CALIPER: MULTIMODAL DISEASE STAGING AND VELOCITY ESTIMATION FROM CLINICAL SNAPSHOTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal disease progression is often observed through sparse clinical snapshots rather than dense trajectories. Alzheimer's disease provides a canonical case, with amyloid (A), tau (T), and neurodegeneration (N) biomarkers measured irregularly and incompletely. Individuals at the same disease time may exhibit different modality-specific stages and velocities: a modality that is further advanced need not be changing faster. Comparing these changes also requires accounting for state-dependent response sensitivity and residual variation. We introduce CALIPER, a generative framework for modality-specific staging and Fisher-calibrated population-velocity readouts from incomplete multimodal snapshots. An anchored observation model learns a shared disease clock, modality-specific stage offsets, and heteroscedastic emissions. A reverse Schrödinger bridge constructs endpoint-conditioned population currents and regime responsibilities from pseudo-ordered snapshots. Differentiating modality-stage maps along these currents yields model-implied local velocities, calibrated using the mean-response component of Fisher information. In model-matched simulations, CALIPER recovers modality-stage ordering and relative calibrated speeds; matched ablations show higher mean relative-speed correlation and leading-modality accuracy with local Fisher scaling than with fixed or omitted scales. In ADNI and SCAN-linked NACC development data, the clock captures clinical severity, and imaging-and-age regime assignments achieve higher mean anchor-recovery accuracy than a same-anchor logistic reference. Matched ADNI follow-up comparisons show modality-pair-dependent ordering gains over uncalibrated stage speed. CALIPER separates disease position from locally standardized change, enabling comparison of heterogeneous multimodal profiles at the same disease time.

1 INTRODUCTION

Multimodal disease cohorts are often built to study processes that unfold over years, but the data they provide are usually sparse snapshots rather than dense individual trajectories. Each measurement modality may reflect a different biological process, operate on a different temporal scale, and become informative at a different stage of disease. In Alzheimer's disease, for example, amyloid (A), tau (T), and neurodegeneration (N) biomarkers, together with clinical assessments, are used to describe complementary aspects of disease progression (Jack et al., 2010; 2013; Jack Jr et al., 2018). These modalities are rarely observed together at every visit, so an encounter-level comparison depends on both their availability and their local interpretation. The task is therefore more than diagnosis or scalar staging. For a subject observed at a clinical encounter, we would like to infer where they lie on the disease clock, which available modality currently carries the clearest progression signal, and which endpoint reference group should be used to interpret that signal. These anchor-defined groups serve as endpoint regimes.

Learning such an encounter-level readout is difficult because cohort data rarely reveal complete disease-course trajectories. Costly or burdensome measurements such as PET and MRI are often collected only for selected subjects and visits, leaving pre-symptomatic and early symptomatic stages especially sparsely observed. As a result, cohorts provide partially overlapping observations along a latent progression axis rather than complete modality trajectories over the disease course. Existing disease-progression methods address important parts of this setting: event-based models estimate abnormality orderings, disease-progression-score methods align subjects to a continuous severity

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.

coordinate, subtype-and-stage models capture heterogeneous progression patterns, and continuous-time ODE/SDE, flow, diffusion, and Schrödinger-bridge models provide flexible tools for irregular dynamics or distributional transport (Fonteijn et al., 2012; Jedynak et al., 2012; Young et al., 2018; Chen et al., 2018; Rubanova et al., 2019; Song et al., 2020; Lipman et al., 2022; De Bortoli et al., 2021; Shi et al., 2023). These methods provide useful disease coordinates and population dynamics, but a shared coordinate alone cannot answer "which modality is leading now?" Subjects at the same disease time can have different modality stages, and a modality that is further advanced need not be changing faster. Comparing raw stage velocities also overlooks differences in response sensitivity and residual variation. A local comparison therefore requires both a population progression direction and a statistical scale for the modality changes it induces.

We introduce CALIPER, a generative framework that separates disease position from calibrated local change. Available imaging defines a shared latent state; a common clock and modality-specific stage offsets use this state together with age. Clinical assessments define source and endpoint anchors that orient the clock and provide reference regimes. On the frozen representation, a reverse Schrödinger bridge constructs endpoint-conditioned population currents and regime responsibilities. Differentiating modality stages along these currents yields local stage velocities. A heteroscedastic emission model supplies the response sensitivities and residual scales used by mean-response Fisher information to calibrate the induced biomarker changes. The leading modality is the observed modality with the largest calibrated change along the fitted population field. Model-matched simulations assess recovery of modality stages and relative calibrated speeds, while matched interventions separate output calibration from current estimation. In 6,935 participants from SCAN-linked NACC (Beekly et al., 2007; Cody et al., 2025) and ADNI (Weiner et al., 2017), the clock captures clinical severity on the development split, and imaging-and-age regime assignments achieve higher mean anchor-recovery accuracy than a same-anchor logistic reference. Matched ADNI follow-up comparisons further show modality-pair-dependent ordering gains from Fisher calibration over uncalibrated stage speed.

Our contributions are:

- **Staging:** a shared disease clock and modality-specific stages learned from incomplete A/T/N snapshots, without cognitive or diagnostic inputs at inference.
- **Regime-level velocity:** an anchor-conditioned reverse Schrödinger bridge that constructs population currents and induces modality-stage velocities per unit of local bridge progress.
- **Calibrated leading modality:** a mean-response Fisher scale for within-query modality-speed comparison, with higher mean relative-speed and leading-modality recovery than fixed or omitted rulers in matched simulations.
- **Interpretable regimes:** imaging- and age-based assignments to clinically defined endpoint reference groups, with higher mean anchor-recovery accuracy than the same-anchor logistic reference in the AD application.

2 RELATED WORK

Disease progression, pseudotime, and generative transport. Latent disease-time and pseudotime models turn cross-sectional or sparsely longitudinal observations into an ordered progression coordinate. Event-based models estimate abnormality orderings, disease-progression-score and semi-parametric models align subjects to a continuous severity axis, and subtype-and-stage models capture heterogeneous progression patterns (Fonteijn et al., 2012; Venkatraghavan et al., 2019; Jedynak et al., 2012; Lorenzi et al., 2019; Donohue et al., 2014; Young et al., 2018; Eshaghi et al., 2021). Related snapshot-ordering ideas also appear in single-cell trajectory inference (Trapnell et al., 2014; Saelens et al., 2019). Continuous-time latent ODE/SDE models, score-based models, flow matching, and Schrödinger bridges provide flexible parameterizations of irregular dynamics or distributional transport when timestamps, paired samples, or source/target marginals constrain the vector field (Chen et al., 2018; Rubanova et al., 2019; Song & Ermon, 2019; Song et al., 2020; Lipman et al., 2022; Tong et al., 2023; De Bortoli et al., 2021; Shi et al., 2023; Zhang et al., 2024). These methods are natural baselines for learning disease coordinates or population dynamics. However, a local encounter-level readout still requires deciding which observed modality carries the clearest progression-aligned signal at a given state, and on what statistical scale modalities should be compared.

Incomplete multimodal and calibrated representations. Multimodal representation learning studies how to combine heterogeneous inputs, learn shared latent spaces, and remain robust when some views are absent (Ngiam et al., 2011; Baltrušaitis et al., 2018; Wu & Goodman, 2018). These mechanisms address availability and predictive fusion under incomplete views, but do not by themselves define a disease-clock estimand or a local modality-lead readout. Complementary work in geometric progression modeling and deep generative models shows that learned coordinates need not be Euclidean: disease trajectories can be modeled on manifolds, decoder geometry can induce Riemannian metrics, and Fisher information defines the local metric of a statistical model (Schiratti et al., 2017; Arvanitidis et al., 2017; Amari, 2016). Heteroscedastic regression, Bayesian uncertainty modeling, and identifiability work in nonlinear location-scale models further motivate separating signal from state-dependent noise (Kendall & Gal, 2017; Stirn et al., 2023; Hyvarinen et al., 2019; Khemakhem et al., 2020; Immer et al., 2023). These lines of work motivate state-dependent observation geometry, but mainly use it for representation learning, uncertainty calibration, or identifiability, rather than for comparing heterogeneous disease measurements as local progression signals.

The problem studied here combines these tasks: aligning incomplete snapshots, conditioning population movement on endpoint references, and comparing the resulting modality changes on a local statistical scale. CALIPER provides a shared clock, modality stages, endpoint-regime assignments, and a calibrated leading-modality readout within one framework.

3 METHOD

3.1 PROBLEM AND TARGET

A clinical snapshot contains regional measurements $x_{n,m} \in \mathbb{R}^{R_m}$ for an observed subset of modalities $m \in \Omega_n \subseteq \mathcal{M}$, together with optional age covariates c_n. Our target is to compare modality progression at the same disease-time location. This requires separating three quantities: a shared clock aligns subjects, a modality's stage describes its current position, and its velocity describes change along a population progression direction. Subjects at the same clock position can have different modality stages and different local velocities.

CALIPER constructs these quantities in two fitting steps. First, an anchored observation model learns the clock, modality-specific stages, and their local response and residual scales. Second, a reverse Schrödinger bridge (SB) on the frozen representation estimates currents conditioned on clinical endpoint regimes. Applying the observation model's local geometry to these currents gives a common rule for comparing modality speeds. Figure 1 summarizes the pipeline. We omit the subject index below except in training sums (notation: Appendix A).

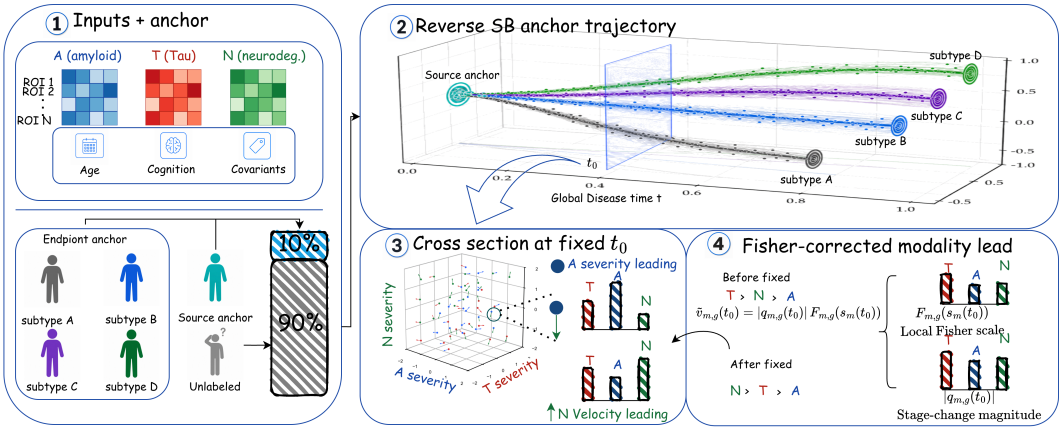

Figure 1: **Overview of CALIPER.** An anchored observation model learns a shared disease clock, modality-specific stages, and local response and residual scales. A reverse Schrödinger bridge on the frozen representation estimates endpoint-conditioned population currents. The induced modality stage velocities are calibrated by local Fisher scales to compare progression across modalities.

3.2 SHARED STAGES AND THE OBSERVATION MODEL

One clock and one stage map per modality. We average the encodings of available modalities to obtain the imaging state z. A sigmoid head produces the shared clock, while bounded modality-specific offsets allow each modality's stage to lie ahead of or behind that clock:

$$\begin{aligned} z &= |\Omega|^{-1} \sum_{m \in \Omega} E_m(x_m), \qquad & t &= \mathrm{sigmoid}(h(z, c)), \\ s_m &= \mathrm{clip}_{[0,1]} \left[t + \delta \tanh \Delta_m(z, c, t) \right]. \end{aligned} \tag{1}$$

Here, $\delta > 0$ is a prespecified bound on modality-specific stage deviations from the shared clock, ensuring $|s_m - t| \leq \delta$.

Missing modalities contribute neither encoder features nor reconstruction terms. The stage maps are shared across endpoint regimes $g \in \mathcal{G}$: changing the reporting regime does not create a different stage axis. Smooth regional response curves relate these stages to the observations, with stage- and regime-dependent residual scales:

$$x_m \mid s_m, g \sim \mathcal{N}(\mu_m(s_m), \Sigma_{m,g}(s_m)), \qquad \Sigma_{m,g}(s) = \mathrm{diag}\big(\sigma_{m,g}(s)^2\big). \tag{2}$$

The mean curve μ_m is also shared across regimes. Thus all regimes retain the same interpretation of modality stage, while their residual variation and population currents may differ.

Learning and orientation. Let $\mathcal{D}_m$ be the training snapshots with modality m observed. Training-time endpoint compatibility weights $\alpha_n(g)$ are obtained from latent prototype and path proximity, and are one-hot for visible anchors. We minimize the observed-modality, regime-weighted conditional loss

$$\begin{aligned} \mathcal{L}_{\mathrm{rec}} &= -\frac{1}{|\mathcal{M}|} \sum_m \frac{1}{|\mathcal{D}_m|} \sum_{n \in \mathcal{D}_m} \sum_g \alpha_n(g) \log p(x_{n,m} \mid s_{n,m}, g), \\ \mathcal{L} &= \mathcal{L}_{\mathrm{rec}} + \eta_a \mathcal{L}_{\mathrm{anchor}} + \eta_t \mathcal{L}_{\mathrm{clock}}. \end{aligned} \tag{3}$$

The anchor term fits Gaussian prototypes for a low-progress reference group (PRE) and the endpoint groups. The clock term combines a PRE clock prior, separation of endpoints from the source, and age-ordered neighbor consistency. Clinical variables define training anchors; query inputs contain imaging and the age covariates. After fitting, we freeze the representation and set the training PRE median to clock zero by a common affine transformation of clock and stage. Subsequent t and s_m denote these reporting coordinates, with response curves transformed consistently. Network specifications, anchor rules, and optimization details are given in Appendix C.

3.3 CALIPER GEOMETRY FOR COMPARABLE MODALITY CHANGES

Local calibration. A stage increment is more informative where a modality's mean response is steep relative to its residual variation. We quantify this using the nonnegative mean-response Fisher ruler and its pull-back to the imaging state:

$$\begin{aligned} F_{m,g}(s)^2 &= \frac{1}{R_m} \mu_m'(s)^\top \Sigma_{m,g}(s)^{-1} \mu_m'(s), \\ J_m(z) &= \frac{\partial \mu_m(s_m(z, c))}{\partial z}, \qquad G_{m,g}(z) = J_m(z)^\top \Sigma_{m,g}(s_m)^{-1} J_m(z). \end{aligned} \tag{4}$$

All derivatives hold observed covariates fixed and include the stage map's dependence on the clock. With diagonal covariance, $F_{m,g}(s)|\mathrm{d}s|$ is the root mean square change in regional means in local residual-standard-deviation units. Equivalently, an imaging-state displacement $\mathrm{d}z$ induces squared mean-response distance $\mathrm{d}z^\top G_{m,g}(z)\mathrm{d}z$. This connects the modality's stage ruler to a geometry in which population movement can be estimated.

We retain mean-response information so that changes in residual dispersion alone do not constitute progression speed. The covariance-derivative term of the full Gaussian Fisher information is omitted. The factor $1/R_m$ sets a per-region comparison convention; covariance assumptions still matter when regions are redundant. The fitted residual variation need not represent measurement error alone. The local statistical interpretation and covariance-aware alternative are in Appendix B.1.

A shared geometry for transport. The same response geometry supplies a fixed metric for bridge fitting. Let D_z contain the training standard deviations of the imaging coordinates on its diagonal and let $\bar{z}$ be their training mean. For N_{tr} training snapshots, we form

$$\Gamma_{\mathrm{raw}} = D_z \left[\frac{1}{N_{\mathrm{tr}}} \sum_{n \in \mathrm{tr}} \sum_{m} \sum_{g} w_g G_{m,g}(z_n) \right] D_z, \qquad y = \Gamma^{1/2} D_z^{-1} (z - \bar{z}), \tag{5}$$

where $w_g = 1/|\mathcal{G}|$ and Γ is a trace-normalized, ridge-regularized version of Γ_{raw}. This single training-average metric is shared across regimes and queries. It makes transport sensitive to changes expressed by the fitted modalities; the final readout retains the local matrix $G_{m,g}(z)$ at the query state.

3.4 Endpoint-conditioned reverse Schrödinger bridging

Each regime has a visible training anchor set $\mathcal{A}_g$ with terminal clock $T_g = \mathrm{median}_{n \in \mathcal{A}_g} t_n$. We draw a source set $\mathcal{S}$ from PRE snapshots with $t \le 0$ and a clock-stratified pool $\mathcal{U}$ of unlabeled, non-PRE snapshots with $t > 0$. For regime g, only pool members satisfying $t \le T_g$ enter its interior support. Sorting these members by clock and dividing them into K windows gives the ladder $X_{g,0} = \mathcal{S}$, $X_{g,k} = W_{g,k}$ for $1 \le k \le K$, and $X_{g,K+1} = \mathcal{A}_g$, at knots $\lambda_k = k/(K+1)$. An early endpoint therefore uses an earlier portion of the population. All terminal anchors are retained with their observed clock spread.

Let R be a reference on regime-labelled node sequences whose transition weights are proportional to $\mathcal{N}(y_j; y_i, \sigma^2 \Delta\lambda_k I)/|X_{g,k+1}|$, where $\Delta\lambda_k = \lambda_{k+1} - \lambda_k$. Writing $M_{g,k}(i)$ for the joint regime/node mass induced by a candidate path law Q, the fitted bridge solves

$$\begin{aligned} Q^* &= \operatorname*{argmin}_{Q} \mathrm{KL}(Q \| R), \\ \text{subject to} \qquad M_{g,0}(i) &= \frac{w_g}{|\mathcal{S}|}, \qquad M_{g,K+1}(a) = \frac{w_g}{|\mathcal{A}_g|}, \\ \sum_{g,k: n \in W_{g,k}} M_{g,k}(n) &= \sum_{g,k: n \in W_{g,k}} \frac{w_g}{|W_{g,k}|}. \end{aligned} \tag{6}$$

The first constraints attach each regime to PRE and its own endpoints. The last constraint couples regimes through each shared pool subject: the optimization determines how that subject's mass is allocated among compatible regime supports. The dual has regime-specific source and terminal potentials and one shared potential per pool subject. Backward messages propagate terminal compatibility toward earlier supports; forward messages impose source consistency. This endpoint conditioning defines *reverse* SB: endpoint compatibility is propagated backward through the ladder, while positive progress remains oriented toward the endpoint.

Normalize the fitted edge masses within regime g to weights $\pi_{g,k,ij}$. On a segment, let $u = (\lambda - \lambda_k)/\Delta\lambda_k$, $a_{ij}(\lambda) = (1-u)y_i + u y_j$, and let $b_k(\lambda) > 0$ be the scalar smoothing variance. The smoothed path density and integrated responsibility are

$$p_g(y \mid \lambda) = \sum_{i,j} \pi_{g,k,ij} \mathcal{N}(y; a_{ij}(\lambda), b_k(\lambda) I)\,, \qquad r_g(y) \;= \frac{w_g \int_0^1 p_g(y \mid \lambda)\,\mathrm{d}\lambda}{\sum_{g'} w_{g'} \int_0^1 p_{g'}(y \mid \lambda)\,\mathrm{d}\lambda}. \tag{7}$$

Reporting combines these imaging-based bridge responsibilities with age evidence as specified below. The weights α in Eq. (3) are used only during representation training.

Let $\omega_{ij}^g(y, \lambda)$ be the posterior mixture weights at the query. The corresponding probability-current velocity is

$$V_g(y, \lambda) = \sum_{i,j} \omega_{ij}^g(y, \lambda) \left[\frac{y_j - y_i}{\Delta\lambda_k} + \frac{b_k'(\lambda)}{2 b_k(\lambda)} \big(y - a_{ij}(\lambda)\big) \right]. \tag{8}$$

The two terms account for the movement of component means and the change of their spread. Summing component fluxes gives $\partial_\lambda p_g + \nabla_y \cdot (p_g V_g) = 0$ within each segment. Thus the current follows from the fitted bridge and its smoothing rule (Appendix B.2).

We locate a query on each ladder by inverting the piecewise-linear map through its mass-weighted source/interior knot clocks $\bar{t}_{g,k} = w_g^{-1} \sum_i M_{g,k}(i) t_i$. Evaluating Eq. (8) at the resulting $\lambda_g(t)$ and

mapping back gives $j_g = D_z \Gamma^{-1/2} V_g(y, \lambda_g(t))$. Query progress is clipped to the source/interior range; the terminal segment contributes to responsibilities but is excluded from velocity readout. The post-fit construction and reporting procedure are summarized in Appendix C.4.

Across support draws, bridge responsibilities are aggregated to $\bar{r}_g(y)$. Reporting uses $r_g^{\mathrm{rep}}(y, a) \propto \bar{r}_g(y) L_g(a)$, where L_g is estimated from training-anchor ages. The AD anchor-recovery analysis uses a minimum SA/NA terminal-clock support cutoff of 0.25. Aggregation, age evidence, and the support adjustment are specified in Appendix C.4.

3.5 Calibrated velocity and the leading modality

The current determines how the shared stage maps change. Pushing it through each modality's stage map and applying the CALIPER ruler yields

$$\begin{aligned} q_{m,g} &= \nabla_z s_m^\top j_g, & v_{m,g} &= |q_{m,g}| F_{m,g}(s_m), \\ v_{m,g}^2 &= \frac{1}{R_m} j_g^\top G_{m,g}(z) j_g = \frac{1}{R_m} \left\| \Sigma_{m,g}(s_m)^{-1/2} J_m(z) j_g \right\|_2^2 . \end{aligned} \tag{9}$$

This identity clarifies the comparison (Appendix B.3): the bridge supplies a population direction, the decoder maps it to regional mean changes, and residual covariance supplies their local scale. The sign of q indicates direction; v measures change magnitude per unit of local bridge progress. These are population-field evaluations at a query's state, so equal disease time does not require equal velocity across subjects.

For the reporting regime and observed modalities, the leading-modality readout is

$$\hat{g} = \operatorname*{argmax}_g r_g^{\mathrm{rep}}(y, a), \qquad m_{\mathrm{lead}} = \operatorname*{argmax}_{m \in \Omega} v_{m,\hat{g}}. \tag{10}$$

We align comparisons by the common disease clock and evaluate each modality at its own stage. For a fixed population field, a differentiable change of local progress with positive derivative rescales all modality speeds at a given state and regime by the same factor. Their ratios and leading-modality readout are consequently invariant (Appendix B.4). Stage describes accumulated position, whereas calibrated speed measures local population-field change in residual-standard-deviation units per unit of bridge progress.

4 Experiments and Results

4.1 Experimental Setup

Tasks. Synthetic experiments evaluate modality-stage ordering, relative calibrated-speed recovery, endpoint assignment, and leading-modality accuracy under known population paths. End-to-end fits estimate representation, current, and calibration from snapshots; controlled experiments with supplied coordinates isolate path geometry and current estimation (Appendices D and G). In the AD application, amyloid PET, tau PET, and a cortical-volume proxy for neurodegeneration provide the A/T/N modalities. Clinical analyses assess the model-defined clock and an auxiliary latent-PC1 readout, endpoint-anchor recovery, and matched longitudinal ordering. Repeated-visit checks and sensitivity analyses are reported in Appendix H.

Datasets. We analyze $6{,}935$ participants from NACC with SCAN-standardized imaging (March 2026 freeze) (Beekly et al., 2007; Cody et al., 2025) and ADNI (Weiner et al., 2017). Amyloid PET, tau PET, and MRI-derived cortical volume are evaluated in shared 31-ROI and native views. Both cohorts contribute to representation training; all visible endpoint anchors come from NACC. Evaluation uses the participant-separated TUNE split, previously inspected during method development. Data preparation, anchor definitions, and evaluation protocols are detailed in Appendices E.1–E.6.

4.2 Synthetic audits of recovery and calibrated readouts

We evaluate CALIPER on synthetic snapshot cohorts under two assumption-matched observation models: local heteroscedasticity and constant noise. Across five replicates per setting, we assess four

targets: modality-stage recovery, relative information-speed recovery, endpoint-regime identification, and leading-modality accuracy. Within each setting, methods use the same training and held-out cohorts, and all fitted readouts use training data only.

Table 1: **Simulation comparison under two noise settings.** Results are reported as mean ± sample SD across five replicates per setting. All metrics are higher-is-better. Bold indicates the best mean within each setting.

	Local heteroscedasticity				Constant noise			
Method	Stage ρ	Rel. vel. ρ	Endpoint BAcc	Leading acc.	Stage ρ	Rel. vel. ρ	Endpoint BAcc	Leading acc.
Majority	–	–	0.250 ± 0.000	–	–	–	0.250 ± 0.000	–
Age only	–	–	0.584 ± 0.033	–	–	–	0.584 ± 0.033	–
Same-anchor logistic	–	–	0.765 ± 0.016	–	–	–	0.772 ± 0.012	–
FA + LR + PROFILE	0.945 ± 0.003	0.654 ± 0.052	$\mathbf{0.867} \pm 0.022$	0.763 ± 0.075	0.946 ± 0.002	0.625 ± 0.055	$\mathbf{0.875} \pm 0.019$	0.740 ± 0.053
Label spreading	–	–	0.660 ± 0.023	–	–	–	0.709 ± 0.018	–
Gradient boosting	–	–	0.842 ± 0.017	–	–	–	0.831 ± 0.012	–
EBM	–	–	0.666 ± 0.032	–	–	–	0.665 ± 0.034	–
DEBM-style	–	–	0.663 ± 0.026	–	–	–	0.670 ± 0.041	–
VAE staging + PROFILE	0.925 ± 0.008	0.177 ± 0.151	0.753 ± 0.052	0.502 ± 0.075	0.931 ± 0.006	0.011 ± 0.084	0.742 ± 0.059	0.468 ± 0.047
Flow matching	0.941 ± 0.001	0.686 ± 0.125	0.764 ± 0.026	0.772 ± 0.066	0.953 ± 0.001	0.802 ± 0.037	0.738 ± 0.026	0.782 ± 0.010
SB-inspired drift	0.942 ± 0.001	0.662 ± 0.068	0.767 ± 0.015	0.768 ± 0.051	0.953 ± 0.002	0.720 ± 0.029	0.748 ± 0.026	0.771 ± 0.026
MoPoE-VAE + PROFILE	0.949 ± 0.002	0.738 ± 0.041	0.802 ± 0.029	0.748 ± 0.049	$\mathbf{0.960} \pm 0.001$	0.788 ± 0.033	0.809 ± 0.024	0.762 ± 0.030
Per-modality burden + PROFILE	0.943 ± 0.002	0.622 ± 0.071	0.753 ± 0.040	0.698 ± 0.068	0.944 ± 0.002	0.631 ± 0.072	0.753 ± 0.042	0.693 ± 0.043
Burden-slope PROFILE	–	0.606 ± 0.065	–	0.681 ± 0.069	–	0.603 ± 0.068	–	0.676 ± 0.049
CALIPER	$\mathbf{0.950} \pm 0.005$	$\mathbf{0.769} \pm 0.031$	0.648 ± 0.058	$\mathbf{0.775} \pm 0.046$	0.955 ± 0.002	$\mathbf{0.808} \pm 0.013$	0.649 ± 0.058	$\mathbf{0.786} \pm 0.044$

Stage and relative velocity report Spearman correlations with true modality stages and within-subject information-speed shares, respectively. Endpoint reports balanced accuracy; leading reports accuracy among evaluable subjects with at least two observed modalities. All readouts are fitted on training data only. Dashes indicate unavailable readouts.

Across both observation models, CALIPER recovers modality-stage ordering ($\rho = 0.950/0.955$) and achieves the highest mean relative-speed correlations among the listed methods ($0.769/0.808$; Table 1). Strong staging alone does not ensure speed-share recovery: under constant noise, VAE staging reaches stage $\rho = 0.931$ but relative-speed $\rho = 0.011$. Likewise, FA+LR achieves endpoint BAcc 0.875 with relative-speed $\rho = 0.625$. These comparisons motivate evaluating disease position, endpoint identity, and relative modality change separately. CALIPER's endpoint BAcc is $0.648/0.649$, compared with $0.765/0.772$ for the task-specific logistic reference.

Matched output-ruler ablations isolate the contribution of calibration. Under local heteroscedasticity, relative-speed correlation is 0.769 with local Fisher, 0.684 with fixed Fisher, and 0.513 with omitted Fisher; corresponding leading accuracies are 0.775, 0.737, and 0.633. Local Fisher has higher relative-speed correlation than both controls in all five paired seeds in each noise setting (Appendix G.1).

Current comparisons reveal dependence on path geometry and the evaluation target. Endpoint secants attain higher mean leading accuracy in the main simulation ($0.825/0.815$ versus $0.775/0.786$ for the full readout). With supplied coordinates and Fisher scales, intermediate bridge constraints improve recovery on curved paths, while the endpoint-linear model performs best on linear paths (Appendix G.2).

4.3 AD APPLICATION: STAGING AND ENDPOINT-CONDITIONED READOUTS

Table 2 evaluates clinical discrimination and severity association using two readouts of the same fitted model. CALIPER (clock) is $t(z, c)$, the imaging-and-age coordinate used for staging and bridge localization. CALIPER (latent PC1) summarizes the imaging state through training-fitted, anchor-oriented PC1. Each readout is fixed across outcomes; together they evaluate the staging coordinate and the clinical information retained in its underlying representation.

On the reused development split, the clock attains CN–dementia AUCs of $0.838/0.865$ in ADNI/NACC, compared with $0.865/0.841$ for MoPoE-VAE. Its clearest advantage is on ADNI CN–MCI discrimination: the clock achieves the highest mean AUC (0.730), compared with 0.649 for the strongest non-CALIPER comparator, gradient boosting. In NACC, latent PC1 achieves the highest mean CN–MCI AUC (0.735), while the clock (0.714) and MoPoE-VAE (0.716) are comparable.

The clinical pattern extends to continuous severity. Clock CDR-SB correlations are $0.465/0.442$ in ADNI/NACC, while latent PC1 achieves the largest means in both cohorts ($0.470/0.460$). ADNI MMSE associations are $-\rho = 0.436$ for the clock and 0.458 for latent PC1, compared with 0.469 for

Table 2: **Clinical discrimination and severity association.** Results on the reused TUNE development split are averaged across common31 and native fits (five seeds per view for seeded models). Entries $x^{+u}_{-\ell}$ report means with 95% CIs $[x-\ell, x+u]$ from 1,000 shared participant-bootstrap draws. Clock is the direct staging readout; latent PC1 is auxiliary. Bold denotes column maxima.

	CN vs. dementia AUC		CN vs. MCI AUC		CDR-SB ρ		MMSE $-\rho$
Method	ADNI	NACC	ADNI	NACC	ADNI	NACC	ADNI
n	275	730	375	827	160	1,005	149
Age-only LR	$0.372^{+0.070}_{-0.066}$	$0.521^{+0.060}_{-0.059}$	$0.496^{+0.059}_{-0.054}$	$0.442^{+0.047}_{-0.045}$	$-0.057^{+0.142}_{-0.148}$	$-0.008^{+0.061}_{-0.061}$	$-0.071^{+0.171}_{-0.170}$
Anchor LR	$0.808^{+0.051}_{-0.049}$	$0.820^{+0.041}_{-0.043}$	$0.642^{+0.053}_{-0.055}$	$0.652^{+0.040}_{-0.042}$	$0.401^{+0.126}_{-0.127}$	$0.355^{+0.049}_{-0.058}$	$0.361^{+0.145}_{-0.147}$
Mean burden	$0.842^{+0.046}_{-0.055}$	$0.778^{+0.048}_{-0.052}$	$0.598^{+0.054}_{-0.062}$	$0.706^{+0.040}_{-0.042}$	$0.443^{+0.117}_{-0.124}$	$0.377^{+0.053}_{-0.060}$	$0.405^{+0.134}_{-0.142}$
PC1	$0.795^{+0.057}_{-0.062}$	$0.760^{+0.053}_{-0.061}$	$0.631^{+0.058}_{-0.059}$	$0.686^{+0.040}_{-0.043}$	$0.459^{+0.116}_{-0.119}$	$0.344^{+0.057}_{-0.062}$	$0.463^{+0.133}_{-0.146}$
Static FA	$0.786^{+0.058}_{-0.060}$	$0.756^{+0.054}_{-0.061}$	$0.633^{+0.058}_{-0.059}$	$0.680^{+0.039}_{-0.045}$	$0.444^{+0.119}_{-0.130}$	$0.337^{+0.056}_{-0.064}$	$\mathbf{0.469}^{+0.129}_{-0.140}$
Label spreading	$0.711^{+0.060}_{-0.063}$	$0.782^{+0.045}_{-0.049}$	$0.575^{+0.051}_{-0.056}$	$0.640^{+0.036}_{-0.042}$	$0.257^{+0.110}_{-0.124}$	$0.319^{+0.056}_{-0.058}$	$0.263^{+0.131}_{-0.144}$
Gradient boosting	$0.771^{+0.053}_{-0.058}$	$0.818^{+0.036}_{-0.037}$	$0.649^{+0.050}_{-0.055}$	$0.642^{+0.037}_{-0.042}$	$0.332^{+0.126}_{-0.132}$	$0.368^{+0.047}_{-0.056}$	$0.303^{+0.177}_{-0.156}$
EBM	$0.707^{+0.051}_{-0.047}$	$0.658^{+0.044}_{-0.046}$	$0.622^{+0.038}_{-0.037}$	$0.600^{+0.027}_{-0.029}$	$0.231^{+0.107}_{-0.120}$	$0.270^{+0.047}_{-0.051}$	$0.216^{+0.123}_{-0.139}$
DEBM-style	$0.745^{+0.056}_{-0.061}$	$0.640^{+0.053}_{-0.055}$	$0.637^{+0.048}_{-0.051}$	$0.609^{+0.037}_{-0.039}$	$0.350^{+0.112}_{-0.119}$	$0.199^{+0.065}_{-0.062}$	$0.360^{+0.117}_{-0.138}$
Multimodal VAE	$0.835^{+0.042}_{-0.048}$	$0.807^{+0.040}_{-0.042}$	$0.608^{+0.050}_{-0.054}$	$0.694^{+0.038}_{-0.037}$	$0.464^{+0.107}_{-0.115}$	$0.388^{+0.048}_{-0.054}$	$0.412^{+0.123}_{-0.130}$
Flow matching	$0.533^{+0.049}_{-0.053}$	$0.565^{+0.040}_{-0.043}$	$0.485^{+0.045}_{-0.045}$	$0.535^{+0.028}_{-0.031}$	$0.160^{+0.089}_{-0.098}$	$0.081^{+0.047}_{-0.049}$	$0.141^{+0.106}_{-0.098}$
SB-inspired drift	$0.580^{+0.051}_{-0.050}$	$0.645^{+0.042}_{-0.041}$	$0.505^{+0.045}_{-0.047}$	$0.582^{+0.028}_{-0.032}$	$0.215^{+0.092}_{-0.095}$	$0.190^{+0.042}_{-0.045}$	$0.208^{+0.101}_{-0.117}$
MoPoE-VAE	$\mathbf{0.865}^{+0.038}_{-0.046}$	$0.841^{+0.043}_{-0.046}$	$0.627^{+0.051}_{-0.059}$	$0.716^{+0.035}_{-0.040}$	$0.465^{+0.102}_{-0.116}$	$0.426^{+0.049}_{-0.053}$	$0.448^{+0.111}_{-0.130}$
CALIPER (clock)	$0.838^{+0.046}_{-0.054}$	$\mathbf{0.865}^{+0.049}_{-0.054}$	$\mathbf{0.730}^{+0.052}_{-0.061}$	$0.714^{+0.038}_{-0.042}$	$0.465^{+0.110}_{-0.114}$	$0.442^{+0.053}_{-0.064}$	$0.436^{+0.127}_{-0.135}$
CALIPER (latent PC1)	$0.856^{+0.047}_{-0.052}$	$0.864^{+0.036}_{-0.042}$	$0.714^{+0.051}_{-0.056}$	$\mathbf{0.735}^{+0.039}_{-0.044}$	$\mathbf{0.470}^{+0.117}_{-0.123}$	$\mathbf{0.460}^{+0.050}_{-0.052}$	$0.458^{+0.124}_{-0.128}$

Static FA. These results connect the learned coordinate to both intermediate clinical status and graded cognitive severity, extending its clinical relevance beyond separation of CN and dementia groups.

With age conditioning and the SA/NA support floor, the bridge attains balanced accuracy $0.748^{+0.028}_{-0.029}$ on 525 development endpoint anchors, compared with $0.677^{+0.036}_{-0.039}$ for same-anchor logistic and $0.656^{+0.023}_{-0.027}$ for age-only classification (Appendix H.2). Reporting combines imaging-based bridge responsibilities with age evidence to assign participants to clinically defined endpoint references. Figure 7 also reports a latent-plus-age logistic comparator.

The matched analysis holds the fitted bridge, reporting responsibilities, participants, and follow-up episodes fixed while changing only the output ruler. Averaged over supported ADNI cells, local Fisher calibration increases ordering agreement over uncalibrated stage speed by $10.06/10.57$ percentage points in common31/native. A–N favors local calibration, whereas T–N favors the standardized-vector readout. Local and fixed Fisher give similar mean ordering agreement. Figure 8 summarizes these descriptive cell-level comparisons (Appendix H.3).

As an external biological consistency check, we examine APOE-$\varepsilon4$ carrier fractions because this allele is a major genetic risk factor for AD (Corder et al., 1993). APOE is excluded from training, anchor definition, and preprocessing. Assigned Y-AD/O-AD groups have higher APOE-$\varepsilon4$ carrier fractions than SA/NA, preserving the qualitative anchor-group gradient (Figure 2).

We further examine regime-specific leading-modality profiles along the learned clock (Figure 3). Y-AD and O-AD share an A-to-T-to-N ordering, but the displayed changes in leading modality occur at different clock locations. Their model-implied profiles therefore combine a common ordering with differences in where A-, T-, and N-leading states appear along the shared coordinate. SA and NA instead exhibit predominantly A- or N-leading profiles over the displayed range. These contrasts illustrate heterogeneity in modality leadership across anchor-defined endpoint regimes. Regional burden atlases provide anatomical context in Appendix I.

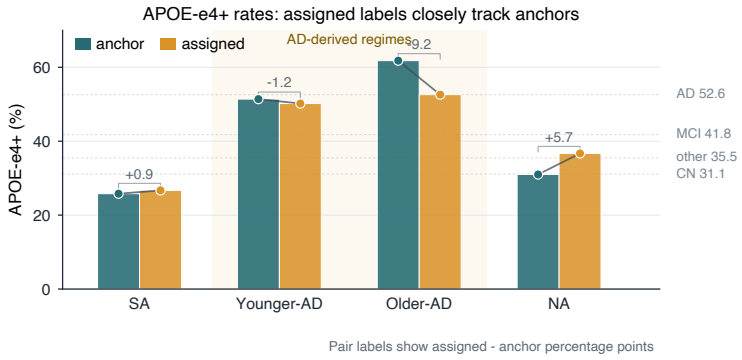

Figure 2: **APOE-$\varepsilon 4$ biological consistency check.** Assigned Y-AD/O-AD groups have higher APOE-$\varepsilon 4$ carrier fractions than SA/NA, preserving the qualitative anchor-group gradient. APOE is excluded from fitting and anchor definition.

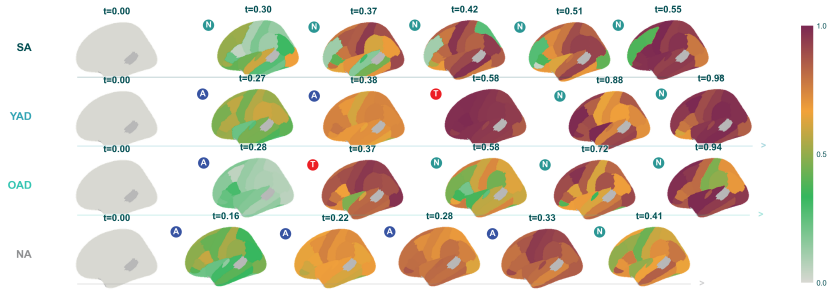

Figure 3: **Endpoint-regime-specific leading-modality profiles.** Rows are anchor-defined endpoint regimes; columns show selected locations along the learned clock. Circle labels identify the modality with the largest model-implied Fisher-calibrated speed.

5 Discussion

CALIPER separates disease position from modality-specific change along an endpoint-conditioned population field. The matched simulation ablations identify local observation geometry as an important component of this comparison: under heteroscedastic noise, local Fisher calibration raises relative-speed correlation from 0.513 with an omitted output ruler to 0.769, with higher correlation in all five paired seeds. The controlled path experiments further show the value of intermediate bridge constraints for curved population paths.

The clinical results connect these readouts to disease severity and endpoint references. The clock achieves the highest mean ADNI CN–MCI AUC (0.730), while imaging-and-age regime assignment reaches mean BAcc 0.748, compared with 0.677 for same-anchor logistic. Matched ADNI follow-up comparisons provide complementary, pair-dependent evidence for calibrated modality ordering. The shared clock, modality stages, population velocities, and endpoint memberships together describe heterogeneous multimodal progression profiles.

Limitations. Population currents depend on anchors, support, and the fitted observation model; velocity units are bridge progress rather than calendar time. Simulations are model-matched. Clinical comparisons reuse a development split, with intervals conditional on fitted models and unadjusted for prior selection. Longitudinal ordering summaries are descriptive, and the separate interval-association analysis was inconclusive (Appendix H).

AI USE STATEMENT

Generative AI tools were used to assist with converting the manuscript's LaTeX source to the ICLR 2027 template, checking typesetting, polishing the writing.

ETHICS STATEMENT

This study uses pre-existing, de-identified, controlled-access ADNI and NACC cohort data under the source cohorts' data-use agreements. Original participant consent and IRB or equivalent approvals were handled by the contributing cohort sites. No new participants were recruited and no interventions or new human-subject data collection were performed for this secondary analysis. The readouts are intended for research on multimodal disease progression; snapshot-derived population coordinates should not be interpreted as validated individual trajectories or deployment-ready clinical decisions. Raw clinical data and identifying subject information are not redistributed.

REPRODUCIBILITY STATEMENT

The appendix provides the theoretical derivations, model components, training objective and algorithm, locked hyperparameters, synthetic simulation settings, data preparation, anchor definitions, evaluation metrics, baseline configuration, computational requirements, and extended results. Proofs appear in Appendix B; model and training details in Appendix C; simulation settings in Appendix D; data and evaluation protocols in Appendix E; and computational requirements in Appendix F.

REFERENCES

Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Duane L Beekly, Erin M Ramos, William W Lee, Woodrow D Deitrich, Mary E Jacka, Joylee Wu, Janene L Hubbard, Thomas D Koepsell, John C Morris, Walter A Kukull, et al. The national alzheimer's coordinating center (nacc) database: the uniform data set. *Alzheimer Disease & Associated Disorders*, 21(3):249–258, 2007.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Karly A. Cody, Andrzej Sokolowski, Lucah Medina Guerra, Philip S. Insel, William J. Jagust, Theresa M. Harrison, and Elizabeth C. Mormino. Comparison of amyloid PET acquired through standardized and unstandardized protocols. *Alzheimer's & Dementia*, 21(9):e70697, 2025. ISSN 1552-5260. doi: 10.1002/alz.70697. URL https://doi.org/10.1002/alz.70697. SCAN: Standardized Centralized Alzheimer's & Related Dementias Neuroimaging, NIH/NIA U24-AG067418.

E H Corder, A M Saunders, W J Strittmatter, D E Schmechel, P C Gaskell, G W Small, A D Roses, J L Haines, and M A Pericak-Vance. Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. *Science*, 261(5123):921–923, 1993. doi: 10.1126/science.8346443.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in neural information processing systems*, 34:17695–17709, 2021.

Michael C Donohue, Hélène Jacqmin-Gadda, Mélanie Le Goff, Ronald G Thomas, Rema Raman, Anthony C Gamst, Laurel A Beckett, Clifford R Jack Jr, Michael W Weiner, Jean-François Dartigues, et al. Estimating long-term multivariate progression from short-term data. *Alzheimer's & Dementia*, 10:S400–S410, 2014.

Arman Eshaghi, Alexandra L Young, Peter A Wijeratne, Ferran Prados, Douglas L Arnold, Sridar Narayanan, Charles RG Guttmann, Frederik Barkhof, Daniel C Alexander, Alan J Thompson, et al. Identifying multiple sclerosis subtypes using unsupervised machine learning and mri data. *Nature communications*, 12(1):2078, 2021.

Hubert M Fonteijn, Marc Modat, Matthew J Clarkson, Josephine Barnes, Manja Lehmann, Nicola Z Hobbs, Rachael I Scahill, Sarah J Tabrizi, Sebastien Ourselin, Nick C Fox, et al. An event-based model for disease progression and its application in familial alzheimer's disease and huntington's disease. *NeuroImage*, 60(3):1880–1889, 2012.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pp. 859–868. PMLR, 2019.

Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In *International Conference on Machine Learning*, pp. 14316–14332. PMLR, 2023.

Clifford R Jack, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

Clifford R Jack, David S Knopman, William J Jagust, Ronald C Petersen, Michael W Weiner, Paul S Aisen, Leslie M Shaw, Prashanthi Vemuri, Heather J Wiste, Stephen D Weigand, et al. Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The lancet neurology*, 12(2):207–216, 2013.

Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M Holtzman, William Jagust, Frank Jessen, Jason Karlawish, et al. Nia-aa research framework: toward a biological definition of alzheimer's disease. *Alzheimer's & dementia*, 14(4):535–562, 2018.

Bruno M Jedynak, Andrew Lang, Bo Liu, Elyse Katz, Yanwei Zhang, Bradley T Wyman, David Raunig, C Pierre Jedynak, Brian Caffo, Jerry L Prince, et al. A computational neurodegenerative disease progression score: method and results with the alzheimer's disease neuroimaging initiative cohort. *Neuroimage*, 63(3):1478–1486, 2012.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217. PMLR, 2020.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Marco Lorenzi, Maurizio Filippone, Giovanni B Frisoni, Daniel C Alexander, Sébastien Ourselin, Alzheimer's Disease Neuroimaging Initiative, et al. Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in alzheimer's disease. *NeuroImage*, 190:56–68, 2019.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In *Icml*, volume 11, pp. 689–696, 2011.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

Jean-Baptiste Schiratti, Stéphanie Allassonnière, Olivier Colliot, and Stanley Durrleman. A bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. *Journal of Machine Learning Research*, 18(133):1–33, 2017.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. *Advances in neural information processing systems*, 36:62183–62223, 2023.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Andrew Stirn, Harm Wessels, Megan Schertzer, Laura Pereira, Neville Sanjana, and David Knowles. Faithful heteroscedastic regression with neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 5593–5613. PMLR, 2023.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, 2014.

Vikram Venkatraghavan, Esther E Bron, Wiro J Niessen, Stefan Klein, Alzheimer's Disease Neuroimaging Initiative, et al. Disease progression timeline estimation for alzheimer's disease using discriminative event based modeling. *NeuroImage*, 186:518–532, 2019.

Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Robert C Green, Danielle Harvey, Clifford R Jack Jr, William Jagust, John C Morris, et al. The alzheimer's disease neuroimaging initiative 3: Continued innovation for clinical trial improvement. *Alzheimer's & Dementia*, 13(5):561–571, 2017.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018.

Alexandra L Young, Razvan V Marinescu, Neil P Oxtoby, Martina Bocchetta, Keir Yong, Nicholas C Firth, David M Cash, David L Thomas, Katrina M Dick, Jorge Cardoso, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nature communications*, 9(1):4273, 2018.

Xi Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis L Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modelling. *Advances in Neural Information Processing Systems*, 37:107198–107224, 2024.

A NOTATION AND DEFINITIONS

Table 3 summarizes the main symbols used throughout the paper. We suppress the subject index n and learned parameter subscripts when no ambiguity arises. Derivatives with respect to the imaging state hold observed covariates fixed and include the stage map's dependence on the shared clock. The diagram notation $\widetilde{v}_{m,g}$ denotes the same calibrated speed as $v_{m,g}$.

Notation / term	Meaning
$\mathcal{M}$, Ω_n	Set of measurement modalities and the subset observed for subject or visit n, respectively.
$x_{n,m} \in \mathbb{R}^{R_m}$	Regional measurement vector for modality m; R_m is the number of regional features.
c_n	Observed age covariates used by the clock and stage heads. Cognition and diagnosis are excluded from query inputs.
E_m	Modality-specific encoder mapping regional measurements into the shared imaging space.
z_n	Imaging state obtained by averaging the available modality encodings: $z_n = \lvert\Omega_n\rvert^{-1} \sum_{m \in \Omega_n} E_m(x_{n,m})$.
t_n	Shared disease clock, oriented toward later progression. Reporting coordinates place the training PRE median at zero.
$s_{n,m}$	Modality-specific stage, shared across endpoint regimes and expressed in the reporting coordinates.
$\Delta_m(z, c, t)$, δ	Modality-specific offset network and prespecified bound in $\delta \tanh \Delta_m$. The bound applies before the reporting-coordinate transformation.
$\mathcal{G}$, g	Set of endpoint regimes and its index. In the AD application, SA, Y-AD, O-AD, and NA are anchor-defined reporting groups.
$\mathcal{S}$, $\mathcal{A}_g$	PRE source support and visible training endpoint-anchor set for regime g, respectively.
$\alpha_n(g)$	Endpoint compatibility weight used during representation training; one-hot for visible endpoint anchors.
T_g	Anchor-median terminal clock of regime g, $T_g = \mathrm{median}_{n \in \mathcal{A}_g} t_n$. The AD anchor-recovery analysis floors the SA/NA support cutoff at 0.25.
λ, $\lambda_g(t)$	Local bridge progress and the query location obtained by inverting the piecewise-linear map through the regime's mass-weighted source/interior knot clocks.
y, D_z, Γ	Transport coordinate $y = \Gamma^{1/2} D_z^{-1}(z - \bar{z})$. Here, $\bar{z}$ is the training mean, D_z contains training coordinate standard deviations, and Γ is the regularized training-average response metric.
$p_g(y \mid \lambda)$	Smoothed bridge density at local progress λ, conditioned on endpoint regime g.
$r_g(y)$	Endpoint-regime responsibility obtained by integrating the bridge density over local progress, weighting by $w_g = 1/\lvert\mathcal{G}\rvert$, and normalizing across regimes.
$r_g^{\mathrm{rep}}(y, a)$, $\hat{g}$	Reporting weights after support-draw aggregation and age conditioning; $\hat{g} = \mathrm{argmax}_g\, r_g^{\mathrm{rep}}(y, a)$.
$V_g(y, \lambda)$	Probability-current velocity in transport coordinates, derived from the fitted bridge and its smoothing rule.
j_g	Current mapped to imaging coordinates: $j_g = D_z \Gamma^{-1/2} V_g(y, \lambda_g(t))$, per unit of local bridge progress.
$\mu_m(s)$, $\mu_m'(s)$	Regional mean-response curve and its derivative with respect to modality stage. Both are shared across endpoint regimes.
$\sigma_{m,g}(s)$, $\Sigma_{m,g}(s)$	Regional residual standard deviations and covariance. The diagonal model uses $\Sigma_{m,g}(s) = \mathrm{diag}(\sigma_{m,g}(s)^2)$.
$F_{m,g}(s)$	Nonnegative local mean-response Fisher scale: $F_{m,g}(s)^2 = R_m^{-1} \mu_m'(s)^\top \Sigma_{m,g}(s)^{-1} \mu_m'(s)$.
$J_m(z)$, $G_{m,g}(z)$	Response Jacobian with respect to the imaging state and its pull-back metric: $G_{m,g}(z) = J_m(z)^\top \Sigma_{m,g}(s_m)^{-1} J_m(z)$.
$q_{m,g}$	Signed stage velocity induced by the population current: $q_{m,g} = \nabla_z s_m^\top j_g$. The gradient holds covariates fixed and includes the stage map's dependence on the clock.
$v_{m,g}$	Fisher-calibrated speed magnitude: $v_{m,g} = \lvert q_{m,g} \rvert F_{m,g}(s_m)$, per unit of local bridge progress.
Leading modality	Observed modality maximizing calibrated speed in the reporting regime: $m_{\mathrm{lead}} = \mathrm{argmax}_{m \in \Omega}\, v_{m,\hat{g}}$.

Table 3: **Notation and definitions.** The disease clock aligns observations, modality stages describe position, and calibrated speeds describe local population movement. Subject indices are suppressed where no ambiguity arises. The diagram notation $\widetilde{v}_{m,g}$ denotes the same calibrated speed as $v_{m,g}$. The speed readout is not an individual's calendar-year rate.

B PROOFS AND THEORETICAL QUALIFICATIONS

This section establishes the local statistical interpretation of the Fisher ruler, the continuity identity for the bridge current, and the coordinate invariance of the modality-lead readout. All statements condition on the fitted representation and bridge. Observed covariates are held fixed when differentiating with respect to the imaging state.

B.1 LOCAL STATISTICAL MEANING OF THE FISHER RULER

Fix a modality m, an endpoint regime g, and a stage s. Assume that μ_m is differentiable and that $\Sigma_{m,g}(s)$ is positive definite. The ruler in Eq. (4) is

$$F_{m,g}(s)^2 = \frac{1}{R_m} \mu'_m(s)^\top \Sigma_{m,g}(s)^{-1} \mu'_m(s). \tag{11}$$

Local mean-response distance. Hold the covariance fixed at its value at s, and define

$$P_\epsilon = \mathcal{N}(\mu_m(s+\epsilon), \Sigma_{m,g}(s)), \qquad P_0 = \mathcal{N}(\mu_m(s), \Sigma_{m,g}(s)).$$

For a small stage displacement ϵ,

$$\frac{2}{R_m} \mathrm{KL}(P_\epsilon \| P_0) = \epsilon^2 F_{m,g}(s)^2 + o(\epsilon^2). \tag{12}$$

Proof. For Gaussian distributions with the same covariance,

$$2\,\mathrm{KL}(P_\epsilon \| P_0) = \Delta\mu^\top \Sigma_{m,g}(s)^{-1} \Delta\mu,$$

where $\Delta\mu = \mu_m(s+\epsilon) - \mu_m(s)$. Differentiability gives $\Delta\mu = \epsilon\mu'_m(s) + o(\epsilon)$. Substitution and division by R_m establish Eq. (12). □

For diagonal covariance, this ruler has the explicit form

$$F_{m,g}(s)^2 = \frac{1}{R_m} \sum_{r=1}^{R_m} \left(\frac{\mu'_{m,r}(s)}{\sigma_{m,g,r}(s)} \right)^2. \tag{13}$$

Thus $F_{m,g}(s)|\epsilon|$ is the leading-order root mean square change in regional means, expressed in local residual standard-deviation units.

Relation to full Gaussian Fisher information. If the covariance is differentiable, the full Fisher information for the stage parameter is

$$\mathcal{I}^{\mathrm{full}}_{m,g}(s) = R_m F_{m,g}(s)^2 + \frac{1}{2} \operatorname{tr}\left[\left(\Sigma_{m,g}(s)^{-1} \Sigma'_{m,g}(s) \right)^2 \right]. \tag{14}$$

Proof. Suppress the modality and regime indices and write $e = x - \mu(s)$. Differentiating the Gaussian log-density gives

$$U_s = \mu'^\top \Sigma^{-1} e + \frac{1}{2} \left[e^\top \Sigma^{-1} \Sigma' \Sigma^{-1} e - \operatorname{tr}(\Sigma^{-1}\Sigma') \right]. \tag{15}$$

The linear and centered quadratic terms are uncorrelated because centered Gaussian third moments vanish. Their variances are $\mu'^\top \Sigma^{-1} \mu'$ and $\frac{1}{2} \operatorname{tr}[(\Sigma^{-1}\Sigma')^2]$, respectively. Taking $\mathbb{E}[U_s^2]$ gives Eq. (14). □

CALIPER uses the mean-response component so that changes in residual dispersion alone do not constitute progression speed. Accordingly, Eq. (12) concerns a local experiment with covariance held fixed. The reciprocal mean-response information is not a general Cramér–Rao bound for the full heteroscedastic model. The residual covariance describes fitted conditional variation and need not represent measurement error alone.

Feature units and covariance assumptions. Under an invertible affine feature transformation $x^\star = Ax + b$, consistent transformation of the mean and full covariance gives

$$(A\mu')^\top (A\Sigma A^\top)^{-1}(A\mu') = \mu'^\top \Sigma^{-1}\mu'. \tag{16}$$

Thus the full-covariance ruler is invariant to invertible affine changes of feature coordinates. The diagonal approximation retains invariance to component-wise rescaling, but not arbitrary mixing of correlated regions. A covariance-aware alternative replaces the diagonal covariance by

$$\Sigma_{m,g}(s) = D_{m,g}(s) C_m D_{m,g}(s), \qquad D_{m,g}(s) = \mathrm{diag}(\sigma_{m,g}(s)),$$

where C_m is a positive-definite, regularized correlation estimate from training residuals.

The factor $1/R_m$ specifies a per-region comparison convention; it does not eliminate arbitrary redundancy among regions. These statements establish a local geometric interpretation. Improvements over fixed or Euclidean rulers remain empirical questions evaluated by the corresponding ablations.

B.2 Probability current of the smoothed bridge

Consider the interior of a bridge segment $[\lambda_k, \lambda_{k+1}]$, with $\Delta\lambda_k = \lambda_{k+1} - \lambda_k > 0$. Let the conditional edge weights $\pi_{g,k,ij}$ be fixed, nonnegative, and sum to one. Define

$$\begin{aligned} u &= \frac{\lambda - \lambda_k}{\Delta\lambda_k}, \\ a_{ij}(\lambda) &= (1-u)y_i + u y_j, \\ \varphi_{ij}(y,\lambda) &= \mathcal{N}\big(y; a_{ij}(\lambda), b_k(\lambda) I\big), \\ p_g(y \mid \lambda) &= \sum_{i,j} \pi_{g,k,ij}\varphi_{ij}(y,\lambda), \end{aligned} \tag{17}$$

where $b_k(\lambda) > 0$ is differentiable within the segment. The posterior component weights are

$$\omega_{ij}^g(y,\lambda) = \frac{\pi_{g,k,ij}\varphi_{ij}(y,\lambda)}{p_g(y \mid \lambda)}. \tag{18}$$

The current velocity in Eq. (8) satisfies

$$\partial_\lambda p_g + \nabla_y \cdot (p_g V_g) = 0 \tag{19}$$

within each segment.

Proof. For an individual Gaussian component, define

$$U_{ij}(y,\lambda) = a'_{ij}(\lambda) + \frac{b'_k(\lambda)}{2b_k(\lambda)}\big(y - a_{ij}(\lambda)\big). \tag{20}$$

Direct differentiation of the Gaussian density gives

$$\partial_\lambda \varphi_{ij} + \nabla_y \cdot (\varphi_{ij} U_{ij}) = 0.$$

Since $a'_{ij}(\lambda) = (y_j - y_i)/\Delta\lambda_k$, the field in Eq. (8) obeys

$$p_g V_g = \sum_{i,j} \pi_{g,k,ij}\varphi_{ij} U_{ij}.$$

Multiplying each component continuity equation by its fixed edge weight and summing proves Eq. (19). □

This construction derives the current from the fitted bridge density and its smoothing rule. It does not require a separately learned drift or denoising-score network. The continuity identity concerns the analytic field within a segment; numerical approximations must be distinguished from this exact identity.

B.3 PUSH-FORWARD TO MODALITY SPEED

At a query's local progress $\lambda_g(t)$, the current in imaging coordinates is

$$j_g = D_z \Gamma^{-1/2} V_g(y, \lambda_g(t)). \tag{21}$$

The signed modality-stage velocity is

$$q_{m,g} = \nabla_z s_m^\top j_g. \tag{22}$$

The gradient here is the total derivative of the composite stage map. Writing $s_m(z, c) = S_m(z, c, t(z, c))$, the chain rule gives

$$\nabla_z s_m = \partial_z S_m|_{c,t} + \partial_t S_m \, \nabla_z t. \tag{23}$$

Thus the clock contribution is determined by $\nabla_z t^\top j_g$; it is not fixed to one. These derivative statements apply wherever the clipped stage map is differentiable.

Because the response mean depends on z through the scalar stage,

$$J_m(z) = \mu'_m(s_m) \nabla_z s_m^\top. \tag{24}$$

Consequently, the calibrated readout satisfies

$$\begin{aligned} v_{m,g}^2 &= q_{m,g}^2 F_{m,g}(s_m)^2 \\ &= \frac{1}{R_m} j_g^\top G_{m,g}(z) j_g \\ &= \frac{1}{R_m} \left\| \Sigma_{m,g}(s_m)^{-1/2} J_m(z) j_g \right\|_2^2. \end{aligned} \tag{25}$$

This follows by substituting $G_{m,g} = J_m^\top \Sigma_{m,g}^{-1} J_m$ and Eq. (22).

The sign of $q_{m,g}$ describes the direction of stage change. The nonnegative quantity $v_{m,g} = |q_{m,g}| F_{m,g}(s_m)$ describes the magnitude of induced regional mean change in local residual units. A modality's current stage and its calibrated speed therefore measure distinct aspects of the fitted progression profile.

B.4 COORDINATE INVARIANCE AND THE REPORTING TARGET

Changing the local progress coordinate. Fix the fitted population field, a query state, and a regime g. For a differentiable reparameterization $\lambda^\star = \psi(\lambda)$ with $\psi'(\lambda) > 0$, the chain rule gives

$$j_g^\star = \frac{j_g}{\psi'(\lambda)}, \qquad q_{m,g}^\star = \frac{q_{m,g}}{\psi'(\lambda)}, \qquad v_{m,g}^\star = \frac{v_{m,g}}{\psi'(\lambda)}. \tag{26}$$

The positive factor is common to all modalities at the same state and regime. Therefore, speed ratios with nonzero denominators, normalized shares with positive total speed, and modality rankings remain unchanged. In particular,

$$\operatorname*{argmax}_{m \in \Omega} v_{m,g}^\star = \operatorname*{argmax}_{m \in \Omega} v_{m,g},$$

with ties preserved.

Changing the modality-stage coordinate. Let $s_m^\star = f_m(s_m)$ be an invertible differentiable reparameterization with $f'_m \neq 0$. Transform the response and covariance by composition:

$$\mu_m^\star = \mu_m \circ f_m^{-1}, \qquad \Sigma_{m,g}^\star = \Sigma_{m,g} \circ f_m^{-1}.$$

Then

$$q_{m,g}^\star = f'_m(s_m) q_{m,g}, \qquad F_{m,g}^\star(s_m^\star) = \frac{F_{m,g}(s_m)}{|f'_m(s_m)|}. \tag{27}$$

The factors cancel, yielding

$$|q_{m,g}^\star| F_{m,g}^\star(s_m^\star) = |q_{m,g}| F_{m,g}(s_m). \tag{28}$$

This includes the common affine PRE-origin transformation of stage and response coordinates used for reporting.

Scope of the invariance statements. These identities concern a fixed fitted field and consistently transformed readout. They do not imply that refitting after a coordinate change produces the same estimator, or that recomputing responsibilities under a different progress prior preserves regime assignments.

The main readout compares modalities within a query state and reporting regime. Its speed unit is local bridge progress, while the shared disease clock locates the query on each regime's ladder. Absolute speeds across regimes require a common progress scale. The readout does not identify an individual's calendar-year rate. Associations with observed follow-up changes provide separate empirical evidence and are not guaranteed by these identities.

C Model and Training Details

C.1 Model components

Inputs and representation. Regional measurements are standardized using training-set statistics and oriented so that larger values indicate greater pathology. For a subject with at least one observed modality, the imaging state is

$$z_n = \frac{1}{|\Omega_n|} \sum_{m \in \Omega_n} E_m(x_{n,m}). \tag{29}$$

Each encoder is a two-layer network with dimensions $R_m \to 64 \to 16$ and a SiLU activation between the linear layers. Missing modalities contribute neither encoder features nor reconstruction terms.

The imaging state z_n is kept distinct from the age covariate block c_n. The clock and stage heads receive both. Age is standardized using observed training ages, with missing age represented by zero after standardization. Clinical information defines the visible training anchors; cognitive scores and diagnosis are excluded from query inputs.

Shared clock and modality stages. To distinguish network outputs from the reporting coordinates, write the untransformed clock and stages as t_n^{raw} and $s_{n,m}^{\mathrm{raw}}$:

$$\begin{aligned} t_n^{\mathrm{raw}} &= \mathrm{sigmoid}(h(z_n, c_n)), \\ s_{n,m}^{\mathrm{raw}} &= \mathrm{clip}_{[0,1]}\left[t_n^{\mathrm{raw}} + \delta \tanh \Delta_m(z_n, c_n, t_n^{\mathrm{raw}})\right]. \end{aligned} \tag{30}$$

The clock head has two hidden layers of width 64. Each offset head has one hidden layer of width 32. The prespecified bound δ limits the raw stage deviation: $|s_{n,m}^{\mathrm{raw}} - t_n^{\mathrm{raw}}| \le \delta$. The stage maps are shared across endpoint regimes.

Regional response curves. Each modality has a vector-valued Bernstein mean curve,

$$\begin{aligned} \mu_m^{\mathrm{raw}}(u) &= \sum_{k=0}^{K_\mu} B_{k,K_\mu}(u)\theta_{m,k}, \qquad \theta_{m,k} \in \mathbb{R}^{R_m}, \\ B_{k,K_\mu}(u) &= \binom{K_\mu}{k} u^k (1-u)^{K_\mu - k}, \qquad K_\mu = 6. \end{aligned} \tag{31}$$

The regional coefficient vectors are learned freely; the mean is not restricted to a scalar trajectory times a fixed regional direction. The initial coefficients interpolate between the observed training PRE mean and the pooled training dementia-anchor mean for that modality. This initialization does not impose an A-to-T-to-N ordering or constrain the fitted curves to remain linear. The mean curve is shared across regimes.

Residual scales. The conditional observation model is

$$x_{n,m} \mid s_{n,m}^{\mathrm{raw}}, g \sim \mathcal{N}\big(\mu_m^{\mathrm{raw}}(s_{n,m}^{\mathrm{raw}}), \Sigma_{m,g}^{\mathrm{raw}}(s_{n,m}^{\mathrm{raw}})\big), \tag{32}$$

with

$$\begin{aligned} \Sigma_{m,g}^{\mathrm{raw}}(u) &= \mathrm{diag}\left(\sigma_{m,g}^{\mathrm{raw}}(u)^2\right), \\ \log \sigma_{m,g}^{\mathrm{raw}}(u) &= \beta_m + \eta_{m,g}(u), \\ \sum_{g \in \mathcal{G}} w_g \eta_{m,g}(u) &= 0, \qquad w_g = |\mathcal{G}|^{-1}. \end{aligned} \tag{33}$$

Here $\beta_m \in \mathbb{R}^{R_m}$ is a regional baseline log-scale. A network with one hidden layer of width 32 maps stage to regime-specific regional deviations. Subtracting their weighted mean enforces the centering constraint. This separates the baseline scale from the regime deviations. The covariance represents fitted conditional residual variation.

C.2 Training compatibility and objective

Training-time endpoint compatibility. The weights $\alpha_n(g)$ are used during representation fitting. For their construction, let $\zeta_n = (c_n, z_n)$ be the covariate-augmented state, and let p_0 and p_g denote the Gaussian source and endpoint prototype means in this space. These prototypes are distinct from the empirical support sets used by the subsequent bridge.

Let $\ell_g(\zeta_n)$ be the Euclidean projection of ζ_n onto the segment joining p_0 and p_g. With $L = \dim(\zeta_n)$ and endpoint standard deviations $d_{g,j}$ floored at 0.35, define

$$\begin{aligned} A_{n,g} &= -\frac{1}{2L}\sum_{j=1}^{L}\left[\frac{(\zeta_{n,j}-p_{g,j})^2}{d_{g,j}^2} + 2\log d_{g,j}\right], \\ B_{n,g} &= -\frac{1}{2L}\sum_{j=1}^{L}\frac{(\zeta_{n,j}-\ell_g(\zeta_n)_j)^2}{d_{g,j}^2}, \\ \alpha_n(g) &= \mathrm{softmax}_g\left[\frac{\log w_g + 0.25A_{n,g} + B_{n,g}}{2}\right]. \end{aligned} \tag{34}$$

Visible endpoint anchors instead receive one-hot weights. These compatibility weights are auxiliary quantities for representation learning. Bridge responsibilities $r_g(y)$ and the final reporting weights $r_g^{\mathrm{rep}}(y,a)$ are constructed after fitting as described in Appendix C.4.

Reconstruction loss. Let $\mathcal{D}_m = \{n : m \in \Omega_n\}$ be the training snapshots with modality m observed. The reconstruction objective is the regime-weighted conditional Gaussian negative log-likelihood:

$$\begin{aligned} \mathcal{L}_{\mathrm{rec}} &= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\frac{1}{|\mathcal{D}_m|}\sum_{n\in\mathcal{D}_m}\sum_{g\in\mathcal{G}}\alpha_n(g)\ell_{n,m,g}, \\ \ell_{n,m,g} &= \frac{1}{2}\sum_{r=1}^{R_m}\left[\frac{\left(x_{n,m,r}-\mu_{m,r}^{\mathrm{raw}}(s_{n,m}^{\mathrm{raw}})\right)^2}{\sigma_{m,g,r}^{\mathrm{raw}}(s_{n,m}^{\mathrm{raw}})^2} + 2\log\sigma_{m,g,r}^{\mathrm{raw}}(s_{n,m}^{\mathrm{raw}})\right]. \end{aligned} \tag{35}$$

Constants independent of the parameters are omitted. Modalities with no observed training targets are excluded from the outer average. The numerical implementation adds 10^{-8} to the variance denominator.

This objective averages conditional negative log-likelihoods under $\alpha_n(g)$. It is distinct from the negative logarithm of a Gaussian-mixture likelihood, $-\log\sum_g \alpha_n(g)p(x_{n,m} \mid s_{n,m}, g)$.

Anchor and clock regularization. The anchor loss is the sum of the average Gaussian negative log-likelihood of PRE states under the source prototype and that of visible endpoint states under their labeled regime prototypes.

Let $\mathcal{P}$ be the training PRE set and $\mathcal{A} = \bigcup_g \mathcal{A}_g$ the visible endpoint set. The clock-orientation term is

$$\begin{aligned} \mathcal{L}_{\mathrm{prior}} = &\frac{1}{|\mathcal{P}|}\sum_{n\in\mathcal{P}}(t_n^{\mathrm{raw}})^2 \\ &+ \frac{1}{|\mathcal{A}|}\sum_{n\in\mathcal{A}}\left[\gamma - (t_n^{\mathrm{raw}} - \bar{t}_{\mathcal{P}}^{\mathrm{raw}})\right]_+^2, \qquad \gamma = 0.1, \end{aligned} \tag{36}$$

where $\bar{t}_{\mathcal{P}}^{\mathrm{raw}}$ is the mean PRE clock and $[u]_+ = \max(u, 0)$.

A second term encourages consistency between nearby training snapshots. For each eligible source, eight nearest neighbors are considered, and at most one later-clock, non-younger neighbor with the

same training-time regime assignment is retained. PRE subjects and subjects with missing ages are excluded from age-based pairing.

A bounded auxiliary clock-rate network advances the source clock over four Euler steps using the age gap. The resulting clock prediction $\widehat{t}_{j|i}$ is compared with the destination clock:

$$\mathcal{L}_{\text{pair}} = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \left(\widehat{t}_{j|i} - t_j^{\text{raw}}\right)^2, \qquad \mathcal{L}_{\text{clock}} = \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{pair}}. \tag{37}$$

The pair term is zero when no eligible pairs exist. The source state, source clock, and source compatibility weights are detached in the pair prediction.

These are cross-sectional pseudo-pairs used to regularize the representation. They are not observed within-person trajectories, and the auxiliary clock-rate network is not the reported SB current. The main snapshot-training objective does not use follow-up outcomes.

The representation objective is

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \eta_a \mathcal{L}_{\text{anchor}} + \eta_t \mathcal{L}_{\text{clock}}. \tag{38}$$

Bridge-current fitting is performed separately after this representation has been frozen.

C.3 Reporting coordinates

After representation fitting, define the training PRE clock median

$$a_0 = \text{median}_{n\in\mathcal{P}}\, t_n^{\text{raw}}.$$

The reporting coordinates are

$$t_n = \frac{t_n^{\text{raw}} - a_0}{1 - a_0}, \qquad s_{n,m} = \frac{s_{n,m}^{\text{raw}} - a_0}{1 - a_0}. \tag{39}$$

The response and covariance functions are transformed consistently:

$$\begin{aligned} \mu_m(s) &= \mu_m^{\text{raw}}\big(a_0 + (1 - a_0)s\big), \\ \Sigma_{m,g}(s) &= \Sigma_{m,g}^{\text{raw}}\big(a_0 + (1 - a_0)s\big). \end{aligned} \tag{40}$$

This transformation places the PRE median at clock zero while retaining a shared stage axis across regimes. It is not followed by another clipping operation. The original offset bound δ applies to the raw coordinates; in reporting coordinates, $|s_{n,m} - t_n| \le \delta/(1 - a_0)$.

For a fixed imaging-coordinate current, the stage gradient is divided by $1 - a_0$ and the Fisher scale is multiplied by $1 - a_0$. Their product in the calibrated-speed readout is therefore unchanged.

C.4 Post-fit bridge and training algorithm

Once the representation is frozen, the response geometry is used to construct the fixed transport metric Γ and transport coordinates y. For each regime, the terminal clock is

$$T_g = \text{median}_{n\in\mathcal{A}_g}\, t_n.$$

The PRE source support, eligible interior windows, and terminal-anchor support are constructed as described in Section 3.4. The bridge is then fitted by minimizing the path-space KL objective subject to the source, endpoint, and shared interior-mass constraints in Eq. (6).

The resulting edge masses determine the smoothed bridge densities, responsibilities $r_g(y)$, and current velocities $V_g(y, \lambda)$. Query progress is obtained from the mass-weighted source/interior knot clocks. The terminal segment contributes to responsibilities but is excluded from velocity readout. The fitting and reporting sequence is summarized in Algorithm 1.

Support-draw aggregation and age-conditioned reporting. Let $r_g^{(b)}(y)$ denote the bridge responsibility from support draw b. The main synthetic and clinical fits use $B = 5$ support draws and aggregate them as

$$\bar{r}_g(y) = \frac{\exp\Big\{B^{-1}\sum_{b=1}^{B} \log r_g^{(b)}(y)\Big\}}{\sum_h \exp\Big\{B^{-1}\sum_{b=1}^{B} \log r_h^{(b)}(y)\Big\}}. \tag{41}$$

Age evidence $L_g(a)$ estimated from training endpoint anchors gives the reporting weights

$$r_g^{\mathrm{rep}}(y,a) = \frac{\bar{r}_g(y)L_g(a)}{\sum_h \bar{r}_h(y)L_h(a)}, \qquad \hat{g} = \operatorname*{argmax}_g r_g^{\mathrm{rep}}(y,a). \tag{42}$$

The age evidence uses Gaussian kernels with bandwidth selected from training endpoint ages. Missing query age contributes $L_g(a) = 1$ for every regime; missing endpoint ages are integrated under the age-evidence model. Age evidence updates reporting weights and the assigned regime while preserving the fitted within-regime currents. Regime-specific currents and induced stage velocities are averaged across support draws before output calibration.

AD endpoint-support adjustment. The AD anchor-recovery analysis applies a minimum terminal-clock cutoff of 0.25 to the SA and NA interior supports. Thus their support cutoffs are $\max(T_g, 0.25)$, while other regimes use T_g; terminal anchors retain their observed clock values. This changes support and current construction, not the frozen clock or stage maps. Appendix H.2 ablates age conditioning and support adjustment. The matched ruler comparison in Appendix H.3 holds the fitted model and its reporting weights fixed and changes only the output ruler.

Algorithm 1 CALIPER fitting and modality-lead readout

Require: Training snapshots $\{x_{n,m} : m \in \Omega_n\}$, age covariates c_n, PRE labels, visible endpoint labels, and query snapshots.
Ensure: Shared clocks, modality stages, bridge responsibilities, reporting regimes, and calibrated modality speeds.

Estimate preprocessing statistics from training data.
Initialize encoders, clock and stage heads, response curves, residual scales, and anchor prototypes.
for each representation-training stage **do**
  Apply the stage-specific prototype and loss-weight schedule.
  for each optimization step **do**
    Compute z_n, t_n^{raw}, $s_{n,m}^{\mathrm{raw}}$, and $\alpha_n(g)$.
    Construct eligible clock-regularization pseudo-pairs.
    Update trainable parameters using $\mathcal{L}_{\mathrm{rec}} + \eta_a \mathcal{L}_{\mathrm{anchor}} + \eta_t \mathcal{L}_{\mathrm{clock}}$.
  end for
end for
Freeze the fitted representation and observation model.
Apply the PRE-origin transformation to clock, stages, and response functions.
Construct Γ, transport coordinates, and regime-specific bridge supports, applying the support adjustment when specified.
Fit the constrained bridge in Eq. (6) for the configured support draws.
for each query **do**
  Compute z, t, $\{s_m\}$, and y.
  Aggregate bridge responsibilities and compute $r_g^{\mathrm{rep}}(y,a)$ using Eq. (42).
  Set $\hat{g} = \operatorname{argmax}_g r_g^{\mathrm{rep}}(y,a)$.
  for each regime g **do**
    Locate $\lambda_g(t)$ on the source/interior ladder.
    Compute $j_g = D_z \Gamma^{-1/2} V_g(y, \lambda_g(t))$.
    Compute $q_{m,g} = \nabla_z s_m^\top j_g$; average over support draws, then set $v_{m,g} = |q_{m,g}| F_{m,g}(s_m)$.
  end for
  Determine $m_{\mathrm{lead}} = \operatorname{argmax}_{m \in \Omega} v_{m,\hat{g}}$.
end for
return t, $\{s_m\}$, $\{r_g^{\mathrm{rep}}(y,a)\}$, $\hat{g}$, $\{v_{m,g}\}$, and m_{lead}.

The stage gradients include the dependence of the clock on the imaging state and hold covariates fixed. A unique leading modality is reported only when the maximum is unique; meaningful cross-modality comparison requires at least two observed modalities.

C.5 CURRICULUM AND HYPERPARAMETERS

Representation learning uses four full-batch Adam stages (Table 4) with 160, 80, 160, and 80 optimization steps, respectively. The learning rate is 3×10^{-3}, and the gradient-norm clipping

threshold is 5. The reconstruction-loss weight is one throughout. The clock and anchor weights are

$$(\eta_t, \eta_a) = \begin{cases} (2, 0), & \text{stage 1,} \\ (0.5, 1), & \text{stages 2–4.} \end{cases} \tag{43}$$

At the start of stage 2, source and endpoint prototype moments are reset from the corresponding training anchors, with a standard-deviation floor of 0.1. They are optimized during stage 2 and frozen from stage 3, when the optimizer is rebuilt. There is no separate preliminary clock-only warmup.

The four stages train the representation and observation model. They do not jointly train the post-fit SB current.

Representation setting	Value
Imaging-state dimension	16
Encoder hidden width	64
Clock-head hidden widths	64, 64
Stage-offset hidden width	32
Residual-scale hidden width	32
Bernstein degree K_μ	6
Raw stage-offset bound δ	0.20
Adam learning rate	3×10^{-3}
Gradient-norm clipping threshold	5
Optimization steps by stage	160, 80, 160, 80
Regime centering weights	$1/\lvert\mathcal{G}\rvert$

Table 4: **Representation-training settings.** The stage-offset bound applies before the PRE-origin reporting transformation. Post-fit bridge construction is summarized separately in Appendix C.4.

D Synthetic Simulation Settings

D.1 Design and population stage paths

We evaluate the complete pipeline in a synthetic snapshot study with two paired observation settings: local heteroscedasticity and constant noise. The study tests recovery under stage-map and observation assumptions compatible with the fitted model. The generator was designed after preliminary simulation analyses; its settings and five new seeds were then fixed before fitting the methods reported here. Compatibility with the observation model does not require the generating population paths to coincide with the fitted entropic bridge.

Each replicate contains $N = 6{,}935$ independent subjects, randomly partitioned into 5,209 training and 1,726 held-out observations. There are three modalities, denoted by A, T, and N, with $R_m = 31$ regional features each, and four equally probable endpoint regimes. The regimes specify schematic population paths rather than validated clinical subtypes.

We use a superscript $\star$ to distinguish generating quantities from fitted quantities. For subject n, we sample

$$\begin{aligned} g_n &\sim \text{Categorical}(1/4, 1/4, 1/4, 1/4), \\ u_n &\sim 0.20\ \text{Beta}(1, 6) + 0.65\ \text{Uniform}(0, 1) + 0.15\ \text{Beta}(6, 1), \\ t_n^\star &= T_{g_n}^\star u_n, \qquad T^\star = (0.40, 0.80, 0.85, 0.65). \end{aligned} \tag{44}$$

The terminal values are disease-clock locations, rather than calendar ages or disease durations.

For regime g and modality m, the generating stage path is

$$s_{g,m}^\star(t) = t + A_{g,m} \sin\left(\frac{\pi t}{2T_g^\star}\right), \qquad 0 \le t \le T_g^\star, \tag{45}$$

where rows index regimes and columns index (A, T, N):

$$A = \begin{pmatrix} -0.10 & -0.08 & -0.06 \\ 0.03 & 0.15 & 0.02 \\ 0.10 & 0.00 & -0.12 \\ -0.07 & 0.00 & 0.12 \end{pmatrix}. \tag{46}$$

The corresponding signed stage velocity is

$$q^\star_{g,m}(t) = \frac{ds^\star_{g,m}(t)}{dt} = 1 + \frac{\pi A_{g,m}}{2T^\star_g} \cos\left(\frac{\pi t}{2T^\star_g}\right). \tag{47}$$

All paths are monotone, remain within $[0, 1]$, and satisfy $|s^\star_{g,m}(t) - t| < 0.2$. Regime membership changes the population path through the shared modality-stage coordinates; it does not define a different stage axis or mean-response function for each regime.

D.2 Regional observations and paired noise settings

Each modality has a shared regional mean-response curve represented by a degree-six Bernstein polynomial:

$$\mu^\star_{m,r}(s) = \sum_{j=0}^{6} \theta_{m,r,j} B_{j,6}(s), \qquad B_{j,6}(s) = \binom{6}{j} s^j (1-s)^{6-j}. \tag{48}$$

The coefficients are generated as

$$\theta_{m,r,j} = \epsilon_m a_{m,r} \,\mathrm{sigmoid}\left[k_{m,r}\left(\frac{j}{6} - o_{m,r}\right)\right], \tag{49}$$

with

$$\begin{aligned} a_{m,r} &\sim \mathrm{Uniform}(0.8, 1.2), \\ k_{m,r} &\sim \mathrm{Uniform}(5, 9), \\ o_{m,r} &\sim o_m + \mathrm{Uniform}(-0.15, 0.15), \\ o = (0.30, 0.50, 0.65), &\qquad \epsilon = (1, 1, -1). \end{aligned} \tag{50}$$

The raw A and T responses increase with stage, whereas the raw N response decreases. The N measurements are subsequently sign-oriented so that larger standardized values indicate greater pathological burden.

Regional observations are generated as

$$x_{n,m,r} = \mu^\star_{m,r}(s^\star_{n,m}) + \sigma^\star_{m,g_n,r}(s^\star_{n,m})\varepsilon_{n,m,r}, \qquad \varepsilon_{n,m,r} \sim \mathcal{N}(0, 1), \tag{51}$$

where $s^\star_{n,m} = s^\star_{g_n,m}(t^\star_n)$. Residuals are independent across regions conditional on stage and regime.

Local heteroscedasticity. Residual standard deviations depend on stage and regime:

$$\sigma^\star_{m,g,r}(s) = c_m a_{m,r} \exp\{b_{m,g,r}(s - 1/2)\}, \qquad c = (0.22, 0.28, 0.35). \tag{52}$$

Let

$$H = \begin{pmatrix} -0.9 & 0.3 & 1.0 & -0.4 \\ 0.4 & 1.0 & -0.9 & -0.5 \\ 0.8 & -0.7 & -0.6 & 0.5 \end{pmatrix}, \tag{53}$$

with rows indexing modalities and columns indexing regimes. For each region, draw $\eta_{m,r} \sim \mathrm{Uniform}(0.8, 1.2)$ and set

$$b_{m,g,r} = \eta_{m,r}\left(H_{m,g} - \frac{1}{4}\sum_{h=1}^{4} H_{m,h}\right). \tag{54}$$

Thus, $\sum_g b_{m,g,r} = 0$, matching the regime-centering constraint of the fitted residual-scale model.

Constant noise. The paired constant-noise setting sets $b_{m,g,r} = 0$:

$$\sigma^\star_{m,g,r}(s) = c_m a_{m,r}. \tag{55}$$

The two settings share generating clocks, stages, regional mean responses, Gaussian residual draws, ages, training splits, modality masks, and visible anchor selections.

Both settings use the complete Fisher-calibrated method. Constant residual noise does not imply a constant Fisher scale, because the mean-response derivative can vary with stage. The settings hold baseline residual scales fixed, rather than matching average residual variance. Their performance difference therefore does not isolate the benefit of Fisher calibration.

D.3 Missing modalities, covariates, and anchors

Every subject has at least one observed modality. In the order A, T, N, AT, AN, TN, ATN, the pattern probabilities are

$$p_\Omega = \big(0.115665,\ 0.030389,\ 0.413899,\ 0.185630,\ 0.131684,\ 0.009658,\ 0.113074\big). \tag{56}$$

These probabilities approximate the selected real-data modality-availability frequencies and define a controlled missingness mechanism.

A noisy age covariate is generated by

$$a_n = \bar{a}_{g_n} - 5(T^\star_{g_n} - t^\star_n) + \xi_n, \qquad \xi_n \sim \mathcal{N}(0, 4^2), \qquad \bar{a} = (80, 58, 76, 75). \tag{57}$$

Age is independently missing with probability $626/5209$. No cognitive scores or longitudinal measurements are supplied to any method.

The training set contains 294 visible endpoint anchors with counts 85, 45, 83, and 81 by regime, sampled from subjects with $u_n > 0.9$. An additional 670 training subjects with $t^\star_n < 0.08$ are labeled as PRE source anchors. Simulated PRE selection uses low generating clock alone, without the clinical age restriction used for real-data anchors.

Each regional feature is standardized using only its observed training values. Age normalization also uses observed training values only. The fitted transformations are applied to held-out subjects, and the two noise settings are standardized separately.

Generating clocks, stages, residual scales, velocities, and held-out regime labels are retained for evaluation. Beyond the designated training anchor labels, these quantities are not supplied to model fitting or query prediction.

D.4 Fitting and reporting configuration

The representation uses a 16-dimensional imaging state, encoder hidden width 64, degree-six regional Bernstein responses, and a bounded stage offset with $\delta = 0.2$. Training uses 480 full-batch updates in four stages of $(160, 80, 160, 80)$ updates, with learning rate 3×10^{-3}. The observation model and representation are then frozen before bridge fitting.

The regime-conditioned bridge uses the training-average response geometry, six interior windows, 64 sampled source observations, and an interior pool of up to 768 observations. Five support draws are fitted using seed offsets $(0, 100, 200, 300, 400)$, with constraint tolerance 10^{-7}. Fitted terminal clocks are determined by training endpoint-anchor clock medians in PRE-centered reporting coordinates.

The numerical current readout uses two quadrature nodes per segment and a linear interpolation rule. Regime-specific currents and induced stage velocities are averaged across the five support draws.

Reporting regime. The simulation uses five support draws and the age-conditioned reporting rule in Eqs. (41)–(42). The SA/NA support floor is an AD-application adjustment and is not part of these synthetic settings.

Baseline protocol. All methods use the same training and held-out datasets, modality masks, visible training labels, and available age. Fitting recipes are fixed before evaluation on these datasets, and generating truth is not used for model selection.

For FA, VAE, MoPoE, and burden-based comparators, PROFILE readouts use local-linear curves fitted on training observations with bandwidth 0.10 and estimated regime weights. Held-out observations enter only as queries. Flow-matching and SB-inspired drift comparators use their learned decoder tangents with training-based readout scaling. Each baseline retains its specified modality-score scale; oracle Fisher scales are not supplied. Methods without a corresponding stage or velocity output are reported as unavailable for that metric.

D.5 GROUND-TRUTH CALIBRATED SPEEDS

The generating mean-response Fisher scale is

$$F^{\star}_{m,g}(s) = \left[\frac{1}{R_m} \sum_{r=1}^{R_m} \left(\frac{\mu^{\star\prime}_{m,r}(s)}{\sigma^{\star}_{m,g,r}(s)} \right)^2 \right]^{1/2}. \tag{58}$$

The true calibrated speed for subject n is

$$v^{\star}_{n,m} = \left| q^{\star}_{g_n,m}(t^{\star}_n) \right| F^{\star}_{m,g_n}(s^{\star}_{n,m}). \tag{59}$$

Feature standardization transforms the mean derivative and residual standard deviation together, leaving this noise-normalized scale unchanged.

The ground-truth leading modality is

$$m^{\star}_n = \operatorname*{argmax}_{m \in \Omega_n} v^{\star}_{n,m}. \tag{60}$$

Leadership therefore depends on stage-change magnitude and local mean-response sensitivity relative to residual variation. This experiment assesses recovery of the specified calibrated estimand under the generating model; it does not independently establish biological modality leadership.

D.6 EVALUATION METRICS

The four main-table columns assess modality-stage recovery, relative calibrated-speed recovery, endpoint classification, and leading-modality identification.

Modality-stage recovery. For each modality, we compute Spearman correlation between estimated and generating stages on held-out subjects for whom that modality is observed:

$$\rho_{\text{stage}} = \frac{1}{3} \sum_{m \in \{A,T,N\}} \text{Spearman}\left(\{\hat{s}_{n,m}\}_{n \in \mathcal{H}_m}, \{s^{\star}_{n,m}\}_{n \in \mathcal{H}_m} \right), \tag{61}$$

where $\mathcal{H}_m$ denotes the observed held-out subset for modality m. This modality-averaged correlation is the stage metric reported in the main table; it is not a shared-clock recovery metric.

Relative calibrated-speed recovery. For CALIPER, the predicted modality score uses the estimated reporting regime:

$$\hat{v}_{n,m} = |\hat{q}_{n,m,\hat{g}_n}| \, \hat{F}_{m,\hat{g}_n}(\hat{s}_{n,m}). \tag{62}$$

Prediction and truth are normalized over the same observed modalities:

$$\hat{p}_{n,m} = \frac{\hat{v}_{n,m}}{\sum_{\ell \in \Omega_n} \hat{v}_{n,\ell}}, \qquad p^{\star}_{n,m} = \frac{v^{\star}_{n,m}}{\sum_{\ell \in \Omega_n} v^{\star}_{n,\ell}}. \tag{63}$$

For a baseline, its nonnegative modality scores replace $\hat{v}_{n,m}$ in the same normalization.

Eligible queries have at least two observed modalities, a supported readout in the estimated regime, finite observed modality scores, and predicted and true normalization denominators exceeding 10^{-10}. Let $\mathcal{Q}$ denote this eligible set and

$$\mathcal{E} = \{(n,m) : n \in \mathcal{Q},\ m \in \Omega_n\}.$$

The reported velocity metric is

$$\rho_{\text{velocity}} = \text{Spearman}\left(\{\hat{p}_{n,m}\}_{(n,m) \in \mathcal{E}}, \{p^{\star}_{n,m}\}_{(n,m) \in \mathcal{E}} \right). \tag{64}$$

Observed subject–modality entries are pooled within each replicate.

A common positive time-rescaling factor cancels within each normalized speed vector. This metric assesses relative modality speeds at a query, rather than absolute calendar-year rates. Generating regime labels enter only the construction of truth; predictions use estimated reporting regimes.

Endpoint balanced accuracy. Endpoint classification is evaluated on held-out subjects with $u_n > 0.9$, using the same eligible subset for all methods. Balanced accuracy is the unweighted mean of the four regime-specific recalls. Training anchors are excluded from this evaluation.

Leading-modality accuracy. For the velocity-eligible queries, the predicted leader is

$$\hat{m}_n = \operatorname*{argmax}_{m \in \Omega_n} \hat{v}_{n,m}. \tag{65}$$

Accuracy is

$$\mathrm{Acc}_{\mathrm{lead}} = \frac{1}{|\mathcal{Q}|} \sum_{n \in \mathcal{Q}} \mathbf{1}\{\hat{m}_n = m_n^\star\}. \tag{66}$$

Prediction and truth use the same observed modality set. Ties are resolved by the fixed modality order (A, T, N), without adding random perturbations.

D.7 Replication and uncertainty

Both noise settings are evaluated using five fixed seeds (9301–9305), with all runs retained. Predictions are saved before evaluation against the generating truth. Each reported entry is the mean $\pm$ sample standard deviation of the five seed-level metric values, computed separately for each noise setting.

E Data, Anchors, and Evaluation Protocol

E.1 Cohorts and preprocessing

The analysis combines ADNIMERGE-linked ADNI imaging with SCAN-linked NACC amyloid PET, tau PET, and structural MRI. The snapshot cohort contains 6,935 participants (Table 5): 5,209 in the training split and 1,726 in the development split. All observations from a participant remain in the same split.

An index snapshot combines available modalities within a trailing 90-day imaging window. Previously fixed index dates are preserved; for newly included participants, the index window is selected to maximize modality availability. A participant is included when at least one modality has an eligible regional measurement vector.

We evaluate two feature views. The common31 view contains 31 regions per modality, whereas the native view contains 68 amyloid-PET, 68 tau-PET, and 62 structural-MRI features. The two views use the same participants, split assignments, and snapshot modality masks.

Regional measurements are standardized using observed training snapshots within each normalization domain. Domains distinguish dataset and modality, with ADNI MRI additionally separated by processing stream. MRI compatibility checks preserve distinctions in FreeSurfer version, field strength, and acquisition. The fitted normalization parameters are applied to development and follow-up observations. MRI volume is sign-oriented so that larger standardized values indicate greater neurodegeneration.

Age is derived from birth-year information and the imaging index date, using a mid-year convention when only birth year is available. Age is standardized using observed training ages. Missing ages retain an explicit observation mask and a zero standardized encoding. Cognitive scores and diagnosis are excluded from query encoding, although they contribute to the clinical definitions of training anchors.

E.2 Clinical anchors and evaluation labels

Training anchors are operational clinical reference groups. For NACC, clinical status is obtained from `NACCUDSD` at the latest eligible visit on or before the imaging index, within one year: normal cognition, MCI, and dementia correspond to codes 1, 3, and 4, respectively.

For anchor construction, measured MMSE is used when available. Otherwise, the existing label rule uses the proxy

$$\mathrm{MMSE}_{\mathrm{proxy}} = \min(\mathrm{MoCA} + 2, 30).$$

Table 5: **Snapshot cohort and modality availability.** Counts refer to index snapshots and are identical across the common31 and native feature views. A, T, and N denote amyloid PET, tau PET, and structural MRI. The final column counts participants with all three modalities.

Cohort	Split	Participants	A	T	N	A+T+N
NACC	Training	3,466	1,499	911	2,474	296
NACC	Development	1,162	510	302	827	102
ADNI	Training	1,743	1,237	762	1,055	260
ADNI	Development	564	377	242	362	95
Total	Training	5,209	2,736	1,673	3,529	556
Total	Development	1,726	887	544	1,189	197
Total	Both	6,935	3,623	2,217	4,718	753

CDR-SB denotes the sum of boxes. The endpoint rules are:

- **SA:** normal cognition, age ≥ 75, MMSE ≥ 28, and CDR-SB ≤ 0.5 or missing.
- **Y-AD:** dementia, age < 65, and CDR-SB ≥ 2.5 or missing.
- **O-AD:** dementia, age ≥ 65, and CDR-SB ≥ 2.5 or missing.
- **NA:** normal cognition, age ≥ 65, $26 \leq \mathrm{MMSE} \leq 29$, CDR-SB ≤ 1.5 or missing, and not SA-eligible.

The names Y-AD and O-AD denote age-defined dementia reference groups in this protocol. The age threshold concerns age at the imaging index, rather than verified age of disease onset, and the labels do not require biomarker-confirmed AD etiology.

The training endpoint pool is balanced by capping each class at twice the size of the smallest class. A fixed random 60% of these labels are made visible, yielding 294 endpoint anchors with counts $(85, 45, 83, 81)$ for $(\mathrm{SA}, \mathrm{Y\text{-}AD}, \mathrm{O\text{-}AD}, \mathrm{NA})$. All visible endpoint anchors are from NACC.

PRE source anchors are training participants without a visible endpoint label who have normal cognition, age < 75, MMSE ≥ 28, and CDR-SB ≤ 0.5 or missing. There are 670 PRE source anchors. The remaining training subjects contribute through their observations without visible endpoint supervision.

Regime-evaluation sets. Regime classification is evaluated on three distinct sets:

- **NACC development anchors:** 408 participants satisfying the index-aligned clinical rules, with class counts $(125, 30, 88, 165)$.
- **ADNI index-rule labels:** 117 development participants, with class counts $(23, 7, 80, 7)$. These labels are constructed for evaluation from diagnosis, MMSE, and CDR-SB records nearest the imaging index within ± 180 days, together with index age. Measured MMSE is used without the NACC MoCA proxy.
- **Hidden training anchors:** 196 training participants whose endpoint labels were withheld, with class counts $(55, 25, 57, 59)$. Their imaging observations entered representation training, so this set measures hidden-label recovery. Of these participants, 40 also satisfy the PRE rule after endpoint-label masking.

Diagnosis and cognition are withheld from query encoding. However, anchor-recovery performance remains dependent on the clinical rules used to define the reference groups.

E.3 Clinical readouts and evaluation metrics

The clinical benchmark evaluates the model-defined clock and the information retained in the imaging representation through discrimination and severity association (Table 2). Endpoint-regime assignment is evaluated separately (Figure 7). These scores are distinct from population-current and Fisher-calibrated velocity readouts.

CALIPER clock staging score. The CALIPER (clock) row in Table 2 uses the fitted clock $t_n = t(z_n, c_n)$ in the PRE-centered reporting coordinates of Eq. (39). This is the same scalar coordinate used to locate a query on the bridge ladders, not a fitted downstream classifier. The same clock score is used for all clinical AUC and cognition-association columns.

CALIPER latent-PC1 staging score. The CALIPER (latent PC1) row in Table 2 is an auxiliary readout of the first principal component of the same frozen imaging representation. PCA is fitted on training latent states only:

$$u_n^{\mathrm{lat}} = \varsigma\, w_1^\top (z_n - \bar{z}_{\mathrm{train}}), \tag{67}$$

where w_1 is the first loading vector. The sign $\varsigma \in \{-1, 1\}$ is chosen so that visible dementia-like anchors have a higher mean score than visible CN-like anchors. The same oriented score is used for all clinical AUC and cognition-association columns. It is a downstream representation readout, distinct from the model's shared clock t_n, and does not replace the clock in the bridge or velocity analysis. The two table rows are not independent models or an outcome-wise selection between scores. The clock receives explicit age covariates; latent PC1 is computed from z_n without an additional query-age term. Their contrast therefore does not isolate the effect of a single model component.

Latent regime classifier. The latent-classifier reference in Figure 7 uses a multinomial logistic readout fitted on the 294 visible training anchors. Its features are standardized imaging latents and age terms:

$$\left[\mathrm{Standardize}_{\mathrm{train}}(z_n),\ a_n^{\mathrm{std}},\ (a_n^{\mathrm{std}})^2,\ \mathbf{1}\{\text{age missing}\}\right]. \tag{68}$$

The classifier uses $C = 1$, the L-BFGS solver, and at most 1,000 iterations. This downstream classifier evaluates the information contained in the representation; it is not the latent-PC1 severity score. Bridge-based reporting weights, rather than this classifier, determine regime assignment in the population-velocity analyses.

Clinical discrimination. CN-versus-dementia AUC is the primary discrimination metric, with CN-versus-MCI AUC reported separately. Both are computed within ADNI and NACC using the same eligible participants for all methods. For NACC, clinical status is taken from the latest eligible visit within one year before the imaging index. For ADNI, diagnosis is obtained from DXSUM records within ± 180 days of the index.

Severity association. We compute Spearman correlation between the staging score and index-aligned CDR-SB or MMSE. CDR-SB correlation is reported as ρ, and MMSE correlation as $-\rho$, so larger values indicate stronger association in the expected pathological direction. These are cross-sectional severity associations.

For this evaluation, NACC MMSE uses measured values only; the MoCA proxy used in anchor construction is not substituted. The main table reports CDR-SB association in both cohorts and MMSE association in ADNI. NACC measured-MMSE association is omitted from the main table because only four eligible measurements are available.

Regime classification. Balanced accuracy is the unweighted mean of class-specific recalls. Macro-F1 averages the four class-specific F1 scores, assigning zero when a class-specific F1 is undefined. The same target labels and participants are used for each method within an evaluation set.

Synthetic metric definitions and the two paired noise settings are given in Appendix D.

E.4 LONGITUDINAL EVALUATION

Longitudinal measurements are reserved for evaluation; results and their scope are discussed in Appendix H. Eligible intervals have positive duration and compatible measurement processing: matched tracer for PET, and matched processing stream and field strength for MRI. Cross-sectional and longitudinal MRI processing streams are not combined within a change interval.

Table 6: **Evaluation targets and metrics.** Clinical clock/representation scores, synthetic recovery, and longitudinal readouts address distinct targets.

Metric	Definition and interpretation
Clinical AUC ↑	Discrimination of CN versus dementia or CN versus MCI using the specified staging score, evaluated separately within each cohort.
CDR-SB ρ ↑; MMSE $-\rho$ ↑	Spearman association with index-aligned clinical severity; MMSE correlation is sign-reversed for reporting.
Regime BAcc / macro-F1 ↑	Four-class prediction performance on NACC development anchors, constructed ADNI index-rule labels, or hidden training labels.
Synthetic stage ρ ↑	Mean of the three observed-modality Spearman correlations between estimated and generating stages.
Synthetic relative-speed ρ ↑	Pooled Spearman correlation between predicted and true within-query modality-speed shares over observed modalities.
Synthetic leading accuracy ↑	Agreement between predicted and generating leading modalities at queries with at least two observed modalities.
Clock monotonicity ↑	Fraction of eligible repeated-visit pairs with a positive change in the learned disease clock, with half credit for ties.
Exploratory interval association	Regime-weighted Spearman correlation between the predicted signed mean-response tangent and observed annualized regional-mean change; its limited evidential scope is discussed in Appendix H.4.
Within-person modality ordering ↑	Agreement between predicted modality-speed ordering and observed-change ordering in temporally matched follow-up episodes.

For each baseline observation, the primary interval uses the first compatible follow-up at least 0.5 years later, without an upper duration cap. The observed modality-specific target is

$$d^{\mathrm{obs}}_{n,m} = \frac{1}{R_m} \sum_{r=1}^{R_m} \frac{x^{\mathrm{followup}}_{n,m,r} - x^{\mathrm{baseline}}_{n,m,r}}{\Delta\tau_n}, \tag{69}$$

where measurements are expressed in the fixed training standardization and $\Delta\tau_n$ is elapsed calendar time.

The corresponding signed model readout is the regional-mean response tangent:

$$d^{\mathrm{pred}}_{n,m,g} = q_{n,m,g} \frac{1}{R_m} \sum_{r=1}^{R_m} \mu'_{m,r}(s_{n,m}). \tag{70}$$

Its association with observed change is evaluated empirically; the bridge current is parameterized by local progress and is not interpreted as an individual calendar-year rate.

Each participant receives unit total weight within a task, divided across their eligible intervals and multiplied by the baseline regime responsibility. A modality–regime cell is evaluated when it contains at least ten positive-weight participants, total responsibility mass of at least five, and Kish effective sample size of at least ten.

For within-person modality ordering, one earliest feasible joint episode is selected per participant and modality combination. Baseline and follow-up scan spans are each at most 90 days, and interval lengths differ by at most 90 days and 20% of their median. Ordering agreement scores one for concordant order, zero for discordant order, and one half for ties. Ruler comparisons retain the same participants, responsibilities, and currents while changing the output calibration.

These comparisons use observed interval changes as validation targets. They do not treat noisy follow-up differences as ground-truth biological modality leadership.

E.5 Baseline configuration

The current comparison includes fourteen baselines. All preprocessing, representation fitting, downstream classifiers, and method-specific parameter selection use training data. Supervised components use the visible training-anchor labels, including their CN-like versus dementia-like collapse where

a binary target is required. Cognitive scores and held-out diagnoses are excluded from prediction inputs.

Missing regional values are handled by the specified method: training-mean imputation with availability indicators, explicit masking, or native missing-value handling. Age terms are included in downstream classifiers where specified.

Table 7 summarizes the configurations. Methods with configurable classical hyperparameters use training-only validation. MoPoE-VAE uses an internal training split for configuration selection. The VAE, flow-matching, and SB-inspired comparators retain their fixed training budgets and final-epoch checkpoints.

Table 7: **Baseline configuration for the clinical benchmark.** All supervised readouts use visible training-anchor labels. Reported outputs follow each method's specified readout; unavailable outputs are not imputed.

Baseline	Inputs, readouts, and configuration
Majority	Predicts the most frequent visible endpoint class; used for regime classification.
Age-only LR	Logistic regression on standardized age, squared age, and a missing-age indicator; separate binary and four-class fits, $C = 1$.
Anchor LR	Logistic regression on standardized regional measurements, modality masks, and age terms; separate binary and four-class fits, $C = 1$.
Mean burden	Mean pathology-oriented measurement over observed regional features; provides a scalar staging score.
PC1	First principal component of training regional measurements after training-mean imputation; sign oriented using visible anchors.
Static FA	Factor analysis on imputed measurements and modality masks. Dimension is selected from $\{1, 2, 3, 5, 8, 10\}$ by training-only likelihood validation. The dominant imaging factor gives the staging score; a logistic head on factors and age terms gives regimes.
Label spreading	A standardized imaging, mask, and age representation with a k-nearest-neighbor graph. Training-only selection uses $k \in \{5, 10, 20, 40, 80\}$ and $\alpha \in \{0.05, 0.2, 0.5, 0.8, 0.95\}$. Dementia-class probability supplies the staging score.
Gradient boosting	Histogram gradient boosting with native missing-value handling. Separate binary and four-class models; training-only selection over learning rate $\{0.03, 0.1, 0.3\}$, leaf count $\{4, 8, 16\}$, and L_2 penalty $\{0, 1\}$.
EBM	Kernel-density event model with anchor-defined CN-like and dementia-like reference groups. Uses 93 regional events in common31 and 18 lobar composites in the native view. The normalized maximum-likelihood event stage supplies the staging score.
DEBM-style	Uses the same biomarker mixtures as EBM, with a weighted-Kemeny central ordering and local insertion search. This is an implementation of the DEBM ordering principle, with normalized event stage as its readout.
Multimodal VAE	Masked multimodal product-of-experts model; hidden width 64, latent dimension 8, 150 epochs, Adam learning rate 10^{-3}. Uses its oriented staging score and a downstream latent-plus-age logistic regime classifier.
Flow matching	Imaging encoder with latent dimension 32 and hidden width 64, a conditional velocity network, and time-conditioned decoders; 200 epochs with Adam learning rate 10^{-3}. Uses trajectory-based staging and a latent-plus-age logistic regime classifier.
SB-inspired drift	Same encoder and decoder backbone as flow matching, with noisy bridge interpolation of scale 0.1 and a learned drift; 200 epochs. Uses trajectory-based staging and a latent-plus-age logistic regime classifier.
MoPoE-VAE	Mixture-of-products-of-experts multimodal VAE. Training-only selection considers latent dimension $\{8, 16\}$ and KL weight $\{0.5, 1\}$, selecting dimension 8 and weight 1 in both views. Final fits use 150 epochs and Adam learning rate 10^{-3}; readouts are anchor-oriented latent PC1 and a latent-plus-age logistic classifier.

For static FA, label spreading, and gradient boosting, supervised hyperparameter selection uses repeated five-fold validation within the visible training anchors. FA dimensionality is selected separately by unsupervised training-data likelihood. MoPoE-VAE configuration selection uses an internal $80/20$ split of the training observations and the validation variational objective.

For EBM and DEBM-style, the native-view analysis uses lobar composites under the prespecified fallback rule after regional mixture fitting failed. The SB-inspired drift comparator is a distinct implementation from CALIPER's fitted multi-marginal bridge. The baseline labels therefore identify the implemented comparators and their readouts.

E.6 EVALUATION SPLITS AND UNCERTAINTY

The development split, denoted TUNE in the evaluation files, was inspected during earlier method development. The reported comparisons reuse this fixed participant partition and do not constitute evaluation on an untouched test cohort. Hidden training anchors provide a separate label-recovery assessment on participants whose observations were available during fitting.

CALIPER, multimodal VAE, flow matching, SB-inspired drift, and MoPoE-VAE each use five seeds (45–49) in each feature view. The remaining baselines contribute one fit per view. These fits share participants and represent variation across optimization seeds and feature views.

For the clinical main table, uncertainty is estimated using $1{,}000$ participant-bootstrap draws shared across methods within each evaluation set. Metrics are first averaged across the relevant fits within each draw; the 2.5th and 97.5th percentiles then define the 95% interval. The displayed form $x^{+u}_{-\ell}$ denotes bounds $x - \ell$ and $x + u$ and preserves interval asymmetry.

Saved per-fit balanced accuracies and macro-F1 values record variation across optimization seeds and feature views. Their medians and interquartile ranges describe between-fit spread, distinct from participant-bootstrap uncertainty. In Figure 7, points summarize means across fits, and whiskers show the saved 95% participant-bootstrap intervals conditional on those fits.

Repeated-visit clock uncertainty uses participant-cluster bootstrap resampling so that intervals from the same participant remain together. Paired bootstrap comparisons reuse participants and draws. Figure 8 instead shows descriptive distributions across overlapping seed–regime–modality-pair cells, not participant-bootstrap confidence intervals. These two forms of summarization must be distinguished. Synthetic results use the separate five-replicate mean $\pm$ sample-standard-deviation convention specified in Appendix D.

F COMPUTATIONAL EFFICIENCY

The reported experiments run on a local Apple M4 Max workstation with 16 CPU cores and 64 GB unified memory, using PyTorch 2.8 on CPU. Representation training uses two computational threads per worker, while bridge fitting uses one thread per worker. Independent view–seed jobs can run concurrently. The implementation operates on regional imaging vectors and requires no pretrained image or language backbone.

Stage-specific runtime. Computation follows the two fitting steps of CALIPER: observation-model and representation training, followed by bridge fitting on the frozen representation. Table 8 summarizes recorded wall-clock times across seeds 45–49, separately for the common31 and native feature views.

Representation training comprises 480 full-batch updates on $5{,}209$ training snapshots; its recorded time includes data loading and boundary checks. Bridge fitting starts from cached training encodings and response geometry and includes all five support draws, each with six interior windows.

The readout stage processes the training and development snapshots, modality-specific longitudinal batches, and matched multimodal follow-up batches. Its recorded time includes prediction export for the snapshot and longitudinal readouts. Preprocessing, initial representation exports, and downstream bootstrap analyses are outside the bridge-fitting and readout timings below.

Table 8: **Recorded computational cost of CALIPER.** Values are wall-clock seconds, reported as mean $\pm$ sample standard deviation across five seeds per feature view. Bridge fitting includes five support draws. Readout and export cover all snapshot and longitudinal batches in the recorded reporting job.

Stage	Common31	Native
Representation training	37.4 ± 0.2	36.3 ± 3.3
Bridge fitting, five support draws	97.2 ± 19.3	110.7 ± 12.2
Batched readout and prediction export	203.0 ± 8.3	193.1 ± 8.2

Model size and reuse. The saved neural module contains fewer than 10^5 parameters in either feature view. Transport supports and fitted bridge variables are stored separately from the neural checkpoint.

Once fitted, the representation, response curves, residual scales, and bridge support are reused across query evaluations. Output-ruler ablations reuse the same representation, regime responsibilities, and currents, changing only the final calibration of modality speeds. Ablations that change representation training or bridge constraints require separate fits.

G EXTENDED QUANTITATIVE RESULTS

This section provides additional evidence for the four readouts of CALIPER: shared staging, regime-conditioned population velocity, Fisher-calibrated modality leadership, and interpretable endpoint regimes. We distinguish readout interventions on the fully learned model from controlled experiments with supplied coordinates. The former assess the fitted pipeline, whereas the latter isolate the roles of the population current and the local Fisher scale.

Unless stated otherwise, simulation results are means $\pm$ sample standard deviations over five seeds. Real-data confidence intervals use participant-level resampling conditional on the fitted models. All real-data analyses retain the development-split qualifications described in Appendix E.

G.1 MATCHED READOUT ABLATIONS ON THE MAIN SIMULATION

We evaluate readout interventions on the same two noise conditions used in the main simulation. Within each fitted model, we hold the learned representation, selected endpoint regime, observed modalities, query subjects, and bridge support fixed. All readouts are evaluated on their common finite support.

The full readout is

$$v_{m,g} = |q_{m,g}| F_{m,g}(s_m).$$

The fixed-Fisher control replaces the local scale by its training-set mean for each modality and regime; the no-Fisher control sets the output scale to one. Both retain the fitted current and transport geometry. Two further controls replace the stage velocity: a constant $q = 1$ retains only the local Fisher scale, while an endpoint secant uses the mean stage difference between visible endpoint anchors and PRE subjects, divided by their mean-clock difference. The secant retains the full model's selected regime and local Fisher scale.

Figure 4 shows that local Fisher calibration improves relative-speed recovery over both fixed and omitted output scales. Under local noise, relative-speed Spearman correlation decreases from 0.769 ± 0.031 to 0.684 ± 0.054 with fixed Fisher and to 0.513 ± 0.124 without Fisher. Under constant noise, the corresponding values are 0.808 ± 0.013, 0.771 ± 0.006, and 0.569 ± 0.077. The full readout has higher correlation than either ruler control in all five paired seeds in both conditions.

Leading-modality accuracy follows the same mean ordering for these ruler comparisons. Under local noise, full, fixed-Fisher, and no-Fisher accuracies are 0.775 ± 0.046, 0.737 ± 0.050, and 0.633 ± 0.100; under constant noise, they are 0.786 ± 0.044, 0.767 ± 0.031, and 0.693 ± 0.074. Constant observation noise does not make the Fisher scale constant, because the response derivative $\mu'_m(s)$ can still vary with stage.

The current controls qualify the scope of this result. The endpoint secant attains leading accuracies of 0.825 ± 0.068 and 0.815 ± 0.036 in the two conditions, exceeding the full readout. The constant-q control also has higher mean leading accuracy. Leading accuracy evaluates only the maximizing modality, whereas speed-share correlation and error assess the relative profile; these metrics need not rank estimators identically. The controls do not establish a uniform advantage of estimating the local current in this main simulation. Thus, these experiments support the benefit of the learned local ruler relative to fixed or omitted rulers, while the advantage of estimating a local current depends on the target metric and path geometry. The interventions are post-hoc readout analyses, not full-model retraining ablations.

G.2 POPULATION-CURRENT ABLATION UNDER CONTROLLED PATH GEOMETRY

To isolate current estimation from representation learning, we use a controlled experiment with supplied clock and stage coordinates and a known Fisher scale. Four equally likely regimes follow

$$s_{g,m}(t) = \beta_{g,m} t + C_{g,m} t(1-t) + \epsilon_m, \qquad q^*_{g,m}(t) = \beta_{g,m} + C_{g,m}(1-2t),$$

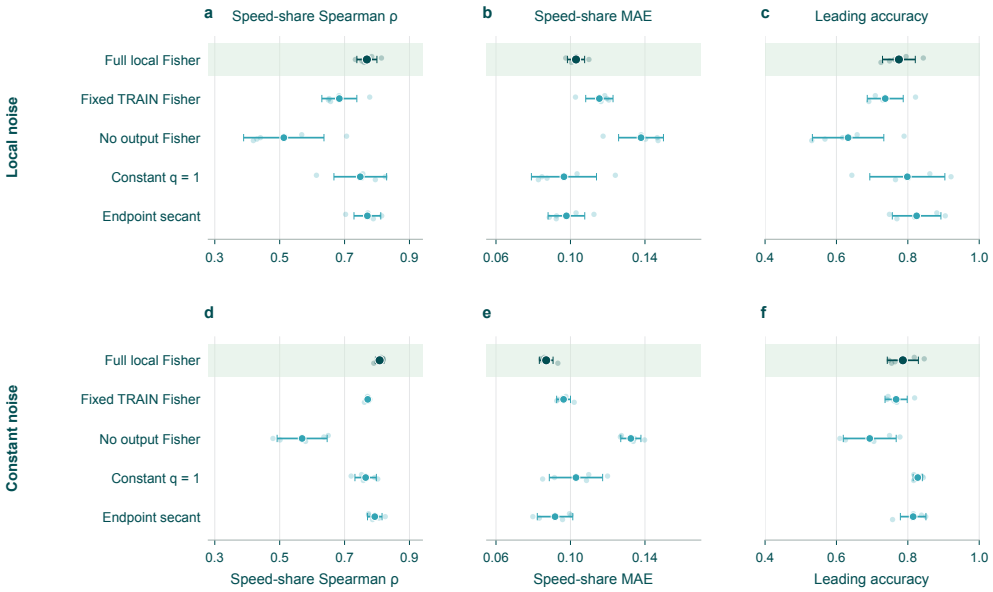

Figure 4: **Matched readout ablations on the main simulation.** Results use both noise conditions and all five seeds per condition. All arms share the learned representation, selected regime, query subjects, observed modalities, and common finite support. Fixed and omitted Fisher scales change only the output ruler. Constant q and endpoint secants replace the stage-velocity readout; the endpoint secant retains the full model's membership and is not a bridge-free refit. Relative-speed correlation and mean absolute error compare normalized speed shares over observed modalities. Large points and whiskers show means $\pm$ sample standard deviations; small points show individual seeds. The full readout improves correlation over fixed and omitted Fisher scales, while simpler current controls can achieve higher leading accuracy.

where $\epsilon \sim \mathcal{N}(0, 0.01^2 I_3)$ is a subject-specific offset that remains constant over time. The coefficient matrices are

$$\beta = \begin{pmatrix} 0.95 & 0.50 & 0.30 \\ 0.35 & 0.95 & 0.60 \\ 0.60 & 0.35 & 0.95 \\ 0.75 & 0.90 & 0.45 \end{pmatrix}, \qquad C = 0.55\,\beta \odot \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ -1 & 1 & 1 \end{pmatrix}.$$

The local ruler is $F_m(s) = (0.4 + b_m s)^{-1}$ with $b = (0.1, 0.8, 0.4)$. Each seed contains 4,096 training snapshots and 1,024 independent queries, with training clocks drawn uniformly from $(0, 1)$ and query clocks from $(0.10, 0.85)$. Source and endpoint anchors have $t < 0.06$ and $t > 0.92$, respectively. The bridge estimates regime membership and currents from these supplied coordinates; it does not learn an encoder or estimate the supplied ruler in this experiment.

On curved paths (Figure 5), the full bridge with supplied coordinates and ruler reaches leading-modality accuracy 0.952 ± 0.010 and calibrated-speed correlation 0.936 ± 0.011. Removing interior bridge constraints lowers speed correlation to 0.890 ± 0.007, replacing the current by an endpoint secant lowers it to 0.776 ± 0.010, and the endpoint-linear model without SB reaches 0.652 ± 0.007. Removing the output Fisher scale lowers leading accuracy to 0.780 ± 0.010. The full bridge exceeds every listed intervention on both metrics in all five paired seeds in this supplied-coordinate experiment.

The linear-path control sets $C = 0$, making endpoint lines correctly specified. In this setting, the endpoint-linear model reaches leading accuracy 0.981 ± 0.005 and speed correlation 1.000 ± 0.000, compared with 0.927 ± 0.019 and 0.942 ± 0.007 for the full bridge. Within these controlled geometries, intermediate constraints improve recovery of curved population paths, while endpoint information suffices for linear paths. This is a component-level test, not an end-to-end robustness result for arbitrary nonlinear trajectories.

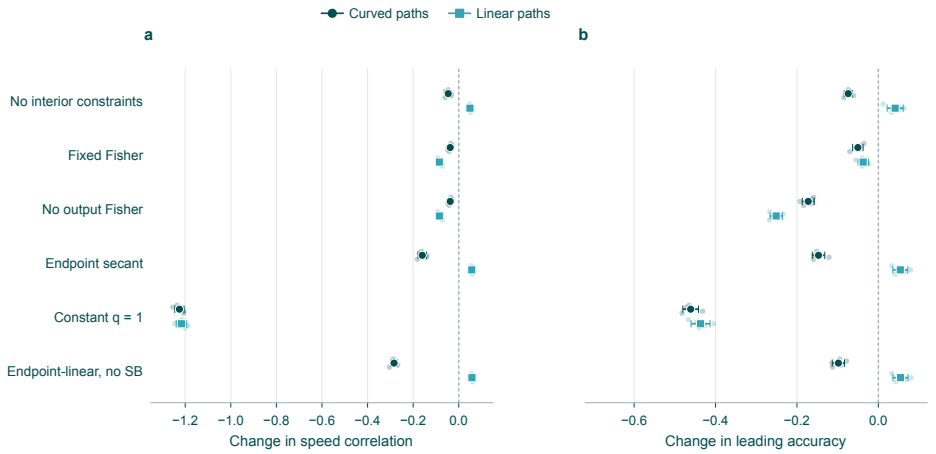

Figure 5: **Current and ruler ablations under curved and linear path geometry.** Coordinates and Fisher scales are supplied, while regime membership and bridge currents are estimated. Points show mean control-minus-full differences; whiskers show ± sample standard deviations over five paired seeds. Speed agreement is the mean of the three within-modality Spearman correlations with true calibrated speed; it differs from the pooled relative-speed-share correlation used in the main simulation. Except for the endpoint-linear model, all controls retain the full bridge's membership. Removing interior constraints also changes the clock-to-progress map; the endpoint-linear control changes both membership and current. Fixed and omitted Fisher scales affect only the output ruler. Both path geometries and all controls are retained.

G.3 Sensitivity to modality availability and evaluation support

We stratify the saved main-simulation predictions by the naturally observed modality combinations AT, AN, TN, and ATN. Figure 6 reports every combination in both noise conditions. Mean leading

accuracy ranges from 0.737 to 0.848 under local noise and from 0.780 to 0.831 under constant noise. Performance varies across modality combinations, and the TN pattern has only 11–20 evaluable subjects per fit, producing greater uncertainty.

These results describe performance under the simulated missingness distribution. They do not constitute an artificial modality-deletion experiment or a test of increasing missingness rates.

We also repeat the main relative-speed comparison on the intersection of evaluable queries across all eight methods that provide velocity readouts. The mean ranking is preserved: CALIPER reaches 0.769 ± 0.031 under local noise, compared with 0.738 ± 0.041 for MoPoE-VAE, and 0.808 ± 0.013 under constant noise, compared with 0.802 ± 0.037 for flow matching. Thus, using a common query set preserves the reported relative-speed ordering. These descriptive comparisons do not establish superiority on the separate staging, endpoint-classification, or leading-accuracy metrics.

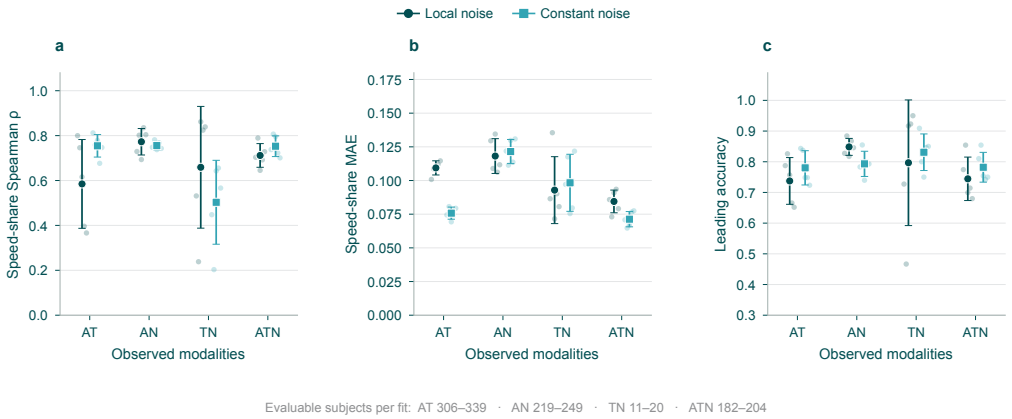

Figure 6: **Sensitivity to observed modality combinations.** Saved predictions from the main simulation are stratified by their observed AT, AN, TN, and ATN patterns. Large points and whiskers show means ± sample standard deviations across five seeds per noise condition; small points show individual seeds. Counts indicate the range of evaluable subjects per fit. The TN pattern has substantially less support than the other patterns. Random leading-modality accuracy is $1/2$ with two observed modalities and $1/3$ with three. The model is not refitted with artificial masking.

H LONGITUDINAL CHECKS AND REAL-DATA SENSITIVITY

We assess whether the learned coordinate remains coherent across repeated visits and examine which components affect endpoint-regime assignment and cross-modality comparison. These analyses use the reused development split and the participant-level evaluation rules in Appendix E.

H.1 REPEATED-VISIT COHERENCE OF THE DISEASE CLOCK

For each follow-up pair, we evaluate whether the learned clock increases, assigning half credit to ties. Participants receive equal total weight within each evaluation cell. The readout is the same learned disease clock reported in the clock row of Table 2, not the auxiliary latent-PC1 score.

Table 9 shows forward-clock probabilities of approximately 0.60–0.65 across the ADNI modality-specific pair sets. For intervals of at least three years, these probabilities increase to approximately 0.68–0.76. The pattern is similar in the common31 and native feature views. NACC MRI pairs also show forward progression, whereas the smaller NACC amyloid and tau samples provide inconclusive estimates.

These results support temporal coherence of the fitted coordinate. Because age is available to the clock and the interval strata contain different participants, they do not isolate an imaging-only effect or a causal effect of follow-up duration.

Table 9: **Repeated-visit coherence of the learned disease clock.** Values are participant-weighted forward-clock probabilities, with half credit for ties. Superscript and subscript offsets give the upper and lower distances to the endpoints of the 95% participant-cluster-bootstrap interval; they are not seed standard deviations. Five fitted seeds are averaged within each of 1,000 bootstrap draws. Participant counts are shown as common31/native. A, T, and N identify the followed modality and do not denote joint clock-and-biomarker worsening events.

Cohort	Modality	Interval	Participants	common31	Native
ADNI	A	All	200/199	$0.617^{+0.047}_{-0.047}$	$0.622^{+0.049}_{-0.048}$
ADNI	T	All	102/102	$0.597^{+0.081}_{-0.068}$	$0.612^{+0.073}_{-0.066}$
ADNI	N	All	308/308	$0.645^{+0.034}_{-0.033}$	$0.630^{+0.033}_{-0.033}$
ADNI	A	≥ 3 years	68/68	$0.760^{+0.084}_{-0.090}$	$0.725^{+0.085}_{-0.088}$
ADNI	T	≥ 3 years	48/48	$0.679^{+0.129}_{-0.123}$	$0.702^{+0.123}_{-0.123}$
ADNI	N	≥ 3 years	23/23	$0.725^{+0.152}_{-0.160}$	$0.733^{+0.145}_{-0.152}$
NACC	A	All	35/35	$0.477^{+0.143}_{-0.149}$	$0.463^{+0.149}_{-0.143}$
NACC	T	All	18/18	$0.450^{+0.194}_{-0.189}$	$0.533^{+0.211}_{-0.222}$
NACC	N	All	153/153	$0.655^{+0.063}_{-0.063}$	$0.629^{+0.066}_{-0.063}$

H.2 Endpoint-regime assignment: age and support sensitivity

We ablate age conditioning and the minimum SA/NA terminal-clock support cutoff of 0.25, yielding the full model and three component controls. The representation and stage maps remain frozen. Age conditioning changes reporting responsibilities without changing the within-regime currents; the terminal-clock floor changes the bridge support and current construction.

Across the 525 development-split endpoint anchors (Figure 7), the full model achieves balanced accuracy $0.748^{+0.028}_{-0.029}$. The corresponding values are $0.669^{+0.030}_{-0.032}$ with age conditioning alone, $0.399^{+0.036}_{-0.038}$ with the support adjustment alone, and $0.328^{+0.036}_{-0.039}$ with neither. The age-only and same-anchor logistic references reach $0.656^{+0.023}_{-0.027}$ and $0.677^{+0.036}_{-0.039}$, respectively. All offsets describe 95% participant-bootstrap intervals.

On the 196 training participants whose endpoint labels were hidden, the full model reaches $0.745^{+0.045}_{-0.048}$, compared with $0.627^{+0.056}_{-0.066}$ for the same-anchor logistic reference. This is a hidden-label assessment within the training population, rather than an independent participant test.

These results support endpoint compatibility from imaging and age and show its sensitivity to both age evidence and endpoint-support construction. Age enters several target-label definitions, so this is joint imaging-and-age reference assignment rather than imaging-only subtype discovery. The full model and all three controls are shown; the reused development comparison is not an untouched-test assessment. The latent classifier in Figure 7 is a separate reference, and its output is not interchangeable with bridge reporting weights or the latent-PC1 severity score.

H.3 Cross-modality ruler ablation on matched follow-up intervals

We compare output rulers using the fitted model and reporting weights from Eq. (42). Each comparison holds the representation, reporting responsibilities, currents, participants, and matched follow-up intervals fixed; only the output ruler changes.

Across supported ADNI cells (Figure 8), the local Fisher ruler improves mean ordering agreement over uncalibrated stage speed by 10.06 and 10.57 percentage points in common31 and native, respectively. Improvements over the full standardized-vector readout are 2.58 and 3.75 points. The largest gains occur for A–N: 13.42 and 15.21 points relative to standardized vectors.

The pair-specific analysis is heterogeneous. For T–N, the corresponding differences are -13.69 and -6.19 points. The local-versus-fixed Fisher differences are also small and inconsistent across views, at -0.45 and $+0.34$ points. The covariance-aware alternative is also pair-dependent: Figure 8d shows lower A–N ordering agreement than the diagonal ruler and gains in some T–N cells. No uniform advantage of the more elaborate covariance model is established. The reported mean gains

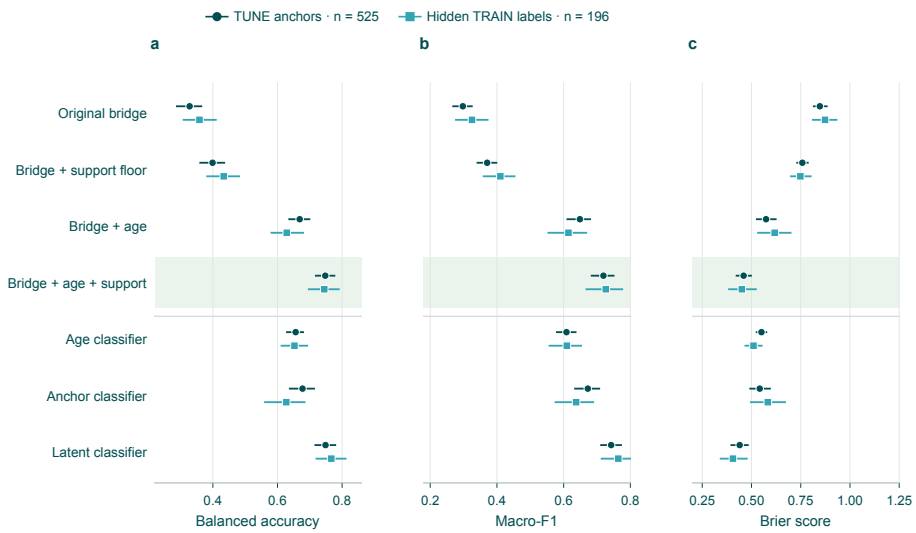

Figure 7: **Age conditioning and endpoint-support sensitivity.** The full model and three controls ablate age conditioning and the minimum SA/NA terminal-clock support cutoff of 0.25, with the representation and stage maps fixed. Points summarize two feature views and five seeds per view; whiskers show the saved 95% participant-bootstrap intervals. The age-only, same-anchor logistic, and latent classifiers are separate reference readouts. Development-split anchors and hidden training labels are shown separately. Differences between the full model and controls describe reporting sensitivity.

are averages over supported cells, not participant-weighted treatment effects or significance tests. Thus the analysis provides descriptive support for some modality comparisons, especially A–N, but no uniform advantage over fixed Fisher.

H.4 Training-time emission-noise ablation

The preceding ruler intervention changes the final readout. A separate matched training ablation evaluates learned heteroscedastic emissions against unit residual noise and learned region-specific constant noise, using the same data, anchors, initialization seeds, and optimization budget.

Figure 9 reports the resulting raw clock and modality-stage discrimination contrasts. For the primary clock-based CN-versus-dementia comparison, the heteroscedastic-minus-unit-noise AUC difference is $+0.010^{+0.036}_{-0.034}$ in ADNI and $-0.044^{+0.026}_{-0.025}$ in NACC, with offsets denoting 95% paired participant-bootstrap intervals. Thus, the training ablation does not establish a uniform clinical discrimination benefit from heteroscedastic emissions. Its outcome is distinct from the improvement in relative-speed comparison obtained by applying the local Fisher ruler to a fixed fitted model.

Scope of longitudinal velocity evidence. Forward-clock coherence and cross-modality ordering assess different properties from quantitative velocity recovery. In the separate clock-stratified analysis of sparse follow-up, the recorded permutation-based criterion does not establish a robust positive regime-level velocity association for the fitted population field. The velocity results therefore support a model-implied population readout, with recovery assessed directly in simulation; they do not establish individual instantaneous rates in calendar time.

I Anchor-Regime Spatial Atlases

Figures 10–13 show spatial atlases for the four anchor-defined endpoint regimes. Each atlas places observed subjects along the learned disease clock and displays regional measurements at representa-

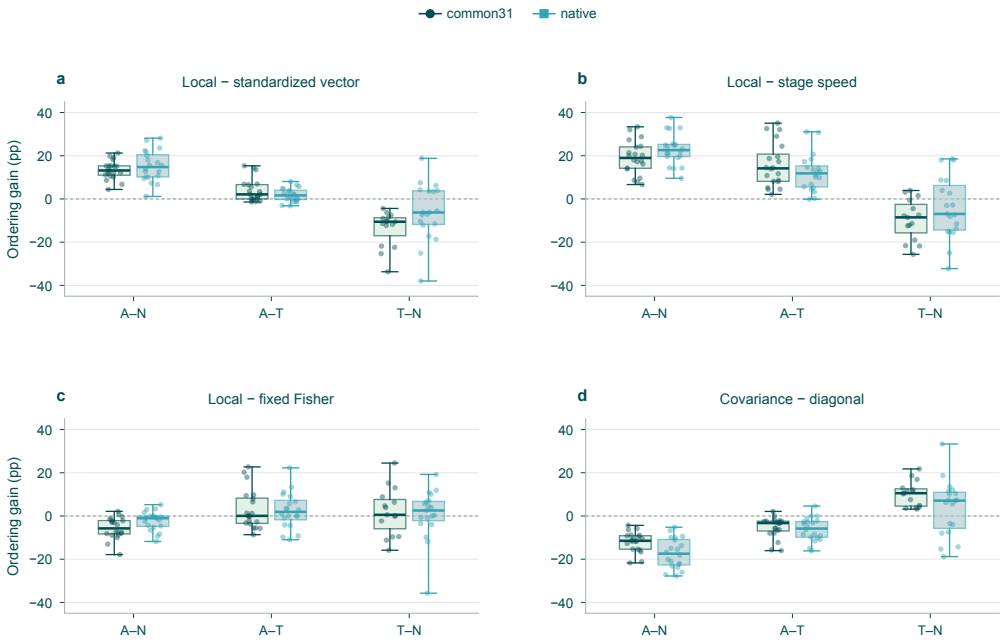

Figure 8: **Matched real-data ruler comparisons.** Output rulers are compared with the fitted model, participants, matched intervals, reporting responsibilities, and currents held fixed. All supported cells are from ADNI; NACC provides no supported matched-modality cells under this protocol. Points are seed-by-regime-by-modality-pair cells. Boxes summarize their median and interquartile range, and whiskers span the full cell range. Cells share participants, so their distribution is descriptive and does not represent independent replicates or a confidence interval. All modality pairs and both feature views are retained.

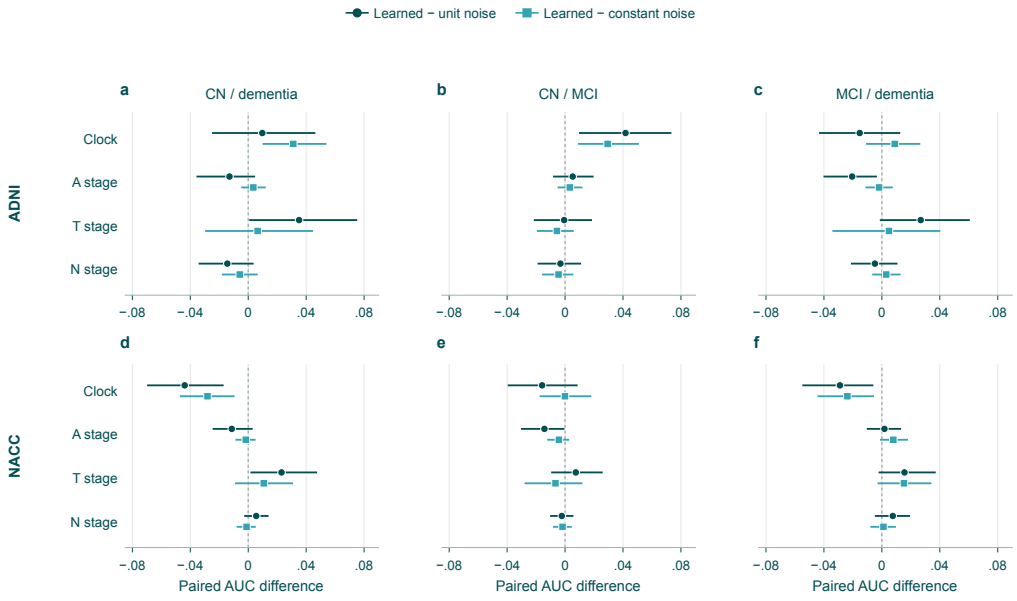

Figure 9: **Training-time emission-noise ablation.** Matched refits compare learned heteroscedastic residual scales with unit noise and learned region-specific constant noise. Points are paired AUC differences summarized over common31 and native views and five seeds per view. Whiskers are paired 95% participant-bootstrap intervals conditional on the fitted models. The figure retains both cohorts, all clinical contrasts, and all raw clock and modality-stage readouts. The clock contrasts concern the direct clock readout in Table 2; the affine reporting transformation preserves AUC. Modality-stage contrasts are separate from both the clock and the auxiliary latent-PC1 score.

tive clock locations. Together, the figures provide anatomical context for the shared coordinate and the modality-specific stage profiles.

The displayed clock locations need not correspond to repeated measurements from the same individual. These atlases describe spatial burden along the learned coordinate; modality leadership is determined separately by the calibrated velocity readout $v_{m,g} = |q_{m,g}|F_{m,g}(s_m)$. Spatial burden alone does not determine which modality has the largest Fisher-calibrated change along the fitted population field.

The Y-AD and O-AD labels denote anchor-defined groups distinguished by age at assessment, rather than verified age at symptom onset. The atlases are therefore interpreted as endpoint-regime visualizations.

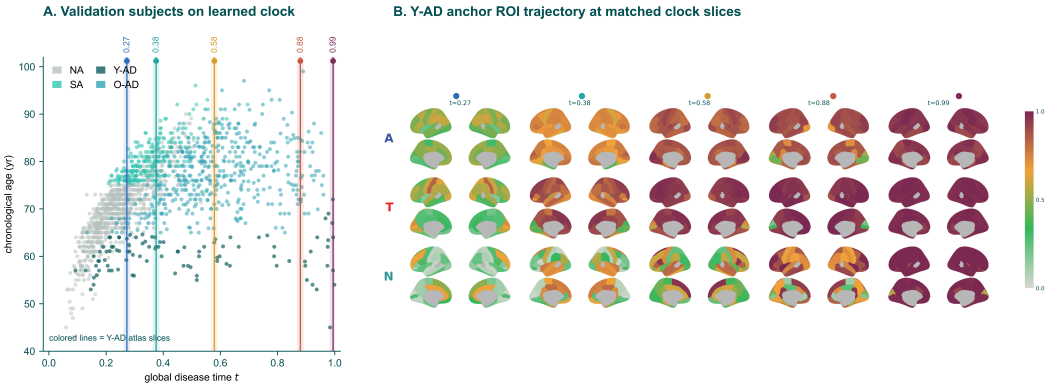

Figure 10: **Y-AD spatial atlas.** The left panel places subjects along the learned disease clock within the Y-AD endpoint regime and marks the clock locations used for visualization. The right panel displays the corresponding real-subject regional maps. Y-AD denotes the younger-at-assessment dementia anchor regime; it does not establish symptom-onset-defined early-onset AD.

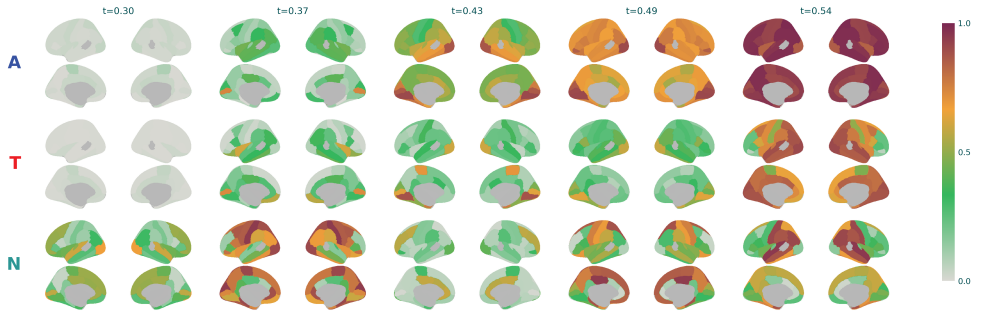

Figure 11: **SA spatial atlas.** Regional A/T/N measurements are displayed at representative locations along the learned clock within the SA endpoint regime. The maps visualize the spatial stage profile of the anchor-defined group. Modality-specific velocity and leadership are evaluated separately from the displayed burden maps.

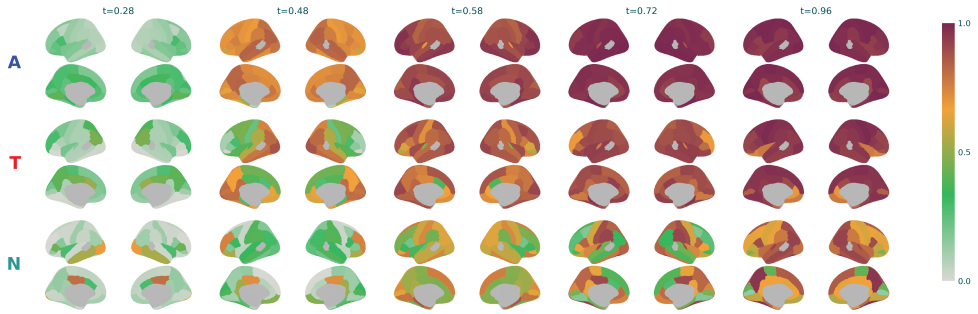

Figure 12: **O-AD spatial atlas.** Regional measurements from subjects assigned to the O-AD endpoint regime are arranged along the learned disease clock. The figure provides spatial context for modality-specific stage differences within this older-at-assessment dementia anchor regime. The displayed sequence is a cross-subject coordinate visualization, not a reconstructed individual longitudinal trajectory.

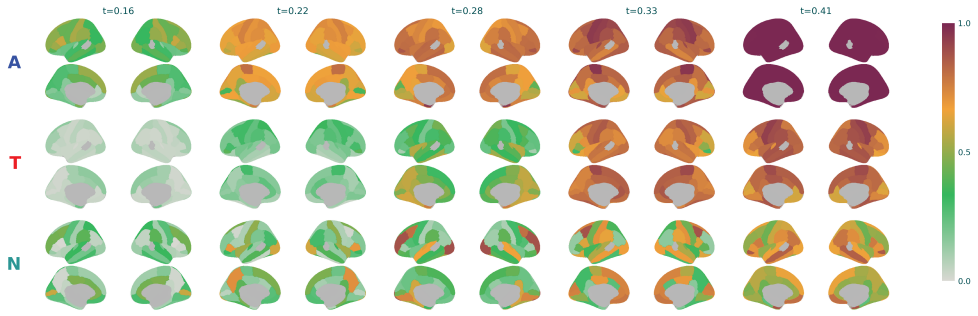

Figure 13: **NA spatial atlas.** Regional A/T/N measurements are shown at representative clock locations within the NA endpoint regime. The maps illustrate the regime's spatial stage profile without imposing a cohort-wide A-to-T-to-N ordering. Leading-modality assignments follow from the calibrated velocity readout and cannot be inferred from the magnitude of spatial burden alone.