MODELING MULTIMODAL SEQUENCES AS HETEROGENEOUS OBSERVATIONS OF CONTINUOUS LATENT DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sequences involve two temporal structures that are often treated as one: the evolution of task-relevant information over time and the modality-dependent grids on which that information is observed. We study a formulation in which temporally evolving multimodal evidence is represented as heterogeneous observations of a shared latent process indexed by chronological time. We instantiate this formulation with Behavioral-Time Semantic Flow (BTS-Flow), a continuous time state estimation model that propagates a latent state between observation events and revises its estimate when new multimodal evidence becomes available. Modality specific observation functions connect the shared latent state to distinct representation spaces, while a decomposition into state related and private components allows modality dependent residual information to remain outside the common dynamics. For acoustic and visual streams, timestamp based local derivatives provide an additional constraint that relates observed representation changes to the pushforward of the latent velocity through the corresponding observation functions. Language remains event valued and enters the latent state only through observation updates. Across CMU-MOSI, CMU-MOSEI, and CH-SIMS-v2, BTS-Flow improves standard predictive metrics over the strongest reported baselines. More importantly, controlled changes to observation schedules and test time temporal sparsification show lower sensitivity to the sampling grid, while held out observation inference and velocity consistency analyses support the temporal behavior implied by the formulation. These results provide empirical evidence that modeling multimodal sequences as heterogeneous observations of continuous latent dynamics is a useful inductive principle for temporally heterogeneous multimodal learning.

1 INTRODUCTION

Multimodal sequence learning involves two temporal structures that need not coincide: the evolution of task-relevant information and the temporal grids on which different modalities are observed. Linguistic information is naturally organized as discrete events, whereas acoustic and visual signals are repeatedly sampled through different sensing and representation mechanisms Zadeh et al. (2017); Meng et al. (2026b). Their observation rates and timestamps can therefore differ substantially across modalities Yang et al. (2024); Meng et al. (2026c). Most multimodal architectures nevertheless construct task representations primarily at observed sequence positions and model interactions among the resulting discrete states Fang et al. (2025). This design is

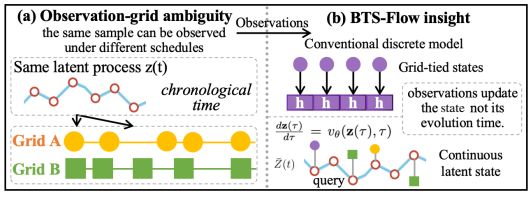

Figure 1: Temporal representation in multimodal sequences. (a) The same latent process may be observed under different modality specific sampling schedules while sharing the same transition law. (b) Discrete models construct states at observed positions, whereas our formulation defines a latent state over chronological time and updates it with multimodal observations.

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.

effective for many prediction tasks, but it leaves
the representation between observations implicit and couples the computational trajectory to a particular temporal discretization. We ask a more basic question: should the temporal representation of a multimodal sequence be defined by the positions at which evidence is observed, or can it instead be associated with a latent process that evolves over chronological time?

Methods for unaligned and asynchronous multimodal sequences model temporal dependencies directly among observed representations using cross-modal attention, graph structures, or modality specific and modality agnostic features Tsai et al. (2019); Yang et al. (2021). Representation oriented approaches further separate shared and modality dependent information or organize multimodal features into structured semantic spaces Hazarika et al. (2020); Yang et al. (2023); Meng et al. (2026d); Wang et al. (2025). Information theoretic and hierarchical methods instead focus on preserving and coordinating task-relevant information across modalities Han et al. (2021); Xiao et al. (2024); Meng et al. (2026a), while recent state space and expert based architectures improve long range temporal modeling or adapt modality contributions to individual inputs He et al. (2025); Fang et al. (2025). Despite these architectural and representational differences, temporal computation remains primarily organized around observed modality states. More related work in Appendix A.

CaReFlow Mai & Han (2026) introduces a different use of continuous vector fields in multimodal learning. It constructs a rectified flow that transports one modality distribution toward another before multimodal fusion. Its independent variable parameterizes progress along this cross-modal transport path. Continuous time sequence models provide another relevant line of work. Neural ODEs Chen et al. (2018), Latent ODEs Rubanova et al. (2019), Neural CDEs Kidger et al. (2020), GRU-ODE-Bayes De Brouwer et al. (2019), and continuous time Transformers model hidden representations over elapsed time and provide general mechanisms for irregularly sampled or sporadically observed sequences Chen et al. (2023). ODE based evolution has also been applied to structured multimodal representations Shou et al. (2025). These developments support using continuous dynamics for sequence modeling. More specifically, can heterogeneous multimodal representations be interpreted as observations of a shared task-relevant process indexed by the sample's chronological time?

We examine this question through a state and observation formulation. Let $z(t)$ denote a task-relevant latent state and let modality m provide an observation $s^m(t) = G_m(z(t)) + \epsilon^m(t)$, where G_m maps the shared latent state to the representation space of that modality. Each G_m may describe a different observation geometry, and a modality may contain additional variation that is not explained by the shared state. If the task-relevant latent state is locally differentiable at the temporal resolution of interest, then its dynamics induce $\dot{s}^m(t) \approx J_{G_m}(z(t))\dot{z}(t)$. The shared structure therefore lies in the latent dynamics that generate state related temporal changes, rather than in equality between multimodal representations. We treat local continuity as an empirically testable modeling hypothesis, assuming only local temporal regularity at the representation level to accommodate complex semantic transitions. Observation times specify evidence availability, whereas chronological time indexes latent-state evolution. As shown in Fig. 1, different observation schedules may reveal different evidence about the same temporal process and thus yield different state estimates, without requiring a different latent transition law. The representation is also available at chronological times with no direct modality observations. Thus, the central question is not whether modalities should be aligned onto a common grid, but how to estimate a latent task state from asynchronously acquired observations.

We instantiate this formulation with Behavioral-Time Semantic Flow (BTS-Flow), where behavioral time denotes the elapsed chronological time within a multimodal episode. BTS-Flow propagates a latent state estimate between observation events using a neural vector field. Here, flow refers to latent-state evolution over chronological time, rather than probability transport used in normalizing flows or flow matching. At each event, modality specific observation functions predict state related evidence from the propagated state; the discrepancy from observed evidence forms an innovation that revises the current estimate. Each modality is decomposed into state related and private components, so residual modality information need not be explained by the common temporal dynamics. For acoustic and visual streams, we estimate local representation derivatives from their timestamps and constrain these observed changes to agree with the pushforward of the shared latent velocity through the corresponding observation functions. Language enters observation updates as event valued evidence without artificial temporal derivatives. Our study makes three main contributions:

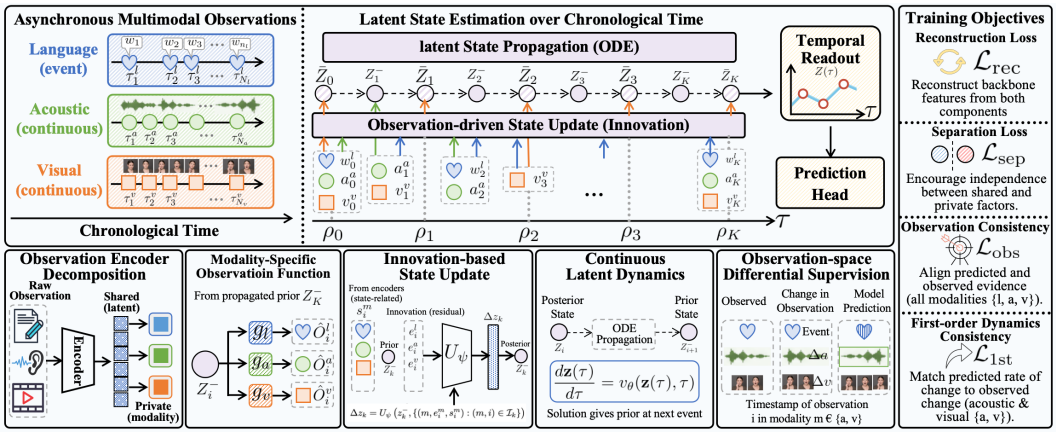

Figure 2: **Overview of BTS-Flow.** BTS-Flow models asynchronous linguistic, acoustic, and visual inputs as heterogeneous observations of a shared latent state defined over chronological time. Each modality is decomposed into a state-related component and a private factor; the latent state is propagated by a neural ODE between observation events and updated at event times through innovation-guided latent-state correction. Acoustic and visual observations further provide differential supervision on the learned dynamics. Training uses $\mathcal{L}_{rec}$, $\mathcal{L}_{sep}$, $\mathcal{L}_{obs}$, and $\mathcal{L}_{1st}$, and the resulting latent trajectory is used for temporal readout and prediction. Schematic event indices start at zero; the text starts at one.

- We formulate temporally heterogeneous multimodal sequence learning as estimation of a task-relevant latent process from modality dependent observations. The model separates latent-state evolution time from modality-specific sampling grids.
- We instantiate this view with BTS-Flow, which combines continuous latent propagation, modality specific observation functions, state related and private representation components, and innovation based updates for asynchronously available multimodal evidence. Building on direct first order dynamics supervision, we extend differential consistency to heterogeneous observation spaces by relating timestamp derived acoustic and visual changes to the pushforward of a shared latent velocity.
- We evaluate the formulation through both downstream prediction and temporal mechanism analyses. Across five datasets, BTS-Flow achieves competitive predictive performance, while controlled changes to observation schedules, temporal sparsification, held out observation inference, and velocity consistency analyses provide complementary evidence for the proposed chronological latent dynamics.

2 Method

2.1 Problem Formulation

For modality $m \in \mathcal{M} = \{l, a, v\}$, let $\mathcal{O}_m = \{(\mathbf{x}_i^m, t_i^m)\}_{i=1}^{N_m}$ denote its observations and timestamps. Sequence lengths and timestamps need not be shared across modalities. We use the normalized chronological coordinate $\tau_i^m = \frac{t_i^m - t_0}{T_s}$, where t_0 denotes the temporal origin of the sample and $T_s > 0$ is a single dataset-level scale estimated from the training split and fixed for all validation and test samples. Unlike per-sample normalization to $[0, 1]$, this transformation preserves differences in elapsed temporal rate across samples. We posit an underlying task-relevant process $\mathbf{z}(\tau) \in \mathbb{R}^{d_z}$ satisfying $\frac{d\mathbf{z}(\tau)}{d\tau} = v_\theta(\mathbf{z}(\tau), \tau)$, where the neural vector field is locally Lipschitz in $\mathbf{z}$, ensuring a well-defined local solution. Modality m provides a heterogeneous observation $\mathbf{s}^m(\tau) = G_m(\mathbf{z}(\tau)) + \boldsymbol{\epsilon}^m(\tau)$. The functions G_m need not share output coordinates or dimensional geometry. The continuous process $\mathbf{z}(\tau)$ should be distinguished from its filtered estimate $\hat{\mathbf{z}}(\tau)$, which is revised whenever new evidence is observed. S_m extracts the state-related observation from raw modality features, while G_m predicts this observation from the latent state. Although both

operate in the same observation space, they are optimized through complementary encoding and prediction objectives.

2.2 Heterogeneous Observation Representation

Each modality is first encoded independently $h_i^m = F_m(x_i^m)$, where F_m preserves the temporal resolution of the observation stream. For acoustic and visual modalities, the features used for derivative estimation are timestamp-preserving, pre-fusion representations rather than sequence-level pooled features. We decompose h_i^m into a state-related observation $s_i^m = S_m(h_i^m)$ and a modality-private factor $p_i^m = P_m(h_i^m)$. The former is the part of the observation allowed to constrain the shared latent state, whereas the latter retains information that need not be explained by the common dynamics.

A modality decoder reconstructs the backbone representation as $\hat{h}_i^m = D_m(s_i^m, p_i^m)$, yielding $\mathcal{L}_{\text{rec}}$. We discourage redundant coding with $\mathcal{L}_{\text{sep}}$, where S^m and P^m collect state-related and private representations within a mini-batch.

$$\mathcal{L}_{\text{rec}} = \sum_{m\in\mathcal{M}} \sum_{i=1}^{N_m} \left\| \hat{h}_i^m - h_i^m \right\|_2^2, \qquad \mathcal{L}_{\text{sep}} = \sum_{m\in\mathcal{M}} \|\text{Cov}(S^m, P^m)\|_F^2 . \tag{1}$$

The observation functions G_m relate the continuous latent state to the corresponding state-related observation spaces. Let $k(i,m)$ index the global event at timestamp τ_i^m. At each observed timestamp, we impose a posterior observation-consistency objective:

$$\mathcal{L}_{\text{obs}} = \sum_{m\in\mathcal{M}} \sum_{i} \left\| G_m(\bar{z}_{k(i,m)}) - \text{sg}(s_i^m) \right\|_2^2, \tag{2}$$

where $\bar{z}_{k(i,m)}$ denotes the posterior state estimate at the corresponding timestamp and sg denotes stop-gradient. This objective encourages the evidence-assimilated state estimate to remain compatible with the state-related observation.

2.3 Continuous State Propagation and Asynchronous Observation Update

Let $\rho_1 < \rho_2 < \cdots < \rho_K$ denote the ordered union of timestamps from all modalities, with simultaneous observations grouped at the same ρ_k. Let $\mathcal{I}_k$ be the set of observations arriving at that timestamp. At the first observation time, the initial state is inferred from the available state-related observations $\bar{z}_1 = I_\phi\left(\{(m, s_i^m) : (m,i) \in \mathcal{I}_1\}\right)$, where I_ϕ is permutation invariant with respect to simultaneously available modalities.

For $k > 1$, suppose $\bar{z}_{k-1}$ is the state estimate after processing all observations at ρ_{k-1}. Before receiving new evidence, BTS-Flow propagates this estimate through the latent dynamics:

$$z_k^- = \Phi_\theta\left(\bar{z}_{k-1}; \rho_{k-1}, \rho_k\right) = \bar{z}_{k-1} + \int_{\rho_{k-1}}^{\rho_k} v_\theta(\hat{z}(\tau), \tau) d\tau, \tag{3}$$

where Φ_θ is evaluated using a numerical ODE solver. The prior state predicts the state-related observation expected in modality m $\hat{s}_i^m = G_m(z_k^-), (m,i) \in \mathcal{I}_k$.

We define the corresponding innovation as $e_i^m = s_i^m - \hat{s}_i^m$. The simultaneously arriving innovations are mapped back to the latent state through:

$$\Delta z_k = U_\psi\left(z_k^-, \{(m, e_i^m, s_i^m) : (m,i) \in \mathcal{I}_k\}\right), \tag{4}$$

where U_ψ is a modality-aware permutation-invariant set function. The posterior state estimate is $\bar{z}_k = z_k^- + \Delta z_k$. State propagation models latent evolution over elapsed time, while posterior correction revises the estimate using new evidence. A discrete correction of the estimate is therefore not interpreted as a discontinuity of the underlying latent process.

Linguistic observations enter the same update mechanism at their timestamps through s_i^l and its innovation, without dense temporal interpolation. The model can also evaluate the state at an unobserved query time $\tau^\star \in (\rho_k, \rho_{k+1})$, $\hat{z}(\tau^\star) = \Phi_\theta\left(\bar{z}_k; \rho_k, \tau^\star\right)$.

2.4 Direct First-Order Observation Supervision

Task and zeroth-order observation supervision constrain the represented state but only indirectly constrain its evolution between observations. We therefore derive an additional constraint from local temporal changes in continuously sampled modalities.

Timestamp-aware derivative estimation. For $m \in \{a, v\}$, consider a timestamp-preserving feature h_i^m and a local neighborhood $\mathcal{W}_i^m$. We approximate the local feature trajectory by an R-th order polynomial $h^m(\tau) \approx \sum_{r=0}^{R} \beta_{i,r}^m (\tau - \tau_i^m)^r$. The coefficients are obtained using the actual timestamps:

$$\left\{\beta_{i,r}^{m*}\right\}_{r=0}^{R} = \arg\min_{\{\beta_{i,r}^m\}} \sum_{j \in \mathcal{W}_i^m} \left\| h_j^m - \sum_{r=0}^{R} \beta_{i,r}^m (\tau_j^m - \tau_i^m)^r \right\|_2^2 . \tag{5}$$

The estimated local derivative is $\dot{h}_i^m = \beta_{i,1}^{m*}$. Because the actual temporal offsets enter Eq. 5, the estimator remains defined for nonuniform observation intervals.

To reduce the influence of poorly determined local derivatives, we compute the fitting residual

$$\varepsilon_i^m = \frac{1}{|\mathcal{W}_i^m|} \sum_{j \in \mathcal{W}_i^m} \left\| h_j^m - \sum_{r=0}^{R} \beta_{i,r}^{m*} (\tau_j^m - \tau_i^m)^r \right\|_2^2, \qquad \xi_i^m = \exp(-\kappa \varepsilon_i^m). \tag{6}$$

We weight each derivative target by its local fitting quality.

Observed first-order change. The state-related observation is $s_i^m = S_m(h_i^m)$. Its local temporal derivative follows from the chain rule:

$$u_i^m = J_{S_m}(h_i^m)\dot{h}_i^m \approx \left. \frac{ds^m(\tau)}{d\tau} \right|_{\tau=\tau_i^m} . \tag{7}$$

We evaluate Eq. 7 using a Jacobian–vector product without explicitly materializing J_{S_m}.

Dynamics-induced first-order change. The observation model provides a second expression for temporal change. Differentiating $G_m(z(\tau))$ yields

$$\frac{dG_m(z(\tau))}{d\tau} = J_{G_m}(z(\tau)) v_\theta(z(\tau), \tau). \tag{8}$$

Eq. 8 works without enforcing a shared representation space across modalities. Their observation spaces may differ through G_m, only the latent process explaining their changes is shared.

To align latent drift with observed changes, we evaluate v_θ at the prior state $z^-_{k(i,m)}$ before the observation update, for events after initialization ($k(i,m) > 1$):

$$\mathcal{L}_{1st} = \sum_{m \in \{a,v\}} \sum_{\substack{1 \le i \le N_m \\ k(i,m) > 1}} \xi_i^m \left\| J_{G_m}\left(z^-_{k(i,m)}\right) v_\theta\left(z^-_{k(i,m)}, \tau_i^m\right) - \mathrm{sg}(u_i^m) \right\|_2^2 \tag{9}$$

2.5 Learning Objective and Downstream Readout

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}} + \lambda_{\text{obs}}\mathcal{L}_{\text{obs}} + \lambda_{\text{1st}}\mathcal{L}_{\text{1st}}. \tag{10}$$

$\mathcal{L}_{\text{task}}$ makes the estimated trajectory informative for the downstream prediction. For a sequence-level downstream task, a task-specific readout $\mathcal{R}$ operates on the estimated state trajectory $\hat{y} = C_{\text{task}}\left(\mathcal{R}(\{\bar{z}_k, \rho_k\}_{k=1}^K)\right)$. $\mathcal{R}$ may be instantiated by the final state or a lightweight temporal pooling operator depending on the downstream protocol. For time-resolved prediction, the same head can be evaluated directly at a query time $\hat{y}(\tau^\star) = C_{\text{task}}\left(\hat{z}(\tau^\star)\right)$.

Table 1: Comparison with existing SOTA. † indicates that the baseline results are from the CaReFlow report Mai & Han (2026). ↑ and ↓ indicate that higher and lower values are better, respectively.

Model	CMU-MOSI					CMU-MOSEI					CH-SIMS-v2				
	Acc-7↑	Acc-2↑	F1↑	MAE↓	Corr↑	Acc-7↑	Acc-2↑	F1↑	MAE↓	Corr↑	Acc-5↑	Acc-2↑	F1↑	MAE↓	Corr↑
MISA	41.27	82.07	82.43	0.786	0.746	51.43	82.03	82.13	0.594	0.724	47.5	78.2	78.3	0.342	0.671
Self-MM†	45.8	84.9	84.8	0.731	0.785	53.0	85.2	85.2	0.540	0.763	53.5	78.7	78.6	0.315	0.691
DMD	44.9	84.3	84.3	0.726	0.788	52.8	84.6	84.6	0.538	0.768	47.7	78.2	78.3	0.335	0.686
EMOE	45.2	84.8	84.8	0.723	0.790	52.5	85.0	85.0	0.542	0.760	47.9	79.1	78.7	0.332	0.694
MAG-B.	43.69	84.43	84.66	0.727	0.782	52.67	84.42	84.71	0.544	0.756	49.2	77.1	77.1	0.346	0.641
AtCAF†	46.5	88.6	88.5	0.650	0.831	55.9	87.0	86.8	0.508	0.785	–	–	–	–	–
ITHP†	47.7	88.5	88.5	0.663	0.856	52.2	87.1	87.1	0.550	0.792	–	–	–	–	–
DLF†	49.4	88.7	88.6	0.669	0.856	55.0	87.5	87.4	0.515	0.774	47.5	78.1	77.9	0.346	0.683
CLCR	49.6	88.1	87.99	0.678	0.821	55.9	87.9	88.0	0.511	0.792	48.6	79.6	78.6	0.334	0.695
MMIM	46.65	86.7	85.98	0.704	0.806	54.26	86.1	85.97	0.528	0.775	50.5	77.8	77.8	0.339	0.641
DEVA†	46.3	86.3	86.3	0.730	0.787	52.3	86.1	86.2	0.541	0.769	48.3	79.6	79.1	0.329	0.697
KuDA†	–	–	–	–	–	–	–	–	–	–	53.1	80.2	80.1	0.289	0.741
CaReFlow†	50.6	89.8	89.7	0.616	0.858	55.7	87.9	88.0	0.504	0.799	57.9	82.9	82.9	0.277	0.745
BTS-Flow	51.8 ±0.18	90.4 ±0.14	90.3 ±0.15	0.604 ±0.005	0.866 ±0.004	56.2 ±0.11	88.4 ±0.08	88.5 ±0.09	0.496 ±0.003	0.808 ±0.003	59.2 ±0.20	83.8 ±0.14	83.7 ±0.15	0.269 ±0.004	0.756 ±0.005

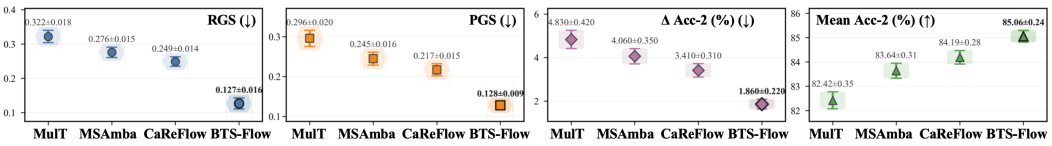

Figure 3: Observation-grid sensitivity on MOSI under matched budget and temporal coverage. RGS and PGS measure representation and prediction sensitivity, respectively (Appendix F). Lower RGS, PGS, and ΔAcc-2 indicate greater robustness, higher Mean Acc-2 better performance.

3 Experiments

3.1 Experimental Setup

Benchmarks. We evaluate BTS-Flow on five multimodal benchmarks: CMU-MOSI Zadeh et al. (2016), CMU-MOSEI Zadeh et al. (2018), and CH-SIMS-v2 Liu et al. (2022) for Multimodal Sentiment Analysis, UR-FUNNY Hasan et al. (2019) for Humor Detection, and MUStARD Castro et al. (2019) for Sarcasm Detection. These tasks involve evolving affective or conversational states, making them suitable for evaluating whether latent trajectories capture temporally distributed multimodal evidence. **Comparison protocol.** Published results are retained only when their data and feature protocols match ours and are used as descriptive reference points without significance claims. Continuous-time baselines are reimplemented under matched unimodal inputs, latent dimensionality, prediction head, data splits, and tuning budget, with five-seed mean±standard deviation reported for controlled comparisons. All hyperparameters are selected on the validation set only. **Implementation.** Controlled comparisons use a two-layer MLP predictor. BTS-Flow uses $d_z = 128$, rescales timestamps by a single training-set time scale shared across samples, and employs RK4 for latent-state propagation. Acoustic and visual derivatives are estimated by timestamp-aware second-order local polynomial regression. We optimize with AdamW. Full metrics and implementation details are provided in Appendix B.

3.2 Comparison with State-of-the-Art Methods

Table 1 compares BTS-Flow with representative state-of-the-art multimodal methods on CMU-MOSI, CMU-MOSEI, and CH-SIMS-v2. BTS-Flow performs best across reported metrics on all three benchmarks. Relative to the strongest flow-based baseline CaReFlow, it gains 1.2 Acc-7 and 0.6 Acc-2/F1 points on CMU-MOSI; 0.5 points in Acc-7, Acc-2, and F1 on CMU-MOSEI; and 1.3 Acc-5, 0.9 Acc-2, and 0.8 F1 points on CH-SIMS-v2, with lower MAE and higher Corr on each dataset. These results suggest gains beyond binary classification, extending to finer-grained prediction and continuous-valued evaluation. The consistent gains over CaReFlow indicate that modeling

Table 2: Results on MHD and MSD Tasks.

UR-FUNNY (Acc. %)		MUStARD (Acc. %)	
MulT	66.65	ITHP	75.30
DMD	68.70	MIL	76.30
Self-MM	68.77	MAG-XLNet	76.47
MMIM	69.53	MuLOT	76.82
DMD+SuCI	70.84	MCL	77.94
AtCAF	72.13	MGCL	77.94
MuLOT	73.97	MO-Sarcation	79.71
CaReFlow	75.75	CaReFlow	82.35
BTS-Flow	**76.66**	**BTS-Flow**	**83.06**

Table 3: Ablation Studies of BTS-Flow.

Variant	CMU-MOSI Acc-2(%)	CMU-MOSI MAE↓	CH-SIMS-v2 Acc-2(%)	CH-SIMS-v2 MAE↓
BTS-Flow	**90.4**	**0.604**	**83.8**	**0.269**
Continuous-Time State Representation				
Discrete-trans.	89.7	0.617	82.9	0.280
Sequence	89.8	0.615	83.0	0.279
Per-[0,1]	90.0	0.611	83.3	0.275
Heterogeneous Observation Modeling				
Across-Modal	89.9	0.612	83.2	0.277
w/o Pri-Factor	90.1	0.608	83.5	0.273
w/o Innovation	89.6	0.619	82.8	0.281
Dense-IL	89.9	0.612	83.2	0.276
Observation-Space Supervision				
w/o $\mathcal{L}_{obs}$	89.8	0.614	83.1	0.278
w/o $\mathcal{L}_{1st}$	90.1	0.609	83.4	0.274
Finite-d.	90.2	0.606	83.6	0.272

multimodal evidence as observations of a chronologically evolving latent state improves upon cross-modal distribution transport under standard benchmark settings.

Following the CaReFlow Mai & Han (2026) protocol, we further evaluate BTS-Flow on Multimodal Humor Detection (MHD) with UR-FUNNY Hasan et al. (2019) and Multimodal Sarcasm Detection (MSD) with MUStARD Castro et al. (2019). Their context–punchline sequences contain linguistic, acoustic, and visual observations, testing transfer across downstream targets. In Table 2, BTS-Flow achieves 76.66% and 83.06% accuracy, respectively, exceeding CaReFlow by 0.91 and 0.71 points. The same state propagation and asynchronous observation-update mechanism can model progression from context to punchline without task-specific changes to the core formulation.

3.3 Ablation Studies

Continuous-Time State Representation. The first group in Table 3 compares chronological time with discrete sequence positions. Replacing ODE propagation with a parameter-matched discrete transition (Discrete-trans.) reduces Acc-2 from $90.4/83.8$ to $89.7/82.9$ and increases MAE from $0.604/0.269$ to $0.617/0.280$ on CMU-MOSI/CH-SIMS-v2. Sequence indices (Sequence) also degrade Acc-2/MAE ($89.8/0.615$ and $83.0/0.279$, respectively), showing that observation order alone is insufficient. Per-sample $[0, 1]$ normalization (Per-[0,1]) retains relative within-sequence timing and outperforms sequence indices but underperforms the shared chronological scale. Together, these results support continuous propagation and preservation of elapsed-time information across samples. **Heterogeneous observation modeling.** Using a shared observation function across modalities (Across-Modal) reduces Acc-2 by $0.5/0.6$ points on CMU-MOSI/CH-SIMS-v2, supporting modality-specific G_m. Removing the private factor (*w/o* Pri-Factor) causes a smaller degradation ($0.3/0.3$ points), whereas removing the innovation update (*w/o* Innovation) yields the largest drop in this group ($0.8/1.0$ points) together with the highest MAE. This contrast indicates that modality-specific residual information is useful, but explicit state revision from new evidence is more important to the estimation process. Dense Interpolation of Linguistic representations (Dense-IL.) also degrades performance, supporting the treatment of language as event-valued observations. **Observation-Space Supervision.** The third group tests constraints connecting latent dynamics with modality observations. Removing $\mathcal{L}_{\text{obs}}$ reduces Acc-2 to $89.8/83.1$ and increases MAE to $0.614/0.278$ on CMU-MOSI/CH-SIMS-v2, demonstrating the importance of zeroth-order observation consistency. Removing $\mathcal{L}_{\text{1st}}$ causes smaller but consistent degradation ($90.1/0.609$ and $83.4/0.274$), indicating complementary supervision for the learned dynamics. Replacing the timestamp-aware local polynomial estimator with finite differences (Finite-d.) yields $90.2/0.606$ and $83.6/0.272$. The smaller effect on task metrics is expected because first-order supervision primarily constrains temporal behavior, evaluated more directly in Sec. 3.6.

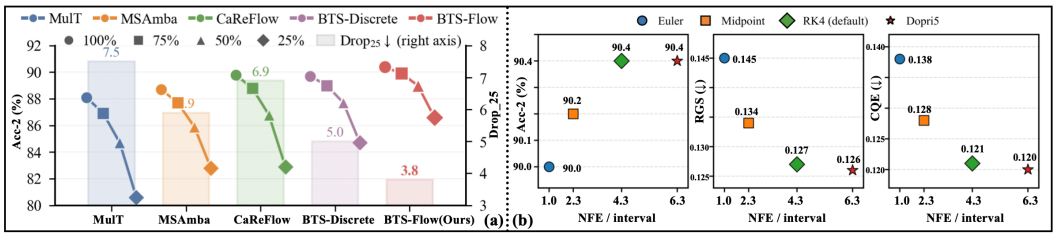

Figure 4: (a) Acc-2 (%) on CMU-MOSI under test-time sparsification of acoustic and visual observations. All models are trained with the full observation rate and evaluated without retraining. Drop_{25} denotes the absolute Acc-2 decrease from 100% to 25% observation rate. Lower Drop_{25} indicates greater robustness. (b) Inference-time solver sensitivity on CMU-MOSI. NFE/interval denotes the average number of vector-field evaluations per latent-state propagation interval.

3.4 Observation-Grid Sensitivity

As shown in Fig. 3, a continuous latent representation should preserve similar latent trajectories when the observation schedule changes but the underlying temporal coverage remains comparable. BTS-Flow reduces RGS from 0.249 to 0.127 and PGS from 0.217 to 0.128 relative to CaReFlow, while reducing the cross-grid Acc-2 range from 3.41 to 1.86 points. Mean Acc-2 simultaneously increases from 84.19 to 85.06, indicating that lower grid sensitivity is not obtained by collapsing predictions to a uniformly conservative solution. Because different schedules may expose genuinely different evidence, these metrics are interpreted as sensitivity rather than a requirement of exact invariance.

3.5 Robustness to Temporal Observation Sparsification

We reduce acoustic and visual observation density to $\{100\%, 75\%, 50\%, 25\%\}$ at test time while preserving temporal coverage. All models are trained only with full observations, and all baselines are evaluated without retraining or schedule-specific adaptation. In Fig. 4(a), BTS-Flow loses 3.8 Acc-2 points from full to 25% density, versus 6.9 for CaReFlow and 7.5 for MulT. Its widening margin over CaReFlow suggests that the advantage extends beyond the standard sampling grid. BTS-Discrete isolates state-estimation structure from continuous propagation. Its gap to BTS-Flow grows from 0.7 points at full density to 1.9 at 25%, while BTS-Discrete increasingly outperforms CaReFlow as observations are removed. These results suggest complementary effects: heterogeneous innovation-based estimation improves assimilation of sparse evidence, whereas continuous propagation becomes more valuable as inter-observation intervals grow.

3.6 Continuous-Time Mechanism Analysis

Standard task metrics do not directly assess whether between-observation trajectories remain meaningful or local dynamics agree with observed temporal changes. We therefore examine held-out state inference, observation-space velocity consistency, and innovation-conditioned state revision on CMU-MOSI.

Held-out intermediate-observation inference. We remove interior acoustic and visual observations before temporal contextualization and query the latent state at their original timestamps. The removed observations serve only as evaluation targets, and modality representations are standardized using training-set statistics. We define Continuous Query Error (CQE) as

$$\mathrm{CQE} = \frac{1}{|\mathcal{M}_c|} \sum_{m \in \mathcal{M}_c} \frac{1}{|\mathcal{Q}_m|} \sum_{q \in \mathcal{Q}_m} \left\| G_m(\hat{z}_{-q}(\tau_q)) - s_q^m \right\|_2^2, \tag{11}$$

where $\mathcal{M}_c = \{a, v\}$, $\mathcal{Q}_m$ is the held-out query set for modality m, and $\hat{z}_{-q}$ is inferred without the query observation. Lower CQE indicates better recovery of unobserved intermediate measurements.

Velocity consistency. For method $\mathcal{A}$, let $\dot{z}_{\mathcal{A}}(\tau)$ denote its instantaneous latent derivative. Its induced observation-space change is $\hat{u}_i^m = J_{G_m}(\bar{z}_{k(i,m)})\dot{z}_{\mathcal{A}}(\tau_i^m)$. We compare $\hat{u}_i^m$ with the

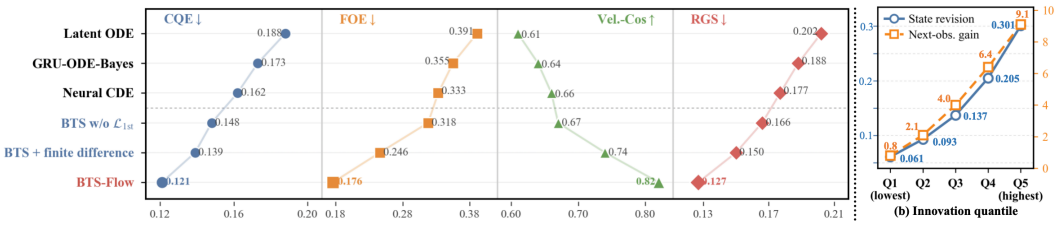

Figure 5: Continuous-time mechanism analysis on CMU-MOSI. (a) CQE evaluates held-out state inference, FOE and Vel.-Cos measure observation-space velocity consistency, and RGS measures observation-grid sensitivity. (b) State-revision magnitude and subsequent prediction gain across innovation quantiles. BTS-Flow yields more consistent continuous dynamics, while larger innovations induce stronger and more useful state corrections.

timestamp-aware estimate u_i^m using

$$\mathrm{FOE} = \frac{\sum_{m,i} \xi_i^m \|\hat{u}_i^m - u_i^m\|_2^2}{\sum_{m,i} \xi_i^m \|u_i^m\|_2^2 + \epsilon}, \tag{12}$$

together with confidence-weighted cosine agreement (Vel.-Cos; Appendix F).

As shown in Fig. 5(a), BTS-Flow achieves the lowest CQE, FOE, and RGS and the highest Vel.-Cos. Compared with Neural CDE, CQE decreases from 0.162 to 0.121 and FOE from 0.333 to 0.176, while Vel.-Cos increases from 0.66 to 0.82. Removing $\mathcal{L}_{1\mathrm{st}}$ or replacing the timestamp-aware derivative with finite differences substantially weakens velocity consistency, confirming that first-order supervision primarily constrains the temporal geometry of the learned trajectory. **Innovation-conditioned state revision.** We group event indices into five quantiles Q by standardized innovation magnitude. Let E_k^{upd} and E_k^{no} denote one-step observation-prediction errors with and without the current update, respectively, evaluated against the same next observation. The relative gain is

$$\mathrm{NextGain}(Q) = 100 \frac{\sum_{k \in Q} (E_k^{\mathrm{no}} - E_k^{\mathrm{upd}})}{\sum_{k \in Q} E_k^{\mathrm{no}} + \epsilon}. \tag{13}$$

Fig. 5(b) shows that larger innovations are associated with stronger state revisions and larger subsequent prediction gains. The gain increases from 0.8% in the lowest quantile to 9.1% in the highest, indicating that the update operator uses unexpected evidence to perform predictive state correction.

3.7 Numerical Stability and Hyperparameter Sensitivity

Fig. 4(b) shows stable BTS-Flow performance across standard integrators. Compared with RK4, Dopri5 changes RGS and CQE by only 0.001 without improving Acc-2, while increasing average NFE/interval from 4.3 to 6.3. We therefore use RK4 by default for its accuracy-efficiency trade-off. Appendix D reports full solver comparisons and sensitivity to $\lambda_{1\mathrm{st}}$, d_z, and local polynomial order.

The Efficiency Analysis is in the Appendix E.

4 Conclusion

We presented BTS-Flow, a state-observation formulation for temporally heterogeneous multimodal sequences. The model separates continuous latent propagation over chronological time from observation-driven state correction, uses modality-specific observation functions to retain heterogeneous representation geometries, and constrains acoustic and visual dynamics through first-order observation-space consistency. Across five multimodal benchmarks, BTS-Flow provides competitive predictive performance, while controlled schedule perturbations and held-out observation analyses show lower sensitivity to temporal sparsification and stronger consistency with the modeled continuous dynamics. These results support chronological latent-state estimation as a useful inductive bias for multimodal temporal learning. Important next steps include evaluating naturally irregular sensing settings and extending the deterministic estimator with explicit uncertainty over latent states and observations.

AI USE STATEMENT

In this work, we used generative AI tools for Paper polishing and grammar proofreading. We have not used generative AI tools for Method design and Experiment. The grammar of the text polished and proofread using LLM has been verified by the authors for accuracy. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies multimodal sequence modeling using existing research benchmarks. Speech, video, and affective annotations can raise privacy concerns and reflect demographic or cultural biases. BTS-Flow is intended for methodological research; its predictions should not be treated as assessments of an individual's psychological state or used alone for consequential decisions.

REFERENCES

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an *Obviously* perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4619–4629. Association for Computational Linguistics, 2019.

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

Yuqi Chen, Kan Ren, Yansen Wang, Yuchen Fang, Weiwei Sun, and Dongsheng Li. ContiFormer: Continuous-time transformer for irregular time series modeling. *Advances in Neural Information Processing Systems*, 36:47143–47175, 2023.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. *Advances in neural information processing systems*, 32, 2019.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. EMOE: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14314–14324. IEEE/CVF, 2025.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14755–14766, 2024.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9192, 2021.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2046–2056. Association for Computational Linguistics, 2019.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-Invariant and - Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, pp. 1122–1131, New York, NY, USA, 2020. Association for Computing Machinery.

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*, pp. 1309–1317, 2025.

Changqin Huang, Jili Chen, Qionghao Huang, Shijin Wang, Yaxin Tu, and Xiaodi Huang. Atcaf: Attention-based causality-aware fusion network for multimodal sentiment analysis. *Information Fusion*, 114:102725, 2025.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. *Advances in neural information processing systems*, 33:6696–6707, 2020.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled Multimodal Distilling for Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6631–6640. IEEE/CVF, 2023.

Dongxiu Liu, Haoyi Niu, Peng Cheng, Yuan Gao, Xirui Kang, Sangli Teng, Koushil Sreenath, and Xianyuan Zhan. Odeworld: A continuous predictive architecture via physical-time flow, 2026.

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 (ACL) (Volume 1: Long Papers)*, pp. 2247–2256. Association for Computational Linguistics, 2018.

Sijie Mai and Shiqin Han. Careflow: Cyclic adaptive rectified flow for multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37799–37809, June 2026.

Chunlei Meng, Guanhong Huang, Rong Fu, Runmin Jian, Zhongxue Gan, and Chun Ouyang. Clcr: Cross-level semantic collaborative representation for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1606–1615. IEEE/CVF, 2026a.

Chunlei Meng, Jiabin Luo, Pengbin Feng, Zhenglin Yan, Chengyin Hu, Zhongxue Gan, and Chun Ouyang. Mitigating shared-private branch imbalance via dual-branch rebalancing for multimodal sentiment analysis, 2026b.

Chunlei Meng, Jiabin Luo, Zhenglin Yan, Zhenyu Yu, Rong Fu, Zhongxue Gan, and Chun Ouyang. Tri-subspaces disentanglement for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8791–8800. IEEE/CVF, 2026c.

Chunlei Meng, Ziyang Zhou, Lucas He, Xiaojing Du, Chun Ouyang, and Zhongxue Gan. Temporal-spatial decouple before act: Disentangled representation learning for multimodal sentiment analysis. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11917–11921. IEEE, 2026d.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2359–2369, 2020.

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

Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. Dynamic graph neural ODE network for multimodal emotion recognition in conversation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 256–268, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

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.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. DLF: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 21180–21188. AAAI Press, 2025.

Sheng Wu, Dongxiao He, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1601–1609. AAAI Press, 2025.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *International Conference on Learning Representations*, pp. 30072–30100, 2024.

Dingkang Yang, Mingcheng Li, Linhao Qu, Kun Yang, Peng Zhai, Song Wang, and Lihua Zhang. Asynchronous multimodal video sequence fusion via learning modality-exclusive and -agnostic representations. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):12360–12375, 2024.

Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1009–1021, 2021.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7617–7630. Association for Computational Linguistics, 2023.

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 (AAAI)*, pp. 10790–10797. AAAI Press, 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. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1103–1114. Association for Computational Linguistics, 2017.

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 (ACL) (volume 1: Long papers)*, pp. 2236–2246. Association for Computational Linguistics, 2018.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 756–767. Association for Computational Linguistics, 2023.

A Related Work

A.1 Multimodal Representation and Temporal Interaction

Multimodal understanding requires modeling both information shared across modalities and information available only through individual observations. Early methods primarily focused on designing explicit multimodal fusion operators Zadeh et al. (2017); Liu et al. (2018); Rahman et al. (2020). Subsequent work increasingly considered temporal heterogeneity among modality streams.

MulT Tsai et al. (2019) introduced directional cross-modal attention that allows one modality to attend to another across different temporal positions, enabling direct modeling of unaligned multimodal sequences. MTAG Yang et al. (2021) represents multimodal temporal relations through a heterogeneous modal-temporal graph, while MEA Yang et al. (2024) explicitly studies asynchronous multimodal sequence fusion through modality-exclusive and modality-agnostic representations.

A complementary line of research focuses on the structure of multimodal representations. MISA Hazarika et al. (2020) separates modality-invariant and modality-specific components, while ConFEDE Yang et al. (2023) performs contrastive feature decomposition. TSD Meng et al. (2026c) further organizes multimodal information into structured semantic subspaces, and DLF Wang et al. (2025) studies task-oriented disentanglement with a language-focused formulation. Other methods focus on preserving and coordinating useful information rather than explicitly aligning complete representations. MMIM Han et al. (2021) employs hierarchical mutual-information objectives, ITHP Xiao et al. (2024) constructs neuro-inspired hierarchical information pathways, and CLCR Meng et al. (2026a) models collaborative representations across semantic levels. Recent architectures additionally explore adaptive temporal computation: MSAmba He et al. (2025) combines state-space sequence modeling with cross-modal interaction, whereas EMOE Fang et al. (2025) adjusts modality contributions through input-dependent expert routing.

These methods substantially improve representations defined at observed sequence positions. BTS-Flow addresses a complementary problem. Representation factorization specifies what an observed feature should encode, and cross-modal interaction specifies how observed features should exchange information. Our formulation instead asks how heterogeneous observations should be related to a separate task-relevant state whose temporal coordinate is defined independently of their sampling grids. Accordingly, representation decomposition in BTS-Flow serves as part of an observation model for a behavioral-time latent process rather than as the final fused representation.

A.2 Cross-Modal Alignment and Distribution Flow

A broad line of multimodal research seeks to reduce or reorganize representation discrepancy through invariant-specific modeling, contrastive feature decomposition, and language-guided adaptive representations Hazarika et al. (2020); Yang et al. (2023); Zhang et al. (2023). These methods improve the compatibility of heterogeneous modality information before or during multimodal integration. CaReFlow Mai & Han (2026) is the most closely related flow-based multimodal method. It reformulates cross-modal alignment as rectified-flow distribution transport. For source and target modality representations X_{m_1} and X_{m_2}, the transport path is defined as

$$X^{\tau}_{m_1,m_2} = (1-\tau)X_{m_1} + \tau X_{m_2}, \qquad \tau \in [0,1], \tag{14}$$

and a learned vector field transports representations along this path. CaReFlow further introduces one-to-many pairing, adaptive relaxed alignment, and cyclic flow to improve distribution coverage while preserving modality-dependent information. The vector fields in CaReFlow and BTS-Flow act on different axes of the problem. In CaReFlow, τ is a synthetic coordinate connecting two modality distributions; the trajectory exists in representation-transport space. In BTS-Flow, t is the elapsed chronological time within one multimodal sample; the trajectory describes a latent state associated with that sample. Moreover, BTS-Flow does not transport all modalities into a designated target distribution. Each modality retains its own observation function, and cross-modal consistency is imposed through the latent process that explains their temporal observations. Distribution compatibility and chronological state evolution are therefore distinct and complementary modeling problems.

A.3 Continuous-Time Representation Learning

Neural ODEs Chen et al. (2018) replace discrete hidden-state transformations with a learnable differential equation and obtain states through numerical integration. Latent ODEs and ODE-RNNs Rubanova et al. (2019) extend this formulation to observations separated by arbitrary temporal intervals. GRU-ODE-Bayes De Brouwer et al. (2019) combines continuous latent evolution with discrete Bayesian updates when sporadic observations become available. Neural CDEs Kidger et al. (2020) instead formulate irregular and partially observed sequences as continuously controlled dynamical systems. ContiFormer Chen et al. (2023) further extends attention-based relation modeling to continuous time. Together, these methods establish important generic baselines for any claim

concerning continuous-time or sampling-grid-independent sequence representations. Unlike Neural CDEs that use the observed path itself as the driving signal, BTS-Flow treats each modality as a heterogeneous observation process of a latent state.

Continuous dynamics have also been introduced into structured multimodal models. DGODE Shou et al. (2025), for example, applies graph Neural ODE evolution to multimodal conversational representations. In predictive world modeling, ODEWorld Liu et al. (2026) models a latent velocity field over physical time and constructs direct first-order latent-dynamics supervision through Jacobian-vector products. The transition dynamics are therefore independent of the sampling schedule, while modality-specific observations only provide state corrections when available.

BTS-Flow builds upon continuous-time representation learning, but differs in how heterogeneous modalities are incorporated into the latent state formulation. The distinction lies in the observation structure imposed on the continuous state: different modalities retain separate observation functions and private factors, asynchronously arriving observations correct a continuously propagated task state, and first-order supervision is imposed through the differential of each modality-specific observation function. BTS-Flow preserves modality-specific observation geometries while coordinating their temporal evolution through a shared latent state.

Recent world models and latent state-space models also study hidden state evolution from partial observations. However, existing approaches primarily focus on single-agent physical environments, while BTS-Flow investigates heterogeneous multimodal observations with modality-specific observation geometries.

B Experimental Setup

Benchmarks and metrics. We evaluate BTS-Flow on five multimodal benchmarks following the feature and task protocols used by CaReFlow: CMU-MOSI Zadeh et al. (2016), CMU-MOSEI Zadeh et al. (2018), and CH-SIMS-v2 Liu et al. (2022) for Multimodal Sentiment Analysis, UR-FUNNY Hasan et al. (2019) for Humor Detection, and MUStARD Castro et al. (2019) for Sarcasm Detection. The first three evaluate the method across English and Chinese sentiment settings, whereas the latter two test whether the same temporal state-estimation formulation transfers to context-punchline reasoning without task-specific changes to the core model.

For CMU-MOSI and CMU-MOSEI, we report seven-class accuracy (Acc-7), binary accuracy (Acc-2), weighted F1, mean absolute error (MAE), and Pearson correlation (Corr). For CH-SIMS-v2, we report five-class accuracy (Acc-5), Acc-2, F1, MAE, and Corr. Following CaReFlow, we report binary accuracy for UR-FUNNY and MUStARD. Higher values are better for accuracy, F1, and Corr, whereas lower MAE is better.

Comparison protocol. For standard benchmark comparisons, previously published values are retained only when their data and feature protocols match ours. Because such entries do not provide seed-level distributions, we treat them as descriptive reference points rather than using them for significance claims. For the continuous-time baselines most directly related to BTS-Flow, we reimplement all methods under a matched protocol with identical unimodal inputs, latent dimensionality, prediction head, training split, and tuning budget, and report five-seed mean$\pm$standard deviation.

Matched architecture. All controlled comparisons use the same unimodal inputs and downstream two-layer MLP predictor. BTS-Flow uses a latent dimension of $d_z = 128$. The state-related encoder S_m, private encoder P_m, observation function G_m, initial-state estimator I_ϕ, and innovation updater U_ψ are implemented as lightweight two-layer networks with hidden width 128. Simultaneously arriving observations are combined through permutation-invariant set aggregation with learned modality embeddings.

Continuous-time implementation. Chronological timestamps are rescaled by a single training-set time scale T_s shared by all samples. We do not independently normalize each sample to $[0, 1]$, because doing so removes absolute differences in temporal rate.

Unless otherwise stated, latent propagation is evaluated using fixed-step RK4. The influence of solver choice is analyzed separately in Sec. 3.7. Acoustic and visual derivatives are estimated using

second-order local polynomial regression over five adjacent observations. The polynomial design matrix uses the actual timestamp differences. The derivative confidence weight is $\xi_i^m = \exp(-\kappa\varepsilon_i^m)$ with $\kappa = 5$.

Optimization. All newly trained configurations use AdamW with an initial learning rate of 1×10^{-4}, weight decay 1×10^{-5}, batch size 32, gradient clipping at 1.0, and at most 50 epochs. Early stopping with patience 8 is performed on the validation split. The default loss weights are

$$\lambda_{\text{rec}} = 0.1, \quad \lambda_{\text{sep}} = 0.01, \quad \lambda_{\text{obs}} = 1.0, \quad \lambda_{\text{1st}} = 0.1.$$

All hyperparameters are selected using the validation set only.

C Model Zoo

We compare BTS-Flow against representative multimodal methods spanning representation decomposition, self-supervised learning, pretrained multimodal adaptation, information-theoretic fusion, dynamic modality modeling, causal fusion, and flow-based alignment.

Representation-oriented methods. MISA Hazarika et al. (2020) decomposes each modality into modality-invariant and modality-specific subspaces, aiming to preserve both cross-modal commonality and modality-dependent information. Self-MM Yu et al. (2021) introduces self-supervised unimodal label generation and jointly optimizes unimodal and multimodal objectives to improve modality-specific representation learning. DMD Li et al. (2023) decouples modality representations into shared and exclusive components and performs adaptive cross-modal knowledge distillation within the resulting subspaces. MMIM Han et al. (2021) preserves task-relevant multimodal information through hierarchical mutual-information maximization between unimodal inputs and fused representations. DLF Wang et al. (2025) further separates modality-shared and modality-specific information and constructs a language-focused representation using complementary non-linguistic cues. CLCR Meng et al. (2026a) explicitly organizes multimodal representations into multiple semantic levels and constrains cross-modal exchange according to their hierarchical semantic structure.

Adaptive and pretrained multimodal methods. MAG-BERT Rahman et al. (2020) augments pretrained language representations with a multimodal adaptation gate conditioned on acoustic and visual information. ITHP Xiao et al. (2024) adopts an information-bottleneck perspective and constructs a hierarchical information pathway to retain compact task-relevant multimodal representations. EMOE Fang et al. (2025) employs modality-specific experts and sample-dependent routing to dynamically adjust modality contributions while preserving unimodal predictive information. KuDA Feng et al. (2024) uses sentiment knowledge to dynamically identify the dominant modality and adapt modality contributions during fusion. DEVA Wu et al. (2025) converts acoustic and visual affective cues into textual descriptions and progressively incorporates them into text-centered multimodal representations.

Causal and flow-based fusion. AtCAF Huang et al. (2025) introduces causality-aware text debiasing and counterfactual cross-modal attention to reduce spurious correlations during multimodal fusion. CaReFlow Mai & Han (2026), our primary baseline, formulates modality discrepancy as rectified-flow distribution transport and introduces one-to-many mapping, adaptive relaxed alignment, and cyclic flow to improve cross-modal distribution alignment while retaining modality-specific information.

These baselines provide complementary reference points for BTS-Flow. Importantly, CaReFlow is the closest flow-based competitor, whereas the remaining methods mainly operate on observed modality representations, representation distributions, or discrete sequence states. BTS-Flow instead models heterogeneous modality observations as measurements of a latent state defined over chronological time.

Table 4: Sensitivity to BTS-Flow-specific hyperparameters on CMU-MOSI. For the derivative-estimator comparison, confidence weights are kept fixed to isolate the effect of the derivative target.

Setting	First-order weight $\lambda_{1\mathrm{st}}$					Latent dimension d_z			Derivative estimator ($\lvert\mathcal{W}\rvert = 5$)		
	0	0.05	0.10	0.20	0.50	64	128	256	Linear	Quadratic	Cubic
Acc-2(%)	90.1	90.3	**90.4**	90.3	90.0	90.1	**90.4**	**90.4**	90.2	**90.4**	90.3
RGS↓	0.166	0.139	**0.127**	0.129	0.143	0.139	0.127	**0.126**	0.142	**0.127**	0.130

D Numerical Stability and Hyperparameter Sensitivity

We examine whether the behavior of BTS-Flow is robust to numerical integration choices rather than being tied to a particular solver. For this analysis, we keep the RK4-trained checkpoints fixed and replace only the inference-time integrator. All fixed-step solvers use the same integration partition, while the adaptive solver uses fixed error tolerances across all samples. We additionally examine sensitivity to the first-order supervision weight, latent-state dimensionality, and local polynomial order. Each hyperparameter configuration is trained independently under the same optimization protocol, with all remaining settings fixed.

Fig. 4(b) shows that the performance of BTS-Flow is stable across standard numerical integrators. Lower-order solvers introduce moderate degradation in both task performance and temporal consistency, whereas RK4 and Dopri5 yield nearly identical results. In particular, Dopri5 changes RGS and CQE by only 0.001 relative to RK4 without improving Acc-2, while requiring 6.3 versus 4.3 vector-field evaluations per propagation interval. We therefore use RK4 as the default solver, which provides a favorable trade-off between numerical accuracy and computational cost. These results indicate that the reported behavior is not specific to a single integration scheme, although the comparison is not intended as an equal-compute evaluation of solver order.

Table 4 further shows that BTS-Flow is stable over a practical range of model-specific hyperparameters. Removing first-order supervision increases RGS from 0.127 to 0.166, while its effect on Acc-2 is comparatively smaller, consistent with its role in constraining temporal dynamics rather than directly increasing classifier capacity. Performance remains similar for $\lambda_{1\mathrm{st}} \in [0.05, 0.20]$, indicating that the method does not depend on a narrowly tuned loss weight. Increasing d_z from 64 to 128 improves both Acc-2 and RGS, whereas further increasing it to 256 yields no additional Acc-2 gain and only a marginal change in RGS; we therefore retain $d_z = 128$ as the default. Finally, quadratic local regression performs better than a linear estimator, while a cubic fit provides no further improvement, supporting the use of a low-order local model for timestamp-aware derivative estimation.

E Efficiency Analysis

We examine whether the temporal gains can be explained by additional model capacity or substantially increased numerical computation. Because continuous-time methods incur solver-dependent computation, parameter count alone is insufficient; we report both model size and inference overhead. The purpose of Table 5 is not to claim that continuous state estimation is computationally free. Rather, it determines whether the observed temporal consistency is obtained with a moderate and transparent solver overhead rather than through a substantially larger network. Together with Fig. 4(b), this separates the representational contribution of BTS-Flow from numerical computation.

F Definitions of Diagnostic Metrics

For observation-grid sensitivity, we evaluate the same N test samples under $L \geq 2$ schedules with matched observation budgets and temporal coverage, holding model parameters and linguistic observations fixed. Let $r_n^{(g)}$ be the sequence representation supplied to the prediction head for sample n under schedule g, and let $\hat{y}_n^{(g)}$ be its continuous sentiment score before discretization. For BTS-Flow, $r_n^{(g)}$ is the output of the temporal readout $\mathcal{R}$. The matrix $R^{(g)}$ stacks these representations as rows in the same sample order for every schedule.

Table 5: Efficiency under the matched CMU-MOSI setting. Published parameter counts for MulT and CaReFlow follow the CaReFlow comparison.

Method	Params (M)	Rel. latency	Peak mem. (GB)	Acc-2
MulT	185.51	1.00	3.04	88.1
CaReFlow	185.38	1.08	3.21	89.8
Neural CDE	186.40	1.42	3.55	89.5
BTS-Flow	188.10	1.31	3.48	**90.4**

Representation-grid sensitivity (RGS). We compare schedules within each model using linear centered kernel alignment (CKA). Let $\widetilde{R}^{(g)} = HR^{(g)}$, where $H = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ centers each feature across samples. We compute

$$\mathrm{CKA}(R^{(g)}, R^{(h)}) = \frac{\|(\widetilde{R}^{(g)})^\top \widetilde{R}^{(h)}\|_F^2}{\max\Big(\|(\widetilde{R}^{(g)})^\top \widetilde{R}^{(g)}\|_F \|(\widetilde{R}^{(h)})^\top \widetilde{R}^{(h)}\|_F, \epsilon\Big)}, \tag{15}$$

and define RGS as one minus the mean similarity over unordered schedule pairs:

$$\mathrm{RGS} = 1 - \frac{2}{L(L-1)} \sum_{g<h} \mathrm{CKA}(R^{(g)}, R^{(h)}). \tag{16}$$

Prediction-grid sensitivity (PGS). We measure the mean absolute shift in continuous predictions, in the original sentiment-score units:

$$\mathrm{PGS} = \frac{2}{NL(L-1)} \sum_{g<h} \sum_{n=1}^{N} \left| \hat{y}_n^{(g)} - \hat{y}_n^{(h)} \right|. \tag{17}$$

Velocity cosine agreement (Vel.-Cos). We use the same predicted changes $\hat{u}_i^m$, estimated changes u_i^m, and confidence weights ξ_i^m as the FOE evaluation in Sec. 3.6. With $\mathcal{M}_c = \{a, v\}$, we define

$$\begin{aligned} c_\epsilon(a, b) &= \frac{a^\top b}{\max(\|a\|_2, \epsilon)\max(\|b\|_2, \epsilon)}, \\ \mathrm{Vel.Cos} &= \frac{\sum_{m\in\mathcal{M}_c} \sum_i \xi_i^m c_\epsilon(\hat{u}_i^m, u_i^m)}{\max\big(\sum_{m\in\mathcal{M}_c} \sum_i \xi_i^m, \epsilon\big)}. \end{aligned} \tag{18}$$

The sums include all valid acoustic and visual observations in the evaluation set, with sample indices suppressed. Linguistic observations do not enter this derivative metric. Here $\epsilon > 0$ is a numerical stabilizer.