MSSD: MULTIMODAL FUSION AS SHARED-STATE DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sequence models commonly organize temporal processing around modality specific representations and introduce cross source interaction through separate fusion mechanisms. This leaves recurrent memory either partitioned by source or shared only after source observations have already been combined. We study an alternative organization in which heterogeneous observations remain distinct at the interface but act on one evolving multimodal memory. Multimodal Shared-State Dynamics (MSSD) realizes this view with a single modality free matrix state shared by all sources. Source conditioned write and read factors preserve distinct observation interfaces, while Dual-Axis dynamics jointly evolve latent feature and temporal memory structure within the same recurrent state. The same formulation accommodates different observation times without changing the shared recurrent organization. Across four text audio visual benchmarks, MSSD performs competitively with general and task specific multimodal models, while controlled comparisons and representation analyses support the roles of shared recurrent storage and Dual-Axis dynamics. MSSD thus formulates multimodal fusion as the evolution of one shared recurrent memory rather than as an auxiliary interaction stage around separate source trajectories.

1 INTRODUCTION

Multimodal sequence models differ in where cross source interaction enters temporal modeling. Early fusion combines source representations before sequence processing, while recurrent fusion, cross modal attention, bottleneck models, and multimodal state space models preserve source organized representations and introduce interaction during or after temporal processing (Zadeh et al., 2017; Liu et al., 2018; Zadeh et al., 2018a; Liang et al., 2018; Tsai et al., 2019; Nagrani et al., 2021). Figure 1 illustrates two recurring organizations. Separate stream models retain modality specific temporal trajectories and connect them through an interaction mechanism, while joint input models combine the sources before a common recurrent backbone. In one case recurrent storage remains indexed by source, while in the other source separation is removed before recurrent evolution.

The closest alternatives already move interaction toward a common memory. Bottleneck and Perceiver models provide shared latent workspaces (Nagrani et al., 2021; Jaegle et al., 2021), RAM Networks maintain a common hidden memory updated through sensor specific recurrent operators (Gehrig et al., 2021), and recent multimodal state space models couple modality specific state chains or combine modality specific processing with shared interaction modules (Li et al., 2024; He et al., 2025; Ma et al., 2025). These designs share information across sources, but the recurrent organization still associates state evolution with modality specific trajectories, modality specific subspaces, or source dependent transition operators. The remaining gap is therefore not whether modalities can interact, but whether source identity must remain part of recurrent storage itself.

Multimodal Shared State Dynamics (MSSD) starts from the opposite choice. Source identity determines how an observation writes to and queries memory, while recurrent storage is a single modality free state governed by common dynamics. We represent this shared memory as a matrix with latent feature and temporal memory axes, allowing information from different sources to interact through the evolution of the recurrent state itself. Dual Axis dynamics jointly evolve both axes of this matrix, combining temporal memory evolution with structured dynamics over the latent feature dimension. The temporal component builds on the rank-R multi input, multi output formulation of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

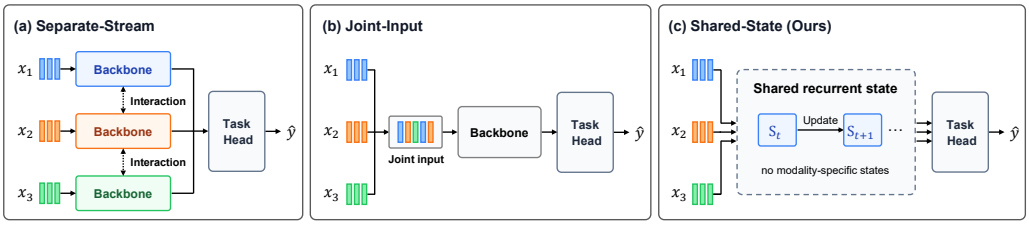


Figure 1: Representative recurrent organizations for multimodal sequence modeling. Separate streams preserve modality-specific temporal trajectories and introduce interaction between them. Joint input models combine sources before a common backbone. MSSD keeps the source interfaces distinct while all sources update one recurrent state without a modality-specific state dimension.

Mamba-3 (Lahoti et al., 2026). MSSD adapts its recurrent interface so that multiple sources retain distinct write and read factors while acting on the same state and common transition. This preserves the MIMO state space machinery while making it compatible with the shared multimodal recurrent organization. Different observation times use the same source interfaces and do not require modality specific recurrent memories.

The empirical study follows the same decomposition. Benchmark comparisons test the shared state backbone across four datasets, controlled architecture comparisons isolate the effect of recurrent organization, and component studies separate shared storage from Dual-Axis dynamics. Source removal and linear probes then examine interaction inside the learned shared representation, while observation schedule perturbations evaluate the held drive extension. MSSD introduces a shared-state formulation of multimodal sequence modeling in which fusion is realized through the evolution of a single modality-free recurrent state, with Dual-Axis dynamics providing its structured state-space realization.

2 Related Work

2.1 Multimodal Sequence Modeling and Fusion

Multimodal sequence models place interaction at different stages of representation and temporal processing. Tensor Fusion Network and Low-rank Multimodal Fusion combine modality representations through explicit high-order operators (Zadeh et al., 2017; Liu et al., 2018). Recurrent approaches such as MFN and RMFN maintain modality-specific temporal streams and introduce cross-view memory or recurrent fusion (Zadeh et al., 2018a; Liang et al., 2018). MulT performs directional cross-modal attention over unaligned sequences (Tsai et al., 2019), while Multimodal Bottleneck Transformers and Perceiver architectures mediate interaction through bottleneck or shared latent representations (Nagrani et al., 2021; Jaegle et al., 2021). These methods establish a broad design space in which cross-source interaction is introduced around modality-organized representations or through intermediate latent spaces.

Recent multimodal state-space models bring recurrent state dynamics directly into this design space. Coupled Mamba couples modality state chains while retaining their intra-modality processes (Li et al., 2024). MSAmba combines intra-modal state-space blocks with a dedicated cross-modal interaction module (He et al., 2025). MSLMamba learns modality-specific subspaces and uses a shared state matrix for cross-modal interaction (Ma et al., 2025). Shared recurrent and latent dynamical states also appear in RAM Networks and multimodal Deep Markov models (Gehrig et al., 2021; Zhi-Xuan et al., 2020). MSSD organizes these roles differently by placing source identity at the write and read interfaces while using one modality-free recurrent state under common transition dynamics.

2.2 State Space and Continuous-Time Sequence Models

Structured state-space models provide efficient recurrent mechanisms for sequence modeling. S4 develops structured linear state-space layers for long sequences (Gu et al., 2022), Mamba introduces input-dependent selective dynamics (Gu & Dao, 2024), and Mamba-2 develops state-space duality

and efficient algorithms for the resulting recurrence (Dao & Gu, 2024). Mamba-3 extends this line with more expressive discretization, complex-valued state updates, and a multi-input, multi-output formulation (Lahoti et al., 2026). MSSD uses this temporal state-space machinery inside a shared multimodal state and adds structured dynamics along its latent-feature axis.

Irregular and incomplete sequences have been modeled through decay-based recurrence, continuous latent dynamics, and controlled differential equations (Che et al., 2018; Rubanova et al., 2019; Kidger et al., 2020). Incomplete multimodal time series have also been modeled through a shared latent dynamical state (Zhi-Xuan et al., 2020). MSSD uses source-conditioned active and held drives inside the same recurrent dynamics that perform multimodal fusion, allowing the shared state to evolve when source observation times differ.

3 Multimodal Shared-State Dynamics

The core modeling choice is to separate source identity from recurrent storage. Source specific interfaces determine how observations write to and query memory, while the recurrent state itself is shared across sources. MSSD instantiates the temporal part of this state with the rank-R multi input, multi output formulation of Mamba-3 (Lahoti et al., 2026). Using the orientation adopted here, a MIMO recurrence can be written schematically as

$$S_t = \bar{A}_t(S_{t-1}) + X_t B_t^\top, \qquad Y_t = S_t C_t, \tag{1}$$

where $S_t \in \mathbb{R}^{P\times N}$, $X_t \in \mathbb{R}^{P\times R}$, and $B_t, C_t \in \mathbb{R}^{N\times R}$. MSSD keeps this low rank write and read interface and changes its multimodal organization so that all sources act on one evolving state.

3.1 Shared-State Formulation

Consider M heterogeneous sources. Source m provides observations

$$\mathcal{X}_m = \{(x_{m,k}, \tau_{m,k})\}_{k=1}^{T_m}. \tag{2}$$

We construct an ordered event sequence from the union of all source timestamps. At event t, $\mathcal{A}_t$ contains sources with a newly arriving observation and $\mathcal{H}_t$ contains previously observed sources whose latest observation remains valid. Reads query the current state without modifying it.

After lightweight input adapters, all sources interact with a shared matrix valued recurrent state

$$S_t \in \mathbb{R}^{H\times P\times N}, \tag{3}$$

where H denotes heads, P the latent feature axis, and N the temporal memory axis. The state has no modality dimension. Let $h_{t,m} = \mathcal{A}_m(x_{t,m}) \in \mathbb{R}^D$ denote the adapted representation of source m. The source interface is parameterized by learned projection maps $\Pi_\bullet$ applied to $h_{t,m}$. Suppressing the head index, we write

$$\begin{aligned} X_{t,m} &= \mathrm{mat}_{P\times R}(\Pi_X h_{t,m}), & B_{t,m} &= \mathrm{mat}_{N\times R}(\Pi_B h_{t,m}),\\ C_{t,m} &= \mathrm{mat}_{N\times R}(\Pi_C h_{t,m}), & z_{t,m} &= \Pi_z h_{t,m},\\ \kappa_{t,m} &= \Pi_\kappa h_{t,m}, & r^\Delta_{t,m} &= \Pi_\Delta h_{t,m}. \end{aligned} \tag{4}$$

Here $X_{t,m}$ and $B_{t,m}$ define the rank-R write, $C_{t,m}$ defines the read direction, $z_{t,m}$ gates the MIMO readout, and $\kappa_{t,m}$ controls endpoint interpolation. The projected temporal control $r^\Delta_{t,m}$ is combined across the valid source interfaces. Let $\mathcal{V}_t$ denote those interfaces and let Γ_Δ be a permutation invariant aggregation operator. We define

$$\bar{r}^\Delta_t = \Gamma_\Delta\left(\{r^\Delta_{t,m} : m \in \mathcal{V}_t\}\right), \qquad d_t = \mathrm{softplus}(\bar{r}^\Delta_t + b_\Delta). \tag{5}$$

The event level recurrence is

$$S_t = \mathcal{T}_t(S_{t-1}) + \sum_{m\in\mathcal{A}_t} W^{\mathrm{act}}_{t,m} + \sum_{m\in\mathcal{H}_t} W^{\mathrm{held}}_{t,m} \tag{6}$$

with one transition $\mathcal{T}_t$ acting on the common state. Sources remain distinct at the observation interface, while their information interacts through changes to the same recurrent memory.

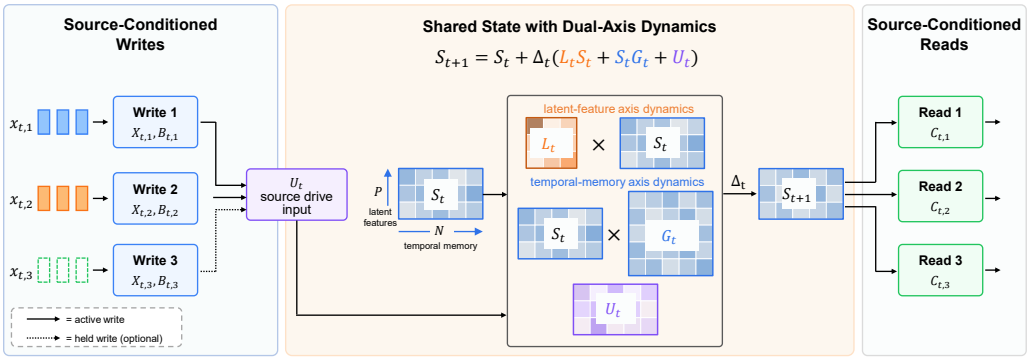


Figure 2: MSSD architecture. Source-conditioned writes drive one shared matrix state. Dual-Axis dynamics evolve the latent-feature and temporal-memory axes of that state, and source-conditioned reads obtain different views of the resulting shared memory. The state equation in the diagram is a first-order schematic of the continuous generator in Eq. 7; the model uses the structured exponential propagation and discrete update defined in Eqs. 11–19. Held writes use the latest valid source observation when no new sample arrives.

3.2 Dual-Axis Shared Dynamics

MSSD represents latent-feature structure and temporal memory along the two axes of the shared matrix state. For one head, the continuous system is

$$\frac{dS(\tau)}{d\tau} = L_t S(\tau) + S(\tau) G_t + U(\tau), \tag{7}$$

where $L_t \in \mathbb{R}^{P\times P}$ acts along the latent-feature axis, $G_t \in \mathbb{R}^{N\times N}$ governs temporal memory, and $U(\tau)$ contains source-driven input. For fixed dynamics over an interval of duration Δ, the homogeneous solution is

$$S(\tau+\Delta) = e^{\Delta L_t} S(\tau) e^{\Delta G_t}. \tag{8}$$

Equivalently, vectorization gives the joint generator

$$\frac{d\ \mathrm{vec}(S)}{d\tau} = \left(I_N \otimes L_t + G_t^\top \otimes I_P\right) \mathrm{vec}(S) + \mathrm{vec}(U), \tag{9}$$

so the two axes define one linear system rather than separate recurrent branches.

The state generators are also input conditioned. Let $q^{\mathrm{time}}_{t,m} = \Pi_{\mathrm{time}} h_{t,m}$ denote temporal transition features and let $q^{\mathrm{lat}}_{t,m} = \Pi_{\mathrm{lat}} h_{t,m}$ denote latent axis features. Their source aggregated controls are

$$\bar{q}^{\mathrm{time}}_t = \Gamma_{\mathrm{time}}\left(\{q^{\mathrm{time}}_{t,m} : m \in \mathcal{V}_t\}\right), \qquad \bar{q}^{\mathrm{lat}}_t = \Gamma_{\mathrm{lat}}\left(\{q^{\mathrm{lat}}_{t,m} : m \in \mathcal{A}_t\}\right). \tag{10}$$

The temporal rate and phase are functions of $\bar{q}^{\mathrm{time}}_t$, while the damping and rotation parameters of the latent generator are functions of $\bar{q}^{\mathrm{lat}}_t$.

The temporal memory component follows the selective state space dynamics of the Mamba lineage (Gu & Dao, 2024; Dao & Gu, 2024; Lahoti et al., 2026). Let $g_t = \tau_t - \tau_{t-1}$ denote the physical interval between consecutive events. The common controller scales that interval through

$$\Delta_t = g_t d_t, \qquad ADT_t = \Delta_t \lambda^{\mathrm{time}}_t, \qquad \alpha_t = e^{ADT_t}, \tag{11}$$

where $\lambda^{\mathrm{time}}_t < 0$ is the selective temporal rate. The temporal rate and phase are produced by input conditioned transition projections and are shared by all source writes at the event. MSSD retains the complex temporal basis of Mamba-3 and incorporates its phase into the MIMO write and read directions (Lahoti et al., 2026).

The second generator acts along the latent feature axis. The P real coordinates are grouped into $J = P/2$ pairs and pair j evolves under

$$L^{(j)}_t = \begin{bmatrix} -a_{t,j} & -\omega_{t,j} \\ \omega_{t,j} & -a_{t,j} \end{bmatrix}, \tag{12}$$

where $a_{t,j} \geq 0$ controls damping and $\omega_{t,j}$ controls rotation. With $Z_{t,j,n} = S_{t,2j,n} + iS_{t,2j+1,n}$, the corresponding complex rate is $\mu_{t,j} = -a_{t,j} + i\omega_{t,j}$. Because both axes evolve over the same effective interval, the historical state propagates as

$$Z^{\text{hist}}_{t,j} = \exp(ADT_t - \Delta_t a_{t,j} + i\Delta_t \omega_{t,j}) \odot Z_{t-1,j}. \tag{13}$$

Temporal decay and latent-feature dynamics thus act jointly on the shared state. For the updates below, let $\mathcal{E}_t(U) = e^{\Delta_t L_t} U$ denote latent-feature propagation over the current interval.

3.3 Source Drives and Shared-State Readout

The projections in Eq. (4) determine how each observation interacts with the shared memory. The write factors $X_{t,m} \in \mathbb{R}^{P\times R}$ and $B_{t,m} \in \mathbb{R}^{N\times R}$ form the rank-R update

$$\mathcal{U}(B, X) = XB^\top = \sum_{r=1}^{R} X_{:,r} B_{:,r}^\top. \tag{14}$$

For a newly observed source, MSSD retains the endpoint discretization used by the Mamba-3 temporal recurrence (Lahoti et al., 2026). With interpolation coefficient $\rho_{t,m} = \sigma(\kappa_{t,m})$, define

$$\beta_{t,m} = (1 - \rho_{t,m})\Delta_t \alpha_t, \qquad \gamma_{t,m} = \rho_{t,m}\Delta_t. \tag{15}$$

Let $\bar{X}_m$ denote the cached input factor and let $B^{\text{prev}}_{t,m}$ and $B^{\text{curr}}_{t,m}$ denote the previous and current write directions in the corresponding temporal bases. The active write is

$$W^{\text{act}}_{t,m} = \mathcal{E}_t\big[\beta_{t,m}\mathcal{U}(B^{\text{prev}}_{t,m}, \bar{X}_m)\big] + \gamma_{t,m}\mathcal{U}(B^{\text{curr}}_{t,m}, X_{t,m}). \tag{16}$$

Every active source contributes to the same recurrent update, and simultaneous observations are summed before the state is advanced.

If source m receives no new observation but its latest value remains valid, the cached observation continues to drive the state over the current interval. Let U_m denote the drive induced by its cached factors and define $\mathcal{A}_t[S] = L_t S + SG_t$. The accumulated contribution is

$$W^{\text{held}}_{t,m} = \int_0^{\Delta_t} e^{(\Delta_t - s)\mathcal{A}_t}[U_m]ds = \Delta_t \phi_1(\Delta_t \mathcal{A}_t)[U_m], \tag{17}$$

where $\phi_1(z) = (e^z - 1)/z$ and $\phi_1(0) = 1$ (Hochbruck & Ostermann, 2010). The structured operators admit this integral without introducing a separate recurrent state for held sources. The complete discrete state update can be written as

$$S^{\text{hist}}_t = \mathcal{E}_t(\alpha_t \odot S_{t-1}), \tag{18}$$

$$S_t = S^{\text{hist}}_t + \sum_{m\in\mathcal{A}_t} W^{\text{act}}_{t,m} + \sum_{m\in\mathcal{H}_t} W^{\text{held}}_{t,m}. \tag{19}$$

This is the concrete form of Eq. (6) used by the structured recurrence.

A source conditioned read contracts the shared state with $C_{t,m}$

$$u_{t,m,r,p} = \sum_{n=1}^{N} S_{t,p,n} C_{t,m,n,r}. \tag{20}$$

The projected gate $z_{t,m} \in \mathbb{R}^P$ modulates this read before rank reduction. Following the Mamba-3 MIMO output path (Lahoti et al., 2026), let $D \in \mathbb{R}^P$ denote the direct path and let $Z^{\text{MIMO}}, O^{\text{MIMO}} \in \mathbb{R}^{R\times P}$ denote learned rank mixing parameters. The direct term is present only when source m is active and the output is

$$y_{t,m,p} = \sum_{r=1}^{R} O^{\text{MIMO}}_{r,p} \left[(u_{t,m,r,p} + \mathbf{1}[m \in \mathcal{A}_t] D_p X_{t,m,p,r}) \,\text{SiLU}\big(z_{t,m,p} Z^{\text{MIMO}}_{r,p}\big)\right]. \tag{21}$$

The source index identifies the requesting interface rather than a modality-specific recurrent state.

3.4 Synchronized and Unimodal Reductions

For synchronized multimodal inputs, all available sources are active at the same event and Eq. (19) sums their writes into one state update without imposing a source order. A unimodal special case recovers the temporal structure inherited from Mamba-3. With $M = 1$, $\mathcal{A}_t = \{1\}$, $\mathcal{H}_t = \varnothing$, unit event spacing, and $L_t = 0$, the aggregation in Eq. (5) reduces to the single source controller and $\mathcal{E}_t$ becomes the identity. Equation (19) then reduces to

$$S_t = \alpha_t \odot S_{t-1} + \beta_t \mathcal{U}(B_t^{\mathrm{prev}}, \bar{X}) + \gamma_t \mathcal{U}(B_t^{\mathrm{curr}}, X_t), \tag{22}$$

with the first endpoint term omitted before a predecessor is available. Together with Eq. (21), this has the rank-R MIMO endpoint form of the Mamba-3 temporal backbone under the state orientation used here. The multimodal contribution is the shared organization of this recurrence together with the additional Dual Axis dynamics.

4 Experiments

4.1 Experimental Setup

We evaluate MSSD on CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018b), CH-SIMS (Yu et al., 2020), and CH-SIMS v2 (Liu et al., 2022) using their standard splits and released text, audio, and visual features. MOSI and MOSEI use aligned inputs, while CH-SIMS and CH-SIMS v2 use the released unaligned settings. We optimize the models with AdamW (Loshchilov & Hutter, 2019) using MSE loss and select hyperparameters on the validation split. Every model trained in our experiments is evaluated over three random seeds with one fixed configuration for each model and dataset. We report MAE, Pearson correlation, and binary classification accuracy in the main tables. The appendix reports the complete metric set, run statistics, and dataset specific configurations. The experiments follow the structure of the method. We first test whether the shared state backbone is competitive across datasets, then isolate recurrent organization under matched settings, and finally test the two mechanism level claims that sources interact through the same recurrent representation and that held drives extend this recurrence across observation gaps.

4.2 Benchmark Effectiveness

Table 1: Benchmark comparison across four datasets. General backbone rows report our three seed means. Task specific rows use the public MMSA benchmark repository (THUIAR, 2020). Ours is repeated within each comparison block for direct reading. Bold values mark the best result within the corresponding block. Acc2 is reported in percent.

	CMU-MOSI			CMU-MOSEI			CH-SIMS			CH-SIMS v2		
Model	MAE	Corr	Acc2	MAE	Corr	Acc2	MAE	Corr	Acc2	MAE	Corr	Acc2
General sequence backbones												
Joint BiLSTM	0.745	**0.792**	83.64	0.545	0.772	84.94	0.415	0.602	80.31	0.301	**0.704**	79.30
Joint Transformer	0.732	0.786	84.00	0.546	0.772	84.75	0.423	0.598	80.23	0.312	0.679	77.76
Joint Mamba-3	0.736	0.789	83.74	0.541	0.771	84.83	0.419	0.597	79.58	0.300	0.695	79.50
Perceiver style	0.741	0.787	83.84	0.544	0.771	84.72	0.418	0.601	79.65	0.311	0.677	77.76
Ours	**0.721**	0.784	**85.26**	**0.535**	**0.774**	**85.91**	**0.406**	**0.616**	**80.38**	**0.298**	0.702	**79.68**
Task specific multimodal models from MMSA												
TFN	0.947	0.673	79.08	0.573	0.714	81.89	0.432	0.591	78.38	0.303	0.707	80.14
LMF	0.950	0.651	79.18	0.576	0.717	83.48	0.441	0.576	77.77	0.367	0.557	74.18
MFN	0.927	0.670	78.87	0.573	0.718	82.86	0.435	0.582	77.90	0.295	0.727	**81.14**
MulT	0.880	0.702	80.98	0.559	0.733	84.63	0.453	0.564	78.56	**0.290**	**0.738**	80.68
Self-MM	**0.708**	0.796	**85.46**	0.531	0.765	85.15	0.425	0.595	80.04	0.311	0.695	79.69
TETFN	**0.708**	**0.798**	85.37	0.537	0.770	86.21	0.420	0.577	**81.18**	0.310	0.695	79.73
CENet	0.725	0.795	85.21	**0.526**	**0.777**	**86.38**	0.471	0.539	77.90	0.310	0.699	79.56
Ours	0.721	0.784	85.26	0.535	0.774	85.91	**0.406**	**0.616**	80.38	0.298	0.702	79.68

Table 1 separates controlled general backbones from public task specific MMSA results (Zadeh et al., 2017; Liu et al., 2018; Zadeh et al., 2018a; Tsai et al., 2019; Yu et al., 2021; Wang et al., 2023a;b).

Table 2: Controlled comparison of recurrent organization under matched dataset settings. Results are means over three seeds.

	CMU-MOSI			CMU-MOSEI			CH-SIMS			CH-SIMS v2		
Model	MAE	Corr	Acc2	MAE	Corr	Acc2	MAE	Corr	Acc2	MAE	Corr	Acc2
Independent + concat	0.739	0.792	83.94	0.545	0.770	85.02	0.410	0.614	79.58	0.294	**0.712**	79.21
Independent + cross attn	0.748	0.790	84.04	0.545	0.770	84.62	0.421	**0.617**	79.65	**0.294**	0.709	**79.69**
Joint Mamba-3	0.736	0.789	83.74	0.541	0.771	84.83	0.419	0.597	79.58	0.300	0.695	79.50
Shared temporal only	0.726	**0.798**	84.35	0.538	0.772	85.34	0.411	0.614	79.94	0.308	0.680	78.53
Full MSSD	**0.721**	0.784	**85.26**	**0.535**	**0.774**	**85.91**	**0.406**	0.616	**80.38**	0.298	0.702	79.68

Table 3: Component analysis on three datasets. The first block compares the shared temporal state with Dual Axis dynamics. The second block compares three treatments of previously observed source drives.

	CMU-MOSI			CMU-MOSEI			CH-SIMS		
Variant	MAE	Corr	Acc2	MAE	Corr	Acc2	MAE	Corr	Acc2
Dual Axis dynamics									
Temporal only	0.726	0.798	84.35	0.538	0.772	85.34	0.411	0.614	79.94
Full MSSD	0.721	0.784	85.26	0.535	0.774	85.91	0.406	0.616	80.38
Held source updates									
Active only	0.723	0.787	84.30	0.535	0.774	85.52	0.421	0.598	79.07
Naive persistence	0.732	0.785	84.70	0.533	0.776	85.64	0.410	0.610	79.94
Analytic held	0.721	0.784	85.26	0.535	0.774	85.91	0.406	0.616	80.38

Against the general backbones, Ours gives the lowest MAE and highest Acc2 on all four datasets. The task specific block places these results in the range of specialized sentiment models, with the strongest relative result on CH-SIMS.

4.3 Isolating Recurrent Organization

Table 2 isolates recurrent organization under matched dataset settings. The variants move from separate modality trajectories, through explicit cross attention and joint recurrence, to a shared temporal state and finally the full Dual Axis model.

Sharing recurrent storage gives the clearest change on MOSI, MOSEI, and CH-SIMS, where the shared temporal variant improves over Joint Mamba-3 and Full MSSD improves again in MAE and Acc2. Explicit cross attention between separate recurrent streams does not reproduce this pattern. CH-SIMS v2 has a different ordering across organizations, while Dual Axis dynamics still improve the shared temporal variant.

4.4 Component Contributions

Dual Axis dynamics improve MAE and Acc2 over the temporal only state on the three displayed datasets and follow the same direction on CH-SIMS v2 in Table 2. The held variants are closer on clean data, which motivates evaluating them under explicit observation gaps in Figure 5.

4.5 Interaction Inside the Shared State

The shared state formulation implies that a write from one source can alter a representation queried through another source interface. We measure this property by removing one source and computing the relative change in every source conditioned read. For removed source s and queried read q, we use

$$I_{s\to q} = \frac{\mathrm{mean}\|h_q(x) - h_q(x_{\setminus s})\|_2}{\mathrm{mean}\|h_q(x)\|_2 + \epsilon}. \tag{23}$$

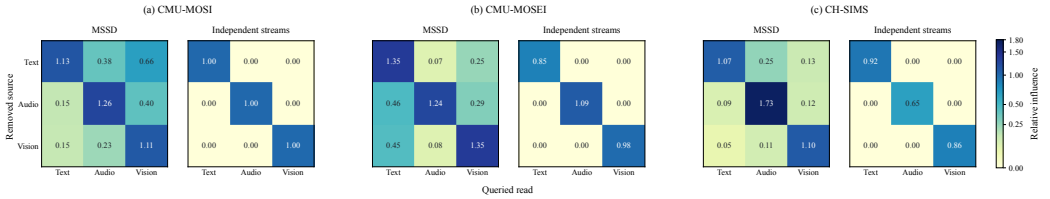


Figure 3: Cross source influence through the recurrent representation on CMU-MOSI, CMU-MOSEI, and CH-SIMS. Rows denote the removed source and columns denote the queried source conditioned read. Each dataset pairs MSSD with an independent stream control before fusion.

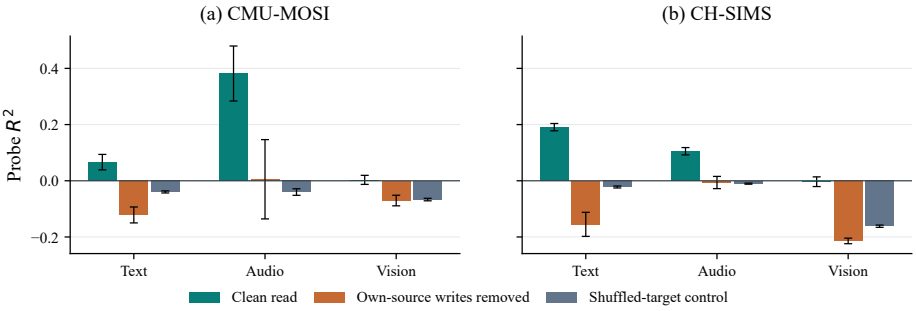


Figure 4: Linear modality probes on CMU-MOSI and CH-SIMS. Held out R^2 is shown for clean reads, reads with the corresponding source writes removed, and shuffled target controls. Error bars show variation across probe runs.

Figure 3 shows nonzero off diagonal influence for every source pair in MSSD across all three datasets, while the independent stream controls remain zero before fusion. The pattern is dataset dependent, with larger effects on MOSI, strong audio and vision influence on the MOSEI text read, and smaller but persistent effects on CH-SIMS. Information written through one interface therefore changes representations queried through another interface of the same recurrent state.

We next examine whether information associated with an individual source remains accessible after shared processing. Linear probes are trained on source conditioned reads and evaluated with held out R^2. Each source is evaluated using the clean representation, a representation obtained after removing its own writes, and a shuffled target control.

Figure 4 shows positive linear probe performance for text and audio in the clean representations of both datasets, while removing the corresponding source writes reduces the scores toward the shuffled target control. Together with Figure 3, the result shows that cross source interaction and source conditioned accessibility coexist in the shared recurrent representation.

4.6 Observation Schedule Extension

We evaluate held updates on MOSI and CH-SIMS under synchronous sparsification, temporal shift, and random observation drop. Analytic held integration is compared with active only updates and naive persistence.

Figure 5 separates the update rules most clearly under sparsification and observation drop. On MOSI, analytic held reaches 0.745 MAE at $k = 8$ versus 0.757 for active only and 0.770 for naive persistence, and CH-SIMS shows the same ordering under the corresponding degradations. Temporal shift produces smaller differences.

Figure 6 shows the corresponding write norms. Active writes occur at new arrivals, held contributions appear between arrivals after caching, and both enter the same evolving shared state.

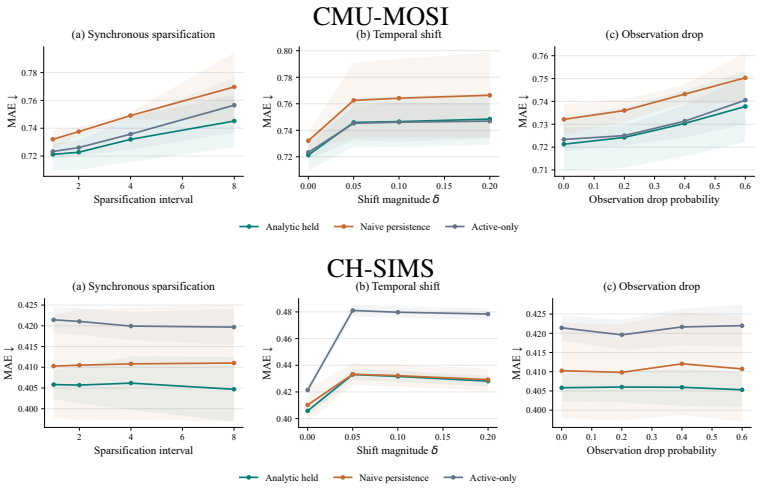


Figure 5: Observation schedule robustness on CMU-MOSI and CH-SIMS under synchronous sparsification, temporal shift, and random observation drop. Shaded regions show variation across runs.

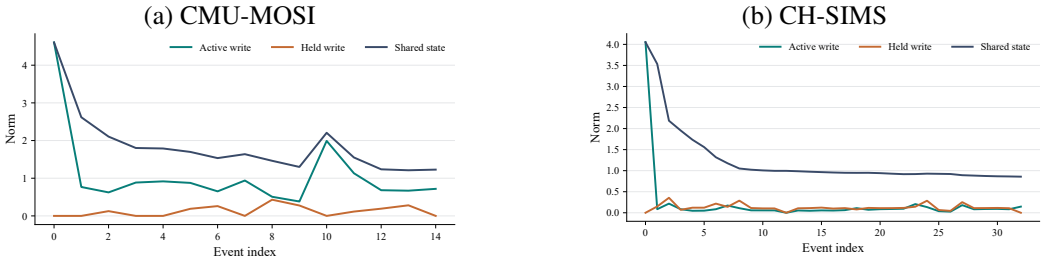


Figure 6: Representative active write, held write, and shared state norms on CMU-MOSI and CH-SIMS. Held contributions appear between new source arrivals and enter the same shared recurrent state.

5 Conclusion

MSSD formulates multimodal fusion as the evolution of one modality free recurrent state with source conditioned write and read interfaces. Controlled comparisons associate the main gain with shared recurrent storage, while Dual Axis dynamics provide a consistent refinement of the shared temporal state. Source removal shows cross source influence across three datasets and linear probes recover text and audio information after shared processing. Held drives extend the same recurrence across observation gaps, with the clearest gains under sparse and missing observations. These results support recurrent organization itself as a useful design axis for multimodal sequence modeling.

AI use statement

Generative AI tools assisted with manuscript drafting and editing, literature and reference checks, mathematical presentation, and interpretation of existing experimental results. The authors reviewed and verified all AI assisted material and take responsibility for the final content of this work.

Ethics statement

This work uses existing multimodal sentiment benchmarks containing human speech, visual behavior, and text annotations and does not introduce new human-subject data collection. These datasets can reflect demographic, cultural, linguistic, and annotation biases. Benchmark performance should therefore be interpreted within the scope of the datasets and tasks used in this study.

Reproducibility statement

Section 4 describes the evaluation protocol and controlled comparison setting. The appendix reports dataset specific model configurations, the complete metric sets for the benchmark, controlled, and held-update experiments, three seed MSSD statistics, additional observation schedule results, and the protocols used for the representation analyses. The submitted source contains the LaTeX tables and figures used to produce the reported results.

References

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 10041–10071. PMLR, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Xilin He, Haijian Liang, Boyi Peng, Weicheng Xie, Muhammad Haris Khan, Siyang Song, and Zitong Yu. Msamba: Exploring multimodal sentiment analysis with state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1309–1317, 2025.

Marlis Hochbruck and Alexander Ostermann. Exponential integrators. *Acta Numerica*, 19:209–286, 2010.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

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

Aakash Sunil Lahoti, Kevin Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations*, 2026.

Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. Coupled mamba: Enhanced multimodal fusion with coupled state space model. *Advances in Neural Information Processing Systems*, 37:59808–59832, 2024.

Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 150–161, 2018.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pp. 247–258, 2022.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2247–2256, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Fei Ma, Shuzhi Li, Feixia Yang, and Guangxian Xu. Mslmamba: Mamba-driven cross-modal fusion network with modality-specific learning for multimodal sentiment analysis. *Neurocomputing*, pp. 132164, 2025.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 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*, volume 32, 2019.

THUIAR. MMSA: A unified framework for multimodal sentiment analysis. GitHub repository, 2020. URL `https://github.com/thuiar/MMSA`. Accessed 2026-09-26.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019.

Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, Lihuo He, and Xuemei Luo. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, 136:109259, 2023a.

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25:4909–4921, 2023b.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 3718–3727, 2020.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10790–10797, 2021.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE intelligent systems*, 31(6):82–88, 2016.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis, 2017. URL `https://arxiv.org/abs/1707.07250`.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018a.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2236–2246, 2018b.

Tan Zhi-Xuan, Harold Soh, and Desmond Ong. Factorized inference in deep markov models for incomplete multimodal time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10334–10341, 2020.

A Training and Model Configurations

All runs use AdamW with weight decay 10^{-4}, BF16 autocast, gradient clipping at 1.0, and a linear warmup followed by cosine annealing. The English datasets use `bert-base-uncased` and the Chinese datasets use `bert-base-chinese` (Devlin et al., 2019). All configurations use $d_{\text{model}} = 128$ and head dimension 64. Table 4 lists the dataset specific settings that vary across experiments.

Table 4: Dataset specific MSSD configurations.

Setting	CMU-MOSI	CMU-MOSEI	CH-SIMS	CH-SIMS v2
d_{state}	128	128	16	16
MIMO rank	2	4	8	8
Layers	3	2	3	3
Backbone LR	4×10^{-4}	1×10^{-4}	1×10^{-4}	2×10^{-4}
BERT LR	3×10^{-5}	1×10^{-5}	2×10^{-5}	2×10^{-5}
Batch size	32	64	32	32
Task dropout	0.2	0.1	0.1	0.1

B Complete Metric Results

The main paper reports MAE, Corr, and Acc2 for readability. This section retains the full standard test metric set for MSSD and for the internal general-backbone, recurrent-organization, and held-update comparisons. The aggregate comparison files contain means, while MSSD additionally reports the supplied three-seed standard deviations. Full task-specific benchmark statistics are available from the public MMSA repository cited in the main paper and are not repeated here.

B.1 MSSD Run Statistics

Table 5: Complete MSSD results on CMU-MOSI and CMU-MOSEI over three seeds. Accuracy and F1 values are percentages.

Metric	CMU-MOSI	CMU-MOSEI
Has0 Acc2	83.48 ± 0.52	81.82 ± 0.34
Has0 F1	83.39 ± 0.50	82.35 ± 0.33
Non0 Acc2	85.26 ± 0.69	85.91 ± 0.38
Non0 F1	85.23 ± 0.67	85.91 ± 0.38
Acc5	52.53 ± 0.72	54.93 ± 0.28
Acc7	45.96 ± 1.17	53.05 ± 0.31
MAE	0.721 ± 0.011	0.535 ± 0.001
Corr	0.784 ± 0.010	0.774 ± 0.001
MSE loss	1.026 ± 0.017	0.501 ± 0.002

Table 6: Complete MSSD results on CH-SIMS and CH-SIMS v2 over three seeds. Accuracy and F1 values are percentages.

Metric	CH-SIMS	CH-SIMS v2
Acc2	80.38 ± 1.46	79.68 ± 0.54
Acc3	67.32 ± 0.77	73.26 ± 0.07
Acc5	42.23 ± 0.58	56.96 ± 0.00
F1	80.41 ± 1.33	79.57 ± 0.38
MAE	0.406 ± 0.004	0.298 ± 0.001
Corr	0.616 ± 0.012	0.702 ± 0.006
MSE loss	0.325 ± 0.018	0.187 ± 0.006

B.2 Full Internal Comparisons

Table 7: Full internal aggregate metrics on CMU-MOSI. Values are means over the supplied runs. Accuracy and F1 values are percentages.

Model	H0 A2	H0 F1	N0 A2	N0 F1	A5	A7	MAE	Corr
General sequence backbones								
Joint BiLSTM	81.92	81.91	83.64	83.68	50.49	45.04	0.745	0.792
Joint Transformer	82.22	82.19	84.00	84.02	51.12	45.48	0.732	0.786
Joint Mamba-3	82.12	82.08	83.74	83.76	50.29	44.37	0.736	0.789
Perceiver style	82.22	82.19	83.84	83.86	51.26	45.19	0.741	0.787
Controlled organization variants								
Independent + concat	82.31	82.29	83.94	83.97	49.47	43.73	0.739	0.792
Independent + cross attn	82.12	82.08	84.04	84.07	48.84	43.15	0.748	0.790
Shared temporal only	82.56	82.53	84.35	84.38	50.05	44.27	0.726	0.798
Held update variants								
Active only	82.46	82.42	84.30	84.32	53.26	45.82	0.723	0.787
Naive persistence	83.09	83.05	84.70	84.72	51.94	45.72	0.732	0.785
MSSD / analytic held	83.48	83.39	85.26	85.23	52.53	45.96	0.721	0.784

Table 8: Full internal aggregate metrics on CMU-MOSEI. Values are means over the supplied runs. Accuracy and F1 values are percentages.

Model	H0 A2	H0 F1	N0 A2	N0 F1	A5	A7	MAE	Corr
General sequence backbones								
Joint BiLSTM	80.56	81.19	84.94	84.98	53.99	52.08	0.545	0.772
Joint Transformer	80.27	80.92	84.75	84.80	53.95	52.06	0.546	0.772
Joint Mamba-3	80.62	81.23	84.83	84.87	54.08	52.28	0.541	0.771
Perceiver style	80.44	81.09	84.72	84.78	53.64	51.76	0.544	0.771
Controlled organization variants								
Independent + concat	80.70	81.32	85.02	85.07	54.85	52.93	0.545	0.770
Independent + cross attn	80.34	80.96	84.62	84.65	54.12	52.16	0.545	0.770
Shared temporal only	81.16	81.73	85.34	85.37	55.14	53.30	0.538	0.772
Held update variants								
Active only	81.16	81.74	85.52	85.53	54.95	53.09	0.535	0.774
Naive persistence	81.62	82.14	85.64	85.62	55.16	53.24	0.533	0.776
MSSD / analytic held	81.82	82.35	85.91	85.91	54.93	53.05	0.535	0.774

Table 9: Full internal aggregate metrics on CH-SIMS. Values are means over the supplied runs. Accuracy and F1 values are percentages.

Model	A2	A3	A5	F1	MAE	Corr
General sequence backbones						
Joint BiLSTM	80.31	67.18	41.50	80.44	0.415	0.602
Joint Transformer	80.23	66.08	41.36	80.17	0.423	0.598
Joint Mamba-3	79.58	64.99	41.28	79.45	0.419	0.597
Perceiver style	79.65	65.57	41.79	79.65	0.418	0.601
Controlled organization variants						
Independent + concat	79.58	66.81	42.09	79.64	0.410	0.614
Independent + cross attn	79.65	66.08	41.14	79.59	0.421	0.617
Shared temporal only	79.94	67.18	41.65	79.98	0.411	0.614
Held update variants						
Active only	79.07	64.77	39.10	79.18	0.421	0.598
Naive persistence	79.94	66.96	42.60	80.22	0.410	0.610
MSSD / analytic held	80.38	67.32	42.23	80.41	0.406	0.616

Table 10: Full internal aggregate metrics on CH-SIMS v2. Values are means over the supplied runs. Accuracy and F1 values are percentages.

Model	A2	A3	A5	F1	MAE	Corr
General sequence backbones						
Joint BiLSTM	79.30	72.92	55.42	79.32	0.301	0.704
Joint Transformer	77.76	72.15	54.26	77.81	0.312	0.679
Joint Mamba-3	79.50	74.18	58.12	79.57	0.300	0.695
Perceiver style	77.76	73.02	55.03	77.82	0.311	0.677
Controlled organization variants						
Independent + concat	79.21	73.31	56.87	79.30	0.294	0.712
Independent + cross attn	79.69	73.89	56.38	79.76	0.294	0.709
Shared temporal only	78.53	73.79	56.96	78.58	0.308	0.680
MSSD	79.68	73.26	56.96	79.57	0.298	0.702

C Representation Analysis Protocol

For the cross source influence analysis, one source is removed from the input and the relative change in every source conditioned read is measured with the influence definition in Section 4.5. The independent stream control applies the same intervention before fusion, where off diagonal influence is structurally zero because recurrent states are separate. For the modality probes, linear predictors are trained from the source conditioned reads to the corresponding source target representation. Probe performance is measured on held out data and compared with source write removal and shuffled target controls. The same trained representation is used across the three probe conditions.

D Additional Observation Schedule Results

MOSEI provides a complementary observation schedule setting in which the three update rules remain closer than on MOSI and CH-SIMS. This variation is consistent with the clean sequence ablation, where naive persistence and analytic held are also close on MOSEI.

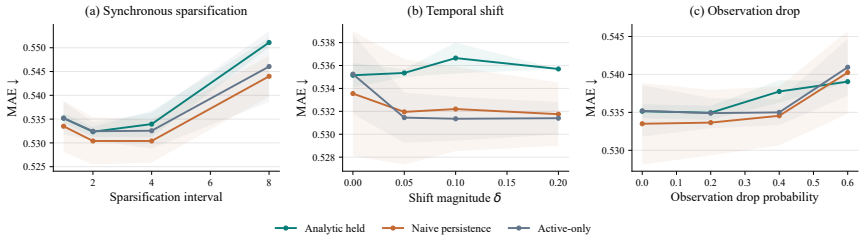
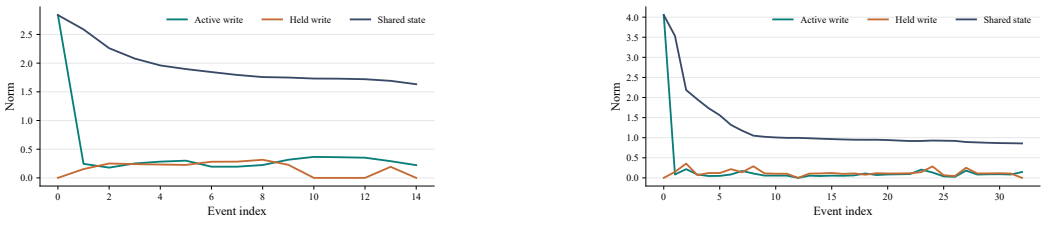


Figure 7: Observation schedule robustness on CMU-MOSEI.



(a) CMU-MOSEI (b) CH-SIMS

Figure 8: Representative active write, held write, and shared state trajectories on two additional datasets.