. Numerical data figures use no generative image model. Source hashes and online-policy checks are retained in the local delivery audit, separately from the anonymous manuscript.

Verification replays 49,152 base-study test predictions, 4,096 posterior-extension predictions, and 4,096 public-prefix predictions against saved checkpoints; checks 24,576 candidate-augmentation records and rebuilds 128 adaptive candidates; recomputes all primary, extension, and same-information comparisons; and passes 36 unit/property tests. The additional hidden-process diagnostics recompute 10,240 trajectories. These are verification counts, not new independent scien-

tific samples. Full benchmark reruns, external hyperparameter searches, hardware experiments, and clean-machine reconstruction are not claimed by the source package.

W HISTORICAL INTERFACE AND MECHANISM VISUALIZATIONS

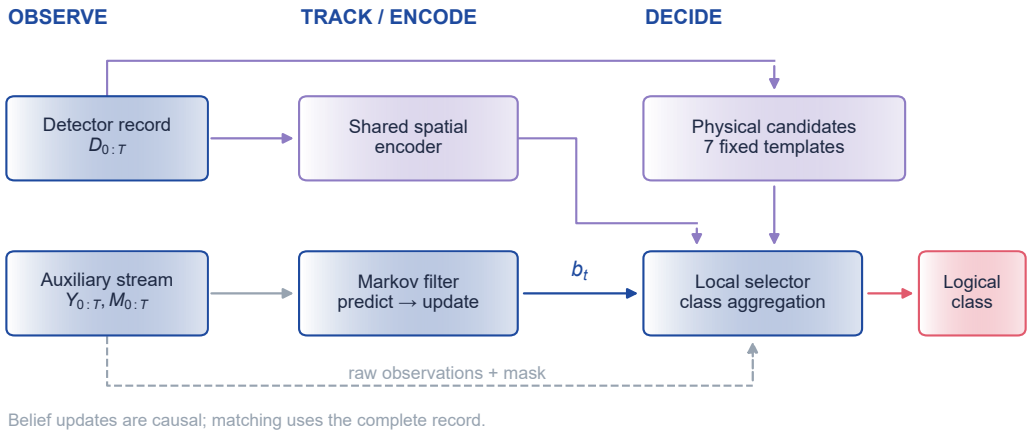

Figure 25: **Historical terminal architecture.** Seven fixed physical candidates and a learned selector operate on a complete detector record. This terminal design is distinct from the rolling decoder of the main text.

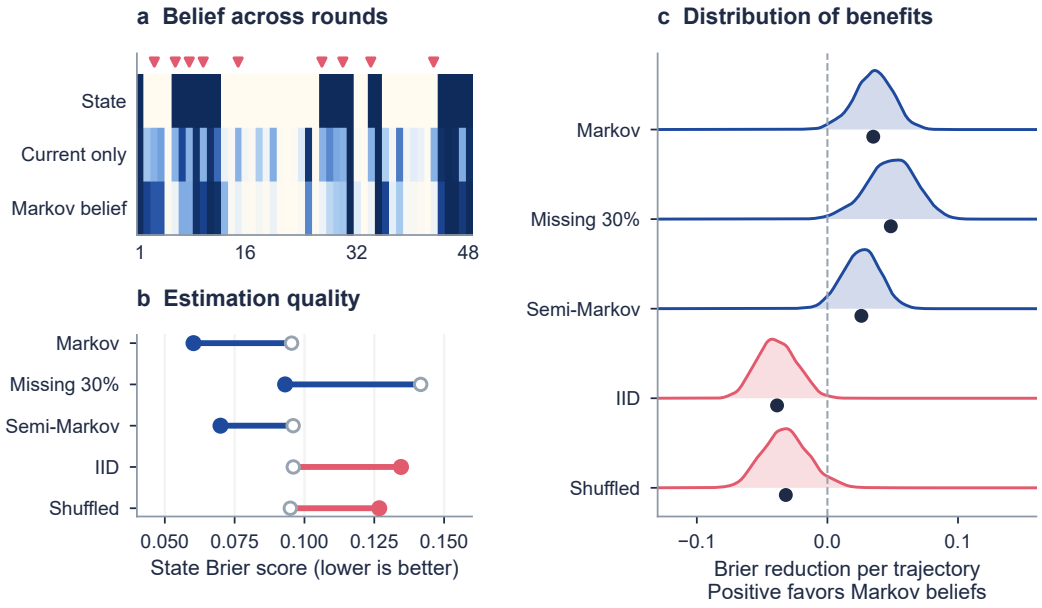

Figure 26: **Historical state-inference controls.** Probability raster, Brier comparisons, and peak-normalized densities of per-trajectory effects from the original five-condition, 10,240-trajectory experiment. The 45-condition study in the main text uses independent data.

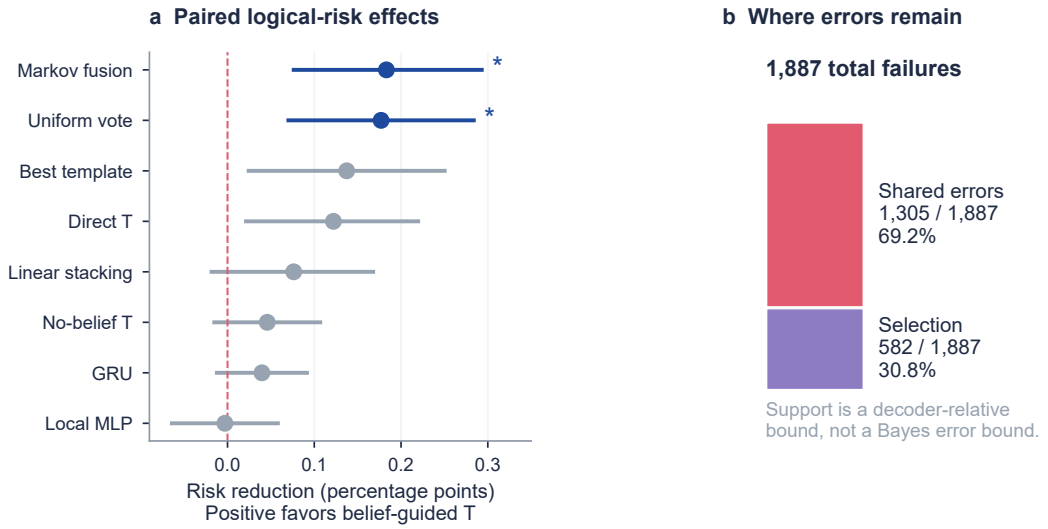

Figure 27: **Historical V0 paired comparisons and error decomposition.** All 32,768 records use the original model. Unadjusted intervals are shown, and stars mark significance under the original eight-comparison Holm family. V2 results appear in Appendix I.

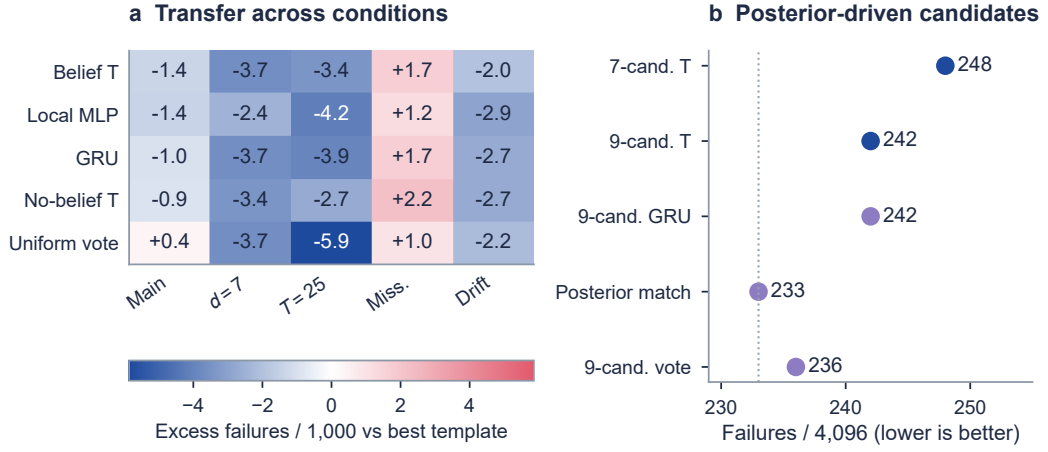

Figure 28: **Historical circuit transfer and nine-candidate extension.** Secondary conditions and the separate 4,096-record posterior-extension study, reported as original results rather than measurements of the rolling implementation.