Weakly Supervised Temporal Modeling of Latent Dynamics in Dyadic Conversations

Tahiya Chowdhury
Colby College & Davis Institute for AI
Waterville, Maine, USA

SIGDIAL 2026, Atlanta, Georgia, USA

Colby

DAI DAVIS AI Institute for Artificial Intelligence

Motivation: Hidden states in Conversations

Understanding Remote collaborative conversation between two people need more focus on the **latent** social cognitive states

- **Cognitive load** How hard the task feels, interaction difficulty
- **Coordination Quality** How well partners align and hand off turns to each others
- **Participation asymmetry** Who holds the floor, who defers to their partners during the conversations

This state is not directly observable

- Yet it shapes how people interact
- How the state changes as the interaction unfolds.
- Uncovering such latent states is promising for dynamic modeling of dialogue.

Prior Work: Dialogue State Tracking in Interaction

Dialogue state tracking

- State = discrete slot-value pairs
- State is updated at each turn
- Semantic, annotated explicitly

Cognitive load during interaction

- Load inferred from speech & video
- Post-hoc, task-level labels
- Predicted from aggregate features

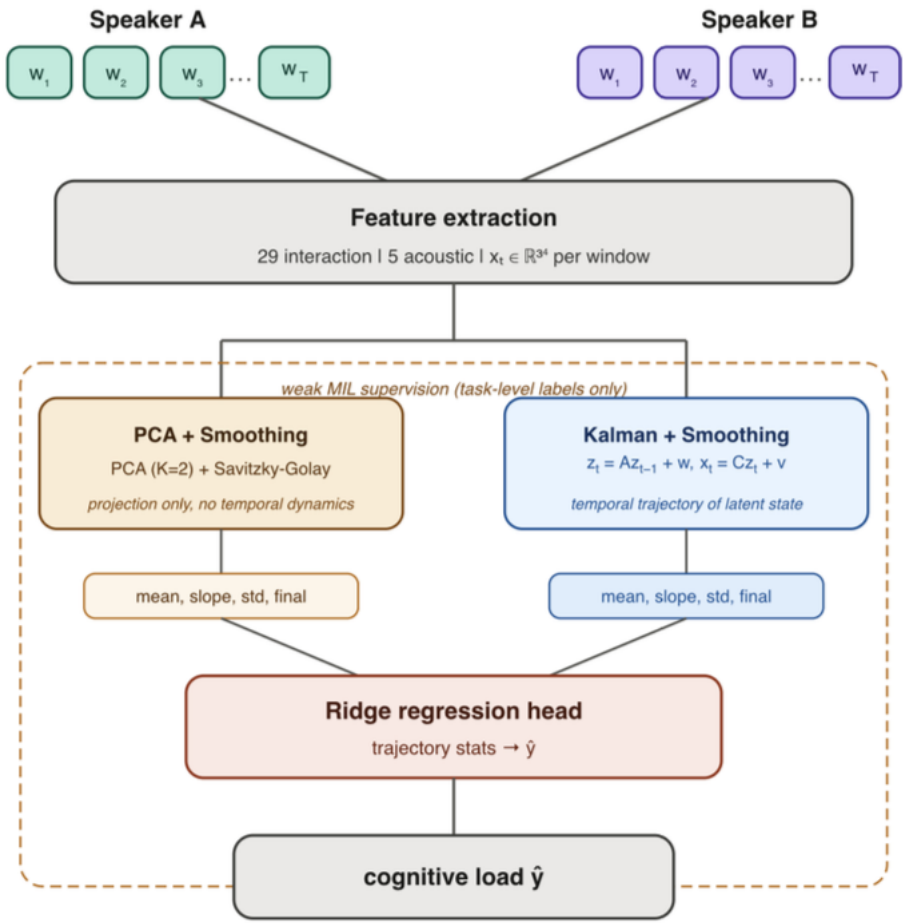
Traditionally, conversational state are viewed as a ***fixed, static endpoint,*** rather than an evolving process that changes over ***time***

Williams & Young. Partially observable Markov decision processes for spoken dialog systems. Computer Speech & Language, 2007.
Young et al. POMDP-based statistical spoken dialogue systems: A review. Proc. IEEE, 2013.

Dyadic dialogue as a partially observable dynamic system

- The social-cognitive state of a dyad: not directly observable
- Changes within a task as the interaction unfolds.
- Temporal evolution carries information that static summary features ignore.

This work: Model dyadic conversation as a **partially observable dynamic system** with **weak, task-level supervision**



Problem Formulation: DST for Continuous State Tracking

Partially observable temporal process to capture time-evolving properties of the cognitive states

The Learning Problem

$X^{(i)}$

Observed features
Extracted at window level

$$X^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_T^{(i)}]$$

$Z^{(i)}$

Latent conversational state from a given window, not annotated

$$Z^{(i)} = [z_1^{(i)}, z_2^{(i)}, \ldots, z_T^{(i)}]$$

$y^{(i)}$

Task-level load score from NASA TLX, **self-reported**

Each task is a bag of windows with a single bag level label

Carbonneau et al. MIL: A survey. Pattern Recognition, 2018.
Song et al. Customer-service dialogues for satisfaction analysis with context-assisted MIL. EMNLP-IJCNLP, 2019.

Data: AVCAffe Dataset (53 dyads, 9 collaborative tasks)

AVCAffe: audio-visual corpus of remote dyadic collaboration over video-conferencing (Sarkar et al., 2023) where tasks divided in 3 levels (low, medium, high)

ID	Task	Category	Dur. (min)	Seqs	Win. (mean)	Mental demand mean ± std	Temporal demand mean ± std
1	Open discussion	Low	7.5	52	11.2	5.1 ± 3.7	5.3 ± 3.8
2	Sharing jokes	Low	7.5	52	11.0	7.2 ± 4.6	4.2 ± 3.2
3	Find differences	Medium	10.0	52	14.8	10.4 ± 3.3	9.2 ± 4.4
6	Reading comp 1	Medium	5.0	52	7.3	10.1 ± 4.0	5.9 ± 4.0
7	Reading comp 2	Medium	5.0	51	7.1	10.9 ± 3.8	9.1 ± 4.6
4	Map matching 1	High	2.5	51	3.7	11.2 ± 4.7	15.8 ± 4.1
5	Lost at Sea	High	10.0	51	15.1	11.1 ± 4.5	14.7 ± 4.9
8	Multi-task	High	10.0	50	14.5	9.4 ± 4.0	7.5 ± 4.5
9	Map matching 2	High	2.5	50	3.6	14.0 ± 3.6	10.3 ± 4.4
	All tasks			461	8.7	9.92 ± 4.67	9.11 ± 5.67

Table 1: AVCAffe dataset statistics by task, grouped by cognitive demand category. *Seqs* = dyad-task sequences (53 dyads × 9 tasks); *Win.* = mean valid windows per sequence after speech-activity masking (1,576 of 5,982 windows excluded, 26.3%); demand scores are on the 0–21 NASA-TLX scale. Low = open-ended tasks; Medium = cognitive tasks; High = time-pressured or problem-solving tasks.

Sarkar, Posen & Etemad. AVCAffe: A large-scale audio-visual dataset of cognitive load and affect for remote work. AAAI, 2023.
Hart & Staveland. NASA-TLX. Advances in Psychology, 1988.

Target Label for Prediction from NASA TLX

Mental demand

degree of mental & cognitive difficulty the task requires

Temporal demand

perceived time pressure due to task urgency

Single Dyad-level Target: Mean of Self-reported score (on a scale of 0-21) of A and B

Rationale

- **Shared context :** one joint conversation experienced by both speakers
- **Features are inherently joint:** floor imbalance, overlap, turn-count ratio cannot be attributed to one speaker

Hart & Staveland. NASA-TLX (Task Load Index). Advances in Psychology, 1988

Features extracted from windowed segments

- Task recording divided into 30s non-overlapping windows (8.7 windows per task)
- Applied to speech activity segments only (SILERO VAD)

29 Interaction features

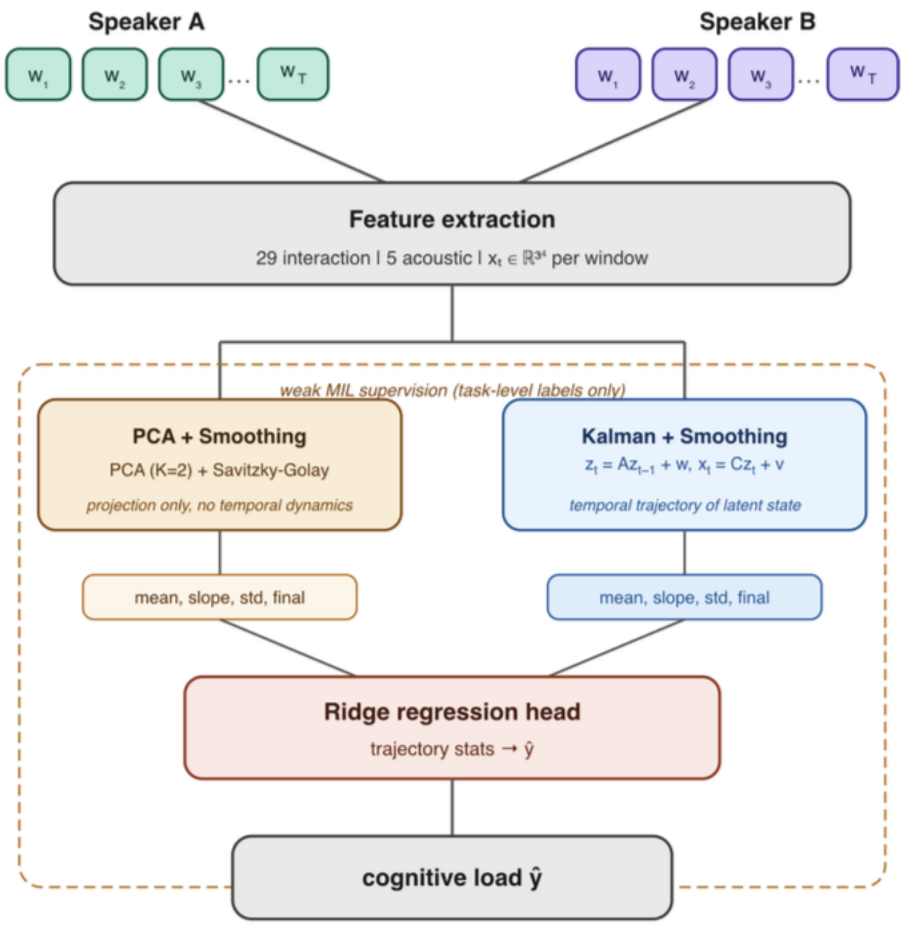
- **Floor control:** talk-time, floor imbalance
- **Turn-taking:** turn counts, mean turn duration
- **Overlap :** simultaneous speech count, duration, & ratio
- **Interruptions:** A→B, B→A, rate, imbalance
- **Response latency:** mean, median, std, max
- **Silences:** long-pause counts, max inter-turn

5 Dyad relative acoustic features

Relative for robustness to automatically applied video-conferencing gain control

- pitch mean, std, range
- voiced segments / min
- voiced duration

Experimental Models



Static

Ridge and Random Forest based Regression

- Mean-pool features from a window
- Predict the task level label
- All temporal order are discarded

Sequential (neural)

Multiple Instance based pooling

- Mean pooling over temporal window sequence: every window is treated the same way
- Attention pooling: learns which windows matter, contextual

Trajectory

Latent state using PCA and Kalman

- Estimate latent 2-D trajectory
- predict from shape statistics (e.g. slope)

Ilse et al. Attention-based deep MIL. ICML, 2018.
Shumway & Stoffer. Time-series smoothing via EM. J. Time Series Analysis, 1982.

Trajectory Based Models

PCA: No model of temporal dynamics

Static projection

- Linear projection to 2-D space, then Savitzky–Golay smoothing
- Window order does not affect the projection
- Shows how the dyad moved through a **static** feature space

Kalman Smoother: Temporal

Learned dynamics

- Linear-Gaussian state space, fit per window via Expectation Maximization
- Transition matrix A makes z_t at time t depend on z_{t-1}
- Explicitly models how state evolves window to window

- 4 scalar statistics calculated from 2-D trajectory (latent state over time)
- Mean (overall state), final value (end of task state), slope (temporal trend), std (variability)

Model Evaluation

Leave one dyad out

- All task from a particular dyad will be the test set
- Separate speakers in train and test

Metric

- Concordance Correlation Coefficients: captures dyad agreement
- Mean Absolute Error: Error

- Reported as mean and std over 53 folds of LODO to capture dyad variability

Results: Predicting Cognitive Load

	Mental Demand		Temporal Demand	
Model	CCC	MAE	CCC	MAE
Ridge (static)	0.163 ± 0.228	3.85	0.126 ± 0.178	4.85
Random Forest	$\mathbf{0.197} \pm 0.221$	**3.71**	$\mathbf{0.254} \pm 0.198$	**4.36**
GRU (mean pool)	-0.002 ± 0.152	4.23	-0.020 ± 0.108	5.10
GRU + attention	0.141 ± 0.258	3.90	0.149 ± 0.231	4.84
PCA + smoothing	0.129 ± 0.154	3.88	0.057 ± 0.113	4.79
Kalman smoother	0.073 ± 0.111	3.88	0.125 ± 0.109	4.68

- Temporal order alone is insufficient. Trajectory models are far more stable.
- Sequential model with right pooling strategy improved performance
- Claim: dynamic modeling improves interpretability, but may not improve accuracy
- High dyad-level variance: learned attention weights do not generalize to unseen dyads

Lin. A concordance correlation coefficient to evaluate reproducibility. Biometrics, 1989

Feature Ablation

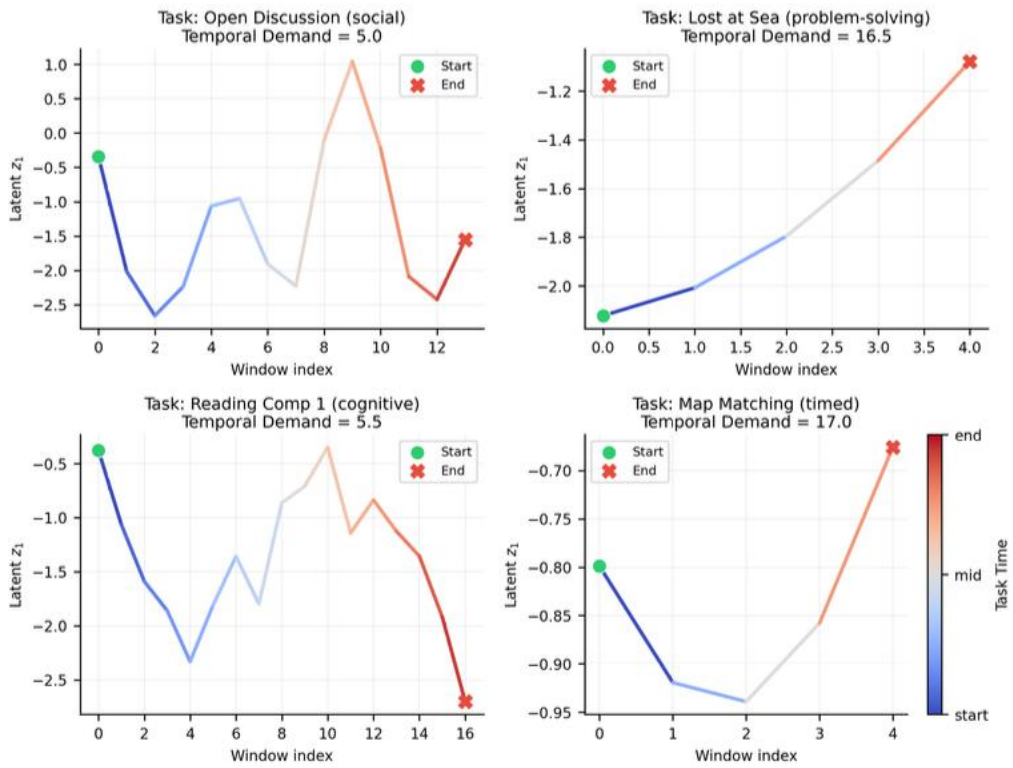
- Dyad heterogeneity reflected by high standard deviation
- Interaction features carry the signal for both targets
- Acoustic features provide minimal signal for mental demand during video-conferencing
- Acoustic helps in temporal demand only

Feature Set	Mental CCC	Temporal CCC
Acoustic (5)	$0.044 \pm .179$	$0.217 \pm .257$
Interaction (29)	$0.197 \pm .250$	$0.209 \pm .201$
Int. + acoustic (34)	$\mathbf{0.197} \pm .221$	$\mathbf{0.254} \pm .198$

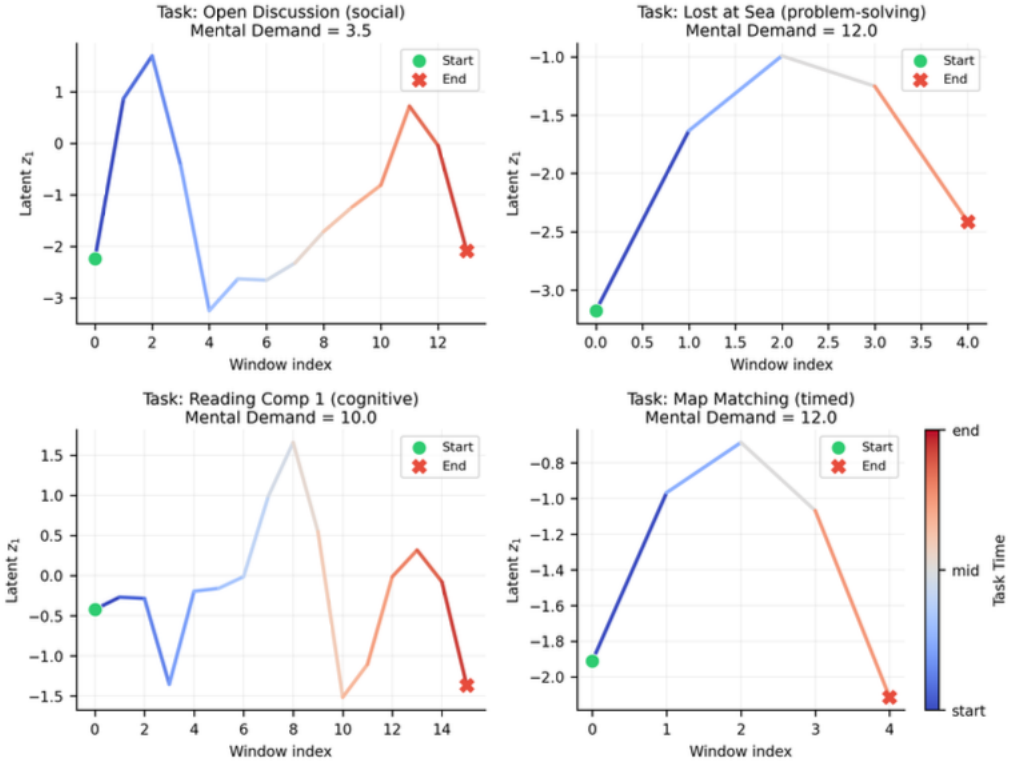
- turn-taking pacing → temporal demand
- participation imbalance → mental demand

Latent State Trajectories: Task Type Matters

Latent state z_1 trajectories over window index for four task categories



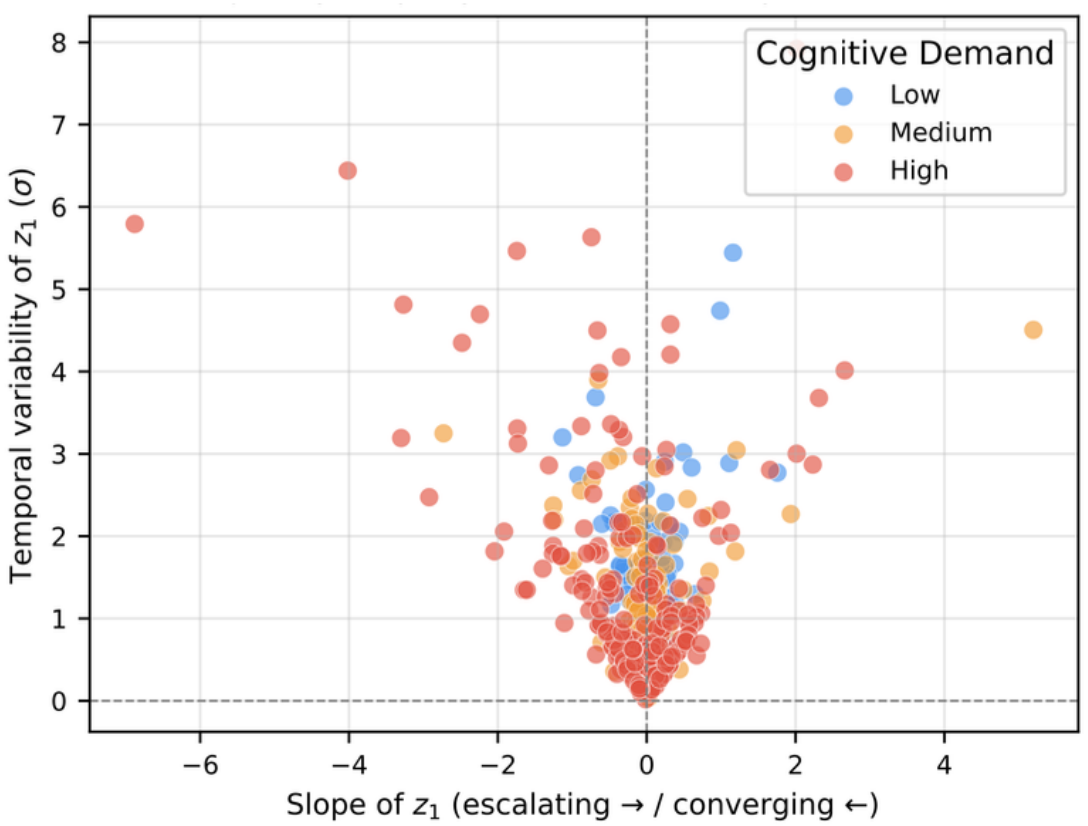
(a) Temporal demand. High-demand tasks (Map Matching, Lost at Sea; label ≥ 16) show monotonically escalating z_1, consistent with accumulating time pressure. Low-demand tasks show flat or declining trajectories.



(b) Mental demand. Trajectory shapes are less systematic; high-demand tasks show higher-amplitude trajectories with greater variability.

Slope is not predictive, but shape is informative

- High-demand tasks (red) span the full slope range and are over-represented at high variability (σ > 3)
- **High demand tasks** have more dynamic and less predictable conversational state trajectories.
- **Low-demand tasks** cluster near the slope ≈ 0 with lower variability, reflecting stable dynamics
- Slope of z_1 is weakly correlated with load labels
- Trajectory shape expose within-task dynamics that static summaries cannot



Takeaways: Toward dynamic dialogue modeling

1. **Pooling is an architectural choice.** Attention Pooling improves accuracy but overfits. In small-dataset MIL, how pooling is performed matters

2. **Interpretability.** Trajectory models don't beat static prediction — but they expose escalation, convergence, and stability that no pooling model can.

3. **Task dependence of trajectory models.**
   - Kalman smoother is well-suited for dynamics and temporal modeling
   - PCA for static diffuse mental demand
   - Choice of trajectory model depends on the type of latent state learned

4. **First step toward dynamic state-aware systems.** If conversational state can be tracked, a dialogue system could use it to trigger adaptive, well-timed responses.

Limitations & Practical Considerations

Limitations

- Single dataset, 53 dyads: limited generalizability
- Post-hoc labels are task-level: Labels do not reflect moment-to-moment state
- Some tasks have short duration, with 3-4 windows: limited temporal resolution
- Speech segments may count laughter / filled pauses as speech, may miscalculate overlap features (e.g. in Sharing Jokes)
- No linguistic or visual signals have been considered (***future work***)

Ethical Considerations

- Turn-taking norms, latency, and overlap vary across languages and cultures.
- Equitable performance across different demographic groups should be ensured before any real-time deployment.

Stivers et al. Universals and cultural variation in turn-taking in conversation. PNAS, 2009.

Thank you

Questions? Please reach out at:
tahiya.chowdhury@colby.edu

Scan for link to paper

Sponsors