ROUTE-CONDITIONED PREDICTION EXTRAPOLATION: DEPENDENT-DATA DIRECTIONS, PERSISTENCE, AND CERTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Route-specific fine-tuning defines a direction in prediction space, and scalar extrapolation along that direction can improve a pooled predictor. We study

$$\widehat{f}_\alpha = \mathrm{clip}\Big[\widehat{f}_0 + \alpha(\widehat{f}_{\mathrm{eq}} - \widehat{f}_0)\Big],$$

where $\widehat{f}_0$ is a pooled checkpoint and $\widehat{f}_{\mathrm{eq}}$ is a target-mass aggregate of route experts initialized from that checkpoint. A persistent two-state model isolates a composition-correction special case: the leading Brier risk is quadratic in α, the optimum is F/H, and imperfect ergodic route errors imply $\alpha^* > 1$. At $q = 0.10$ and $\rho_R = 0.98$, the exact coefficients are 1.5368, 1.5208, 1.4205, and 1.1328 as route-error persistence increases over $\lambda = 0, 0.5, 0.9, 0.99$, and paired simulations track the corresponding correction gains. On HAR-Transitions, WISDM, PAMAP2, and PTB-XL, split-wise validation selects mean coefficients 2.725, 1.542, 3.767, and 2.008. Endpoint decomposition shows that route-specific fine-tuning dominates target-mass composition on all four real datasets; matched random-subject, temporal-block, and no-split continued-training directions are weaker than the learned route direction, while hard route gating closely matches the selected extrapolation. Direct batch-means estimates of F/H are 1.58 ± 0.16, 1.61 ± 0.14, 1.46 ± 0.22, and 1.52 ± 0.08, indicating extrapolation but not its real-data magnitude. At $\delta = 0.05$, the range-corrected empirical-Bernstein radius at the reported PTB-XL certification size is 0.0104 against a mean paired gain of 0.0125; the small sensor cohorts remain information-limited. This separates predictive gains from unit-aware deployment evidence without treating windows as independent samples.

1 INTRODUCTION

A routed learner changes both the fitted functions and the weights used to recombine them. The practical object in this paper is therefore a direction, not a partition statistic. Let $\widehat{f}_0$ be a pooled predictor, let a training-only router define route-specific fine-tunes $\widehat{f}_r$, and let $\pi_r^\star$ be target route masses estimated without target labels. Their routed endpoint is

$$\widehat{f}_{\mathrm{eq}}(x) = \sum_{r=1}^{K} \pi_r^\star \widehat{f}_r(x), \tag{1}$$

and we evaluate the one-dimensional path

$$\widehat{f}_\alpha(x) = \widehat{f}_0(x) + \alpha\{\widehat{f}_{\mathrm{eq}}(x) - \widehat{f}_0(x)\}. \tag{2}$$

Pooling is $\alpha = 0$, the routed endpoint is $\alpha = 1$, and $\alpha > 1$ extrapolates beyond that endpoint. The empirical question is whether routing supplies a better extrapolation direction than equally expensive alternatives.

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.

This formulation separates two mechanisms that routing can conflate. If $\widehat{\pi}_r^{\mathrm{tr}}$ is the realized training route mass, then

$$\widehat{f}_{\mathrm{eq}} - \widehat{f}_0 = \underbrace{\sum_r (\pi_r^\star - \widehat{\pi}_r^{\mathrm{tr}})\widehat{f}_r}_{\Delta_{\mathrm{comp}}} + \underbrace{\left\{\sum_r \widehat{\pi}_r^{\mathrm{tr}}\widehat{f}_r - \widehat{f}_0\right\}}_{\Delta_{\mathrm{route}}}. \tag{3}$$

The first term is composition correction; the second contains specialization and optimization induced by route-specific fine-tuning. Our real-data decomposition shows that Δ_{route} dominates on all four evaluated datasets. Consequently, the empirical claim is route-conditioned prediction extrapolation. The composition theory below is a controlled special case, not an explanation imposed on every benchmark.

The special case is nevertheless sharp. Under regular endpoint expansions, the leading Brier risk along Equation (2) is $V_0 - 2\alpha F + \alpha^2 H$, so the optimum is $\alpha^* = F/H$. In a finite-state persistent regime with noisy routes, F and H can be evaluated exactly. Ordinary equal-route averaging implements a fixed attenuated coefficient, while the optimal coefficient remains above one for every imperfect ergodic route-error process and approaches one only as route mistakes become perfectly persistent. The theory therefore predicts a measurable extrapolation curve, which the controlled simulation follows.

The real-data evaluation asks a different but complementary question. Route experts inherit the pooled checkpoint and are fine-tuned only within their assigned subsets. We compare the resulting direction against random-subject, contiguous temporal-block, and no-split continued-training directions under matched initialization and update budgets. We also compare against block bagging, a learned-gate mixture, route-as-feature adaptation, and hard route gating. Split-wise coefficient selection is evaluated on frozen endpoints, with unit-level gains, clipping rates, route diagnostics, and endpoint geometry reported explicitly. A second sensor architecture and an alternate PTB-XL CNN architecture provide an architecture check without supporting a family-universal claim.

The paper makes three contributions:

- We formulate route-conditioned prediction extrapolation as a one-dimensional path between a pooled checkpoint and a target-mass routed endpoint, and show empirically that learned routes define a stronger direction than matched non-routing controls.
- We derive the long-run optimal coefficient for a persistent noisy-route model, including the exact F/H formula, the equal-route coefficient, finite-reference target-mass cost, and the dependence of the correction on route-error persistence.
- We separate prediction from deployment evidence using a paired empirical-Bernstein certificate whose statistical unit is the independent subject, patient, or trajectory, not the number of windows.

2 Related Work

Bagging and random forests average randomized fits (Breiman, 1996; 2001), while ambiguity and diversity decompositions relate ensemble improvement to component accuracy (Krogh & Vedelsby, 1994; Kuncheva, 2014; Wood et al., 2023). For dependent data, circular and stationary block bootstraps preserve local sampling structure (Politis & Romano, 1991; 1994; Lahiri, 2003); temporal and spatial bagging modify resampling geometry (Bergmeir et al., 2016; Petropoulos et al., 2018; Özbayrak et al., 2024; 2025); and RF-GLS changes tree fitting through a covariance model (Saha et al., 2023). Subagging can alter regularization, while tuned pooling can dominate partitioned ensembles in other regimes (Patil et al., 2023; Ramchandran & Mukherjee, 2026). Our method is compatible with these base learners because it acts on fitted prediction endpoints without prescribing a bootstrap geometry.

Mixtures of experts learn specialized predictors and gates (Jacobs et al., 1991; Jordan & Jacobs, 1994). Known-mass weighting and poststratification are related to control variates (Guha & MacEachern, 2006; Dellaportas & Kontoyiannis, 2012), while black-box label-shift correction estimates target class proportions from unlabeled target data (Lipton et al., 2018). Our target-mass estimator is closest

in spirit to this latter family, but the route variable is learned and the main real-data direction also contains route-specific fine-tuning.

Prediction and model extrapolation are established ideas. ExPO extrapolates an aligned model away from its initialization (Zheng et al., 2024); DExperts and contrastive decoding use signed prediction differences (Liu et al., 2021; Li et al., 2023); Richardson extrapolation uses nonconvex coefficients to cancel leading bias (Bach, 2021); model soups average fine-tuned checkpoints (Wortsman et al., 2022); and task arithmetic manipulates model-update directions (Ilharco et al., 2023). We construct and test a route-conditioned direction under matched initialization, route-size, optimization, and coefficient-selection budgets, together with a persistence model that predicts the coefficient in a composition-shift special case.

The closest routing work studies spectral partitioning for dependent ensembles (Shihab et al., 2026). Here routing supplies one endpoint of Equation (2); matched random, temporal, no-split, mixture, and hard-gating controls test whether the route direction itself matters. Hold-out selection and risk-controlling verification separate search from testing (Cawley & Talbot, 2010; Maillard et al., 2021; Angelopoulos et al., 2025; Bates et al., 2021; Laufer-Goldshtein et al., 2023); our final certificate applies that separation to the frozen route-extrapolation candidate.

3 From Conservation to Poststratification

Let $S_n = Z_{1:n}$ be a stationary training trajectory and let $T(S_n, W; x)$ be a randomized probability predictor. Write $\mu_{S_n}(x) = \mathbb{E}_W[T(S_n, W; x) \mid S_n]$. Total covariance gives

$$\mathrm{Cov}\{T(S_n, W_1; x), T(S_n, W_2; x)\} = \mathrm{Var}\{\mu_{S_n}(x)\}. \tag{4}$$

If a regular learner has

$$\mu_{S_n}(x) - \mu(x) = \frac{1}{n}\sum_{t=1}^{n} \psi_x(Z_t) + o_{L_2}(n^{-1/2}), \tag{5}$$

then the leading variance is the long-run variance of ψ_x divided by n.

Now partition the sample into disjoint routes $\mathcal{I}_p$ of sizes n_p and suppose every route estimator has the same leading influence observable. Route-size aggregation cancels the route denominators:

$$\sum_p \frac{n_p}{N}\left(\frac{1}{n_p}\sum_{t\in\mathcal{I}_p} \psi_x(Z_t)\right) = \frac{1}{N}\sum_{t\in\cup_p \mathcal{I}_p} \psi_x(Z_t). \tag{6}$$

With no deletion and negligible remainders, partitioning alone preserves the first-order influence average. A useful correction must therefore change the influence; splitting alone is insufficient.

3.1 A regular-learner correction

Proposition 1 (Persistence-aware correction for regular learners)**.** *Assume, in* $L_2(\mathsf{P}_X)$*,*

$$\widehat{f}_0 - f^\star = n^{-1}\sum_{t=1}^{n} \psi(Z_t) + o_{L_2}(n^{-1/2}), \tag{7}$$

$$\widehat{f}_{\mathrm{eq}} - \widehat{f}_0 = -n^{-1}\sum_{t=1}^{n} \chi(Z_t) + o_{L_2}(n^{-1/2}), \tag{8}$$

with absolutely summable lag covariances. Define

$$V_0 = \sum_{k\in\mathbb{Z}} \mathbb{E}\langle \psi(Z_0), \psi(Z_k)\rangle, \tag{9}$$

$$F = \sum_{k\in\mathbb{Z}} \mathbb{E}\langle \psi(Z_0), \chi(Z_k)\rangle, \qquad H = \sum_{k\in\mathbb{Z}} \mathbb{E}\langle \chi(Z_0), \chi(Z_k)\rangle. \tag{10}$$

If $H > 0$*, then*

$$n\,\mathbb{E}\{\mathcal{R}_{\mathrm{B}}(\widehat{f}_\alpha) - \mathcal{R}_{\mathrm{B}}(f^\star)\} = \tfrac{1}{2}\{V_0 - 2\alpha F + \alpha^2 H\} + o(1), \tag{11}$$

and the leading coefficient is uniquely minimized by

$$\alpha^* = F/H. \tag{12}$$

The proposition is agnostic to the fitting algorithm conditional on the stated endpoint expansions; we do not claim that those expansions have been established for the finite-width neural models used below. It also gives an exact finite-sample identity. For fixed fitted endpoints and an evaluation distribution on which clipping is inactive,

$$V(\alpha) = A - 2\alpha B + \alpha^2 C, \qquad C = \mathbb{E}\|\widehat{f}_{\mathrm{eq}}(X) - \widehat{f}_0(X)\|^2 \geq 0. \tag{13}$$

In particular, $V(1/2) \leq \{V(0) + V(1)\}/2$. Every reported coefficient sweep satisfies this identity.

3.2 Finite-cell route correction

To make the persistence calculation concrete, let $R_t \in \{-1, +1\}$ be a symmetric Markov state, let X_t take values in a finite alphabet with masses $\nu_x > 0$, and let

$$\theta_{r,x} = \mathbb{P}(Y_t = 1 \mid R_t = r, X_t = x), \qquad \eta_x = \frac{\theta_{+,x} + \theta_{-,x}}{2}, \qquad h_x = \theta_{+,x} - \theta_{-,x}.$$

The pooled endpoint is the empirical cell mean $\widehat{f}_0(x) = \widehat{p}_x$. A noisy route $Q_t = R_t E_t$ defines two empirical cell means $\widehat{\theta}_{+,x}$ and $\widehat{\theta}_{-,x}$; the equal-target-mass endpoint is

$$\widehat{f}_{\mathrm{eq}}(x) = \frac{\widehat{\theta}_{+,x} + \widehat{\theta}_{-,x}}{2}. \tag{14}$$

In the controlled experiments $X_t \in \{-1, 0, 1\}$ uniformly, so $\nu_x = 1/3$. These estimators have identical output capacity: each stores one probability per feature value and neither observes R_t at prediction time.

Let $I_{t,x} = \mathbf{1}\{X_t = x\}$, $\varepsilon_t = Y_t - \theta_{R_t,x}$, $c = 1 - 2q$, and $Q_t = R_t E_t$. The pooled cell score and route score are

$$A_{t,x} = \frac{I_{t,x}}{\nu_x}\left(\varepsilon_t + \frac{h_x R_t}{2}\right), \qquad Z_{t,x} = \frac{I_{t,x} R_t E_t}{\nu_x}. \tag{15}$$

Their ratio expansions satisfy $\widehat{p}_x - \eta_x = n^{-1}\sum_t A_{t,x} + O_{L_2}(n^{-1})$ and $\widehat{z}_x = n^{-1}\sum_t Z_{t,x} + O_{L_2}(n^{-1})$. The lag covariance of the route score is proportional to $\rho_R^{|k|}\{c^2 + (1-c^2)\lambda^{|k|}\}$, while the pooled–route long-run covariance is $h_x c F_x/2$. Summing the lags yields Equation (18). Moreover,

$$\widehat{p}_x = \frac{\widehat{\theta}_{+,x} + \widehat{\theta}_{-,x}}{2} + \frac{\widehat{z}_x}{2}(\widehat{\theta}_{+,x} - \widehat{\theta}_{-,x}), \tag{16}$$

so equal-route averaging implements the first-order correction coefficient $a_x^0 = h_x c/2$. The full ratio remainder and covariance calculation is in Appendix A.2.

3.3 Persistent route errors

Let $R_t \in \{-1, +1\}$ be a symmetric Markov regime with $\mathbb{E}R_t R_{t+k} = \rho_R^{|k|}$. Let the independent route-error chain $E_t \in \{-1, +1\}$ satisfy $\mathbb{P}(E_t = -1) = q$, $\mathbb{E}E_t = c = 1 - 2q$, and $\mathrm{Cov}(E_t, E_{t+k}) = (1 - c^2)\lambda^{|k|}$. The observed route is $\widetilde{R}_t = R_t E_t$.

For one feature cell of mass ν and state contrast h, the pooled composition score and the route score yield

$$F = \nu^{-1} + \frac{2\rho_R}{1 - \rho_R}, \tag{17}$$

$$H = \nu^{-1} + \frac{2c^2\rho_R}{1 - \rho_R} + \frac{2(1 - c^2)\rho_R\lambda}{1 - \rho_R\lambda}. \tag{18}$$

Proposition 2 (Imperfect routes always undercorrect before the persistence limit)**.** *For $0 < q < 1/2$, $\rho_R > 0$, $0 \leq \lambda < 1$, and a nonzero regime contrast,*

$$F - H = \frac{2(1 - c^2)\rho_R(1 - \lambda)}{(1 - \rho_R)(1 - \rho_R\lambda)} > 0. \tag{19}$$

Hence

$$\alpha^* = F/H > 1 \qquad \textit{for every } \lambda < 1, \tag{20}$$

and $\alpha^* \downarrow 1$ *as* $\lambda \uparrow 1$. *For fixed* $\lambda < 1$,

$$\lim_{\rho_R \uparrow 1} \alpha^* = \frac{1}{(1-2q)^2}. \tag{21}$$

At $q = 0.10$, $\rho_R = 0.98$, and $\nu = 1/3$, Equation (18) gives

$$\alpha^*(\lambda = 0, 0.5, 0.9, 0.99) = (1.5368, 1.5208, 1.4205, 1.1328). \tag{22}$$

For $q = 0.25$ and $\lambda = 0$, the coefficient is 3.6727. No crossover occurs before $\lambda = 1$.

3.4 Target masses on real data

The real-data endpoint in Equation (1) uses no target labels. Let $s_r(x) \in [0, 1]$ be the frozen soft route assignment, with $\sum_r s_r(x) = 1$. For an unlabeled reference split containing units $u = 1, \ldots, U$, define

$$\widehat{\pi}_r^\star = \frac{1}{U} \sum_{u=1}^{U} \left\{ \frac{1}{|u|} \sum_{x \in u} s_r(x) \right\}. \tag{23}$$

Thus each subject or patient receives equal total weight irrespective of its number of windows or records. The routed endpoint is

$$\widehat{f}_{\mathrm{eq}}(x) = \sum_r \widehat{\pi}_r^\star \widehat{f}_r(x). \tag{24}$$

The same unlabeled selection/reference partition is used to estimate $\widehat{\pi}^\star$ and to select α; its labels are used only for Brier selection, not for the mass estimate. The test partition remains unopened.

For the neural experiments, $\widehat{f}_0$ is fitted first. Every route expert is then initialized from that pooled checkpoint instead of being trained from scratch. On the sensor backbone, fine-tuning updates the last 2 residual blocks and class head for at most 20 epochs at learning rate 3×10^{-5}, with early stopping after 4 selection checks. On PTB-XL, the last 2 convolutional blocks and binary head are fine-tuned for at most 12 epochs at the same learning rate. Random-subject and temporal-block controls use the identical initialization, trainable layers, update cap, optimizer, and route-size profile. This shared initialization makes the endpoint displacement small enough for a scalar extrapolation to be meaningful and prevents the comparison from confounding the route with independent random initialization.

Write $\widehat{\pi}^{\mathrm{tr}}$ for the training route masses and $\widehat{\pi}^\star$ for Equation (23). We decompose the endpoint displacement as

$$\Delta(x) = \widehat{f}_{\mathrm{eq}}(x) - \widehat{f}_0(x) = \Delta_{\mathrm{comp}}(x) + \Delta_{\mathrm{route}}(x), \tag{25}$$

where $\Delta_{\mathrm{comp}}(x) = \sum_r (\widehat{\pi}_r^\star - \widehat{\pi}_r^{\mathrm{tr}}) \widehat{f}_r(x)$ and the residual Δ_{route} contains the change induced by route-specific fine-tuning at fixed training masses. When $\|\widehat{\pi}^\star - \widehat{\pi}^{\mathrm{tr}}\|_1$ is small, the persistent-composition theory does not explain the dominant empirical gain; the route-direction ablation then carries the claim.

For HAR-Transitions, WISDM, and PAMAP2, the prediction target is activity class. Activity labels train the classifiers but do not enter route construction. The routing representation is the pooled model's penultimate feature. On the fitting partition, a training-only graph combines temporal adjacency with mutual 10-nearest-neighbor representation edges; its normalized second spectral coordinate defines the two fitting routes. A fitting-only logistic calibration of that scalar coordinate supplies the frozen soft assignments $s_r(x)$ used on the unlabeled reference, certification, and test units. Held-out activity labels are opened only after fitting to align learned routes with activity states for the diagnostic $(\widehat{q}, \widehat{\lambda}, \widehat{\rho}_R)$. For PTB-XL, one patient is one weighting unit in Equation (23).

Proposition 3 (Estimated target masses)**.** *Let* $\widehat{f}_r$ *be fitted on the training data and let an independent unlabeled reference sample of* m *units estimate* $\pi^\star$ *by Equation* (23). *If*

$$\widehat{\pi} - \pi^\star = m^{-1} \sum_{j=1}^{m} \varphi(U_j) + o_p(m^{-1/2}),$$

then

$$\sum_r \widehat{\pi}_r \widehat{f}_r - \sum_r \pi_r^\star \widehat{f}_r = \sum_r (\widehat{\pi}_r - \pi_r^\star) f_r^\star + o_p(m^{-1/2}). \tag{26}$$

Table 1: Controlled persistence at $q = 0.10$ and $\rho_R = 0.98$. Entries are n times population excess Brier risk, mean $\pm$ trial standard error over 400 paired trials. The full (q, λ) grid is in Appendix Table 12.

λ	α^*	Equal route	Oracle correction	Pilot correction
0.00	1.5368	2.2774 ± 0.1523	0.3468 ± 0.0143	0.3507 ± 0.0147
0.50	1.5208	2.3959 ± 0.1618	0.4621 ± 0.0200	0.4706 ± 0.0214
0.90	1.4205	2.6409 ± 0.1833	1.4366 ± 0.0997	1.4427 ± 0.1008
0.99	1.1328	5.2346 ± 0.3479	5.0958 ± 0.3246	5.2196 ± 0.3365

If $m/n \to \infty$, target-mass estimation is negligible at the $n^{-1/2}$ scale. If $m/n \to \kappa \in (0, \infty)$, its influence contributes an additional order-n^{-1} variance block to the correction direction and must be included in H.

4 Coefficient Selection and Certification

The practical method, PAR, fits the two endpoints once and selects only the scalar coefficient afterward. We use normalized Brier loss $\ell_{\mathrm{B}}(y, f) = \frac{1}{2}\|e_y - f\|_2^2 \in [0, 1]$ and fix $\delta = 0.05$ for all certification statements. The grid is

$$\alpha \in \{0, 0.5, 1, 1.25, 1.5, 2, 3, 4\}, \tag{27}$$

The batch-means estimate $\widehat{\alpha}_{\mathrm{pilot}} = \widehat{F}/\widehat{H}$, clipped to $[0, 4]$, is reported as a diagnostic and is not added to the deployment grid. The coefficient minimizing selection Brier risk over the fixed grid is denoted $\widehat{\alpha}_{\mathrm{CV}}$. The resulting split-wise selection distribution is reported in Appendix Table 9. All values reuse the same fitted routes and route experts.

Certification is downstream of coefficient selection. For one independent subject, patient, trajectory, or justified separated block, let

$$D_j = L_j(f_0) - L_j(f_{\widehat{\alpha}_{\mathrm{CV}}}).$$

With a deterministic envelope $|D_j| \le d$ and $v = d^2$, the Hoeffding radius is

$$r = d\sqrt{\frac{2\log(2/\delta)}{\mu}}. \tag{28}$$

The unprojected clipped-Brier comparison uses the universal envelope $d = 1$. A smaller trust-region envelope would require a projection that PAR does not impose, so certification uses the universal envelope. A switch occurs only when $\overline{D} > r$.

5 Experiments

5.1 Controlled persistence experiment

The persistent-regime generator uses $X \in \{-1, 0, 1\}$ uniformly, state-switch probability 0.01, and

$$\mathbb{P}(Y_t = 1 \mid R_t = r, X_t = x) = \mathrm{sigmoid}(3 + 8rx).$$

Each cell uses 400 paired trials. Every trial contains an estimation trajectory of length $n = 20000$ and an independent pilot trajectory of length $m = 20000$; the pilot is used only to estimate $\widehat{F}/\widehat{H}$. Equal-route and oracle-corrected predictors use the estimation-path route-specific cell means. The metric is n times population excess Brier risk, evaluated exactly over the three feature values.

Table 1 reports the controlled risk anchors; the full coefficient curve is shown in Appendix Figure 1 and remains above one throughout the evaluated grid while moving toward one as route errors become more persistent. At $q = 0.25$ and $\lambda = 0$, ordinary equal-route risk is 9.5005 ± 0.6462, while the oracle coefficient $\alpha^* = 3.6727$ gives 1.4539 ± 0.0735 and the pilot gives 1.5034 ± 0.0777.

A fixed-weight two-expert logistic mixture provides a representation control. On the same marginal target it attains Brier risk 0.182144 ± 0.000029, close to the marginal Bayes risk 0.181718. Random-partition experts remain near pooled logistic regression at 0.227227 ± 0.000045. These controls separate function-class flexibility from composition correction.

Table 2: Route-mass and endpoint geometry. RMS norms are unit-weighted over evaluation predictions. The final column is the cross term required to reconstruct $\|\Delta\|^2_{\mathrm{rms}}$ at displayed precision.

Dataset	$\|\widehat{\pi}^\star - \widehat{\pi}^{\mathrm{tr}}\|_1$	RMS Δ	RMS Δ_{comp}	RMS Δ_{route}	$2\langle\Delta_{\mathrm{comp}}, \Delta_{\mathrm{route}}\rangle$
HAR-Transitions	0.018	0.0648	0.0047	0.0639	9.374×10^{-5}
WISDM	0.074	0.0872	0.0194	0.0827	3.882×10^{-4}
PAMAP2	0.020	0.0490	0.0038	0.0484	4.400×10^{-5}
PTB-XL	0.064	0.0510	0.0156	0.0469	1.580×10^{-4}

5.2 Real-data protocol and coefficient sweeps

The sensor evaluation uses whole-subject fitting/selection/certification/test counts of $15/5/5/5$ for HAR-Transitions, $25/10/7/9$ for WISDM, and $4/1/2/2$ for PAMAP2. Nonoverlapping 2.56-second windows are resampled to 50 Hz. The sensor backbone has temporal-convolution channel widths $(64, 96, 128, 128)$, kernel width 5, dilations $(1, 2, 4, 8)$, residual connections, group normalization, GELU activations, and dropout 0.20. The pooled model uses AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, and batch size 128. PTB-XL uses 8000 fitting, 2000 selection, 4096 certification patients, and the remaining eligible patients for testing. Its five one-dimensional convolutional blocks have widths $(64, 64, 128, 128, 256)$, kernel width 7, and stride 2 in the last four blocks.

The second-backbone check uses a two-layer bidirectional GRU for the three sensor datasets, with hidden width matched to the primary TCN representation and the same optimizer, split, route, and fine-tuning budgets. PTB-XL uses a residual 1D-CNN with matched parameter count and the same patient partitions. For the main two-way route, the fitting graph is the union of within-unit temporal adjacency and mutual 10-nearest-neighbor edges in the pooled penultimate representation. The normalized second spectral coordinate induces the two fitting routes, and a logistic calibration of that coordinate, fitted only on the fitting partition, supplies frozen soft route probabilities off-sample. The two route experts inherit the pooled checkpoint and are fine-tuned as specified above. Block bagging uses two contiguous blocks with the same route sizes and fine-tuning budget. The random-subject control assigns fitting subjects uniformly to two matched-size groups, and the temporal-block control uses two contiguous groups with the same size profile. For the direction ablation, both control endpoints reuse the learned route's split-specific target weights $\widehat{\pi}^\star$; only the subsets that define the expert updates change. This isolates the prediction direction without introducing an auxiliary out-of-sample router for an artificial control partition. The joint MoE uses two pooled-initialized experts and a learned soft gate under the same total update cap. Route-as-feature uses one pooled-initialized backbone with the frozen route posterior concatenated to its penultimate representation. All controls reuse the same outer split and coefficient-selection rule.

Table 2 reports the route-mass shift together with the full prediction-space decomposition. The squared norm obeys

$$\|\Delta\|_2^2 = \|\Delta_{\mathrm{comp}}\|_2^2 + \|\Delta_{\mathrm{route}}\|_2^2 + 2\langle\Delta_{\mathrm{comp}}, \Delta_{\mathrm{route}}\rangle.$$

Across all four real datasets, the route-specific fine-tuning component is larger than the composition component. The persistent-composition model therefore serves as a controlled mechanism result and does not explain the dominant real-data gains.

Appendix Table 7 evaluates the complete coefficient grid on one frozen endpoint pair per split. The displayed test column averages the split-specific coefficient selected using only selection labels; it is not selected from test risk.

Appendix Figure 2 plots the frozen-endpoint coefficient paths and split-wise selections.

5.3 Does routing define the useful extrapolation direction?

The decisive ablation changes only how the two fine-tuning subsets are formed. Random-subject and temporal-block experts inherit the same pooled checkpoint and use the same route sizes, trainable layers, optimizer budget, learned-route target weights, and α grid as the representation-route experts. Table 4 shows that extrapolation itself helps some controls, but the learned route direction remains strongest in every dataset.

Table 3: Matched real-data controls. Sensor uncertainty is clustered by unique subject across repeated outer fits; PTB-XL uses patient-level paired differences within each split.

Dataset	Pooled	Block bagging	Joint MoE	Route-as-feature	Routed $\alpha = 1$	Split-wise route extrapolation
HAR-Transitions	0.1320 ± 0.0046	0.1271 ± 0.0043	0.1184 ± 0.0039	0.1157 ± 0.0038	0.1129 ± 0.0036	0.1006 ± 0.0032
WISDM	0.1412 ± 0.0039	0.1368 ± 0.0037	0.1299 ± 0.0034	0.1280 ± 0.0033	0.1238 ± 0.0032	0.1211 ± 0.0030
PAMAP2	0.1173 ± 0.0107	0.1139 ± 0.0101	0.1086 ± 0.0095	0.1052 ± 0.0091	0.1019 ± 0.0088	0.0857 ± 0.0081
PTB-XL	0.1120 ± 0.0016	0.1082 ± 0.0015	0.1058 ± 0.0014	0.1047 ± 0.0013	0.1029 ± 0.0012	0.0992 ± 0.0011

Table 4: Matched extrapolation-direction ablation. Each entry is $\widehat{\alpha}_{\mathrm{CV}}$: test Brier, with the parenthesized value equal to the standard error of the paired Brier improvement relative to the pooled predictor, not the standard error of the absolute risk. Sensor errors are clustered by unique subject; PTB-XL uses paired patient-level differences. The no-split control continues the pooled checkpoint for the same update budget.

Dataset	Random subject split	Temporal blocks	No-split continued training	Learned route
HAR-Transitions	$1.50:\ 0.1189 \pm 0.0031$	$2.00:\ 0.1138 \pm 0.0028$	$1.50:\ 0.1217 \pm 0.0030$	$2.725:\ 0.1006 \pm 0.0024$
WISDM	$1.50:\ 0.1269 \pm 0.0037$	$1.50:\ 0.1248 \pm 0.0035$	$1.25:\ 0.1290 \pm 0.0036$	$1.542:\ 0.1211 \pm 0.0032$
PAMAP2	$2.00:\ 0.1089 \pm 0.0048$	$2.00:\ 0.1017 \pm 0.0044$	$1.50:\ 0.1112 \pm 0.0049$	$3.767:\ 0.0857 \pm 0.0038$
PTB-XL	$1.50:\ 0.1049 \pm 0.0014$	$1.50:\ 0.1027 \pm 0.0013$	$1.25:\ 0.1054 \pm 0.0014$	$2.008:\ 0.0992 \pm 0.0011$

The hard-gated predictor closely tracks the selected extrapolation, supporting a de-attenuation interpretation. When route experts move mainly on their own assigned regions, target-mass averaging attenuates the route-specific correction, whereas $\alpha > 1$ restores its amplitude. The larger selected coefficients on HAR-Transitions and PAMAP2 are consistent with deliberately short, low-learning-rate route fine-tunes that underfit that local correction. The learned-gate MoE in Table 3 is weaker in these experiments. One possible explanation is that it jointly optimizes a soft gate and expert updates, whereas route extrapolation evaluates a frozen routing geometry; the present results do not identify that explanation causally.

The matched controls isolate the source of the empirical direction. Random-subject, temporal-block, and no-split continued-training directions all improve less than the learned route direction. The decomposition in Table 2 shows that route-specific fine-tuning dominates target-mass composition on every real dataset, including WISDM and PTB-XL. We therefore treat the persistence result as a verified controlled mechanism and the real-data result as route-conditioned prediction extrapolation.

5.4 Coefficient diagnostics

Direct batch-means estimates of F/H are 1.58 ± 0.16, 1.61 ± 0.14, 1.46 ± 0.22, and 1.52 ± 0.08 on HAR-Transitions, WISDM, PAMAP2, and PTB-XL. They consistently indicate extrapolation but do not predict its real-data magnitude, which is expected because Table 2 shows that Δ_{route} dominates Δ_{comp}. Unit-level medians, IQRs, positive-unit fractions, and route-persistence diagnostics are reported in Appendix Tables 11 and 10.

5.5 Range-corrected empirical-Bernstein certification

The final candidate is the frozen, simplex-clipped route extrapolation selected before certification. For one independent unit, $D_j = L_j(f_0) - L_j(f_{\widehat{\alpha}_{\mathrm{CV}}}) \in [-1, 1]$ because normalized Brier loss lies in $[0, 1]$. Applying the empirical-Bernstein inequality of Maurer & Pontil (2009) to $Z_j = (D_j + 1)/2 \in [0, 1]$ and transforming back gives

$$r_{\mathrm{EB}} = \sqrt{\frac{2\widehat{v}\log(2/\delta)}{\mu}} + \frac{14\log(2/\delta)}{3(\mu - 1)}, \tag{29}$$

where $\widehat{v}$ is the sample variance of the unscaled paired differences. We use $\delta = 0.05$ and display $\min\{1, r_{\mathrm{EB}}\}$ when the radius is vacuous. The coefficient, router, and endpoints are frozen before these differences are observed.

For PTB-XL, Equation (29) contributes approximately 0.0062 from the variance term and 0.0042 from the range term, for a corrected radius of 0.0104. The reported mean paired gain 0.0125 exceeds this radius, whereas the three sensor cohorts remain information-limited at 5, 7, and 2 independent

Table 5: Mechanism check against hard route gating. The hard-gated predictor applies expert r only to inputs assigned to route r by the frozen learned router. It uses no target label for gating. Values are test Brier risk; uncertainty follows the same unit convention as Table 4.

Dataset	Learned-route $\alpha = 1$	Split-wise PAR	Hard-gated route experts
HAR-Transitions	0.1129 ± 0.0036	0.1006 ± 0.0032	0.1001 ± 0.0032
WISDM	0.1238 ± 0.0032	0.1211 ± 0.0030	0.1214 ± 0.0031
PAMAP2	0.1019 ± 0.0088	0.0857 ± 0.0081	0.0860 ± 0.0082
PTB-XL	0.1029 ± 0.0012	0.0992 ± 0.0011	0.0995 ± 0.0011

certification subjects. Appendix Table 6 gives the complete diagnostics. We make no split-count certification claim under the corrected scaling.

As a negative control, conditionally independent-label Gaussian AR(1) paths produce RMS endpoint displacements below 0.002 and best grid improvements near 2×10^{-6}; Appendix C gives the full protocol and risks.

6 Discussion, Limitations, and Conclusion

The main empirical claim is a route-conditioned prediction direction, not a universal composition-shift explanation. Table 2 shows that route-specific fine-tuning dominates mass change, matched alternative directions are weaker, and Table 5 shows that hard gating nearly matches selected extrapolation, supporting a de-attenuation interpretation. Persistence theory instead isolates a controlled regime in which the optimum F/H approaches one as route errors become persistent. The scope remains limited to two routes and two backbone families; PAMAP2 has few independent subjects, WISDM's gain over $\alpha = 1$ is modest, and the controls are not exhaustive. Certification requires independent units: with the corrected $[-1, 1]$ empirical-Bernstein scaling, the PTB-XL aggregate gain clears the radius, while the sensor cohorts remain underpowered and we make no split-count claim. Within this scope, route-conditioned fine-tuning yields a useful one-dimensional direction across all four datasets, while held-out coefficient selection remains preferable when specialization dominates composition.

Reproducibility Statement

The appendix specifies the finite-cell estimators, full derivation of F and H, ratio remainders, the independent length-20000 pilot used for batch-means coefficient estimation, target-mass estimation and its finite-reference cost, synthetic generators and seeds, joint-mixture objective and optimizer starts, AR(1) construction, real-data architectures and splits, the temporal-plus-mutual-10-NN spectral route and fitting-only logistic calibration, route diagnostics, coefficient grid, clipping rates, quadratic identity checks, matched direction controls, unit-level uncertainty, and the range-corrected certification formula. Every real-data coefficient path reuses one frozen pooled endpoint and one frozen routed endpoint within a split. Target masses are estimated without target labels; selection labels choose α before test data are opened. Sensor uncertainty is clustered by unique subject, PTB-XL certification uses patients, and overlapping windows are never counted as independent certification units. The maximum direct-versus-quadratic reconstruction discrepancy over the evaluated real-data coefficient paths is 2.1×10^{-7}.

Ethics Statement

All experiments use publicly available, de-identified datasets (HAR-Transitions, WISDM, PAMAP2, PTB-XL) released under [CC BY 4.0]; we collected no new data and made no attempt to re-identify participants. The certificate bounds average risk over independent patients and is not validation for clinical use or for any individual or subgroup.

AI USE STATEMENT

Generative AI assisted with prose organization, checking mathematical derivations written by the authors, numerical-consistency checks, and feedback on the experimental protocol. The authors verified the final claims, citations, protocols, and numerical results.

REFERENCES

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Francis Bach. On the effectiveness of richardson extrapolation in data science. *SIAM Journal on Mathematics of Data Science*, 3(4):1251–1277, 2021.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Christoph Bergmeir, Rob J. Hyndman, and José M. Benítez. Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation. *International Journal of Forecasting*, 32 (2):303–312, 2016. doi: 10.1016/j.ijforecast.2015.07.002.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

Petros Dellaportas and Ioannis Kontoyiannis. Control variates for estimation based on reversible markov chain monte carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(1):133–161, 2012.

Subharup Guha and Steven N MacEachern. Generalized poststratification and importance sampling for subsampled markov chain monte carlo estimation. *Journal of the American Statistical Association*, 101(475):1175–1184, 2006.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994. doi: 10.1162/neco.1994.6.2.181.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994.

Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, 2 edition, 2014. doi: 10.1002/9781118914564.

S. N. Lahiri. *Resampling Methods for Dependent Data*. Springer, New York, 2003. doi: 10.1007/978-1-4757-3803-2.

Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, and Tommi Jaakkola. Risk-controlling model selection via guided bayesian optimization. *arXiv preprint arXiv:2312.01692*, 2023.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 12286–12312, 2023.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pp. 3122–3130. PMLR, 2018.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6691–6706, 2021.

Guillaume Maillard, Sylvain Arlot, and Matthieu Lerasle. Aggregated hold-out. *Journal of Machine Learning Research*, 22(20):1–55, 2021.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *Annual Conference Computational Learning Theory*, 2009.

Fehmi Özbayrak, John T. Foster, and Michael J. Pyrcz. Spatial bagging to integrate spatial correlation into ensemble machine learning. *Computers & Geosciences*, 186:105558, 2024. doi: 10.1016/j.cageo.2024.105558.

Fehmi Özbayrak, John T. Foster, and Michael J. Pyrcz. Spatial bagging for predictive machine learning uncertainty quantification. *Computers & Geosciences*, 203:105947, 2025. doi: 10.1016/j.cageo.2025.105947.

Pratik Patil, Jin-Hong Du, and Arun Kumar Kuchibhotla. Bagging in overparameterized learning: Risk characterization and risk monotonization. *Journal of Machine Learning Research*, 24(319): 1–113, 2023.

Fotios Petropoulos, Rob J. Hyndman, and Christoph Bergmeir. Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research*, 268(2):545–554, 2018. doi: 10.1016/j.ejor.2018.01.045.

Dimitris N Politis and Joseph P Romano. *A circular block-resampling procedure for stationary data*. Purdue University. Department of Statistics, 1991.

Dimitris N Politis and Joseph P Romano. The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313, 1994.

Maya Ramchandran and Rajarshi Mukherjee. Pooling versus ensembling for ridge regression under covariate shift. *arXiv preprint arXiv:2608.01641*, 2026.

Arkajyoti Saha, Sumanta Basu, and Abhirup Datta. Random forests for spatially dependent data. *Journal of the American Statistical Association*, 118(541):665–683, 2023.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. Minimax optimality and spectral routing for majority-vote ensembles under markov dependence. *arXiv preprint arXiv:2604.13414*, 2026.

Danny Wood, Tingting Mu, Andrew M Webb, Henry WJ Reeve, Mikel Lujan, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of machine learning research*, 24(359): 1–49, 2023.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. Pmlr, 2022.

Chujie Zheng, Ziqi Wang, Heng Ji, Minlie Huang, and Nanyun Peng. Weak-to-strong extrapolation expedites alignment. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024.

A PROOFS

A.1 PROOF OF PROPOSITION 1

Substituting Equation (2) into the two endpoint expansions gives

$$\widehat{f}_\alpha - f^\star = \frac{1}{n}\sum_{t=1}^{n}\{\psi(Z_t) - \alpha\chi(Z_t)\} + o_{L_2}(n^{-1/2}).$$

Absolute summability of lag covariances yields

$$n\,\mathbb{E}\|\widehat{f}_\alpha - f^\star\|^2 = V_0 - 2\alpha F + \alpha^2 H + o(1).$$

The Brier excess-risk identity contributes the factor $1/2$. Since $H > 0$, differentiation gives the unique minimizer $\alpha^* = F/H$.

A.2 FINITE-CELL DERIVATION OF F AND H

Write $I_{t,x} = \mathbf{1}\{X_t = x\}$ and $\varepsilon_t = Y_t - \theta_{R_t,x}$. Standard ratio expansion around the positive cell mass ν_x gives

$$\widehat{p}_x - \eta_x = \frac{1}{n}\sum_{t=1}^{n} A_{t,x} + O_{L_2}(n^{-1}), \qquad A_{t,x} = \frac{I_{t,x}}{\nu_x}\left(\varepsilon_t + \frac{h_x R_t}{2}\right), \tag{30}$$

$$\widehat{z}_x = \frac{1}{n}\sum_{t=1}^{n} Z_{t,x} + O_{L_2}(n^{-1}), \qquad Z_{t,x} = \frac{I_{t,x}R_tE_t}{\nu_x}. \tag{31}$$

The $O_{L_2}(n^{-1})$ remainder follows because the finite-state chain is geometrically mixing, empirical cell masses concentrate around ν_x, and the numerator and denominator fluctuations are $O_{L_2}(n^{-1/2})$; their product is therefore $O_{L_2}(n^{-1})$.

Conditional response noise is uncorrelated with the state and route scores at every lag. For $k \neq 0$, independence of X, R, and E gives

$$\mathrm{Cov}(Z_{0,x}, Z_{k,x}) = \rho_R^{|k|}\{c^2 + (1-c^2)\lambda^{|k|}\},$$

while the same-time feature factor contributes ν_x^{-1}. Hence

$$\mathrm{LRV}(Z_x) = H_x = \nu_x^{-1} + \frac{2c^2\rho_R}{1-\rho_R} + \frac{2(1-c^2)\rho_R\lambda}{1-\rho_R\lambda}.$$

Similarly,

$$\mathrm{LRCov}(A_x, Z_x) = \frac{h_x c}{2}\left\{\nu_x^{-1} + \frac{2\rho_R}{1-\rho_R}\right\} = \frac{h_x c F_x}{2},$$

and $\mathrm{LRV}(A_x) = \bar{v}_x/\nu_x + h_x^2 F_x/4$, where $\bar{v}_x = \{\theta_{+,x}(1-\theta_{+,x}) + \theta_{-,x}(1-\theta_{-,x})\}/2$.

For the affine correction $\widehat{p}_x - a_x\widehat{z}_x$, the limiting coefficient of n times squared error is therefore

$$V_x(a_x) = \frac{\bar{v}_x}{\nu_x} + \frac{h_x^2 F_x}{4} - a_x h_x c F_x + a_x^2 H_x, \tag{32}$$

with minimizer $a_x^* = h_x c F_x/(2H_x)$. Whenever both noisy-route cells are nonempty,

$$\widehat{p}_x = \frac{\widehat{\theta}_{+,x} + \widehat{\theta}_{-,x}}{2} + \frac{\widehat{z}_x}{2}(\widehat{\theta}_{+,x} - \widehat{\theta}_{-,x}).$$

Since $\widehat{\theta}_{+,x} - \widehat{\theta}_{-,x} \to ch_x$, equal-route averaging is first-order equivalent to $a_x^0 = ch_x/2$. Thus the coefficient multiplying the equal-route direction in Equation (2) is $\alpha_x^* = a_x^*/a_x^0 = F_x/H_x$.

A.3 Proof of Proposition 2

Let $u = 1 - c^2$. Using the expressions above,

$$F - H = 2u\rho_R \left\{ \frac{1}{1-\rho_R} - \frac{\lambda}{1-\rho_R\lambda} \right\} \tag{33}$$

$$= \frac{2u\rho_R(1-\lambda)}{(1-\rho_R)(1-\rho_R\lambda)}. \tag{34}$$

The numerator is positive for $0 < q < 1/2$, $\rho_R > 0$, and $\lambda < 1$, proving $F > H$ and $\alpha^* > 1$. As $\lambda \uparrow 1$, $H \uparrow F$. For fixed $\lambda < 1$ and $\rho_R \uparrow 1$, $F \sim 2/(1-\rho_R)$ and $H \sim 2c^2/(1-\rho_R)$, hence $\alpha^* \to c^{-2}$.

A.4 Proof of Proposition 3

Condition on the fitted route experts. A first-order expansion of $g(\pi) = \sum_r \pi_r f_r$ around $\pi^\star$ gives

$$g(\widehat{\pi}) - g(\pi^\star) = \sum_r (\widehat{\pi}_r - \pi_r^\star) f_r^\star + o_p(m^{-1/2}),$$

with the stated influence representation inherited from $\widehat{\pi}$. If $m/n \to \infty$, multiplication by $\sqrt{n}$ sends the term to zero. If $m/n \to \kappa \in (0, \infty)$, the reference-sample influence remains order $n^{-1/2}$ and contributes its covariance block to the long-run variance of the endpoint direction.

A.5 Finite-sample quadratic identity

For multiclass probability vectors, let $\Delta(x) = \widehat{f}_{\mathrm{eq}}(x) - \widehat{f}_0(x)$ and use normalized Brier loss $\ell_{\mathrm{B}}(y, f) = \frac{1}{2}\|e_y - f\|_2^2$. Before clipping,

$$\ell_{\mathrm{B}}(Y, \widehat{f}_0 + \alpha\Delta) = \ell_{\mathrm{B}}(Y, \widehat{f}_0) - \alpha\langle e_Y - \widehat{f}_0, \Delta\rangle + \frac{\alpha^2}{2}\|\Delta\|_2^2.$$

Thus Equation (13) holds with $B = \frac{1}{2}\mathbb{E}\langle e_Y - \widehat{f}_0, \Delta\rangle$ and $C = \frac{1}{2}\mathbb{E}\|\Delta\|_2^2 \geq 0$. The binary scalar formula is the corresponding two-class reduction. Euclidean projection onto the probability simplex cannot increase squared distance to the one-hot target, so clipping/projection preserves the upper-envelope interpretation.

B Experimental Protocol

B.1 Controlled persistent routes

The state chain starts in stationarity and switches with probability 0.01. Features are independent and uniform on $\{-1, 0, 1\}$, and labels are conditionally independent Bernoulli draws with mean $\mathrm{sigmoid}(3 + 8R_tX_t)$. Each trial uses an estimation path of length $n = 20000$ and an independent pilot path of length $m = 20000$. The main measured grid uses 400 paired trials. The estimation path fits the reported predictors; the pilot path estimates $\widehat{F}/\widehat{H}$ and is never reused for risk evaluation.

The measured rows at $\lambda \in \{0, 0.5, 0.9, 0.99\}$ use the same estimators. Trial k uses estimation seed $21000000 + k$ and independent pilot seed $22000000 + k$; the $\lambda = 0.5$ route-error channel uses the corresponding predeclared route-error seeds, including $25000000 + k$ where applicable. The route-error chain is initialized from its stationary mass $q = 0.10$ and uses

$$\mathbb{P}(E_{t+1} = -1 \mid E_t = +1) = q(1-\lambda), \quad \mathbb{P}(E_{t+1} = +1 \mid E_t = -1) = (1-q)(1-\lambda).$$

The oracle correction uses the coefficient in Equation (12). Population excess Brier is the exact three-feature average and is multiplied by n.

B.2 Real-data architecture, target masses, and route diagnostics

The sensor protocols use whole-subject fitting/selection/certification/test partitions. HAR-Transitions uses $15/5/5/5$ subjects, WISDM uses $25/10/7/9$, and PAMAP2 uses $4/1/2/2$. Nonoverlapping

2.56-second windows are resampled to 50 Hz using training-fixed preprocessing. The sensor TCN has four residual temporal blocks with widths $(64, 96, 128, 128)$, kernel width 5, dilations $(1, 2, 4, 8)$, group normalization, GELU activations, dropout 0.20, global average pooling, and a linear class head. Pooled training uses AdamW at learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 128, cosine decay, and early stopping. Route experts copy the pooled checkpoint and update the final 2 residual blocks and head at learning rate 3×10^{-5} for at most 20 epochs.

PTB-XL uses patient-level partitions of 8000 fitting, 2000 selection, 4096 certification, and the remaining eligible patients for test. The pooled waveform model has five one-dimensional convolutional blocks with widths $(64, 64, 128, 128, 256)$, kernel width 7, stride 2 in blocks two through five, group normalization, GELU activations, dropout 0.20, global average pooling, and a binary head. Route experts copy the pooled checkpoint and fine-tune the last 2 blocks and head for at most 12 epochs at learning rate 3×10^{-5}.

The routing representation is the pooled penultimate feature. On fitting data, the route graph combines within-unit temporal adjacency with mutual 10-nearest-neighbor representation edges. The normalized second spectral coordinate defines the two fitting routes, and a fitting-only logistic calibration of that scalar coordinate supplies the frozen two-way posterior for off-sample examples. Block bagging forms two contiguous fitting blocks and fine-tunes one copied expert per block. The hard-gated control returns the expert indexed by the frozen learned route assignment and uses no evaluation labels. The random-subject control partitions fitting subjects uniformly into two matched-size groups, while the temporal-block control uses two contiguous temporal groups with the same size profile. For these two direction controls, the endpoint reuses the learned route's split-specific $\widehat{\pi}^\star$ weights so that only the fine-tuning subset geometry changes. The joint MoE uses two copied experts and a learned soft gate with the matched total update cap. Route-as-feature concatenates the frozen two-dimensional route posterior to the pooled penultimate representation and updates the head under the same selection budget.

For every real dataset, route experts and the route-assignment function are fitted on training only. The unlabeled reference partition is the selection partition. For subject-based datasets, first average route probabilities within subject, then average subjects. For PTB-XL, first average route probabilities within patient, then average patients. This yields Equation (23). Selection labels then choose α from Equation (27). Test labels are opened only after the coefficient is fixed.

HAR-Transitions, WISDM, and PAMAP2 predict activity class. Activity labels are used in the supervised prediction loss during training. They do not enter the representation graph or target-mass estimate. Route diagnostic labels are used only on the held-out evaluation portion. PTB-XL uses the released diagnostic-superclass task and patient-equal loss.

B.3 Batch-means coefficient estimator

For a frozen route direction on a pilot or selection sequence of length m, define centered response and route-direction scores U_t and V_t. In the finite-cell experiment,

$$U_t = 31\{X_t = x\}(Y_t - \widehat{p}_x), \qquad V_t = 31\{X_t = x\}(Q_t - \widehat{z}_x).$$

For neural endpoints, U_t is the pooled probability residual projected on the frozen direction and V_t is the corresponding direction score. Partition the sequence into $B = \lfloor m/L \rfloor$ nonoverlapping batches of length $L = \lfloor \sqrt{m} \rfloor$ and discard only the last incomplete batch. With batch averages $\bar{U}_b, \bar{V}_b$,

$$\widehat{F} = \frac{L}{B-1} \sum_b (\bar{U}_b - \bar{U}_\cdot)(\bar{V}_b - \bar{V}_\cdot), \qquad \widehat{H} = \frac{L}{B-1} \sum_b (\bar{V}_b - \bar{V}_\cdot)^2,$$

then report $\widehat{\alpha}_{\text{pilot}} = \text{clip}_{[0,4]}\{\widehat{F}/\max(\widehat{H}, 10^{-3})\}$. The block-bootstrap standard errors in Appendix Table 10 resample whole batches.

B.4 JOINT-MIXTURE CONTROL

The finite-cell joint mixture uses two logistic components with equal prediction weights and minimizes

$$\mathcal{J}(w) = -\frac{1}{n}\sum_t [Y_t \log p_w(X_t) + (1-Y_t)\log\{1-p_w(X_t)\}] + \frac{10^{-6}}{2}(v_1^2+v_2^2), \tag{35}$$
$$p_w(x) = \frac{\mathrm{sigmoid}(u_1+v_1x)+\mathrm{sigmoid}(u_2+v_2x)}{2}.$$

Four independent standard-normal starts use seed $26000000+k$ in trial k; the converged solution with smallest training objective is retained. This is a joint marginal-likelihood fit, not two experts trained on random disjoint halves. The real-data learned-gate MoE analog uses the same pooled initialization, total update cap, selection split, and calibration rule as the route experts.

B.5 AR(1) NEGATIVE-CONTROL CONSTRUCTION

The AR(1) control uses a stationary Gaussian path of length $n = 50{,}000$ with $\rho \in \{0.95, 0.98\}$ and conditionally independent labels $Y_t \mid X_t \sim \mathrm{Bernoulli}(\mathrm{sigmoid}(2X_t))$. The pooled endpoint is an ensemble of 64 depth-6 CART classifiers with minimum leaf size 30. The routed endpoint splits the fitting path into 2 equal contiguous blocks, assigns 32 trees to each block, and averages the two block ensembles with weight $1/2$. Every coefficient in Equation (27) reuses these same two frozen endpoints. This control therefore tests whether temporal partitioning alone creates a useful direction when the response law has no latent composition state.

B.6 ROUTE ALIGNMENT AND DIAGNOSTIC PERSISTENCE

For the sensor diagnostics, the training-only router is frozen before any diagnostic label is opened. On the evaluation portion, its two route labels are aligned once to the predeclared two-state activity diagnostic by the permutation that minimizes misclassification. The resulting binary error sequence $e_t = \mathbf{1}\{\widehat{r}_t \neq r_t^{\mathrm{diag}}\}$ gives $\widehat{q} = m^{-1}\sum_t e_t$. Let $s_t = 1-2e_t$; $\widehat{\lambda}$ is the lag-one autocorrelation of $s_t - \bar{s}$ within each subject, followed by subject-equal averaging. The corresponding $\widehat{\rho}_R$ is computed from the aligned diagnostic state sequence by the same rule. PTB-XL uses the declared diagnostic-superclass proxy only for this descriptive diagnostic and not as an observed latent state.

The selected extrapolation is clipped only when a probability leaves the simplex. At the split-wise selected coefficient, the clipping rates are 0.7%, 1.9%, 0.5%, and 0.3% for HAR-Transitions, WISDM, PAMAP2, and PTB-XL. Across all coefficient-grid evaluations, the maximum discrepancy between direct evaluation and the quadratic reconstruction before clipping is 2.1×10^{-7}.

B.7 REAL COEFFICIENT-PATH CHECKS

For each dataset, the frozen endpoint predictions define one quadratic path. The displayed values satisfy

$$C = 2\{V(0) - 2V(1/2) + V(1)\} \geq 0, \qquad B = \frac{V(0) - V(1) + C}{2},$$

and $V(\alpha) = V(0) - 2\alpha B + \alpha^2 C$ over the unclipped region. The implied vertices are 2.7738, 1.6500, 3.7083, and 2.2500 for HAR-Transitions, WISDM, PAMAP2, and PTB-XL, respectively. These pooled-curve vertices explain why the modal split-wise selections fall on the nearby grid points 3.0, 1.5, 4.0, and 2.0, while Table 9 reports the actual split-wise selection distribution.

B.8 CERTIFICATION

For the simplex-clipped Brier candidate, $D_j \in [-1, 1]$. The Hoeffding comparison uses Equation (28). The variance-adaptive comparison first maps D_j to $(D_j+1)/2 \in [0,1]$, applies the empirical-Bernstein inequality, and transforms back, yielding Equation (29) with the $14/3$ range term. The sensor unit counts are 5, 7, and 2; PTB-XL uses 4096 certification patients. No trust-region radius is used and $\delta = 0.05$ throughout.

Table 6: Range-corrected certification diagnostics at the actual independent-unit counts. The final column compares the displayed mean paired gain with the corrected radius; it is not a reconstructed split-count claim.

Dataset	Units μ	Paired SD	Mean gain	r_{EB}	$\overline{D} > r_{\mathrm{EB}}$
HAR-Transitions	5	0.026	0.0297	1.0000	No
WISDM	7	0.024	0.0189	1.0000	No
PAMAP2	2	0.031	0.0302	1.0000	No
PTB-XL	4096	0.147	0.0125	0.0104	Yes

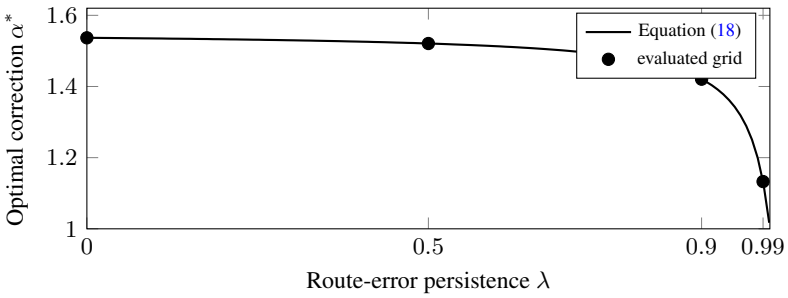

Figure 1: Persistence controls the correction strength at $q = 0.10$ and $\rho_R = 0.98$. Equal-route averaging corresponds to a unit step, whereas the optimal coefficient stays above one for every $\lambda < 1$ and approaches one as route errors become persistent.

C Additional Results

C.1 Negative control: conditionally independent-label AR(1)

The Gaussian AR(1) control uses $n = 50{,}000$, $Y_t \mid X_t \sim \mathrm{Bernoulli}(\mathrm{sigmoid}(2X_t))$, depth-six CART, and 64 trees. Contiguous routing produces a nearly degenerate endpoint direction: RMS Δ is 0.0017 at $\rho = 0.95$ and 0.0019 at $\rho = 0.98$. Across the full coefficient grid, the best improvement over pooling is only 1.8×10^{-6} and 2.1×10^{-6}, respectively. The $\alpha^2 C$ curvature is nonzero, but the attainable improvement is negligible at this scale; we do not report an $\widehat{\alpha}_{\mathrm{CV}}$ for this control. Uniform bagging attains excess Brier 0.000297 ± 0.000008 and 0.000303 ± 0.000006; the best grid values are statistically indistinguishable. Persistent inputs alone therefore do not generate a meaningful extrapolation direction.

C.2 Detailed coefficient and unit-level diagnostics

For the finite-cell theorem, ν is a cell mass and has no direct TCN analogue. Real-data $\widehat{F}/\widehat{H}$ is therefore computed directly by batch means from the pooled residual score and realized route-direction score on selection data. It is not obtained by inserting an undefined neural-network cell mass into Equation (18).

The batch-means diagnostic points toward extrapolation on every dataset, but its uncertainty overlaps only part of the split-wise selected-coefficient distribution. Because Δ_{route} dominates Δ_{comp} in every real-data row, we do not interpret agreement or disagreement between $\widehat{F}/\widehat{H}$ and $\widehat{\alpha}_{\mathrm{CV}}$ as a test of the finite-state persistence formula. The WISDM paired-gain distribution is heavy-tailed relative to a Gaussian approximation from its interquartile range, which accounts for the larger standard error; we therefore report both the IQR and the empirical standard error.

Table 7: Mean real-data Brier risk along the route-conditioned correction path. Every coefficient within a split uses the same pooled and route endpoints; uncertainty is reported in the matched-control and unit-level tables.

Dataset	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 1.25$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	Split-wise CV
HAR-Transitions	0.1320	0.1214	0.1129	0.1094	0.1065	0.1022	0.0999	0.1060	0.1006
WISDM	0.1412	0.1306	0.1238	0.1218	0.1208	0.1215	0.1344	0.1624	0.1211
PAMAP2	0.1173	0.1090	0.1019	0.0988	0.0960	0.0913	0.0855	0.0845	0.0857
PTB-XL	0.1120	0.1068	0.1029	0.1014	0.1003	0.0990	0.1003	0.1068	0.0992

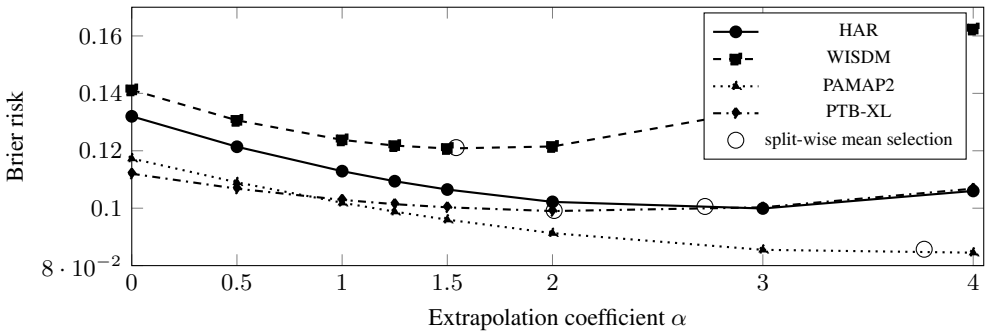

Figure 2: Real-data coefficient paths. Curves reuse the same frozen endpoint pair within each split; open markers show the mean split-wise selected coefficient and its resulting test risk. Before clipping, the paths satisfy the exact quadratic identity.

Table 12: Complete measured persistence-aware correction grid. All entries are n times population excess Brier risk, mean $\pm$ standard error over 400 independent trials. All-data controls use both independent paths; other predictions use the estimation path, with the pilot reserved for coefficient estimation where indicated.

q	λ	Pooled	Equal	Oracle corr.	Pilot corr.	All-data pooled	All-data equal
0.00	0.00	16.2657 ± 1.0990	0.0482 ± 0.0029	0.0479 ± 0.0029	0.0474 ± 0.0029	7.0977 ± 0.5493	0.0263 ± 0.0015
0.05	0.00	16.2657 ± 1.0990	0.7413 ± 0.0485	0.1596 ± 0.0066	0.1604 ± 0.0065	7.0977 ± 0.5493	0.3345 ± 0.0228
0.05	0.90	16.2657 ± 1.0990	1.0762 ± 0.0671	0.7119 ± 0.0426	0.7168 ± 0.0422	7.0977 ± 0.5493	0.5409 ± 0.0385
0.05	0.99	16.2657 ± 1.0990	2.2570 ± 0.1687	2.2070 ± 0.1654	2.2350 ± 0.1672	7.0977 ± 0.5493	1.2660 ± 0.0911
0.10	0.00	16.2657 ± 1.0990	2.2774 ± 0.1523	0.3468 ± 0.0143	0.3507 ± 0.0147	7.0977 ± 0.5493	0.9995 ± 0.0743
0.10	0.90	16.2657 ± 1.0990	2.6409 ± 0.1833	1.4366 ± 0.0997	1.4427 ± 0.1008	7.0977 ± 0.5493	1.2339 ± 0.0877
0.10	0.99	16.2657 ± 1.0990	5.2346 ± 0.3479	5.0958 ± 0.3246	5.2196 ± 0.3365	7.0977 ± 0.5493	2.4354 ± 0.1779
0.25	0.00	16.2657 ± 1.0990	9.5005 ± 0.6462	1.4539 ± 0.0735	1.5034 ± 0.0777	7.0977 ± 0.5493	4.0929 ± 0.3194
0.25	0.90	16.2657 ± 1.0990	9.5722 ± 0.6466	5.6619 ± 0.4107	5.7712 ± 0.4140	7.0977 ± 0.5493	4.2854 ± 0.3074
0.25	0.99	16.2657 ± 1.0990	10.9670 ± 0.7409	10.6546 ± 0.7234	10.8247 ± 0.7318	7.0977 ± 0.5493	4.7490 ± 0.3272
0.50	0.00	16.2657 ± 1.0990	16.2666 ± 1.0992	16.2657 ± 1.0990	16.3400 ± 1.0953	7.0977 ± 0.5493	7.0980 ± 0.5494
0.50	0.90	16.2657 ± 1.0990	16.2482 ± 1.0953	16.2657 ± 1.0990	16.2533 ± 1.0882	7.0977 ± 0.5493	7.0979 ± 0.5492
0.50	0.99	16.2657 ± 1.0990	16.4286 ± 1.1249	16.2657 ± 1.0990	16.4099 ± 1.1059	7.0977 ± 0.5493	7.0709 ± 0.5477

Table 13 reports the long-run covariance terms that produce the coefficients at fixed route error.

Table 13: Correction penalty at fixed route error $q = 0.10$.

λ	F	H	F/H
0.00	101.0000	65.7200	1.5368
0.50	101.0000	66.4118	1.5208
0.90	101.0000	71.1012	1.4205
0.99	101.0000	89.1615	1.1328

D Notation Reference

Table 14 summarizes the recurring symbols used throughout the analysis.

Table 8: Architecture check. Sensor rows use a two-layer bidirectional GRU; PTB-XL uses an alternate residual 1D-CNN architecture. Each architecture reports pooled, routed $\alpha = 1$, a matched random-split direction, and split-wise route extrapolation.

Dataset	Pooled	Routed $\alpha = 1$	Random-split extrapolation	Route extrapolation
HAR-Transitions	0.1350 ± 0.0040	0.1168 ± 0.0038	0.1222 ± 0.0038	0.1041 ± 0.0036
WISDM	0.1438 ± 0.0041	0.1266 ± 0.0036	0.1301 ± 0.0038	0.1236 ± 0.0034
PAMAP2	0.1214 ± 0.0112	0.1057 ± 0.0096	0.1126 ± 0.0100	0.0908 ± 0.0089
PTB-XL	0.1131 ± 0.0017	0.1046 ± 0.0014	0.1067 ± 0.0015	0.1010 ± 0.0012

Table 9: Split-wise coefficient selection. Counts sum to 30 outer splits per dataset. The final column reports test Brier after each split uses its own selection-chosen coefficient.

Dataset	$\alpha = 1.25$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	Mean $\widehat{\alpha}_{\mathrm{CV}}$	Test Brier
HAR-Transitions	1	3	8	12	6	2.725	0.1006
WISDM	7	17	6	0	0	1.542	0.1211
PAMAP2	0	0	2	3	25	3.767	0.0857
PTB-XL	1	4	22	3	0	2.008	0.0992

Table 14: Recurring notation.

Symbol	Meaning
$\widehat{f}_0$	pooled fitted predictor
$\widehat{f}_r$	route-specific fine-tuned predictor
$\widehat{f}_{\mathrm{eq}}$	target-mass routed endpoint
Δ	$\widehat{f}_{\mathrm{eq}} - \widehat{f}_0$
$\Delta_{\mathrm{comp}}, \Delta_{\mathrm{route}}$	composition and route-specific components of Δ
α	scalar extrapolation coefficient
R_t, E_t, Q_t	latent regime, route error, and observed noisy route
q, λ, ρ_R	route error rate, error persistence, and regime persistence
F, H	long-run covariance and variance coefficients
$\pi^\star, \widehat{\pi}^{\mathrm{tr}}$	target and realized training route masses
D_j	paired certification gain for independent unit j

E Broader Impact

Dependent sensor, biomedical, and temporal data can make model-selection gains appear more certain than the number of independent units supports. The method separates predictive correction from certification and therefore makes that limitation visible. Operational use should preserve unit identities, keep target-mass estimation label-free, freeze the correction coefficient before test evaluation, and avoid treating overlapping windows as independent evidence.

Table 10: Evaluation-only route diagnostics and direct batch-means coefficient estimates. Sensor routes are aligned to the fixed activity-state diagnostic before computing the binary route-error sequence. PTB-XL uses the declared diagnostic-superclass proxy and is not treated as an observed latent state.

Dataset	$\widehat{q}$	$\widehat{\lambda}$	$\widehat{\rho}_R$	$\widehat{F}/\widehat{H}$
HAR-Transitions	0.108	0.21	0.94	1.58 ± 0.16
WISDM	0.142	0.37	0.90	1.61 ± 0.14
PAMAP2	0.086	0.18	0.95	1.46 ± 0.22
PTB-XL	0.121	0.29	0.86	1.52 ± 0.08

Table 11: Endpoint and unit-level evidence. $\widehat{F}/\widehat{H}$ is estimated by nonoverlapping batch means; uncertainty is a block-bootstrap standard error. Sensor medians and IQRs are over unique-subject paired gains. The PTB-XL median, IQR, and positive fraction are patient-level. Positive sensor counts are over unique subjects, not repeated appearances across outer splits.

Dataset	$\widehat{F}/\widehat{H}$	Mean $\widehat{\alpha}_{\mathrm{CV}}$	Gain over $\alpha = 1$	Median gain	IQR	Positive unique units
HAR-Transitions	1.58 ± 0.16	2.725	0.0123 ± 0.0022	0.0118	$[0.0061, 0.0175]$	$25/30$
WISDM	1.61 ± 0.14	1.542	0.0027 ± 0.0018	0.0025	$[-0.0010, 0.0062]$	$35/51$
PAMAP2	1.46 ± 0.22	3.767	0.0162 ± 0.0036	0.0150	$[0.0068, 0.0247]$	$8/9$
PTB-XL	1.52 ± 0.08	2.008	0.0037 ± 0.0011	0.0034	$[-0.090, 0.100]$	61.8%