CAMOS: COUPLED OSCILLATORY STATE-SPACE MODEL FOR MULTIMODAL CLINICAL TIME-SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Longitudinal clinical cohorts are multimodal, irregularly sampled and pervasively incomplete: in ADNI, positron emission tomography and cerebrospinal fluid assays are absent from roughly half of all visits. Linear state-space models handle irregular sampling gracefully but treat a missing modality by masking the *input*, leaving the transition operator untouched. We prove that this is a representational limitation: the latent state of any linear state-space layer whose transition operator does not depend on the availability pattern is an *additive* function of the availability indicators, so no such layer can represent an interaction between two modalities being jointly present or jointly absent. We propose CAMOS, which gives each modality a bank of second-order oscillators coupled through a matrix that sits *inside* the differential equation and is gated by availability, so the transition operator itself becomes a function of which measurements were taken. Coupling invalidates the analysis of uncoupled oscillatory models, and we restore it: a per-channel Gershgorin budget makes the effective stiffness positive definite uniformly over all 2^M availability patterns and all gaps, an energy argument charges amplification to availability transitions rather than sequence length, and a channel factorization preserves exact associative parallel scans. On ADNI, CAMOS outperforms uncoupled oscillatory state-space models and clinical fusion models on same-visit staging, landmark prediction and longitudinal forecasting, and under zero-shot transfer to OASIS-3 it is the only model that avoids collapse to the majority class.

1 INTRODUCTION

State-space models (SSMs) have been widely adopted for learning on long sequences (Gu et al., 2022; Smith et al., 2023; Orvieto et al., 2023; Gu & Dao, 2024), and they suit clinical time series because they discretize a continuous-time system and so treat a variable step size as a parameter rather than an approximation. Oscillatory SSMs sharpen this further, deriving stability from a nonnegative diagonal state matrix alone and admitting associative parallel scans in logarithmic depth (Rusch & Rus, 2025; Boyer et al., 2025; Blelloch, 1990). However, clinical cohorts pose a second difficulty alongside irregular sampling, i.e., measurements arrive as several modalities ordered by clinical need, cost and invasiveness, so each has its own gaps, while the underlying processes are coupled. In Alzheimer's disease, amyloid deposition precedes tau pathology, then neurodegeneration, then cognitive decline (Hardy & Selkoe, 2002; Jack et al., 2010; 2013), yet in ADNI (Petersen et al., 2010) imaging and cognitive testing occur at most visits whereas amyloid PET and cerebrospinal fluid assays are far rarer, so the cascade should be tracked through a missingness process that is neither uniform nor ignorable (Rubin, 1976).

Existing models treat that missingness as an input phenomenon. An absent measurement is handled by zeroing or masking the forcing term, sometimes with an appended indicator (Che et al., 2018), while the transition operator, which propagates latent state across time, is left untouched. Unrolling such a layer shows the cost: the latent state is a sum of per-modality, per-step contributions, each scaled by its own availability indicator, so the effect of one modality being present cannot depend on whether another was present at a different step. Joint availability is nevertheless informative, since whether an unverified amyloid estimate should influence a tau prediction depends on whether amyloid was measured, not merely on its value. Multimodal fusion methods provide a solution to this additive class (Hayat et al., 2022; Zhang et al., 2022a; Wang et al., 2023; Yao et al., 2024), but

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.

they fuse after the temporal model has been applied, so nothing lets one modality shape another's trajectory across a long gap.

In this paper, we place availability inside the dynamics. Each modality receives a bank of second-order oscillators, and modalities drive one another through a coupling matrix in the differential equation itself, gated by availability. Because the gate multiplies a matrix in the homogeneous part, the transition operator becomes a function of which measurements were taken, and the additivity obstruction disappears. However, coupling invalidates the analysis of uncoupled oscillatory models in three places: i) stability is a property of the effective stiffness, which no condition on the individual coupling blocks alone controls; ii) parallel evaluation relies on a diagonal state matrix, which coupling destroys; and iii) a per-step spectral radius bound no longer implies boundedness once the operator varies over time. Our contributions, each addressing one of them are:

- We introduce CAMOS, a coupled oscillatory state-space model whose transition operator is gated by the availability pattern and factored channel-wise, preserving exact associative parallel scans.
- We prove that every availability-independent linear state-space layer is additive in the availability indicators while CAMOS is not.
- We perform empirical evaluation on ADNI and OASIS-3 where CAMOS outperforms uncoupled oscillatory SSMs and clinical fusion models on same-visit staging, landmark prediction and longitudinal forecasting.

2 Related Work

Structured, oscillatory and coupled state-space models. S4, S5, LRU and Mamba established structured and selective state matrices solved by FFT or associative scans (Gu et al., 2020; 2022; Smith et al., 2023; Orvieto et al., 2023; Gu & Dao, 2024; Dao & Gu, 2024). LinOSS (Rusch & Rus, 2025) instead discretizes a second-order oscillator system, obtaining stability from a nonnegative diagonal state matrix alone and universality within continuous causal operators; its two variants, LinOSS-IM and LinOSS-IMEX, differ in whether the oscillator system is integrated by an implicit or a symplectic implicit-explicit scheme, and D-LinOSS (Boyer et al., 2025) decouples damping from frequency. These are the direct parents of CAMOS and our primary baselines. All of them take a single monolithic input, so concatenating modalities leaves the state matrix shared and the modality partition invisible to the dynamics, and no object in the model represents one modality driving another; they correspond to a zero coupling with a dense input matrix. Coupled Mamba (Li et al., 2024) is the closest on the oscillator side, coupling the hidden-state chains of several modalities, but it couples a discrete first-order recurrence rather than a continuous-time second-order system, so an irregular gap has no natural meaning.

Irregular sampling and missing modalities. Irregular sampling is handled by GRU-D, ODE-based models, controlled differential equations, mTAN and Raindrop (Che et al., 2018; Chen et al., 2018; Rubanova et al., 2019; De Brouwer et al., 2019; Kidger et al., 2020; Walker et al., 2024; Shukla & Marlin, 2021; Zhang et al., 2022b), all of which are sequential in the sequence length and none of which gates a coupled dynamical system by availability. Clinical fusion methods instead handle missing modalities at the level of *representations* (Ma et al., 2021; Zhang et al., 2022a; Wang et al., 2023; Yao et al., 2024; Hayat et al., 2022; Joze et al., 2020; Pölsterl et al., 2021), fusing after the temporal model has be applied. MedFuse (Hayat et al., 2022) learns a representation of the missing modality inside an LSTM fusion, and DrFuse (Yao et al., 2024) disentangles shared from modality-specific representations so that the shared component survives an absent modality. Both are nonlinear in the availability pattern, which places them outside the additive class, and both are our second family of baselines.

3 Problem Formulation

Let us consider a subject contribute N visits at strictly increasing times $t_1 < \cdots < t_N$ with gaps $\Delta t_n := t_n - t_{n-1} > 0$ in years, where $\Delta t_1 := t_1 - t_0$ and t_0 is the cohort baseline. At visit n, each of the M modalities $k \in [M] := \{1, \ldots, M\}$ contributes a feature vector $u_k^n \in \mathbb{R}^{p_k}$ and a binary availability indicator $a_k^n \in \{0, 1\}$, equal to one exactly when modality k was measured,

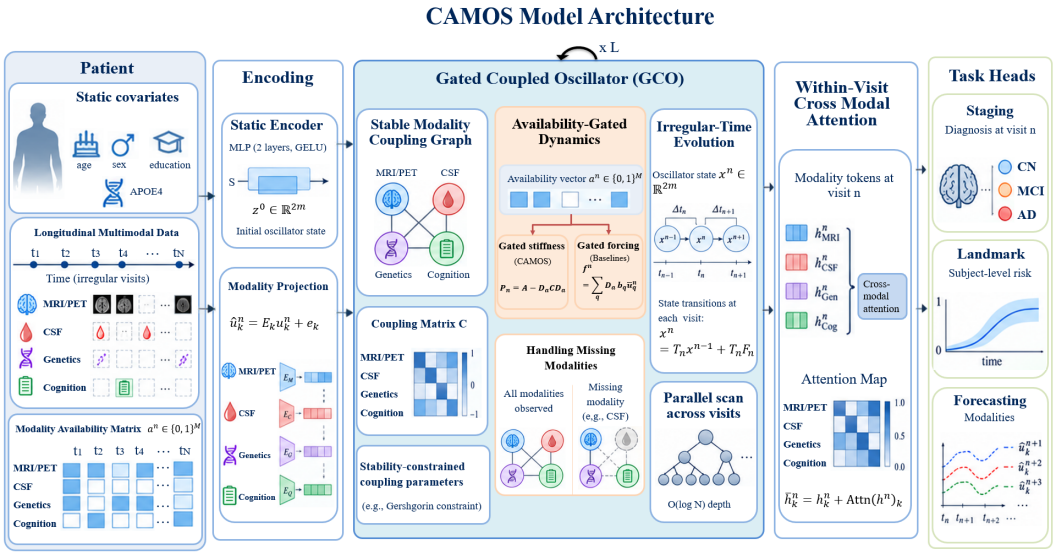


Figure 1: CAMOS architecture consists of static and longitudinal multimodal patient data through availability-gated coupled oscillators, irregular-time dynamics, and cross-modal attention.

with u_k^n undefined and never read when $a_k^n = 0$. We write $a^n = (a_1^n, \ldots, a_M^n) \in \{0,1\}^M$ for the *availability pattern*, $a^{1:N}$ for its sequence, $m := Md$ for the stacked latent width, and $D_a := \mathrm{blkdiag}(a_1 I_d, \ldots, a_M I_d) \in \mathbb{R}^{m \times m}$ for the gate matrix. Each subject also has static covariates $s \in \mathbb{R}^{p_s}$. The three prediction targets are: i) *same-visit staging*: a visit-level ordinal label $\ell^n \in \{\mathrm{CN}, \mathrm{MCI}, \mathrm{AD}\}$ predicted as $\hat{p}^n$ in the 2-simplex; ii) *landmark prediction*: a subject-level binary label $q \in \{0,1\}$ for progression from mild cognitive impairment (MCI) to Alzheimer's disease (AD) within three years, predicted as $\hat{q} \in [0,1]$; and iii) *longitudinal forecasting*: the values u_k^{n+h} at horizon $h \geq 1$, predicted as $\hat{u}_k^{n+h}$ and scored only where $a_k^{n+h} = 1$. The model is a map: $\Phi : \big(\{u_k^n\}, a^{1:N}, \{t_n\}, s\big) \longmapsto \big(\{\hat{p}^n\}, \hat{q}, \{\hat{u}_k^{n+h}\}\big)$, which uses $a^{1:N}$ as a genuine input rather than merely as a mask on u.

4 Proposed CAMOS Model

CAMOS is a stack of L gated coupled oscillator (GCO) blocks between an encoder and three task heads, $\Phi = \{\mathrm{Head}_\tau\}_{\tau \in \{1,2,3\}} \circ \mathrm{Attn} \circ \mathrm{GCO}^{(L)} \circ \cdots \circ \mathrm{GCO}^{(1)} \circ \mathrm{Enc}$. Everything except the GCO recurrence is applied pointwise in n and is therefore trivially parallel. The GCO recurrence is evaluated by an associative scan in $O(\log N)$ depth. All proofs are in Appendix C.

4.1 Input Encoders

Static covariates $s \in \mathbb{R}^{p_s}$ set the initial condition $x^0 = (z^0, y^0)^\top \in \mathbb{R}^{2m}$ of the oscillator bank through a two-layer perceptron with GELU activation (Hendrycks & Gimpel, 2016), rather than adding an offset at the readout, which encodes the hypothesis that baseline demographics modulate the phase and amplitude of the cascade rather than shifting outcomes additively. Modalities are measured on incommensurable scales, so each is projected affinely into a shared width h,

$$\tilde{u}_k^n = E_k\, u_k^n + e_k \in \mathbb{R}^h, \qquad E_k \in \mathbb{R}^{h \times p_k},\ e_k \in \mathbb{R}^h, \tag{1}$$

in parallel across visits and evaluated only when $a_k^n = 1$; the value of $\tilde{u}_k^n$ is irrelevant when $a_k^n = 0$ because the forcing is gated in equation 6.

4.2 COUPLED MULTIMODAL OSCILLATORS, AND WHY THE COUPLING MUST BE BUDGETED

Each modality k carries a bank of d oscillators obeying

$$z_k'(t) = -A_k\, y_k(t) + \sum_{j \neq k} C_{kj}\, y_j(t) + B_k\, \tilde{u}_k(t), \qquad y_k'(t) = z_k(t), \tag{2}$$

with $A_k, C_{kj} \in \mathbb{R}^{d\times d}$ and $B_k \in \mathbb{R}^{d\times h}$. Stacking $y = (y_1^\top, \dots, y_M^\top)^\top \in \mathbb{R}^m$ and likewise for z, equation 2 becomes $y'' = -Py + B\tilde{u}$ with

$$\begin{aligned} P &:= A - C, \qquad A := \mathrm{blkdiag}(A_1, \dots, A_M), \\ [C]_{kj} &:= C_{kj} \quad (k \neq j), \qquad [C]_{kk} := 0, \qquad B := \mathrm{blkdiag}(B_1, \dots, B_M). \end{aligned} \tag{3}$$

Stability, discretization and cost are properties of the *effective stiffness* P rather than of the individual blocks, which invalidates the natural first guess.

Remark 1 (Block-level conditions do not give stability). Parameterizing each block as $C_{kj} = Q_{kj}Q_{kj}^\top \succeq 0$ does *not* make $P \succeq 0$. Under Assumption 1 the assembled C is symmetric with zero diagonal blocks, so $\mathrm{tr}(C) = 0$ and C has a strictly negative eigenvalue whenever $C \neq 0$. The only free bound is $\lambda_{\min}(P) \geq \lambda_{\min}(A) - \|C\|_2$, which can be negative whenever $\|C\|_2 > \lambda_{\min}(A)$; then $I + \Delta t^2 P$ can be singular, the implicit solve fails, and the transition operator can have spectral radius exceeding one. A *joint* condition on A and C is required.

Channel factorization. We take A_k and C_{kj} diagonal, $A_k = \mathrm{diag}(\alpha_{k,1}, \dots, \alpha_{k,d})$ and $C_{kj} = \mathrm{diag}(c_{kj,1}, \dots, c_{kj,d})$, so coupling mixes modalities within each latent channel but not across channels. With Π the permutation $\Pi\, e_{(k-1)d+r} = e_{(r-1)M+k}$, reindexing from (modality-major, channel-minor) to (channel-major, modality-minor),

$$\Pi\, P\, \Pi^\top = \mathrm{blkdiag}\big(P^{(1)}, \dots, P^{(d)}\big), \qquad \big[P^{(r)}\big]_{kk} = \alpha_{k,r}, \quad \big[P^{(r)}\big]_{kj} = -\,c_{kj,r} \;\; (k \neq j), \tag{4}$$

so the coupled system decomposes into d independent networks of M oscillators, each carrying its own weighted modality graph $P^{(r)} \in \mathbb{R}^{M\times M}$. Appendix E shows that a dense C_{kj} raises the per-step cost from $O(dM^3)$ to $O(d^3M^3)$, and that a low-rank C_{kj} does not help, because low-rank-plus-diagonal structure is not closed under the scan combine.

Symmetry and Gershgorin budget. We impose two assumptions.

Assumption 1 (Symmetry)**.** $c_{kj,r} = c_{jk,r}$ for all $k, j \in [M]$, $r \in [d]$, so every $P^{(r)}$ is symmetric and $P = P^\top$.

Assumption 2 (Gershgorin budget)**.** There is $\epsilon \in (0, 1]$ such that $\sum_{j\neq k} |c_{kj,r}| \leq (1-\epsilon)\,\alpha_{k,r}$ for all $k \in [M]$, $r \in [d]$, with $\alpha_{k,r} > 0$.

Both hold by construction under the following parameterization, which needs no projection step, no power iteration, and is differentiable everywhere. Let $\hat{\alpha}_{k,r}$ and $\hat{c}_{kj,r} = \hat{c}_{jk,r}$ be free parameters (parameterize the strict upper triangle and mirror it) and set, with $\delta > 0$ a small constant,

$$\alpha_{k,r} = \mathrm{softplus}(\hat{\alpha}_{k,r}), \quad \sigma_{k,r} = \sum_{l\neq k} |\hat{c}_{kl,r}|, \quad c_{kj,r} = \frac{(1-\epsilon)\,\min\{\alpha_{k,r}, \alpha_{j,r}\}}{\max\{\sigma_{k,r},\, \sigma_{j,r},\, \delta\}}\, \hat{c}_{kj,r}. \tag{5}$$

Symmetry is immediate because the prefactor is symmetric in (k, j), and the budget follows termwise. Writing $\alpha_{\min} := \min_{k,r} \alpha_{k,r}$, $\alpha_{\max} := \max_{k,r} \alpha_{k,r}$ and $\gamma := \max_r \|C^{(r)}\|_2$ for the coupling strength, we set $\mu := \epsilon\,\alpha_{\min}$ and $L_P := (2-\epsilon)\,\alpha_{\max}$, and Assumption 2 gives $\gamma \leq (1-\epsilon)\alpha_{\max}$.

Lemma 2 (Uniform positive definiteness)**.** *For every $a \in [0,1]^M$ the matrix $P(a) = A - D_a C D_a$ is symmetric and $\mu I_m \preceq P(a) \preceq L_P I_m$. In particular $P(a) \succ 0$ simultaneously for all 2^M binary availability patterns, $I_m + \Delta t^2 P(a)$ is symmetric positive definite for every $\Delta t \in \mathbb{R}$, and its Cholesky factorization exists.*

The proof is a per-channel Gershgorin argument using only $|a_k a_j| \leq 1$, so gating can shrink but never enlarge an off-diagonal row sum: the budget is imposed once, on the *ungated* coupling, and gating preserves it automatically, so no availability pattern can destabilize the model.

4.3 AVAILABILITY GATING AND WHAT INPUT MASKING CANNOT REPRESENT

Availability enters in exactly two places, the stiffness and the forcing:

$$P_n := A - D_{a^n}\, C\, D_{a^n}, \qquad f^n := \sum_{k=1}^{M} a_k^n\, \iota_k\, B_k\, \tilde{u}_k^n \in \mathbb{R}^m, \tag{6}$$

where $\iota_k : \mathbb{R}^d \to \mathbb{R}^m$ embeds into the k-th modality block. Since $[D_{a^n} C D_{a^n}]_{kj} = a_k^n a_j^n C_{kj}$, the coupling gate is the *symmetric product* $g_{kj}^n = a_k^n a_j^n \in \{0,1\}$, so $[P_n^{(r)}]_{kj} = -a_k^n a_j^n c_{kj,r}$ off the diagonal. Two properties motivate this over the asymmetric alternative $g_{kj}^n = a_j^n$, which gates only by the source modality: P_n remains symmetric, which Assumption 1 and every result below require, and an unobserved modality is removed from the graph entirely.

Proposition 3 (Exact isolation of unobserved modalities)**.** *If $a_j^n = 0$, then P_n is block diagonal with respect to the splitting $\{j\} \cup ([M] \setminus \{j\})$, hence so are S_n and $\mathcal{T}_n$ of equation 8. Consequently, at visit n the latent state of modality j has exactly zero influence on every other modality and vice versa, while modality j continues to evolve under its own A_j with $S_n^{jj} = (I_d + \Delta t_n^2 A_j)^{-1}$. Under the asymmetric gate $g_{kj} = a_j$, P_n is block* triangular *rather than block diagonal, so the isolation is one-sided and P_n is no longer symmetric, invalidating Lemma 2.*

Without the gate, the one-step cross-modal influence is $\left|[S_n^{(r)}]_{kj}\right| = \Delta t_n^2\, |c_{kj,r}| + O\big(\Delta t_n^4 L_P^2\big)$ at short gaps, peaks at intermediate gaps and decays as Δt_n^{-2} at long ones (Proposition 11), so the gate matters most in the intermediate regime and for slow channels. The structural point is that P_n sits in the homogeneous part of the dynamics and is a function of a^n, and the following theorem is the formal consequence.

Definition 4 (Input-gated linear layers)**.** Let $\mathcal{A}$ be the class of layers $x^n = T_n x^{n-1} + \sum_{k=1}^{M} a_k^n B_k^n \tilde{u}_k^n$ with $x^0 = 0$ and $o^n = R x^n$, in which T_n, B_k^n and R may depend arbitrarily on the parameters, on $\{t_m\}$ and on $\{\Delta t_m\}$, but *not* on the availability pattern $a^{1:N}$.

$\mathcal{A}$ contains the state-space layers of S4, S5, LRU, LinOSS and D-LinOSS applied to concatenated multimodal inputs with masked or zero-filled missing entries. It excludes selective models such as Mamba when the availability vector is appended to the input, since selection then makes T_n depend on a.

Theorem 5 (Availability additivity)**.** *For every layer in $\mathcal{A}$ and every n, $o^n = \sum_{m \le n} \sum_k a_k^m \psi_{k,m}$ with $\psi_{k,m} := R T_n \cdots T_{m+1} B_k^m \tilde{u}_k^m$, so o^n is a multilinear polynomial of degree one in $a^{1:n}$ and every mixed availability difference of order two or higher vanishes identically, $\Delta_{(k,m)} \Delta_{(j,m')} o^n \equiv 0$ for all $(k,m) \neq (j,m')$, where $\Delta_{(k,m)} F(a) := F(a \,|\, a_k^m{=}1) - F(a \,|\, a_k^m{=}0)$. CAMOS does not lie in $\mathcal{A}$: already for $M = 2$, $d = 1$, $L = 1$,*

$$[S_n]_{12} = \frac{\Delta t_n^2\, g\, c}{(1 + \Delta t_n^2 \alpha_1)(1 + \Delta t_n^2 \alpha_2) - \Delta t_n^4 g^2 c^2}, \qquad g = a_1^n a_2^n, \tag{7}$$

which vanishes when $g = 0$ and is nonzero when $g = 1$ for any $c \neq 0$, giving $\Delta_{(1,n)} \Delta_{(2,n)} o^n \neq 0$.

Theorem 5 concerns a *single linear layer*. It does not say that deep networks built from such layers cannot represent availability interactions, since a nonlinear readout applied to an additive statistic is not itself additive, nor does it separate CAMOS from representation-level fusion such as MedFuse or DrFuse, whose encoders are nonlinear in a. It says that in class $\mathcal{A}$ the interaction must be reconstructed by the readout from an additive summary, whereas CAMOS represents it inside the recurrence, in the part of the model that is scannable and carries the stability guarantees above (Remark 8).

4.4 STABLE DISCRETIZATION ON THE IRREGULAR VISIT GRID

We discretize equation 2 on $\{t_n\}$ with the fully implicit (backward Euler) scheme applied to the first-order form, $z^n = z^{n-1} + \Delta t_n(-P_n y^n + f^n)$ and $y^n = y^{n-1} + \Delta t_n z^n$. Collecting $x^n := (z^n, y^n)^\top \in \mathbb{R}^{2m}$ and inverting through the Schur complement $S_n := (I_m + \Delta t_n^2 P_n)^{-1}$, which exists for every $\Delta t_n \in \mathbb{R}$ by Lemma 2, gives

$$x^n = \mathcal{T}_n\, x^{n-1} + \mathcal{T}_n F_n, \qquad \mathcal{T}_n = \begin{pmatrix} S_n & -\Delta t_n P_n S_n \\ \Delta t_n S_n & S_n \end{pmatrix}, \qquad F_n = (\Delta t_n f^n, 0)^\top. \tag{8}$$

Under equation 4, S_n is block diagonal with d blocks $S_n^{(r)} = (I_M + \Delta t_n^2 P_n^{(r)})^{-1}$, each costing $O(M^3)$ by Cholesky. The step size enters as a per-visit scalar, so irregular sampling requires no modification: the model integrates the same continuous system over whatever gap actually occurred, and stability is uniform in that gap.

Proposition 6 (Strict spectral contraction, uniform in Δt_n and a^n)**.** *Let $\nu_1, \dots, \nu_m > 0$ be the eigenvalues of P_n. The $2m$ eigenvalues of $\mathcal{T}_n$ are $\lambda_i^{\pm} = (1 \pm \mathrm{i}\Delta t_n \sqrt{\nu_i})/(1 + \Delta t_n^2 \nu_i)$, so $|\lambda_i^{\pm}| = (1 + \Delta t_n^2 \nu_i)^{-1/2} \le (1 + \Delta t_n^2 \mu)^{-1/2} < 1$ for every $\Delta t_n > 0$ and every availability pattern.*

The proof does not follow from the uncoupled case, because P_n is not diagonal; it orthogonally diagonalizes the symmetric P_n and reduces to m scalar 2×2 problems. Proposition 6 bounds each $\mathcal{T}_n$ individually, which is not sufficient for boundedness of the trajectory, since the $\mathcal{T}_n$ vary with n and are not normal, and finite products of matrices with spectral radius below one can grow transiently. This gap does not arise in the constant-coefficient setting, and we close it with an energy argument.

Theorem 7 (Switching bound governed by availability transitions)**.** *Let $E_n := \frac{1}{2}(\|z^n\|^2 + (y^n)^\top P_n y^n)$ with $f^n = 0$, and let $V_N := \#\{\, n \le N : a^n \ne a^{n-1} \,\}$ count the visits at which the availability pattern changes. Then*

$$E_N \le E_0 \, \exp\!\Big(\tfrac{1}{\mu} \textstyle\sum_{n=1}^{N} \|P_n - P_{n-1}\|_2\Big) \le E_0 \, \exp\!\big(2\gamma V_N / \mu\big), \tag{9}$$

and E_n is non-increasing on every maximal run of visits over which a^n is constant, with exact dissipation $E_n - E_{n-1} = -\frac{1}{2}\Delta t_n^2 ((y^n)^\top P^2 y^n + (z^n)^\top P z^n)$ when $P_n \equiv P$ (Proposition 9).

The bound is length-independent: amplification is charged to *availability transitions*, not to the number of visits or the size of the gaps, and it makes ϵ quantitative through $\gamma/\mu \le \frac{1-\epsilon}{\epsilon}\, \alpha_{\max}/\alpha_{\min}$. The choice of scheme is not free. The implicit-explicit symplectic discretization used by uncoupled oscillatory SSMs keeps unit-modulus eigenvalues only while $\Delta t_n^2 \nu \le 4$ for every eigenvalue ν of P_n, and otherwise has a real eigenvalue pair with product one, so one of them exceeds unity and the recurrence diverges. In ADNI, gaps range over roughly $[0.5, 4]$ years, so enforcing that condition uniformly would require $L_P \le 0.25$, hence $\alpha_{\max} \lesssim 0.14$, an order of magnitude below our initialization range and enough to collapse the representable frequency band.

4.5 Deep blocks, within-visit attention and task heads

GCO block. Following the block design of oscillatory SSMs (Rusch & Rus, 2025), block ℓ maps $\{\tilde{u}^{(\ell-1),n}\}_{n=1}^N \subset \mathbb{R}^{Mh}$ to $\{\tilde{u}^{(\ell),n}\}_{n=1}^N$ by scanning equation 8, forming $w^{(\ell),n} = C_{\mathrm{ro}}^{(\ell)} y^{(\ell),n} + D_{\mathrm{ro}}^{(\ell)} \tilde{u}^{(\ell-1),n}$ with $C_{\mathrm{ro}} \in \mathbb{R}^{Mh \times m}$, and applying a residual GLU (Dauphin et al., 2017) with dropout 0.1 (Appendix B). Each block owns its own $\hat{\alpha}, \hat{c}, B, C_{\mathrm{ro}}, D_{\mathrm{ro}}$, while the pattern $a^{1:N}$ and the gaps $\{\Delta t_n\}$ are shared. We write $h_k^n := \iota_k^\top y^{(L),n} \in \mathbb{R}^d$.

Within-visit cross-modal attention. GCO layers model how modalities drive one another *over time*. Contemporaneous associations at a single visit are captured by one multi-head attention block over the M modality tokens at each visit, applied in parallel across n (Vaswani et al., 2017), with an additive availability bias $-\beta(1 - a_j^n)$ on the logit of key j, where $\beta = \mathrm{softplus}(\hat{\beta}) \ge 0$ is learnable. Taking $\beta \to \infty$ recovers the hard availability mask of Yao et al. (2024); a finite β down-weights but does not erase an unobserved modality, consistent with its latent state continuing to evolve under Proposition 3. We use $H = 4$ heads with a residual connection and layer normalization, and write $\bar{h}_k^n := h_k^n + \mathrm{Attn}(h^n)_k$ and $\bar{h}^n := \|_k \bar{h}_k^n$.

Task heads and objective. All heads are affine in $\bar{h}$: same-visit staging is $\hat{p}^n = \mathrm{softmax}(W_{\mathrm{st}} \bar{h}^n + b_{\mathrm{st}})$; landmark prediction is $\hat{q} = \sigma(w_{\mathrm{cv}}^\top \sum_n \omega_n \bar{h}^n + b_{\mathrm{cv}})$, pooled over visits by learned weights ω_n so that subjects with different numbers of visits are handled without padding artefacts; and longitudinal forecasting is $\hat{u}_k^{n+h} = W_{\mathrm{fc},k}^{(h)} \bar{h}_k^n + b_{\mathrm{fc},k}^{(h)}$ for $h \le H_{\max} = 3$. The objective is

$$\mathcal{L} = \lambda_{\mathrm{st}} \mathcal{L}_{\mathrm{st}} + \lambda_{\mathrm{cv}} \mathcal{L}_{\mathrm{cv}} + \lambda_{\mathrm{fc}} \mathcal{L}_{\mathrm{fc}} + \lambda_{\mathrm{reg}} \sum_{\ell,k,j,r} \big|\hat{c}_{kj,r}^{(\ell)}\big|, \tag{10}$$

Table 1: Same-visit CN/MCI/AD staging on ADNI and OASIS-3. † denotes that CAMOS is significantly better than the best baseline ($p < 0.01$, two-sided Welch's t-test).

Models	Accuracy ↑	Macro F1 ↑	Macro Prec. ↑	Macro Recall ↑	Macro Spec. ↑
ADNI					
LinOSS-IM	$0.6593_{\pm 0.0041}$	$0.6815_{\pm 0.0039}$	$0.6804_{\pm 0.0045}$	$0.6864_{\pm 0.0037}$	$0.8147_{\pm 0.0021}$
LinOSS-IMEX	$0.7206_{\pm 0.0058}$	$0.7234_{\pm 0.0061}$	$0.7394_{\pm 0.0055}$	$0.7424_{\pm 0.0059}$	$0.8513_{\pm 0.0030}$
MedFuse	$0.6648_{\pm 0.0072}$	$0.6697_{\pm 0.0076}$	$0.6564_{\pm 0.0069}$	$0.7139_{\pm 0.0070}$	$0.8284_{\pm 0.0034}$
DrFuse	$0.7322_{\pm 0.0049}$	$0.7354_{\pm 0.0046}$	$0.7488_{\pm 0.0044}$	$0.7683_{\pm 0.0042}$	$0.8599_{\pm 0.0024}$
CAMOS	$\mathbf{0.7824}^{\dagger}_{\pm 0.0016}$	$\mathbf{0.7790}^{\dagger}_{\pm 0.0018}$	$\mathbf{0.7970}^{\dagger}_{\pm 0.0022}$	$\mathbf{0.7786}^{\dagger}_{\pm 0.0025}$	$\mathbf{0.8836}^{\dagger}_{\pm 0.0013}$
OASIS-3					
LinOSS-IM	$0.7271_{\pm 0.0017}$	$0.3253_{\pm 0.0053}$	$0.5334_{\pm 0.0082}$	$0.3541_{\pm 0.0025}$	$0.6773_{\pm 0.0014}$
LinOSS-IMEX	$0.7231_{\pm 0.0020}$	$0.3027_{\pm 0.0064}$	$0.4222_{\pm 0.0094}$	$0.3414_{\pm 0.0028}$	$0.6756_{\pm 0.0016}$
MedFuse	$0.6682_{\pm 0.0032}$	$0.3176_{\pm 0.0059}$	$0.2871_{\pm 0.0091}$	$0.3558_{\pm 0.0026}$	$0.6844_{\pm 0.0019}$
DrFuse	$0.7263_{\pm 0.0023}$	$0.2805_{\pm 0.0081}$	$0.2421_{\pm 0.0086}$	$0.3333_{\pm 0.0000}$	$0.6667_{\pm 0.0000}$
CAMOS	$\mathbf{0.7600}^{\dagger}_{\pm 0.0014}$	$\mathbf{0.5557}^{\dagger}_{\pm 0.0028}$	$\mathbf{0.6280}^{\dagger}_{\pm 0.0031}$	$\mathbf{0.5261}^{\dagger}_{\pm 0.0030}$	$\mathbf{0.7682}^{\dagger}_{\pm 0.0015}$

with $\mathcal{L}_{\text{st}}$ class-weighted cross-entropy over visits with a recorded stage label, $\mathcal{L}_{\text{cv}}$ class-weighted binary cross-entropy, and $\mathcal{L}_{\text{fc}}$ a horizon-discounted ℓ_1 error masked by a_k^{n+h} and normalized by the number of observed targets (Appendix B). The mask is essential: with roughly half of all modality-visit pairs unobserved, scoring against imputed targets would train the model on its own imputation error. The forecasting term also acts as a physical regularizer, forcing the latent oscillators to track measurable quantities, and the ℓ_1 penalty on the raw couplings encourages a sparse modality graph without interfering with the budget equation 5. We use $(\lambda_{\text{st}}, \lambda_{\text{cv}}, \lambda_{\text{fc}}, \lambda_{\text{reg}}) = (1, 0.5, 1, 10^{-3})$.

5 Experiments

Datasets. We evaluate on ADNI (Petersen et al., 2010) with $M = 4$ modalities grouped by acquisition route: *imaging* ($p_1 = 8$; intracranial-volume-normalized volumes of hippocampus, entorhinal cortex, fusiform gyrus, middle temporal gyrus, whole brain and ventricles, plus FDG and AV45 PET SUVRs), *CSF* ($p_2 = 3$; A$\beta_{1\text{-}42}$, total tau, phosphorylated tau), *genetics* ($p_3 = 3$; *APOE* $\varepsilon 4$ allele count, genotype indicators, polygenic risk summary) and *cognition* ($p_4 = 5$; MMSE, ADAS-Cog-13, CDR-SB, RAVLT immediate recall, FAQ), with baseline age, sex and years of education as static covariates ($p_s = 3$). We retain subjects with at least three diagnosed visits and one fully observed modality per visit, and mark a modality available only when all of its features are recorded. We also evaluate on OASIS-3 (LaMontagne et al., 2019), which carries the three modalities (*imaging*, *genetics*, *cognition*) and three targets, refitting only the normalization statistics.

Baselines. LinOSS-IM and LinOSS-IMEX (Rusch & Rus, 2025) are uncoupled oscillatory SSMs, the direct parents of CAMOS, and isolate the effect of coupling. Having no native mechanism for missing modalities, they receive concatenated features with unobserved entries zero-filled, the availability vector a^n as M extra channels, and Δt_n both as an input channel and as the per-step discretization interval. MedFuse (Hayat et al., 2022) and DrFuse (Yao et al., 2024) handle missing modalities at the representation level; we replace their modality encoders by per-modality gated recurrent encoders over the visit sequence, so they receive the same temporal information as the SSMs. Appendix I contains results for all the different families of baselines.

6 Results

Same-visit Staging. On ADNI, CAMOS outperforms all the baselines (Table 1), improving over DrFuse, by 0.0502 in accuracy ($+6.9\%$ relative) and 0.0436 in macro F1 ($+5.9\%$). LinOSS-IMEX outperforms LinOSS-IM, since the implicit-explicit scheme is stable only while $\Delta t_n^2 \nu \leq 4$. This suggests that the tuned baseline learned frequencies low enough to stay in the stable regime. On OASIS-3, evaluated zero-shot with only the normalization statistics refitted, accuracy is weakly informative because one class dominates. DrFuse attains macro recall of exactly $1/3$ and macro

Table 2: 3-year MCI-AD landmark prediction on ADNI and OASIS-3. CAMOS is significantly better than the best baseline at $^{\dagger}p < 0.01$ or $^{*}p < 0.05$ (two-sided Welch's t-test).

	ADNI		OASIS-3	
Models	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
LinOSS-IM	$0.9223_{\pm 0.0031}$	$0.7102_{\pm 0.0087}$	$0.9533_{\pm 0.0024}$	$0.3588_{\pm 0.0142}$
LinOSS-IMEX	$0.9307_{\pm 0.0027}$	$0.7389_{\pm 0.0074}$	$0.9560_{\pm 0.0029}$	$0.4185_{\pm 0.0163}$
MedFuse	$0.9134_{\pm 0.0045}$	$0.6900_{\pm 0.0102}$	$0.8467_{\pm 0.0058}$	$0.2594_{\pm 0.0121}$
DrFuse	$0.9386_{\pm 0.0022}$	$0.7689_{\pm 0.0069}$	$0.8787_{\pm 0.0047}$	$0.3183_{\pm 0.0135}$
CAMOS	$\mathbf{0.9611}^{\dagger}_{\pm \mathbf{0.0014}}$	$\mathbf{0.8687}^{\dagger}_{\pm \mathbf{0.0041}}$	$\mathbf{0.9600}^{*}_{\pm \mathbf{0.0018}}$	$\mathbf{0.7894}^{\dagger}_{\pm \mathbf{0.0093}}$

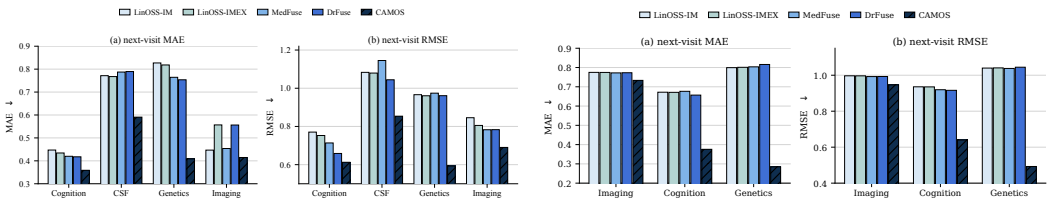


Figure 2: Next-visit ($h = 1$) forecasting on ADNI (left) and OASIS (right) per modality.

specificity of exactly $2/3$, the signature of predicting a single class, so its accuracy of 0.7263 equals the majority-class rate. The other baselines are close to this solution, with macro F1 between 0.2805 and 0.3253. CAMOS alone retains classification among the three stages, raising macro F1 from 0.3253 to 0.5557 and macro recall from 0.3558 to 0.5261, while exceeding the majority-class accuracy. Since these relative gains are measured against near-collapsed baselines, the absolute macro F1 of CAMOS is the informative quantity for transfer.

Landmark Prediction. Our CAMOS improves AUROC from 0.9386 to 0.9611 ($+2.4\%$) and AUPRC from 0.7689 to 0.8687 ($+13.0\%$) over DrFuse on ADNI dataset (Table 2). The asymmetry between the two metrics is the main finding. AUROC is compressed near its ceiling, spanning only 4.8 points across the five models (0.9134 to 0.9611), whereas AUPRC, which is sensitive to the minority positive class, spans 17.9 points (0.6900 to 0.8687). This matters practically because landmark prediction is used for trial enrichment, a precision-limited application in which AUPRC is the relevant summary. On OASIS-3 the pattern is sharper. AUROC barely separates CAMOS from the oscillatory baselines (0.9600 vs. 0.9560 for LinOSS-IMEX, $+0.4\%$), while AUPRC rises from 0.4185 to 0.7894 ($+88.6\%$), with all baselines between 0.2594 and 0.4185.

Longitudinal forecasting. Figure 2 reports next-visit ($h = 1$) error per modality. On ADNI, CAMOS attains the lowest MAE and RMSE on all four modalities. The largest reductions are on the sparsely observed genetics (MAE $0.7535 \to 0.4088$, -45.8%; RMSE -38.3%) and CSF (MAE $0.7676 \to 0.5906$, -23.1%; RMSE -18.3%), where the four baselines are tightly clustered (MAE 0.7535 to 0.8270 and 0.7676 to 0.7895), so uncoupled oscillatory dynamics and representation-level fusion perform alike there while CAMOS separates clearly. Gains are smaller on cognition (MAE -14.1%, RMSE -7.0%) and imaging (MAE -7.2%, RMSE -11.8%), the only modality on which the baselines split, with LinOSS-IM and MedFuse (MAE 0.4468, 0.4538) ahead of LinOSS-IMEX and DrFuse (0.5567, 0.5562). The same ordering holds zero-shot on OASIS-3: CAMOS reduces MAE most on genetics ($0.7875 \to 0.2847$, -63.9%), then on clinical measures ($0.5775 \to 0.3756$, -35.0%), and least on imaging ($0.7664 \to 0.7333$, -4.3%).

Relative improvements over baselines. All relative improvements in Figure 3 are positive, ranging from $+1.3\%$ (macro recall) to $+45.8\%$ (genetics MAE) on ADNI and from $+0.4\%$ (AUROC) to $+88.6\%$ (AUPRC) on OASIS-3, where margins are largest on metrics for which baselines collapse to the majority class (F1 $+70.8\%$, recall $+47.9\%$) and smallest on the imbalance-saturated accuracy and AUROC. The ADNI ablations trace these gains to three design choices. The implicit discretization, stable for every gap, matters most: switching to the implicit-explicit scheme, unstable once $\Delta t_n^2 \nu > 4$ as the 0.5 to 4 year gaps allow, lowers macro F1 from 0.7790 to 0.653 and AUPRC from

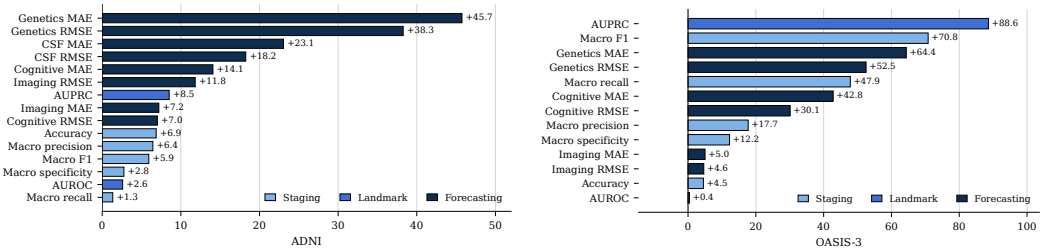


Figure 3: Relative improvement of CAMOS over the best baseline on each of the reported metrics.

Table 3: Ablation study evaluating the contribution of coupling, gating, and discretisation choices.

	Staging	Landmark Prediction		Longitudinal Forecasting	
Variants	Macro F1 ↑	AUROC ↑	AUPRC ↑	Sparse MAE ↓	Dense MAE ↓
w/o coupling ($C \equiv 0$)	$0.7561_{\pm 0.0034}$	$0.9530_{\pm 0.0021}$	$0.8193_{\pm 0.0068}$	$0.5491_{\pm 0.0057}$	$0.4354_{\pm 0.0032}$
w coupling ungated ($g^n_{kj} \equiv 1$)	$0.7694_{\pm 0.0029}$	$0.9478_{\pm 0.0026}$	$0.7916_{\pm 0.0081}$	$0.5668_{\pm 0.0063}$	$0.4487_{\pm 0.0038}$
w asymmetric gate ($g^n_{kj} = a^n_j$)	$0.7532_{\pm 0.0041}$	$0.9562_{\pm 0.0019}$	$0.8384_{\pm 0.0059}$	$0.5315_{\pm 0.0049}$	$0.4176_{\pm 0.0027}$
w IMEX discretization	$0.6530_{\pm 0.0062}$	$0.9546_{\pm 0.0024}$	$0.8063_{\pm 0.0074}$	$0.7127_{\pm 0.0089}$	$0.5057_{\pm 0.0046}$
CAMOS	$\mathbf{0.7790}_{\pm \mathbf{0.0018}}$	$\mathbf{0.9611}_{\pm \mathbf{0.0014}}$	$\mathbf{0.8687}_{\pm \mathbf{0.0041}}$	$\mathbf{0.5097}_{\pm \mathbf{0.0036}}$	$\mathbf{0.4029}_{\pm \mathbf{0.0022}}$

0.8687 to 0.8063, and raises forecasting MAE by 39.8% (sparse) and 25.5% (dense). Coupling lets each modality's state absorb cross-modal information at co-observed visits before an unobserved interval; removing it raises MAE by 7.7% (sparse) and 8.1% (dense) and lowers AUPRC to 0.8193, the two regions of largest margin. The availability gate acts selectively: removing it lowers AUROC to 0.9531 and AUPRC to 0.8296 but leaves staging (F1 0.7804) and forecasting (MAE 0.5016 sparse, 0.4014 dense) essentially unchanged. It also explains the transfer result, since the CSF oscillator is not observed in OASIS-3 and stays isolated instead of entering as a zero-filled input block.

Ablation studies. Table 3 evaluates different variants of CAMOS on ADNI. Removing cross-modal coupling ($C \equiv 0$) degrades every task, lowering staging macro F1 by 2.3 points and landmark prediction AUPRC by 4.9 points while increasing forecasting MAE by 7.7% (sparse) and 8.1% (dense), which shows that information flow between modalities is beneficial. Coupling alone is not sufficient, without gating ($g^n_{kj} \equiv 1$), staging improves slightly over the uncoupled variant, but landmark prediction and forecasting become worse than having no coupling at all (AUPRC 0.7916, MAE $+11.2\%$ sparse and $+11.4\%$ dense), plausibly because unobserved or unreliable states propagate indiscriminately across modalities. An asymmetric gate ($g^n_{kj} = a^n_j$) recovers much of this loss and is the strongest ablation on landmark prediction and forecasting, yet it yields the lowest staging F1 among the non-IMEX variants (0.7532), so only the full gating performs well on all tasks at once. Replacing the implicit discretization with an IMEX scheme causes the largest degradation, increasing forecasting MAE by 39.8% on sparse and 25.5% on dense modalities; the stronger effect on sparse modalities is consistent with their irregular observation gaps.

7 Conclusion

In this paper, we argue that missingness in multimodal clinical sequences belongs in the transition operator rather than only in the forcing term. We propose CAMOS, a coupled oscillatory state-space model whose transition operator is gated by availability. Making coupling admissible required three new results: a per-channel Gershgorin budget that keeps the effective stiffness positive definite over all 2^M availability patterns and all gaps, an energy bound that charges amplification to availability transitions rather than sequence length, and a channel-block matrix family that keeps the associative scan exact at $O(M^2)$ overhead. On ADNI, CAMOS outperforms uncoupled oscillatory SSMs and clinical fusion models on same-visit staging, landmark prediction and longitudinal forecasting. On zero-shot OASIS-3, it is the only model that avoids collapse to the majority class. The ablations attribute these gains to the implicit discretization and the coupling.

AI USE STATEMENT

Generative AI tools were used for language editing of the manuscript. They were not used to generate experimental results, numerical values, or the statements and proofs of the theoretical results, all of which were derived by the authors. The authors take full responsibility for the final content of this work, including all text, claims and artifacts.

ETHICS STATEMENT

CAMOS predicts conversion to Alzheimer's disease and assigns clinical stages, outputs that could influence trial enrolment or a person's understanding of their own prognosis. We regard it as a research instrument rather than a decision-support tool. The results here are obtained on volunteer research cohorts that are not representative of the general population. The predicted probabilities are not calibrated, and the availability gate makes the model sensitive to which tests were ordered, a variable that reflects access to care as much as clinical need. Subgroup evaluation, calibration, and an audit of performance conditional on the availability pattern are therefore prerequisites for any clinical claim. All data are used under the existing data use agreements of ADNI and OASIS-3, and no new human-subject data were collected. The authors declare no competing interests.

REPRODUCIBILITY STATEMENT

The model is fully specified: Table 4 for every hyperparameter and optimizer setting, Appendix F the algorithm in pseudocode, Appendix G the complete configuration table, and Appendix H the exact feature definitions, inclusion criteria, preprocessing and label construction. The assumptions underlying the theory are isolated as Assumptions 1 and 2, and every proof appears in Appendix C. An anonymized code repository accompanies this submission, containing the model implementation, the cohort extraction and preprocessing scripts, the training and evaluation pipeline for CAMOS and all baselines. All datasets are available to qualified researchers from their original providers.

REFERENCES

Guy E. Blelloch. Prefix sums and their applications. *Technical Report CMU-CS-90-190, School of Computer Science, Carnegie Mellon University*, 1990.

Jared Boyer, T. Konstantin Rusch, and Daniela Rus. Learning to dissipate energy in oscillatory state-space models. *arXiv preprint arXiv:2505.12171*, 2025. URL https://arxiv.org/abs/2505.12171.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018. doi: 10.1038/s41598-018-24271-9.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International Conference on Machine Learning (ICML)*, pp. 933–941, 2017.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2312.00752.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. HiPPO: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 1474–1487, 2020.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

John Hardy and Dennis J. Selkoe. The amyloid hypothesis of Alzheimer's disease: Progress and problems on the road to therapeutics. *Science*, 297(5580):353–356, 2002. doi: 10.1126/science.1072994.

Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 479–503. PMLR, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

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. doi: 10.1016/S1474-4422(09)70299-6.

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, Timothy G. Lesnick, Vernon S. Pankratz, Michael C. Donohue, and John Q. Trojanowski. Tracking pathophysiological processes in Alzheimer's disease: An updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, 12(2):207–216, 2013. doi: 10.1016/S1474-4422(12)70291-0.

Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. MMTM: Multimodal transfer module for CNN fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13289–13299, 2020.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6696–6707, 2020.

Peter M. Kogge and Harold S. Stone. A parallel algorithm for the efficient solution of a general class of recurrence equations. *IEEE Transactions on Computers*, C-22(8):786–793, 1973.

Pamela J. LaMontagne, Tammie L.S. Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medRxiv*, 2019. doi: 10.1101/2019.12.13.19014902.

Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. Coupled mamba: Enhanced multimodal fusion with coupled state space model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. URL https://openreview.net/forum?id=UXEo3uNNIX.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021.

Eric Martin and Chris Cundy. Parallelizing linear recurrent neural nets over sequence length. *arXiv preprint arXiv:1709.04057*, 2017.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning (ICML)*, pp. 26670–26698, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Ronald C. Petersen, Paul S. Aisen, Laurel A. Beckett, Michael C. Donohue, Anthony C. Gamst, Danielle J. Harvey, Clifford R. Jack, William J. Jagust, Leslie M. Shaw, Arthur W. Toga, John Q. Trojanowski, and Michael W. Weiner. Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology*, 74(3):201–209, 2010. doi: 10.1212/WNL.0b013e3181cb3e25.

Sebastian Pölsterl, Tom Nuno Wolf, and Christian Wachinger. Combining 3D image and tabular data via the dynamic affine feature map transform. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 688–698. Springer, 2021.

Yulia Rubanova, Ricky T. Q. Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

T. Konstantin Rusch and Daniela Rus. Oscillatory state-space models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=GRMfXcAAFh.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=4c0J6lwQ4_.

Jimmy T.H. Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=Ai8Hw3AXqks.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Benjamin Walker, Andrew D. McLeod, Tiexin Qin, Yichuan Cheng, Haoliang Li, and Terry Lyons. Log neural controlled differential equations: The Lie brackets make a difference. In *International Conference on Machine Learning (ICML)*, 2024.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15878–15887, 2023.

Wenfang Yao, Kejing Yin, William K. Cheung, Jia Liu, and Jing Qin. DrFuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16416–16424, 2024. doi: 10.1609/aaai.v38i15.29578.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Zhao, and Junfeng Wang. M3Care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2418–2428, 2022a.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations (ICLR)*, 2022b. URL https://openreview.net/forum?id=Kwm8I7dU-l5.

A Appendix: Notation

Symbol	Meaning	Symbol	Meaning
M	number of modalities ($M = 4$)	d	oscillators per modality
N	visits for a subject	$m = Md$	total latent width
h	shared projection width	L	number of GCO blocks
p_k	raw dimension of modality k	p_s	static covariate dimension
t_n	time of visit n (years)	Δt_n	$t_n - t_{n-1} > 0$
$u_k^n \in \mathbb{R}^{p_k}$	raw features	$\tilde{u}_k^n \in \mathbb{R}^h$	projected
$a_k^n \in \{0,1\}$	availability indicator	$a^n \in \{0,1\}^M$	availability pattern
D_a	$\mathrm{blkdiag}(a_1 I_d, \dots, a_M I_d)$	$g_{kj}^n = a_k^n a_j^n$	coupling gate
$y(t), z(t) \in \mathbb{R}^m$	position, velocity	$x = (z, y)^\top$	full state, $\mathbb{R}^{2m}$
$A_k = \mathrm{diag}(\alpha_{k,\cdot})$	natural freqs.	$C_{kj} = \mathrm{diag}(c_{kj,\cdot})$	coupling
$P = A - C$	effective stiffness	$P_n = A - D_{a^n} C D_{a^n}$	gated stiffness
$P^{(r)} \in \mathbb{R}^{M \times M}$	channel-r block	Π	channel permutation
$S_n = (I + \Delta t_n^2 P_n)^{-1}$	inverse Schur cmpl.	$\mathcal{T}_n$	transition operator
ϵ	Gershgorin budget (0.2)	δ	numerical floor (10^{-6})
$\mu = \epsilon \alpha_{\min}$	stiffness lower bound	$L_P = (2-\epsilon)\alpha_{\max}$	upper bound
$\gamma = \max_r \|C^{(r)}\|_2$	coupling strength	$H_{\max}$	forecast horizons
E_n	discrete energy	V_N	availability transitions
$\mathcal{F}$	channel-block matrix family	$\mathcal{A}$	input-gated linear layers

B Model Component Details

Static encoder. The initial condition of Section 4.1 is

$$x^0 = (z^0, y^0)^\top = W_2^s \,\mathrm{GELU}(W_1^s s + b_1^s) + b_2^s \in \mathbb{R}^{2m}, \qquad W_1^s \in \mathbb{R}^{2m \times p_s},\ W_2^s \in \mathbb{R}^{2m \times 2m}. \tag{11}$$

GCO block. Block ℓ maps $\{\tilde{u}^{(\ell-1),n}\}_{n=1}^N$ to $\{\tilde{u}^{(\ell),n}\}_{n=1}^N$ by

$$\{x^{(\ell),n}\}_n = \textsc{Scan}_\bullet\big(\{\mathcal{T}_n^{(\ell)}\}_n, \{\mathcal{T}_n^{(\ell)} F_n^{(\ell)}\}_n\big), \tag{12}$$

$$w^{(\ell),n} = C_{\mathrm{ro}}^{(\ell)}\, y^{(\ell),n} + D_{\mathrm{ro}}^{(\ell)}\, \tilde{u}^{(\ell-1),n}, \qquad C_{\mathrm{ro}} \in \mathbb{R}^{Mh \times m}, \tag{13}$$

$$\tilde{u}^{(\ell),n} = \tilde{u}^{(\ell-1),n} + \mathrm{Drop}\Big(\mathrm{GLU}\big(\mathrm{GELU}(w^{(\ell),n})\big)\Big), \tag{14}$$

with $\mathrm{GLU}(v) = \sigma(W_1 v) \odot W_2 v$ (Dauphin et al., 2017) and dropout 0.1.

Within-visit attention. With H heads and $d_H = d/H$,

$$\mathrm{Attn}(h^n)_k = W_O \Big\|_{\eta=1}^{H} \sum_{j=1}^{M} \frac{\exp\big(\langle W_Q^{(\eta)\top} h_k^n, W_K^{(\eta)\top} h_j^n \rangle / \sqrt{d_H} - \beta(1 - a_j^n)\big)}{\sum_{l=1}^{M} \exp\big(\langle W_Q^{(\eta)\top} h_k^n, W_K^{(\eta)\top} h_l^n \rangle / \sqrt{d_H} - \beta(1 - a_l^n)\big)}\, W_V^{(\eta)\top} h_j^n, \tag{15}$$

with $\beta = \mathrm{softplus}(\hat{\beta}) \geq 0$ learnable and initialized at $\beta_0 = 2$.

Task heads and forecasting loss. The three heads are

$$\hat{p}^n = \mathrm{softmax}\big(W_{\mathrm{st}} \bar{h}^n + b_{\mathrm{st}}\big), \qquad W_{\mathrm{st}} \in \mathbb{R}^{3 \times m}, \tag{16}$$

$$\hat{q} = \sigma\Big(w_{\mathrm{cv}}^\top \textstyle\sum_n \omega_n \bar{h}^n + b_{\mathrm{cv}}\Big), \qquad \omega_n = \frac{\exp(v^\top \bar{h}^n)}{\sum_{n'} \exp(v^\top \bar{h}^{n'})}, \tag{17}$$

$$\hat{u}_k^{n+h} = W_{\mathrm{fc},k}^{(h)}\, \bar{h}_k^n + b_{\mathrm{fc},k}^{(h)}, \qquad h \in \{1, \dots, H_{\max}\}, \tag{18}$$

and the masked forecasting term of equation 10 is

$$\mathcal{L}_{\mathrm{fc}} = \frac{\sum_{h=1}^{H_{\max}} \gamma_h \sum_{n,k} a_k^{n+h} \big\|\hat{u}_k^{n+h} - u_k^{n+h}\big\|_1}{\sum_{h=1}^{H_{\max}} \gamma_h \sum_{n,k} a_k^{n+h}\, p_k}, \qquad \gamma_h = 2^{-(h-1)}. \tag{19}$$

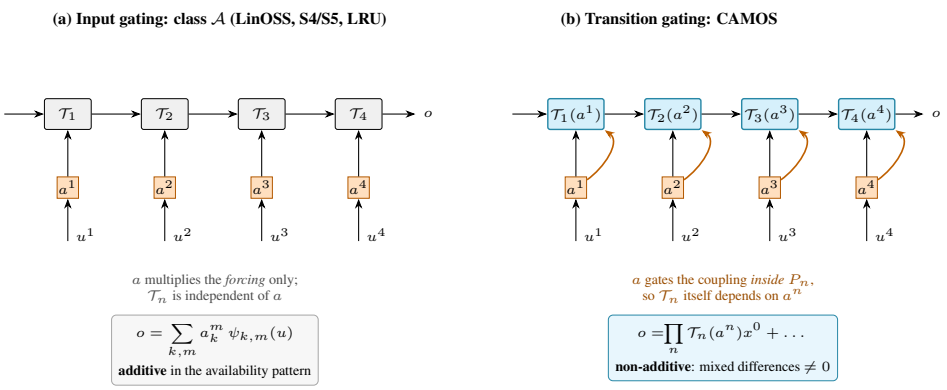


Figure 4: Input gating versus transition gating. In (a) the indicators multiply the forcing only, so the state unrolls into a sum of per-modality, per-visit contributions and every mixed availability difference vanishes. In (b) the gate enters the homogeneous part, so the state is a product of availability-dependent operators and contains genuine interactions among availability indicators.

C Proofs

C.1 Derivation of the implicit recurrence equation 8

Scheme equation 4.4 reads $\mathcal{M}_n x^n = x^{n-1} + F_n$ with $\mathcal{M}_n = \left(\begin{smallmatrix} I & \Delta t_n P_n \\ -\Delta t_n I & I \end{smallmatrix}\right)$. Direct multiplication gives

$$\begin{pmatrix} I & \Delta t_n P_n \\ -\Delta t_n I & I \end{pmatrix} \begin{pmatrix} S_n & -\Delta t_n P_n S_n \\ \Delta t_n S_n & S_n \end{pmatrix} = \begin{pmatrix} (I + \Delta t_n^2 P_n) S_n & 0 \\ 0 & (I + \Delta t_n^2 P_n) S_n \end{pmatrix} = I_{2m}, \tag{20}$$

since $(I + \Delta t_n^2 P_n) S_n = I_m$ by definition of S_n. Hence $\mathcal{T}_n = \mathcal{M}_n^{-1}$ has the stated form. □

C.2 Proof of Lemma 2

Fix $a \in [0,1]^M$ and a channel r. By Assumption 1 the matrix $P^{(r)}(a) \in \mathbb{R}^{M \times M}$ with $[P^{(r)}(a)]_{kk} = \alpha_{k,r}$ and $[P^{(r)}(a)]_{kj} = -a_k a_j c_{kj,r}$ for $k \neq j$ is symmetric, so its eigenvalues are real. By the Gershgorin circle theorem (Horn & Johnson, 2012, Thm. 6.1.1) every eigenvalue ν satisfies $|\nu - \alpha_{k,r}| \leq R_k$ for some k, with

$$R_k = \sum_{j \neq k} \left| a_k a_j c_{kj,r} \right| \leq \sum_{j \neq k} \left| c_{kj,r} \right| \leq (1 - \epsilon)\, \alpha_{k,r}, \tag{21}$$

the first inequality using $a_k, a_j \in [0,1]$ and the second Assumption 2. Hence $\nu \in [\epsilon \alpha_{k,r}, (2-\epsilon)\alpha_{k,r}]$ for some k, so $\nu \geq \mu$ and $\nu \leq L_P$. Since $\Pi P(a) \Pi^\top = \mathrm{blkdiag}(P^{(1)}(a), \dots, P^{(d)}(a))$ and permutation conjugation preserves the spectrum, the same bounds hold for $P(a)$. The eigenvalues of $I_m + \Delta t^2 P(a)$ are $1 + \Delta t^2 \nu \geq 1 + \Delta t^2 \mu > 0$, so it is symmetric positive definite and admits a Cholesky factorization. □

C.3 Proof of Proposition 3

If $a_j^n = 0$ then for every $k \neq j$ the (k,j) and (j,k) blocks of $D_{a^n} C D_{a^n}$ are $a_k^n a_j^n C_{kj} = 0$ and $a_j^n a_k^n C_{jk} = 0$. Since A is block diagonal, P_n is block diagonal with respect to the splitting $\{j\} \cup ([M] \setminus \{j\})$. The inverse of a block diagonal matrix is block diagonal, so S_n is block diagonal, and each of the four blocks of $\mathcal{T}_n$ in equation 8 is a product of S_n, P_n and scalars, hence block diagonal. Restricting to the j block gives $P_n|_{jj} = A_j$ and $S_n|_{jj} = (I_d + \Delta t_n^2 A_j)^{-1}$. Under the asymmetric gate $g_{kj} = a_j$, the (k,j) blocks vanish for all $k \neq j$ but the (j,k) blocks equal $-a_k^n C_{jk}$, which need not vanish, so P_n is block triangular and in general non-symmetric. □

C.4 Proof of Theorem 5

Let $T_{n:m+1} := T_n T_{n-1} \cdots T_{m+1}$ with $T_{n:n+1} := I$. Unrolling the recurrence of Definition 4 from $x^0 = 0$,

$$x^n = \sum_{m=1}^{n} T_{n:m+1} \sum_{k=1}^{M} a_k^m B_k^m \tilde{u}_k^m = \sum_{m=1}^{n} \sum_{k=1}^{M} a_k^m \, T_{n:m+1} B_k^m \tilde{u}_k^m, \tag{22}$$

and $o^n = Rx^n$ gives the additive form. The map $a^{1:n} \mapsto o^n$ is affine in each coordinate with no products of distinct coordinates, so for any $(k,m) \neq (j,m')$, $\Delta_{(k,m)}\Delta_{(j,m')} o^n = (\psi_{k,m} + \psi_{j,m'}) - \psi_{k,m} - \psi_{j,m'} = 0$. For the second claim, take $M = 2$, $d = 1$, $L = 1$ and consider the $(1,2)$ entry of $S_n = (I_2 + \Delta t_n^2 P_n)^{-1}$ with $P_n = \left(\begin{smallmatrix} \alpha_1 & -gc \\ -gc & \alpha_2 \end{smallmatrix}\right)$, $g = a_1^n a_2^n$, $c \neq 0$. The 2×2 inverse gives equation 7, which vanishes iff $g = 0$. Choosing an input supported only on modality 1 at visit n and reading out coordinate 2 of y^n, o^n is a nonzero multiple of $[S_n]_{12}$ when $a_1^n = a_2^n = 1$ and zero if either indicator is zero, so $\Delta_{(1,n)}\Delta_{(2,n)} o^n \neq 0$ and CAMOS is not in $\mathcal{A}$. □

Remark 8 (Scope of Theorem 5). The theorem concerns a single *linear state-space layer*. It does not say that deep networks built from such layers cannot represent availability interactions, since a nonlinear readout applied to an additive statistic is not itself additive; it says that in class $\mathcal{A}$ the interaction must be reconstructed by the readout from an additive summary, whereas CAMOS represents it inside the recurrence, in the part of the model that is scannable and carries the guarantees of Lemma 2 and Theorem 7. Nor does it separate CAMOS from representation-level fusion such as MedFuse or DrFuse, whose encoders are nonlinear in a. We state it narrowly on purpose.

C.5 Proof of Proposition 6

Since P_n is symmetric, write $P_n = U \Lambda U^\top$ with U orthogonal and $\Lambda = \mathrm{diag}(\nu_1, \ldots, \nu_m)$, $\nu_i \geq \mu > 0$ by Lemma 2. Put $\mathcal{U} = \mathrm{blkdiag}(U, U)$, which is orthogonal. Then $\mathcal{U}^\top \mathcal{M}_n \mathcal{U} = \left(\begin{smallmatrix} I & \Delta t_n \Lambda \\ -\Delta t_n I & I \end{smallmatrix}\right)$, which decouples into m independent 2×2 problems $\left(\begin{smallmatrix} 1 & \Delta t_n \nu_i \\ -\Delta t_n & 1 \end{smallmatrix}\right)$, each with inverse $s_i \left(\begin{smallmatrix} 1 & -\Delta t_n \nu_i \\ \Delta t_n & 1 \end{smallmatrix}\right)$, $s_i = (1 + \Delta t_n^2 \nu_i)^{-1}$. The eigenvalues of this inverse are $s_i(1 \pm \mathrm{i} \Delta t_n \sqrt{\nu_i})$, so $|\lambda_i^{\pm}|^2 = s_i^2(1 + \Delta t_n^2 \nu_i) = s_i \leq (1 + \Delta t_n^2 \mu)^{-1} < 1$. Since $\mathcal{T}_n = \mathcal{M}_n^{-1}$ and orthogonal conjugation preserves the spectrum, the claim follows. The argument uses only symmetry and positivity of P_n, not diagonality, which is why the uncoupled proof does not transfer. □

Proposition 9 (Exact dissipation at fixed availability)**.** *Suppose $f^n = 0$ and $P_n \equiv P$. Then $E_n := \frac{1}{2}(\|z^n\|^2 + (y^n)^\top P y^n)$ satisfies, for every $\Delta t_n > 0$,*

$$E_n - E_{n-1} = -\tfrac{1}{2}\Delta t_n^2 \big((y^n)^\top P^2 y^n + (z^n)^\top P z^n\big) \leq -\tfrac{\mu}{2}\Delta t_n^2 \big(\mu \|y^n\|^2 + \|z^n\|^2\big) \leq 0. \tag{23}$$

Proof. From equation 4.4, $z^{n-1} = z^n + \Delta t_n P y^n$ and $y^{n-1} = y^n - \Delta t_n z^n$, so

$$\|z^{n-1}\|^2 = \|z^n\|^2 + 2\Delta t_n \langle z^n, P y^n \rangle + \Delta t_n^2 (y^n)^\top P^2 y^n, \tag{24}$$

$$(y^{n-1})^\top P y^{n-1} = (y^n)^\top P y^n - 2\Delta t_n \langle z^n, P y^n \rangle + \Delta t_n^2 (z^n)^\top P z^n. \tag{25}$$

Adding and halving, the cross terms cancel and $E_{n-1} = E_n + \frac{1}{2}\Delta t_n^2 ((y^n)^\top P^2 y^n + (z^n)^\top P z^n)$. The inequality uses $P \succeq \mu I$ and hence $P^2 \succeq \mu^2 I$. □

The dissipation is negligible for the slowest mode at clinical sequence lengths, since $\prod_n (1 + \Delta t_n^2 \mu)^{-1/2} \geq \exp(-\frac{\mu}{2} \sum_n \Delta t_n^2)$ exceeds 0.90 for $N \approx 10$, $\Delta t_n \approx 1$ year and $\mu \approx 0.02$; high-frequency modes are damped more strongly at long gaps, which is the price of unconditional stability.

C.6 Proof of Theorem 7

Write $E_n^{(Q)} := \frac{1}{2}(\|z^n\|^2 + (y^n)^\top Q y^n)$, so $E_n = E_n^{(P_n)}$. Applying Proposition 9 with P_n frozen over the step from $n-1$ to n gives $E_n^{(P_n)} \leq E_{n-1}^{(P_n)}$, and

$$E_{n-1}^{(P_n)} = E_{n-1}^{(P_{n-1})} + \tfrac{1}{2}(y^{n-1})^\top (P_n - P_{n-1}) y^{n-1} \leq E_{n-1}^{(P_{n-1})} + \tfrac{1}{2}\|P_n - P_{n-1}\|_2 \, \|y^{n-1}\|^2. \tag{26}$$

By Lemma 2, $\|y^{n-1}\|^2 \le (2/\mu)E_{n-1}^{(P_{n-1})}$, so $E_n \le E_{n-1}\exp(\frac{1}{\mu}\|P_n - P_{n-1}\|_2)$, and telescoping gives the first inequality. For the second, if $a^n = a^{n-1}$ then $P_n = P_{n-1}$ and the factor is one, which also proves monotonicity; if $a^n \neq a^{n-1}$ then

$$\|P_n - P_{n-1}\|_2 \le \|D_{a^{n-1}}CD_{a^{n-1}}\|_2 + \|D_{a^n}CD_{a^n}\|_2 \le 2\gamma, \tag{27}$$

using $\|D_a\|_2 \le 1$ and $\|C\|_2 = \max_r \|C^{(r)}\|_2 = \gamma$. Summing over the V_N transition steps completes the proof. □

Proposition 10 (Input-to-state bound)**.** *With forcing f^n and $P_n \equiv P$, $\sqrt{E_n} \le \sqrt{E_{n-1}} + \sqrt{2}\,\Delta t_n\|f^n\|$.*

Proof. Repeating the computation of Proposition 9 with $z^{n-1} = z^n + \Delta t_n(Py^n - f^n)$ gives $2E_n \le 2E_{n-1} + 2\Delta t_n\|z^n\|\|f^n\|$. With $A_* = \sqrt{2E_n}$, $B_* = \sqrt{2E_{n-1}}$, $c = \Delta t_n\|f^n\|$ and $\|z^n\| \le A_*$, we get $A_*^2 \le B_*^2 + 2cA_*$, hence $A_* \le c + \sqrt{c^2 + B_*^2} \le B_* + 2c$. Dividing by $\sqrt{2}$ gives the claim. □

Proposition 11 (Cross-modal leakage and its dependence on the gap)**.** *Fix a channel r, an availability vector $a \in [0,1]^M$ (the ungated case is $a = \mathbf{1}$), and $k \neq j$. Write $s := \Delta t_n^2$ and $S_n^{(r)} = (I_M + sP_n^{(r)})^{-1}$ with $P_n^{(r)} = \mathrm{diag}(\alpha_{\cdot,r}) - D_aC^{(r)}D_a$.*

(i) Short gaps. $\big|[S_n^{(r)}]_{kj} - s\,a_ka_j\,c_{kj,r}\big| \le s^2L_P^2$, *so the leakage grows quadratically in Δt_n while $sL_P \ll 1$.*

(ii) Uniform bound. *For every $\Delta t_n > 0$,*

$$\big|[S_n^{(r)}]_{kj}\big| \le \beta(s) := \frac{1}{2}\Big(\frac{1}{1+s\mu} - \frac{1}{1+sL_P}\Big) = \frac{s\,(L_P - \mu)}{2(1+s\mu)(1+sL_P)}. \tag{28}$$

β increases for $s < (\mu L_P)^{-1/2}$ and decreases afterwards, attains its maximum $\frac{1}{2}(\sqrt{L_P} - \sqrt{\mu})/(\sqrt{L_P} + \sqrt{\mu}) < \frac{1}{2}$, and satisfies $\beta(s) \le (L_P - \mu)/(2s\mu L_P)$.

(iii) Long gaps. $\big|[S_n^{(r)}]_{kj} - s^{-1}[(P_n^{(r)})^{-1}]_{kj}\big| \le s^{-2}\mu^{-2}$, *so the leakage decays as Δt_n^{-2}.*

(iv) Exact form for $M = 2$. *With $g = a_1a_2$ and $D := \alpha_1\alpha_2 - g^2c^2 > 0$, equation 7 reads*

$$[S_n]_{12} = \frac{s\,gc}{1 + s(\alpha_1 + \alpha_2) + s^2D}, \tag{29}$$

which is unimodal in Δt_n with peak at $\Delta t_\star = D^{-1/4}$, peak value $|gc|/(\alpha_1 + \alpha_2 + 2\sqrt{D})$, and $[S_n]_{12} \sim gc/(\Delta t_n^2 D)$ as $\Delta t_n \to \infty$.

Cross-modal transfer of the previous state is therefore largest at intermediate gaps and vanishes at long ones.

Proof. Write $P := P_n^{(r)}$ and $S := S_n^{(r)}$. By Lemma 2, applied channel-wise, P is symmetric with spectrum in $[\mu, L_P]$, so S is symmetric with spectrum in $[(1+sL_P)^{-1}, (1+s\mu)^{-1}]$ and $\|S\|_2 \le 1$.

(i) From $(I + sP)S = I$ we get $S = I - sP + s^2P^2S$. For $k \neq j$ the identity contributes nothing and $[-sP]_{kj} = s\,a_ka_jc_{kj,r}$, while $|[s^2P^2S]_{kj}| \le s^2\|P\|_2^2\|S\|_2 \le s^2L_P^2$.

(ii) If X is symmetric with spectrum in $[\lambda_-, \lambda_+]$, then for $k \neq j$ and $m := (\lambda_- + \lambda_+)/2$, $|X_{kj}| = |e_k^\top(X - mI)e_j| \le \|X - mI\|_2 = (\lambda_+ - \lambda_-)/2$. Applying this to S gives equation 28. Differentiating, $\mathrm{sign}\,\beta'(s) = \mathrm{sign}\big(1 - s^2\mu L_P\big)$, which gives the monotonicity and the maximizer $s = (\mu L_P)^{-1/2}$; substituting yields the stated maximum. The decay bound follows from $(1+s\mu)(1+sL_P) \ge s^2\mu L_P$.

(iii) From $(I+sP)S = I$ we also get $S = s^{-1}P^{-1} - s^{-1}P^{-1}S$, and $\|P^{-1}S\|_2 \le \mu^{-1}(1+s\mu)^{-1} \le s^{-1}\mu^{-2}$.

(iv) Expanding the denominator of equation 7 gives equation 29. Assumption 2 gives $|c| \le (1-\epsilon)\min\{\alpha_1, \alpha_2\}$, so $g^2c^2 < \alpha_1\alpha_2$ and $D > 0$. With $f(s) = sgc/Q(s)$ and $Q(s) = 1 + s(\alpha_1 + \alpha_2) + s^2D$, we have $\mathrm{sign}\,f'(s) = \mathrm{sign}(gc)\cdot\mathrm{sign}\big(Q(s) - sQ'(s)\big) = \mathrm{sign}(gc)\cdot\mathrm{sign}(1 - s^2D)$, so $|f|$ peaks at $s = D^{-1/2}$, where $Q = 2 + (\alpha_1 + \alpha_2)D^{-1/2}$, giving the stated value. The asymptotic form follows because $Q(s) \sim s^2D$. □

Proposition 12 (Normal modes and frequency shift)**.** *Fix a channel r and an availability pattern a, and let $P^{(r)}(a) = U^{(r)}\Lambda^{(r)}U^{(r)\top}$ with $\Lambda^{(r)} = \mathrm{diag}(\nu_1^{(r)}, \dots, \nu_M^{(r)})$. Then the unforced system equation 2 restricted to channel r is a superposition of M decoupled harmonic modes with angular frequencies $\sqrt{\nu_i^{(r)}}$ and mode shapes the columns of $U^{(r)}$; the frequencies are pinned to the uncoupled ones by Weyl's inequality, $|\nu_i^{(r)} - \alpha_{(i),r}| \le \gamma \le (1-\epsilon)\alpha_{\max}$; and the channel transfer function is $H^{(r)}(s) = (s^2 I_M + P^{(r)}(a))^{-1}$, whose cross-modal entry $[H^{(r)}(s)]_{kj}$ vanishes for $k \neq j$ when modalities k and j lie in different connected components of the gated graph $P^{(r)}(a)$, and is nonzero for generic couplings when they lie in the same component. The entries $|c_{kj,r}|$ therefore define d weighted modality graphs whose connected components are the sets of modalities that exchange information in that frequency band.*

Proof. (i) With $f = 0$, substituting $y = U^{(r)}\zeta$ into the channel-r restriction of equation 3 gives $\zeta'' = -\Lambda^{(r)}\zeta$, which is M decoupled scalar oscillators with angular frequency $\sqrt{\nu_i^{(r)}}$, all $\nu_i^{(r)} > 0$ by Lemma 2. (ii) $P^{(r)}(a) = \mathrm{diag}(\alpha_{\cdot,r}) - D_a C^{(r)} D_a$ is a symmetric perturbation of a diagonal matrix by a symmetric matrix of spectral norm at most γ, so Weyl's inequality (Horn & Johnson, 2012, Thm. 4.3.1) gives the stated bound. (iii) Laplace transforming $y'' = -P^{(r)}(a)y + \tilde{f}$ with zero initial conditions gives $H^{(r)}(s) = (s^2 I + P^{(r)}(a))^{-1}$. If k and j are disconnected in $P^{(r)}(a)$, then $s^2 I + P^{(r)}(a)$ is block diagonal and so is its inverse, giving $[H]_{kj} = 0$; conversely, if they are connected by a path $k = i_0, \dots, i_q = j$, the Neumann series $(s^2 I + P)^{-1} = s^{-2}\sum_{q\ge 0}(-s^{-2}P)^q$ contains the monomial $P_{i_0 i_1} \cdots P_{i_{q-1} i_q} \neq 0$ at order q, and for generic couplings the entry is a nonzero rational function of s. □

D Parallel Evaluation, Complexity and Expressivity

Equation equation 8 is a linear recurrence with time-varying coefficients, so it is solvable by the associative scan with binary operator

$$(T_1, v_1) \bullet (T_2, v_2) := \big(T_2 T_1,\ T_2 v_1 + v_2\big), \tag{30}$$

which is associative (Kogge & Stone, 1973; Blelloch, 1990; Martin & Cundy, 2017; Smith et al., 2023). Applying its inclusive scan to $[(\mathcal{T}_1, \mathcal{T}_1 F_1), \dots, (\mathcal{T}_N, \mathcal{T}_N F_N)]$ returns $x^1, \dots, x^N$ in the second components in $\lceil \log_2 N \rceil$ sequential rounds. Associativity is immediate; what is not immediate, and what the channel factorization buys, is that the scan never leaves a cheap structural family.

Definition 13 (Channel-block family)**.** Let $\mathcal{F} \subset \mathbb{R}^{2m\times 2m}$ be the set of matrices whose four $m \times m$ blocks T_{ab} satisfy $\Pi T_{ab} \Pi^\top = \mathrm{blkdiag}(T_{ab}^{(1)}, \dots, T_{ab}^{(d)})$ with $T_{ab}^{(r)} \in \mathbb{R}^{M\times M}$.

Lemma 14 (Closure)**.** *$\mathcal{F}$ is closed under addition and multiplication, contains I_{2m}, and contains $\mathcal{T}_n$ for every n and every availability pattern. A product of two elements of $\mathcal{F}$ costs $8dM^3$ multiply-accumulate operations, against $8m^3 = 8d^3M^3$ for dense $2m \times 2m$ blocks.*

Proof. Let $T, T' \in \mathcal{F}$. Conjugating by $\mathrm{blkdiag}(\Pi, \Pi)$ maps each of the four $m \times m$ blocks of T and T' to block-diagonal matrices with d blocks of size $M \times M$. Block-diagonal matrices with a common partition are closed under addition and multiplication, and each block of the product TT' is a sum of two products of such matrices, hence again block diagonal with the same partition; therefore $TT' \in \mathcal{F}$. Additivity and $I_{2m} \in \mathcal{F}$ are immediate. That $\mathcal{T}_n \in \mathcal{F}$ follows from equation 8: S_n and P_n are channel-block-diagonal by equation 4 and Lemma 2, and every block of $\mathcal{T}_n$ is a scalar multiple of S_n or of $P_n S_n$. For the cost, forming TT' requires 8 products of $m \times m$ channel-block-diagonal matrices, each of which is d independent $M \times M$ products at M^3 multiply-accumulates, giving $8dM^3$; for dense blocks the same 8 products cost $m^3 = d^3M^3$ each. □

Proposition 15 (Complexity and parameter count)**.** *One GCO layer requires $\Theta(NdM^3)$ arithmetic work and $\lceil \log_2 N \rceil$ sequential depth, with $\Theta(NdM^2)$ memory for the scan carries. An uncoupled oscillatory layer of the same total state width $m = Md$ requires $\Theta(Nm)$ work, so the overhead of coupling is a factor $\Theta(M^2)$, independent of N and d, and the within-visit attention adds $\Theta(NM^2d)$ at constant depth. Of the $\Theta(M^2h^2 + M^2dh)$ parameters per layer, only $\binom{M}{2}d$ belong to the coupling: for $M = 4$ and $d = h = 64$ that is 384 out of 2.8×10^5, about 0.14%.*

Proof. By equation 8 each $S_n^{(r)}$ is an $M \times M$ symmetric positive definite inverse costing $\Theta(M^3)$ by Cholesky, and over d channels and N visits this is $\Theta(NdM^3)$. Each $\mathcal{T}_n$ lies in $\mathcal{F}$ by Lemma 14, whose products cost $\Theta(dM^3)$; the associative scan performs $\Theta(N)$ such operations at depth $\lceil \log_2 N \rceil$, giving $\Theta(NdM^3)$ work and $\Theta(\log N)$ depth. Each scan carry stores $4d$ blocks of size $M \times M$, hence $\Theta(NdM^2)$ memory. For the uncoupled model, $C = 0$ makes every $S_n^{(r)}$ diagonal, so all products cost $\Theta(dM) = \Theta(m)$ and the total is $\Theta(Nm)$. The attention block equation 15 computes $M \times M$ scores at each of N visits with d-dimensional values, hence $\Theta(NM^2d)$ at constant depth.

For the parameter count, the free parameters of one GCO layer are $\hat{\alpha} \in \mathbb{R}^{M \times d}$; $\hat{c}$ on the strict upper triangle in (k, j) for each channel, giving $\binom{M}{2}d$; the forcing matrices $B_k \in \mathbb{R}^{d \times h}$ for $k \in [M]$, giving Mdh; the readout $C_{\mathrm{ro}} \in \mathbb{R}^{Mh \times m}$ and skip $D_{\mathrm{ro}} \in \mathbb{R}^{Mh \times Mh}$, giving $Mhm + M^2h^2$; and the two GLU matrices in $\mathbb{R}^{Mh \times Mh}$, giving $2M^2h^2$. Since $m = Md$, we have $Mhm = M^2dh$, so the total is $Md + \binom{M}{2}d + M^2dh + Mdh + 3M^2h^2 = \Theta(M^2h^2 + M^2dh)$. Substituting $M = 4$ and $d = h = 64$ gives $256 + 384 + 65536 + 16384 + 196608 = 279{,}168 \approx 2.8 \times 10^5$, of which the $\binom{4}{2} \cdot 64 = 384$ coupling parameters are 0.14%. □

Expressivity is not lost. Setting $\hat{c} \equiv 0$ in equation 5 gives $C = 0$ and reduces a CAMOS layer to an uncoupled oscillatory layer with strictly positive diagonal state matrix, so CAMOS blocks inherit universality within continuous causal operators. This certifies only that Assumptions 1 and 2 cost nothing in expressivity, and it is weak here because a symmetric coupled system is orthogonally equivalent to an uncoupled one when everything is observed. The representational gain is therefore due to gating (Theorem 5), since the gate acts in the modality basis and does not commute with the diagonalizing rotation $U^{(r)}$ of Proposition 12.

E Dense and Low-Rank Couplings

If $C_{kj} \in \mathbb{R}^{d \times d}$ is dense rather than diagonal, P does not block-diagonalize under Π and three things change.

Stability. Assumption 2 is replaced by a spectral condition. If $C = C^\top$ and $\|C\|_2 \leq (1 - \epsilon)\lambda_{\min}(A)$, then for every $a \in [0, 1]^M$, $\lambda_{\min}(P(a)) \geq \lambda_{\min}(A) - \|C\|_2 \geq \epsilon\lambda_{\min}(A) > 0$ using $\|D_a\|_2 \leq 1$, and Lemma 2 and everything downstream go through with $\mu := \epsilon\lambda_{\min}(A)$. Enforcing this requires spectral normalization of C, hence a power iteration at every training step rather than the closed form equation 5. It also interacts badly with the standard $\alpha \sim \mathcal{U}([0, 1])$ initialization, since $\lambda_{\min}(A)$ is then near zero and the admissible coupling collapses.

Cost. S_n becomes a dense $m \times m$ inverse costing $\Theta(d^3M^3)$, and the scan combine leaves $\mathcal{F}$ and requires dense $2m \times 2m$ products at the same order. The overhead relative to the channel-factored model is $\Theta(d^2)$, about a factor of 4000 per layer at $d = 64$.

Why low rank does not help. The parameterization $C_{kj} = Q_{kj}Q_{kj}^\top$ with $Q_{kj} \in \mathbb{R}^{d \times \rho}$ reduces the parameter count from d^2 to $d\rho$ per block but not the scan cost, because low-rank-plus-diagonal structure is not closed under multiplication: if $T = D_1 + U_1V_1^\top$ and $T' = D_2 + U_2V_2^\top$ with $\operatorname{rank} U_i = \rho$, then $TT' = D_1D_2 + D_1U_2V_2^\top + U_1V_1^\top D_2 + U_1(V_1^\top U_2)V_2^\top$ has rank up to 3ρ in its non-diagonal part, so the carried operator's rank triples at every combine and reaches $N^{\log_2 3}\rho$ after a full scan. Woodbury inversion is available at each individual step but does not survive the scan. This is worth stating because a low-rank PSD coupling is the natural first choice, and it is where intuition from static fusion layers fails in a sequential model.

Intermediate option. Partitioning the d channels into d/ρ groups with dense coupling inside each group interpolates between the extremes: $\mathcal{F}$ becomes block-diagonal matrices with d/ρ blocks of size ρM, the cost becomes $\Theta(Nd\rho^2M^3)$, and $\rho = 1$ recovers the channel-factored model.

Algorithm 1 One gated coupled oscillator (GCO) layer, forward pass

Require: projected inputs $\{\tilde{u}_k^n\}$, availability $\{a_k^n\}$, gaps $\{\Delta t_n\}$, initial state x^0, parameters $\hat{\alpha}, \hat{c}, \{B_k\}$, budget ϵ, floor δ
Ensure: states $\{x^n = (z^n, y^n)\}_{n=1}^N$
$\alpha_{k,r} \leftarrow \mathrm{softplus}(\hat{\alpha}_{k,r})$ for all k, r ▷ $\Theta(Md)$
$\sigma_{k,r} \leftarrow \sum_{l \neq k} |\hat{c}_{kl,r}|$; $c_{kj,r} \leftarrow \dfrac{(1-\epsilon)\min\{\alpha_{k,r}, \alpha_{j,r}\}}{\max\{\sigma_{k,r}, \sigma_{j,r}, \delta\}}\, \hat{c}_{kj,r}$ ▷ Assumptions 1, 2 hold by construction
for $n = 1, \ldots, N$ **in parallel do**
  for $r = 1, \ldots, d$ **in parallel do**
    $[P_n^{(r)}]_{kk} \leftarrow \alpha_{k,r}$; $[P_n^{(r)}]_{kj} \leftarrow -\,a_k^n a_j^n\, c_{kj,r}$ for $k \neq j$ ▷ symmetric product gate
    $S_n^{(r)} \leftarrow \textsc{CholSolve}\big(I_M + \Delta t_n^2 P_n^{(r)},\, I_M\big)$ ▷ SPD by Lemma 2; $\Theta(M^3)$
  end for
  assemble $\mathcal{T}_n$ from $\{S_n^{(r)}\}, \{P_n^{(r)}\}$ by equation 8
  $f^n \leftarrow \sum_k a_k^n\, \iota_k B_k \tilde{u}_k^n$; $v_n \leftarrow \mathcal{T}_n(\Delta t_n f^n,\, 0)^\top$
end for
$\{x^n\} \leftarrow \textsc{AssocScan}_\bullet\big([(\mathcal{T}_1, v_1 + \mathcal{T}_1 x^0), (\mathcal{T}_2, v_2), \ldots, (\mathcal{T}_N, v_N)]\big)$ ▷ $\lceil \log_2 N \rceil$ depth, $\Theta(NdM^3)$ work
return $\{x^n\}$

Algorithm 2 CAMOS forward pass for one subject

Require: $\{u_k^n\}$, $\{a_k^n\}$, $\{t_n\}$, static s
Ensure: $\{\hat{p}^n\}$, $\hat{q}$, $\{\hat{u}_k^{n+h}\}$
$\Delta t_n \leftarrow t_n - t_{n-1}$ for all n
$\tilde{u}_k^{(0),n} \leftarrow E_k u_k^n + e_k$ if $a_k^n = 1$, else 0 ▷ equation 1, parallel in n
$x^0 \leftarrow W_2^s\, \mathrm{GELU}(W_1^s s + b_1^s) + b_2^s$ ▷ equation 11
for $\ell = 1, \ldots, L$ **do**
  $\{x^{(\ell),n}\} \leftarrow$ Algorithm 1 with inputs $\{\tilde{u}^{(\ell-1),n}\}$ and layer-ℓ parameters
  $w^{(\ell),n} \leftarrow C_{\mathrm{ro}}^{(\ell)} y^{(\ell),n} + D_{\mathrm{ro}}^{(\ell)} \tilde{u}^{(\ell-1),n}$
  $\tilde{u}^{(\ell),n} \leftarrow \tilde{u}^{(\ell-1),n} + \mathrm{Drop}(\mathrm{GLU}(\mathrm{GELU}(w^{(\ell),n})))$
end for
$h_k^n \leftarrow \iota_k^\top y^{(L),n}$; $\bar{h}_k^n \leftarrow h_k^n + \mathrm{Attn}(h^n)_k$ ▷ equation 15, parallel in n
$\hat{p}^n, \hat{q}, \hat{u}_k^{n+h} \leftarrow$ heads equation 16, equation 17, equation 18

F Algorithms

G Implementation Details

All models are implemented in PyTorch 2.3 (Paszke et al., 2019) and trained on a single NVIDIA A100 40GB GPU, with the associative scan implemented as a Blelloch up-sweep and down-sweep over batched $2M \times 2M$ channel blocks and the Cholesky solves of equation 8 batched over (n, r) in float32. Table 4 lists the complete configuration; values marked "grid" are selected on the validation split, and the identical grid and trial budget are applied to every baseline so that no model receives a larger search. The validation selection criterion is the unweighted mean of staging macro-F1, landmark prediction AUROC and one minus the normalized forecasting MAE. Staging is scored by accuracy and macro-averaged F1, precision, recall and specificity; landmark prediction by AUROC and AUPRC; forecasting by per-modality MAE and RMSE in the normalized feature space, averaged over observed targets only.

The frequency shift $\alpha_0 > 0$ in Table 4 matters: the standard oscillatory initialization $\alpha \sim \mathcal{U}([0, 1])$ places mass near zero, and the admissible coupling in equation 5 is proportional to $\min\{\alpha_{k,r}, \alpha_{j,r}\}$, so a large fraction of channels would begin with no coupling budget and never recover.

Table 4: Complete CAMOS configuration.

Group	Setting	Value
Architecture	GCO blocks L	grid $\{2, 4, 6\}$
	oscillators per modality d	grid $\{32, 64, 128\}$
	projection width h	grid $\{32, 64, 128\}$
	attention heads H	4
	dropout	0.1
	static encoder	2-layer MLP, width $2m$, GELU
Coupling	budget ϵ	0.2
	numerical floor δ	10^{-6}
	gate	hard, $g^n_{kj} = a^n_k a^n_j$
	structure	channel-factored, symmetric
Initialization	$\alpha_{k,r}$	$\mathcal{U}([\alpha_0, \alpha_0 + 1])$, $\alpha_0 = 0.1$
	$\hat{c}_{kj,r}$	$\mathcal{N}(0, 1/M)$, symmetrized
	attention bias β_0	2.0, learnable via $\mathrm{softplus}$
	remaining weights	PyTorch defaults
Optimization	optimizer	AdamW, $\beta = (0.9, 0.999)$, $\varepsilon = 10^{-8}$
	weight decay	10^{-2}, excluding $\hat{\alpha}$, $\hat{c}$, biases
	peak learning rate	grid $\{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}\}$
	schedule	cosine, 5% linear warmup
	batch size	64 subjects
	gradient clipping	global norm 1.0
	epochs / early stopping	max 200, patience 20
Objective	$(\lambda_{\mathrm{st}}, \lambda_{\mathrm{cv}}, \lambda_{\mathrm{fc}})$	$(1,\ 0.5,\ 1)$
	λ_{reg}	10^{-3} (ℓ_1 on $\hat{c}$)
	forecasting horizons	$H_{\max} = 3$, discount $\gamma_h = 2^{-(h-1)}$
	class weighting	inverse frequency on the training split
Protocol	split	subject-level 70/15/15, stratified
	seeds	5
	selection criterion	mean of staging macro-F1, landmark prediction AUROC, $1-$ normalized forecasting MAE
	precision	float32 throughout, including Cholesky

Numerical notes. The Cholesky factorizations in equation 8 are the only place where conditioning could be a concern. By Lemma 2, $I_M + \Delta t_n^2 P_n^{(r)}$ has condition number at most $(1 + \Delta t_n^2 L_P)/(1 + \Delta t_n^2 \mu)$, bounded above by $L_P/\mu = (2 - \epsilon)\,\mathrm{cond}(A)/\epsilon$ independently of Δt_n; with $\epsilon = 0.2$ and the initialization of Table 4 this is below 200, comfortably inside float32. This is a practical dividend of the budget: an unconstrained coupling would allow $I + \Delta t^2 P$ to become near-singular precisely at long gaps, where the model is most needed.

Baseline adaptation. LinOSS-IM and LinOSS-IMEX receive the concatenation $[u^n_1; \ldots; u^n_M] \in \mathbb{R}^p$ with unobserved entries set to zero, the availability vector a^n appended as M channels, and Δt_n appended as one channel and used as the per-step discretization interval. MedFuse and DrFuse are given per-modality gated recurrent encoders over the visit sequence, so that they see the same temporal information as the SSMs rather than a single pooled vector.

H Dataset Construction

ADNI extraction. From the ADNIMERGE table we retain the columns listed in Section 5 together with the visit code, examination date, baseline diagnosis and current diagnosis. Visits are ordered by examination date and t_n is expressed in years from the subject's baseline visit. A subject is included if it has at least three visits with a recorded diagnosis and, at each retained visit, at least one modality fully observed. Volumetric MRI measures are divided by intracranial volume before normalization, which removes head-size variation that would otherwise dominate the leading principal direction. All continuous features are then robust z-scored, $u \mapsto (u - \mathrm{med})/\mathrm{IQR}$, with statistics computed

on the training split only and applied unchanged to validation and test; the robust form is preferred because several ADNI biomarkers, CSF amyloid, are censored at assay limits and have heavy tails. Categorical static covariates are one-hot encoded, and *APOE4* is kept as its integer allele count in $\{0, 1, 2\}$, which encodes the known dose response.

Table 5: Cohort statistics. M: number of modalities; $p = \sum_k p_k$: total raw feature dimension; missing rate: percentage of modality-visit pairs unobserved.

Dataset	Role	M	p	Subjects	Visits	Mean visits/subj.	Max visits/subj.	Missing (%)
ADNI	train / val / test	4	19	2749	7432	2.7	15	32.5
OASIS-3	zero-shot	3	19	1377	8522	6.2	20	28

Availability convention. $a_k^n = 1$ if and only if every feature of modality k is recorded at visit n. Per-feature masking was rejected because acquisition is organized by procedure: a lumbar puncture yields the whole CSF panel or none of it, and a PET session yields the whole tracer panel or none of it. Modelling availability at the procedure level is therefore both more faithful and what makes the modality-level gate of Section 4.3 meaningful.

Label construction. Staging labels are the recorded diagnosis at each visit, mapped to $\{\mathrm{CN}, \mathrm{MCI}, \mathrm{AD}\}$; visits without a recorded diagnosis contribute to the forecasting loss but not to $\mathcal{L}_{\mathrm{st}}$. For landmark prediction, the index visit is the first visit at which a subject is diagnosed MCI; the label is 1 if AD is recorded at any visit within three years of the index visit, and 0 if the subject has at least three years of subsequent follow-up without an AD diagnosis. Subjects censored before three years without converting are excluded from the landmark prediction task only, so no conversion label is ever imputed. Reversions from MCI to CN are retained in the staging task with their recorded label and do not affect the conversion label unless a later AD diagnosis occurs within the window.

OASIS-3 mapping. For OASIS-3 (LaMontagne et al., 2019) we use the same modalities and, where a feature has no exact ADNI counterpart, the nearest available measure of the same construct, refitting only the normalization statistics; the model is otherwise transferred without retraining, so this is a zero-shot evaluation.

I Extended Results

In this section, we report all the baselines on both cohorts, across the four families: uncoupled oscillatory SSMs (LinOSS-IM, LinOSS-IMEX, D-LinOSS), general and coupled SSMs (S5, Mamba, Coupled Mamba), irregular-time and missing-value models (GRU-D, mTAN, Raindrop, ODE-RNN), and clinical fusion models (MedFuse, DrFuse). All entries are means over five seeds on the subject-level splits of Appendix H.

I.1 Staging with all the baselines

Table 6 extends the staging comparison of Section 6. On ADNI the strongest baselines come from three different families, DrFuse (0.7322 accuracy), LinOSS-IMEX (0.7206) and Mamba (0.7187), so no single design axis accounts for baseline performance, and CAMOS exceeds all twelve on every metric. On OASIS-3 the picture changes: accuracy is compressed between 0.6263 and 0.7525 because one class dominates, while macro F1 separates the models from 0.2805 to 0.5557. Every baseline falls below macro F1 0.44 and macro recall 0.44, with DrFuse at exactly $1/3$ recall and $2/3$ specificity, the signature of predicting a single class. CAMOS is the only model that retains discrimination among the three stages under transfer.

I.2 Landmark prediction with all the baselines

Table 7 extends Table 2, and Figure 5 shows the corresponding ROC and precision-recall curves. The two metrics behave differently in each cohort. On ADNI, AUROC spans 0.8795 to 0.9611 while

Table 6: Same-visit CN/MCI/AD staging on ADNI (in-distribution) and OASIS-3 (zero-shot).

Models	Accuracy ↑	Macro F1 ↑	Macro Prec. ↑	Macro Recall ↑	Macro Spec. ↑
ADNI					
LinOSS-IM	$0.6593_{\pm 0.0041}$	$0.6815_{\pm 0.0039}$	$0.6804_{\pm 0.0045}$	$0.6864_{\pm 0.0037}$	$0.8147_{\pm 0.0021}$
LinOSS-IMEX	$0.7206_{\pm 0.0058}$	$0.7234_{\pm 0.0061}$	$0.7394_{\pm 0.0055}$	$0.7424_{\pm 0.0059}$	$0.8513_{\pm 0.0030}$
D-LinOSS	$0.6820_{\pm 0.0324}$	$0.6882_{\pm 0.0271}$	$0.6912_{\pm 0.0294}$	$0.6975_{\pm 0.0281}$	$0.8392_{\pm 0.0152}$
S5	$0.6562_{\pm 0.0226}$	$0.6696_{\pm 0.0196}$	$0.6731_{\pm 0.0207}$	$0.6804_{\pm 0.0196}$	$0.8215_{\pm 0.0118}$
Mamba	$0.7187_{\pm 0.0203}$	$0.7304_{\pm 0.0163}$	$0.7386_{\pm 0.0174}$	$0.7412_{\pm 0.0169}$	$0.8561_{\pm 0.0097}$
Coupled Mamba	$0.7002_{\pm 0.0282}$	$0.7052_{\pm 0.0209}$	$0.7127_{\pm 0.0233}$	$0.7169_{\pm 0.0214}$	$0.8447_{\pm 0.0126}$
GRU-D	$0.6192_{\pm 0.0356}$	$0.6251_{\pm 0.0391}$	$0.6318_{\pm 0.0372}$	$0.6402_{\pm 0.0365}$	$0.8043_{\pm 0.0198}$
mTAN	$0.6931_{\pm 0.0182}$	$0.7037_{\pm 0.0165}$	$0.7094_{\pm 0.0171}$	$0.7158_{\pm 0.0163}$	$0.8489_{\pm 0.0092}$
Raindrop	$0.6776_{\pm 0.0174}$	$0.6902_{\pm 0.0172}$	$0.6953_{\pm 0.0182}$	$0.7021_{\pm 0.0176}$	$0.8374_{\pm 0.0101}$
ODE-RNN	$0.6619_{\pm 0.0194}$	$0.6682_{\pm 0.0172}$	$0.6729_{\pm 0.0184}$	$0.6817_{\pm 0.0179}$	$0.8268_{\pm 0.0105}$
MedFuse	$0.6648_{\pm 0.0072}$	$0.6697_{\pm 0.0076}$	$0.6564_{\pm 0.0069}$	$0.7139_{\pm 0.0070}$	$0.8284_{\pm 0.0034}$
DrFuse	$0.7322_{\pm 0.0049}$	$0.7354_{\pm 0.0046}$	$0.7488_{\pm 0.0044}$	$0.7683_{\pm 0.0042}$	$0.8599_{\pm 0.0024}$
CAMOS	$\mathbf{0.7824}_{\pm 0.0016}$	$\mathbf{0.7790}_{\pm 0.0018}$	$\mathbf{0.7970}_{\pm 0.0022}$	$\mathbf{0.7786}_{\pm 0.0025}$	$\mathbf{0.8836}_{\pm 0.0013}$
OASIS-3					
LinOSS-IM	$0.7271_{\pm 0.0017}$	$0.3253_{\pm 0.0053}$	$0.5334_{\pm 0.0082}$	$0.3541_{\pm 0.0025}$	$0.6773_{\pm 0.0014}$
LinOSS-IMEX	$0.7231_{\pm 0.0022}$	$0.3027_{\pm 0.0064}$	$0.4222_{\pm 0.0094}$	$0.3414_{\pm 0.0028}$	$0.6756_{\pm 0.0016}$
D-LinOSS	$0.7022_{\pm 0.0343}$	$0.3754_{\pm 0.0376}$	$0.4127_{\pm 0.0412}$	$0.3896_{\pm 0.0284}$	$0.6948_{\pm 0.0163}$
S5	$0.7525_{\pm 0.0290}$	$0.4047_{\pm 0.0396}$	$0.4518_{\pm 0.0437}$	$0.4102_{\pm 0.0315}$	$0.7051_{\pm 0.0172}$
Mamba	$0.7345_{\pm 0.0262}$	$0.4312_{\pm 0.0492}$	$0.4863_{\pm 0.0521}$	$0.4389_{\pm 0.0406}$	$0.7194_{\pm 0.0208}$
Coupled Mamba	$0.7237_{\pm 0.0636}$	$0.3594_{\pm 0.0667}$	$0.3924_{\pm 0.0703}$	$0.3781_{\pm 0.0452}$	$0.6890_{\pm 0.0231}$
GRU-D	$0.7259_{\pm 0.0494}$	$0.3731_{\pm 0.0363}$	$0.4035_{\pm 0.0398}$	$0.3852_{\pm 0.0317}$	$0.6926_{\pm 0.0159}$
mTAN	$0.7371_{\pm 0.0376}$	$0.2931_{\pm 0.0110}$	$0.2846_{\pm 0.0137}$	$0.3409_{\pm 0.0068}$	$0.6705_{\pm 0.0034}$
Raindrop	$0.6263_{\pm 0.0377}$	$0.3411_{\pm 0.0298}$	$0.3627_{\pm 0.0312}$	$0.3597_{\pm 0.0243}$	$0.6798_{\pm 0.0121}$
ODE-RNN	$0.7154_{\pm 0.0403}$	$0.3330_{\pm 0.0124}$	$0.3468_{\pm 0.0154}$	$0.3502_{\pm 0.0097}$	$0.6751_{\pm 0.0048}$
MedFuse	$0.6682_{\pm 0.0032}$	$0.3176_{\pm 0.0059}$	$0.2871_{\pm 0.0091}$	$0.3558_{\pm 0.0026}$	$0.6844_{\pm 0.0019}$
DrFuse	$0.7263_{\pm 0.0023}$	$0.2805_{\pm 0.0081}$	$0.2421_{\pm 0.0086}$	$0.3333_{\pm 0.0000}$	$0.6667_{\pm 0.0000}$
CAMOS	$\mathbf{0.7600}_{\pm 0.0014}$	$\mathbf{0.5557}_{\pm 0.0028}$	$\mathbf{0.6280}_{\pm 0.0031}$	$\mathbf{0.5261}_{\pm 0.0030}$	$\mathbf{0.7682}_{\pm 0.0015}$

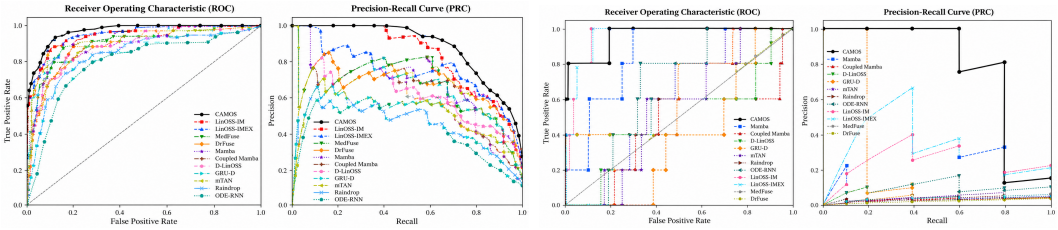


Figure 5: ROC and PR curves for 3-year MCI-to-AD landmark prediction on ADNI (left) and OASIS-3 (right).

AUPRC spans 0.6187 to 0.8687, and the ordering is not preserved: mTAN is the strongest baseline on AUPRC (0.7750) but only third on AUROC. On OASIS-3 the gap widens, with baseline AUPRC between 0.0801 and 0.5982 at AUROC between 0.7896 and 0.9560, indicating few positive cases in the transfer cohort. Baseline standard deviations there reach ± 0.188 on AUPRC, against ± 0.009 for CAMOS, so the transfer margin should be read alongside that variance.

I.3 Longitudinal forecasting with all the baselines

Tables 8 and 9 extend Figure 2 to all baselines. On ADNI, the baselines are tightly clustered on the two sparse modalities, within 0.112 MAE on CSF and, aside from S5, Mamba and Coupled Mamba, within 0.091 on genetics, while CAMOS separates by 0.115 and 0.106 from the best of them. On OASIS-3 the clustering is tighter still: eleven of twelve baselines lie within 0.041 MAE of each other on imaging and within 0.039 on clinical measures, which suggests they largely reproduce the last observed value. CAMOS departs from that cluster on clinical measures (0.3756 against a baseline

Table 7: 3-year MCI-to-AD conversion on ADNI (in-distribution) and OASIS-3 (zero-shot).

	ADNI		**OASIS-3**	
Models	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
LinOSS-IM	$0.9223_{\pm 0.0031}$	$0.7102_{\pm 0.0087}$	$0.9533_{\pm 0.0024}$	$0.3588_{\pm 0.0142}$
LinOSS-IMEX	$0.9307_{\pm 0.0027}$	$0.7389_{\pm 0.0074}$	$0.9560_{\pm 0.0029}$	$0.4185_{\pm 0.0163}$
D-LinOSS	$0.9142_{\pm 0.0187}$	$0.7022_{\pm 0.0721}$	$0.8236_{\pm 0.0412}$	$0.1134_{\pm 0.0481}$
S5	$0.9087_{\pm 0.0163}$	$0.6925_{\pm 0.06201}$	$0.9412_{\pm 0.0265}$	$0.5982_{\pm 0.1889}$
Mamba	$0.8795_{\pm 0.0204}$	$0.6187_{\pm 0.0686}$	$0.8418_{\pm 0.0537}$	$0.1260_{\pm 0.1104}$
Coupled Mamba	$0.8913_{\pm 0.0142}$	$0.6463_{\pm 0.0423}$	$0.8307_{\pm 0.0581}$	$0.1177_{\pm 0.1189}$
GRU-D	$0.9128_{\pm 0.0196}$	$0.7014_{\pm 0.0678}$	$0.8654_{\pm 0.0476}$	$0.1825_{\pm 0.1054}$
mTAN	$0.9346_{\pm 0.0172}$	$0.7750_{\pm 0.0667}$	$0.8372_{\pm 0.0503}$	$0.1224_{\pm 0.1031}$
Raindrop	$0.9061_{\pm 0.0158}$	$0.6909_{\pm 0.0584}$	$0.7896_{\pm 0.0348}$	$0.0801_{\pm 0.0322}$
ODE-RNN	$0.9215_{\pm 0.0211}$	$0.7250_{\pm 0.0752}$	$0.8519_{\pm 0.0392}$	$0.1308_{\pm 0.0676}$
MedFuse	$0.9134_{\pm 0.0045}$	$0.6902_{\pm 0.0102}$	$0.8467_{\pm 0.0058}$	$0.2594_{\pm 0.0121}$
DrFuse	$0.9386_{\pm 0.0022}$	$0.7689_{\pm 0.0069}$	$0.8787_{\pm 0.0047}$	$0.3183_{\pm 0.0135}$
CAMOS	$\mathbf{0.9611_{\pm 0.0014}}$	$\mathbf{0.8687_{\pm 0.0041}}$	$\mathbf{0.9600_{\pm 0.0018}}$	$\mathbf{0.7894_{\pm 0.0093}}$

Table 8: Next-visit ($h = 1$) forecasting on ADNI per modality.

	Cognition ↓		CSF ↓		Genetics ↓		Imaging ↓	
Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
LinOSS-IM	0.4471	0.7704	0.7715	1.0832	0.8270	0.9662	0.4468	0.8457
LinOSS-IMEX	0.4343	0.7526	0.7676	1.0795	0.8183	0.9607	0.5567	0.8053
D-LinOSS	0.5579	0.8736	0.7651	1.0724	0.7802	0.9731	0.6017	0.8612
S5	0.4973	0.8012	0.7591	1.0689	0.5147	0.9712	0.5945	0.8547
Mamba	0.4932	0.7894	0.7056	1.0180	0.5964	0.9480	0.5646	0.8273
Coupled Mamba	0.4976	0.7948	0.7139	1.0518	0.6166	0.9688	0.5764	0.8394
GRU-D	0.5695	0.8891	0.7349	1.0612	0.8011	0.9804	0.5867	0.8481
mTAN	0.5728	0.8824	0.7517	1.0703	0.8000	0.9779	0.5668	0.8325
Raindrop	0.6561	0.9967	0.8179	1.1386	0.7823	0.9856	0.6781	0.9412
ODE-RNN	0.5578	0.8649	0.7471	1.0655	0.7360	0.9694	0.5771	0.8436
MedFuse	0.4199	0.7136	0.7873	1.1452	0.7644	0.9743	0.4538	0.7828
DrFuse	0.4176	0.6586	0.7895	1.0443	0.7535	0.9610	0.5562	0.7832
CAMOS	**0.3588**	**0.6124**	**0.5906**	**0.8537**	**0.4088**	**0.5930**	**0.4146**	**0.6902**

minimum of 0.6569) and genetics (0.2847 against 0.7711), and only marginally on imaging (0.7333 against 0.7705).

Table 9: Next-visit ($h = 1$) forecasting on OASIS-3 per modality.

	Imaging ↓		Cognition ↓		Genetics ↓	
Model	MAE	RMSE	MAE	RMSE	MAE	RMSE
LinOSS-IM	0.7748	0.9968	0.6719	0.9356	0.7993	1.0398
LinOSS-IMEX	0.7745	0.9966	0.6712	0.9352	0.8011	1.0408
D-LinOSS	0.7721	0.9937	0.6789	0.9387	0.7904	1.0341
S5	0.8116	1.0342	0.6916	0.9512	0.7840	1.0296
Mamba	0.7866	1.0081	0.6867	0.9458	0.7711	1.0213
Coupled Mamba	0.7874	1.0094	0.6959	0.9547	0.8026	1.0419
GRU-D	0.7846	1.0057	0.6854	0.9441	0.8054	1.0437
mTAN	0.7705	0.9921	0.6749	0.9369	0.8048	1.0429
Raindrop	0.8968	1.1273	0.8194	1.0836	0.9028	1.1542
ODE-RNN	0.7902	1.0126	0.6775	0.9372	0.7974	1.0385
MedFuse	0.7721	0.9925	0.6765	0.9194	0.8035	1.0377
DrFuse	0.7729	0.9929	0.6569	0.9161	0.8158	1.0446
CAMOS	**0.7333**	**0.9464**	**0.3756**	**0.6401**	**0.2847**	**0.4927**

Table 10: Ablation study evaluating the contribution of coupling, gating, and discretisation choices on OASIS-3.

	Staging	Landmark Prediction		Longitudinal Forecasting		
Variants	Macro F1 ↑	AUROC ↑	AUPRC ↑	Img. MAE ↓	Clin. MAE ↓	Gen. MAE ↓
w/o coupling ($C \equiv 0$)	$0.5422_{\pm 0.0285}$	$0.9678_{\pm 0.0203}$	$0.7677_{\pm 0.0802}$	$0.7647_{\pm 0.0197}$	$0.3947_{\pm 0.0097}$	$0.3214_{\pm 0.0056}$
w coupling ungated ($g^n_{kj} \equiv 1$)	$0.5341_{\pm 0.0339}$	$0.9602_{\pm 0.0350}$	$0.7324_{\pm 0.0991}$	$0.7798_{\pm 0.0201}$	$0.4075_{\pm 0.0116}$	$0.3392_{\pm 0.0076}$
w asymmetric gate ($g^n_{kj} = a^n_j$)	$0.5493_{\pm 0.0341}$	$0.9661_{\pm 0.0361}$	$0.7586_{\pm 0.1121}$	$0.7702_{\pm 0.0201}$	$0.3989_{\pm 0.0116}$	$0.3298_{\pm 0.0076}$
w IMEX discretization	$0.4880_{\pm 0.0425}$	$0.9535_{\pm 0.0308}$	$0.7069_{\pm 0.1782}$	$0.8173_{\pm 0.0255}$	$0.6110_{\pm 0.0265}$	$0.7032_{\pm 0.0131}$
CAMOS	$\mathbf{0.5573}_{\pm \mathbf{0.0260}}$	$\mathbf{0.9707}_{\pm \mathbf{0.0210}}$	$\mathbf{0.7777}_{\pm \mathbf{0.0626}}$	$\mathbf{0.7614}_{\pm \mathbf{0.0190}}$	$\mathbf{0.3913}_{\pm \mathbf{0.0111}}$	$\mathbf{0.3210}_{\pm \mathbf{0.0045}}$

I.4 Ablation studies on OASIS-3

Table 10 repeats the ablations of Section 6 under zero-shot transfer, and the ordering matches ADNI. The IMEX discretization is the most damaging change, lowering macro F1 from 0.5573 to 0.4880 and raising forecasting MAE by 7.3%, 56.1% and 119.1% on imaging, clinical measures and genetics. Removing the gate is the next most damaging, and removing coupling or replacing the symmetric product by the asymmetric gate costs less. Two qualifications apply. The margins between the full model and the three non-IMEX variants are smaller than the seed standard deviations, which reach ± 0.034 on macro F1 and ± 0.112 on AUPRC, so only the discretization effect is resolved at this sample size. The asymmetric gate breaks the symmetry required by Lemma 2, but we observed no divergence, so that guarantee acts as a safeguard rather than a practical necessity at this scale.